School of Information Systems

Trustworthiness in Social Big Data Incorporating Semantic Analysis, Machine Learning and Distributed Data Processing

Bilal Ahmad Abdal Rahman Abu Salih

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

April 2018

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Bilal Ahmad Abdal Rahman Abu Salih

06 April 2018

Abstract

The proliferation of Online Social Networks (OSNs) has opened new horizons and brought profound changes to various aspects of human, cultural, intellectual, and social life. These significant Big Data (BD) tributaries have further transformed the businesses processes by establishing convergent and transparent dialogues between businesses and their customers. Therefore, analysing the flow of Social Big Data (SBD) content is necessary in order to enhance business practices, to increase brand awareness, to develop insights on target markets, to detect and identify positive and negative customer sentiments, etc., thereby achieving the hoped-for added value. However, due to the vast amount of information produced by these platforms, in conjunction with the lack of a gatekeeper for those sites, it is difficult to verify the credibility of their content and users. Therefore, the OSNs are hijacked, and their otherwise valuable tools are used to spread chaos and false news. Hence, it is essential to have an accurate understanding of the contextual content of social users, in order to establish a ground for measuring their social credibility accordingly. Further, it is important to classify users and their content into appropriate categories prior to undertaking further business analytics.

Considerable achievements have been made in SBD analytics motivated by the need for efficient and effective social data analytics solutions. In particular, several studies have been carried out in the areas of social trust, semantic analysis, machine learning and social data classification within the context of SBD. However, the efforts in these areas have shown shortcomings in terms of: (i) the lack of domain-based trustworthiness approaches; (ii) their inability to manage and extract high-level domains from the textual content of SBD; and (iii) the lack of domain-based approaches for dual classification of the textual content of SBD at the user level and post level.

This thesis presents several state-of-the-art approaches for social data analytics in order to address the aforementioned research problems. The frameworks have been constructed for the purpose of studying the trustworthiness of users in OSNs platforms, deriving concealed knowledge from their textual content, and classifying and predicting the domain knowledge of users and their content. The contribution of this thesis is threefold: (i) an effective and an efficient credibility framework for users of OSNs addressing the key features of BD, and incorporating semantic analysis and the temporal factor, (ii) a semantic analysis-based approach to extract knowledge captured from the textual content of SBD at user level and post level, and (iii) an integrated framework incorporating domain knowledge discovery tools and machine-learning-based data classification techniques in the quest for domain-based discovery, classification and prediction. The developed approaches are refined through proof-of-concept experiments, several benchmark comparisons, and appropriate and rigorous evaluation metrics to verify and validate their effectiveness and efficiency, and hence, those of the applied frameworks.

Acknowledgements

First and foremost, I would like to thank God Almighty for giving me the acuity, knowledge and enthusiasm to undertake my PhD degree. Without His blessings, the accomplishment of this work would not have been possible.

I would like to express my genuine and sincere gratitude to my supervisor, Dr Ponnie Clark. Her tremendous insights, guidance, and advice throughout my PhD journey have been priceless. Her constant support and encouragement definitely helped me to grow as a researcher. This is evident through her supervision and direction in respect of papers accepted in several scholarly research avenues which has improved my research track record. I would like to express my sincere thanks to my associate-supervisors Dr Kit Yan Chan, Dr Amit Rudra, Dr Dengya Zhu and Dr Amin Beheshti for their encouragement, insightful guidance, technical help and support. Further, I extend my gratitude to Dr Vidyasagar Potdar who generously agreed to supervise me during the final stage of my study. His splendid guidance and generous support have been extremely valuable.

My particular and deepest gratitude goes to my wife, Bushra Bremie, for giving me an immense internal strength, patience and capability to accomplish this thesis. My heartfelt thanks to my gorgeous babies, Razan and Rashid, who have been an endless source of inspiration and love. Words are inadequate to express my gratitude to my fathers, brothers and sisters for their constant caring and encouragement.

I am highly indebted to the University of Jordan for providing me with the financial support to pursue my doctorate studies. Finally, heaps of thanks to my fellow students at Enterprise Unit 4, Technology Park who made my PhD journey an unforgettable and enjoyable experience.

Dedication

To the loving memory of my late mother, Fayzeh Awajan, and to my Father, Ahmad Abu Salih, who always believed in my ability to succeed academically.
To my wife Bushra Bremie,
To my babies, Razan and Rashid
To my brothers and sisters, Heba, AbdulRahman, Aman, AbdAllah, Batool, AbdulAziz, and AbdulFattah

List of Publications

Journal Articles

B. Abu-Salih, P. Wongthongtham, K. Y. Chan. 2018. "Twitter Mining for Ontology-based Domain Discovery Incorporating Machine Learning", *Journal of Knowledge Management (JKM)*. Vol. 22 Issue: 5, pp.949-981, https://doi.org/10.1108/JKM-11-2016-0489, **(ABDC- Rank : A)**.

B. Abu-Salih, P. Wongthongtham, K. Y. Chan, Z. Dengya. 2018. "CredSaT: Credibility Ranking of Users in Big Social Data incorporating Semantic Analysis and Temporal Factor", *Journal of Information Science (JIS),* https://doi.org/10.1177/0165551518790424 , **(ERA2010- Rank: A)**.

P. Wongthongtham, and **B. Abu-Salih**, "Ontology-based Approach for Identifying the Credibility Domain in Social Big Data", *Journal of Organizational Computing and Electronic Commerce (JOCEC), Inpress - Accepted Mar 2018,* **(ABDC- Rank: A).**

P. Wongthongtham, K. Y. Chan, V. Potdar. 2018. **B. Abu-Salih**, S. Giakwad, J. Pratima, “State-of-the-Art Ontology Annotation for Personalised Teaching and Learning and Prospects for Smart Learning Recommender Based on Multiple Intelligence and Fuzzy Ontology”. *International Journal of Fuzzy Systems (IJFS).* Vol. 20 Issue: 4, pp. 1357-1372, https://doi.org/10.1007/s40815-018-0467-6.

B. Abu-Salih, P. Wongthongtham, Z. Dengya, SH.Alqrainy. 2015. "An Approach for Time-Aware Domain-Based Analysis of Users’ Trustworthiness in Big Social Data", *Services Transactions on Big Data (STBD)*, Vol. 2 Issue: 1, pp. 41-56, https://doi.org/10.29268/stbd.2015.2.1.4.

Conferences Articles

B. Abu-Salih, P. Wongthongtham, S.-M.-R. Beheshti, and Z. Dengya, "A Preliminary Approach to Domain-based Evaluation of Users' Trustworthiness in Online Social Networks," in IEEE International Congress on Big Data (BigData Congress-2015), New York, USA, 2015.

B. Abu-Salih, P. Wongthongtham, S.-M.-R. Beheshti, and B. Zajabbari, "Towards A Methodology for Social Business Intelligence in the era of Big Social Data incorporating Trust and Semantic Analysis" in Second International Conference on Advanced Data and Information Engineering (DaEng-2015), ed. Bali, Indonesia: Springer, 2015.

P. Wongthongtham, **B. Abu-Salih,** "Ontology and trust-based data warehouse in new generation of business intelligence: State-of-the-art, challenges, and opportunities", in IEEE 13th International Conference on Industrial Informatics (INDIN) 2015: 476-483.

R. Nabipourshiri**, B. Abu-Salih,** P. Wongthongtham, "Tree-based Classification to Users' Trustworthiness in OSNs", in 10th International Conference on Computer and Automation Engineering (ICCAE 2018) 2018, Brisbane, Australia: ACM, 2018

J. Kaur, P. Wongthongtham, **B. Abu-Salih,** S. Fathy, "Analysis of Scientific Production of IoE Big Data Research", in the 32nd IEEE International Conference on Advanced Information Networking and Applications Workshops (WAINA), Krakow, Poland, 2018, pp. 715-720. doi: 10.1109/WAINA.2018.00173.

Table of Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence
BD	Big Data
BI	Business Intelligence
CredSAT	Credibility incorporating Semantic Analysis and Temporal factor
DC	Dublin Core
DSRM	Design Science Research Methodology
DW	Data Warehouse
FLSA	Fuzzy Latent Semantic Analysis
FOAF	Friend-of-a-Friend
HDFS	Hadoop Distributed File System
HITS	Hyperlink-Induced Topic Search
GB-DT	Gradient Boosting-based Decision Tree
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
LOD	Linked Open Data
LSA	Latent Semantic Analysis
LR	Logistic Regression
ML	Machine Learning
NDGC	Normalized Discounted Cumulative Gain
NLP	Natural Language Processing
OLAP	Online Analytical Processing
OSNs	Online Social Networks

OWL	Web Ontology Language
RDF	Resource Description Framework
RF-DT	Random Forest-based Decision Tree
SBD	Social Big Data
SIOC	Semantically-Interlinked Online Communities
SKOS	Simple Knowledge Organization System
SBI	Social Business Intelligence
SVM	Support Victor Machine
SW	Semantic Web
TD-DT	Top-Down inducing based Decision Tree
TF-IDF	Term Frequency – Inverse Document Frequency
UGC	User Generated Content
UMBEL	Upper Mapping and Binding Exchange Layer
URI	Universal Resource Identifier
VoC	Voice of the Customer
VoM	Voice of the Market
YAGO	Yet Another Great Ontology

Chapter 1 Introduction

1.1 Overview

Since the mid-1990s, modern technological developments have made a qualitative leap and effected a real revolution in the world of communication. The Internet has spread throughout the globe, connecting most parts of this vast world and paving the way for several societies to converge and to exchange views, ideas and desires. The Online Social Networks (OSNs), positioned on the throne of cyberspace, are continuing to spread exponentially by providing social communication services to their affiliated members. The services offered by these sites have expanded, providing their consumers with broad possibilities for exchanging information in the fields of education, health, culture, sports and other domains of knowledge.

In modern enterprises, OSNs are used as part of the infrastructure for a number of emerging applications such as recommendation and reputation systems. In such applications, trust is one of the most important factors in decision making. In this context, social trust analysis is an emerging task and combines disciplines such as social network analysis, semantic discovery, and big data computing. As the massive amounts of data are derived from a variety of sources, it is essential to determine the reputation of these sources and provide flexibility to the analysts, so that the trust value of each source can be determined. Another important consideration is the semantics of extracted textual data from which meaningful information can be derived. This provides a solid foundation for several advanced analytics such as classification and prediction.

This chapter presents a brief introduction to this thesis; first, the importance and challenges of the exponentially increasing social data and the significance of deriving knowledge and measuring the credibility of the content of the online social platforms are discussed. Second, the motivation for an approach to address the social, big data problem is particularised by demonstrating the importance of determining the domains of interest of users and their content which leads to improving the forecasting of their future interest(s). Third, the list of concerns arising from the propagation of massive social data and the needs to be addressed in this thesis are condensed. Fourth, the overall objectives of this research are summarised. Fifth, the key contributions to the body of knowledge are outlined followed by the practical and theoretical significance of this research. Last, the general structure of this thesis is described.

1.1.1 Social Big Data (SBD)

Since the emergence and proliferation of Web 2.0, the role of web browsers has changed to enable users to send and receive content by means of several online tools that commenced with e-mail applications, chat, and chat forums that evolved into more recent and revolutionary electronic platforms such as OSNs. OSNs such as Facebook®, Twitter®, LiveBoon®, Orkut®, Pinterest®, Vine®, Tumblr®, Google Plus®, Instagram ® etc, which enable users to share videos, photos, files and instant conversations. These platforms provide an important means by which communities can grow and consolidate, allowing individuals or groups to share concepts and visions with others. Moreover, in addition to playing an active and distinctive role as effective media of social interaction, these OSNs allow users to become acquainted with and understand the cultures of different peoples(SAWYER and Guo-Ming 2012).

OSNs are relevant sources of data for SBD which include shared textual content, picture, videos, etc. The vast amount of social data has spread to many different areas in everyday life such as e-commerce (Kaplan and Haenlein 2010), education (Tess 2013), health (Salathé et al. 2013), to name a few. For example, several modern computing applications such as online education, weight loss and public health, music and entertainment rely on the content generated by OSNs (Althoff, Jindal, and Leskovec 2017). This is evident in the dramatic increase in the use of these platforms for networking and communication. The Pew Research Center reported that 70% of American adults in Nov 2016 used OSNs for social interactions compared to 5% usage by the same user category in 2005 (News 2017). In Australia, the statistics for OSNs usage in Jan 2017 indicated around 2.8 million Twitter active users, 14.8 million visits to YouTube, 4.0 million Snapchat active users (News 2017). Such a dramatic connectivity with online social platforms has established a common ground that brings together people with shared interests, ideas and goals. Furthermore, the widespread use of OSNs has established several communication channels between business firms with their current and potential customers; hence, "many marketing researchers believe that social media analytics presents a unique opportunity for businesses to treat the market as a 'conversation' between businesses and customers" (Chen, Chiang, and Storey 2012).

1.1.1.1 Data Overload

The ability to exploit the ever-growing amounts of business-related data will allow to comprehend what is emerging in the world. In this context, Big Data (BD) is one of the current major buzzwords (Dumbill 2012). There is a notable consensus among the research communities that the traditional tools used for collecting, storing and analysing BD are no longer able to cope efficiently with such massive amounts of generated data (Marz and Warren 2015). Advanced, unconventional and adaptable analytics are required to address the challenges of managing and analysing a wide variety of BD islands (Emrouznejad 2016) which are expanding exponentially as a result of the huge amount of data being generated by tracking sensors, social media, transaction records, and metadata to name just a few of the many sources of data. For example, every one second there occur 7,630 new Tweets, 41,644 GB usage of internet traffic, 58,528 Google searches, etc . Furthermore, as depicted in Figure 1-1, the digital universe data is expected to grow by a factor of 10 by 2020s (Gantz and Reinsel 2012).

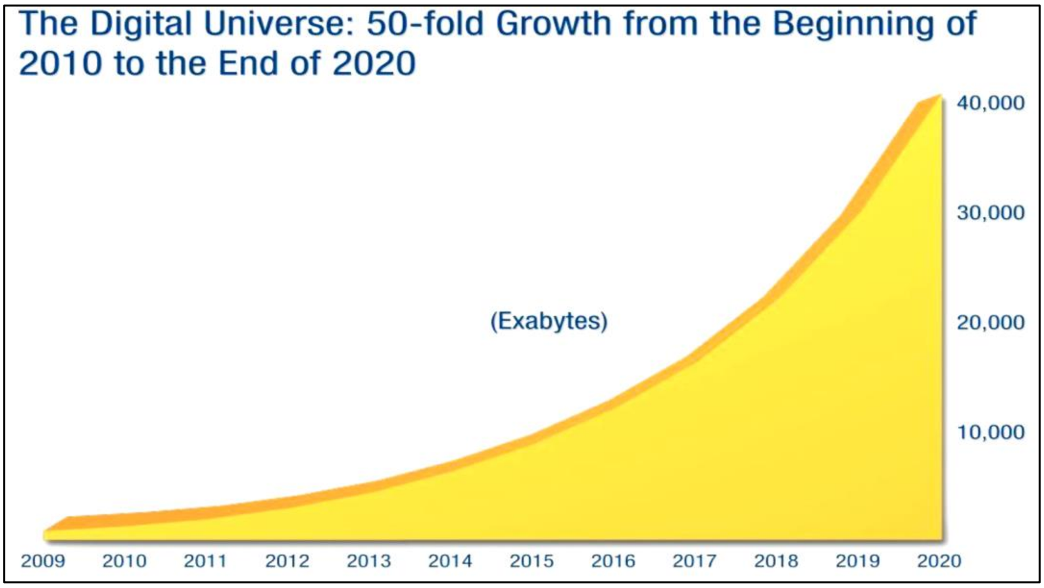


Figure 1-1: Digital Growth[1]

These massive amounts of generated data have implications for businesses and how they operate (Hudy 2015, Lammerant and De Hert 2016). In particular, business firms have over-indulged in providing information which stimulates the decision makers to adapt to the rapid increase in data volume. However, amongst the 85 percent of those that have aspired to become data-driven companies, only 37

[1] Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

percent have proven to be successful (Partners 2017). The information overload[2] (a.k.a. infoxication (Chamorro-Premuzic 2014) or infobesity (Rogers, Puryear, and Root 2013)) has made the process of decision making much more difficult. This has led to the need for a close scrutiny of internal business processes, and a review of the tools used to collect, transfer, store and analyse the flood of data generated by a company's internal and external data sources.

1.1.1.2 More Unstructured Data and Less Structured Data

Data is everywhere, presented in various formats, and is collected from many heterogeneous resources. Business Intelligence (BI) applications are more focused on structured data and support decision-makers by providing meaningful information from extracted data mainly coming from day-to-day operational information systems and structured external data sources (Chaudhuri, Dayal, and Narasayya 2011a). With the rapid increase in the amount of unstructured data, traditional data warehouses cannot be the sole source of data analytics. If 20 percent of data available in an organisation is mainly structured data (Gantz and Reinsel 2010), the unstructured data accounts for 80 percent of the total data that the organisation encounters. Examples of unstructured data include free text, emails, images, audio files, streaming videos, and many other data types (Joa et al. 2012).

The size of large data can range from several terabytes to petabytes and even exabytes (Tien 2013). Thus, there is an inevitable need to develop platforms that deal with this large scale of data. This has driven many companies to examine their current information and communication technology infrastructure, to strengthen the expertise

[2] The term "Information overload" has been popularized by Alvin Toffler in his book, Future Shock (Toffler 1971),

of their employees, and to prepare them to benefit from the systems of sophisticated unstructured data analysis correctly and effectively (Kitchin 2014).

1.1.2 Credibility of Data

The changing role of online users from information consumers to information producers has caused a noticeable variance in the quality of published content. In fact, quality of content is considered as a key difference between the content generated before and after the revolution of the Web 2.0 (Agichtein et al. 2008). In this context, OSNs have been extensively used as a powerful tool to promote diffusion of information in several domains (Stieglitz and Dang-Xuan 2013). Given such an impact, an understanding and comprehension of the content of OSNs has been an essential interest of various research avenues (Guille et al. 2013). In particular, identifying, reviewing, inferring and interpreting reputable social content consume a significant amount of time and effort (Chang, Diaz, and Hung 2015), yet have attracted wide interest due to the significance of obtaining and applying high quality content in many disciplines such as politics (Johnson and Kaye 2014), e-commerce (Hajli 2014), e-learning (Akbari et al. 2016), and health care (Grajales III et al. 2014).

As discussed previously, data are no longer generated only by transactional/structured and limited external sources; the global environment is now producing data in the form of news, economic factors etc., and from Voice of the Market (VoM) (Johne 1994) and Voice of Customer (VoC) (Griffin and Hauser 1993) through social networks, web blogs, etc. However, all external data sources do not have the same level of reputation. Data-users rely on reliable, reputable, and high-quality data and data sources. Likewise, unreliable and/or inaccurate data, such as data generated by suspicious and untrustworthy sources negatively impact on a company's operations and the decision making (Immonen, Paakkonen, and Ovaska 2015).

The quality of the data, which depends on whether it is collected from a reputable or an untrustworthy source, affects the quality of the perceived knowledge. For instance, in dramatic natural disasters such as the earthquake in Haiti and the tsunami in Japan, people used OSNs to report injury, share urgent and vital information, report damage, and provide firsthand observations (Castillo, Mendoza, and Poblete 2011b, Abbasi et al. 2012, Alexander 2014, Ghahremanlou, Sherchan, and Thom 2014, Yin et al. 2012). However, while OSNs provide platforms for legitimate and genuine users, they also enable spammers and other untrustworthy users to publish and spread their content, taking advantage of the open environment and fewer restrictions which these platforms facilitate. This might lead some users to abuse OSNs platforms and hijack events such as emergency situations by spreading rumours, and false and misleading information (Kumar and Geethakumari 2014). Hence, studying users' behaviour in OSNs will lead to a better understanding of their published content. The users' behaviour comprises several social activities such as establishing new friendships, posting new content or replying to another user's content, messaging, browsing and discovery (Jin et al. 2013). Furthermore, an analysis of the users' behaviour helps to determine and understand users' main topic(s) of interest (Bhattacharya et al. 2014), to mine their sentiments (Colace et al. 2015), and to know their needs and demands.

1.1.3 Domain of Interest in Social Data

Many individuals use OSNs to seek and connect with like-minded people. This homophily results in building homogenous personal networks in term of behaviours, interests, feelings, etc.(McPherson, Smith-Lovin, and Cook 2001). In particular, OSNs provide a medium for content makers to express and share their thoughts, beliefs, and domains of interest. This gives individuals access to a wider audience which positively affects their social status and would assist them to obtain, for instance, political support (Rainie and Wellman 2012). Therefore, the cornerstone of the users' online social profiles is an accurate understanding of their domains of interest.

The domain of knowledge is a particular area of people's work, expertise, or specialisation within the scope of subject-matter knowledge (e.g. Sports, Politics, Information Technology, Education, Art and Entertainments, etc.). (Hjørland and Albrechtsen 1995a). In OSNs, the domains of interest can be determined at the user level and at the post level. In other words, the overall published content of the user is analysed, and the domain(s) of interest is inferred. Likewise, the user's posts can be analysed separately to extract the domain(s) of each post. The factual grasp of the users' domain(s) of interest facilitates understanding the domain(s) conveyed from a short text message such as a *tweet*.

1.1.4 Social Data Classification

The rapid growth of enterprise needs correlated with such an increase in the volume of modern data repositories on the one hand, and the nature of the data that can be stored on the other hand, have made traditional statistical methods inadequate to meet all data analysis requirements. This has necessitated the development of advanced data analytics to extract useful knowledge from such a vast volume of data.

In the light of the general perception of the advanced data analytics, a question arises about the benefits that some organisations can acquire from adopting these techniques. One of the professional sectors that has started to benefit from this notion is healthcare (Koh and Tan 2011, Obenshain 2004). With the increase in electronic health records, health care providers and researchers can mine the immense stores of data to detect previously unknown cognitive patterns and then use this information to build predictive models to improve diagnosis and health care outcomes.

In this context, companies incorporate advanced social data analytics to build effective marketing strategies by leveraging the interactivity enabled by OSNs. Thus, to create the required interaction with their customers, companies use many modern means of communication to attract customers and visitors to their online social platforms. Consequently, it is necessary for companies to analyse the customers' social content and classify the customers into appropriate categories, then deliver the right message to the right category. Segmentation (Wu and Lin 2005) is the first step towards effective marketing, and is intended to classify customers according to their interests, needs, geographical locations, purchasing habits, lifestyle, financial status and level of brand interaction. If companies succeed in building effective clusters of customers and determining the basic criteria for each cluster in making their buying decisions, companies will be able to establish goals and take appropriate actions to achieve them. For example, companies can identify the most optimal products/services captured for each segment of customers. This fine-grained analysis can maximise customer satisfaction as companies can then design and manufacture not only one standard product, but several segment-oriented products.

Deep insights into SBD create a substantial difference, driven by the significance of extracting useful knowledge to benefit multiple applications. Accordingly, several related issues have motivated this research.

1.2 Motivation of Study

Since the emergence of OSNs, the propagation of social data has encouraged researchers to develop state-of-the-art techniques for social data analytics. Given the unstructured and uncertain nature of massive social data, understanding the customers' needs and responding to their enquiries, comments, feedback or complaints is a major purpose of any business firm. However, it is not easy to accomplish all these customer-centric tasks. Hence, this research is motivated by the need to address the following BD-related issues that have emerged.

1.2.1 Poor Quality Data

The dissemination of information via the World Wide Web is no longer a monopoly; anyone can be a producer and a publisher of information. Given the abundance of data, it is hard for those who receive such quantities of information to distinguish the accurate from the inaccurate, the good from the poor. With the development of OSNs, and the huge amount of social data that is generated, the quality of social data has not improved; rather, all types of false information have permeated these platforms. Rumours, for instance, are an example of bad quality data (Castillo, Mendoza, and Poblete 2011a). Rumours are a negative social phenomenon that is prevalent in societies. It is also one of the most serious psychological and moral wars that are raging in an atmosphere charged with various economic, political and social factors (Resnick et al. 2014). Spam is another category of low-quality content. Social spam content such as fake accounts, bulk messaging (*sending the same post many times in a relatively short period of time*), malicious links, fake reviews, etc. have degraded the quality of experience obtained by the social community members

(Lee, Caverlee, and Webb 2010).

Data credibility varies according to the reputation of the data producer. For example, in OSNs, all users' posts do not have the same level of reputation; a tweet from a verified user who has established a broad audience of followers has more impact than a tweet from a new user or a user with a small number of followers. Producers of bad quality social data provide their content via text, sound, image, and video which allow them to proliferate, especially since they can do so with anonymity and impunity. Due to the huge amount of information flowing to its recipients, in conjunction with the lack of a gatekeeper for those sites, it is difficult to verify the content, thereby making it easier for others to perform the task of disinformation (Hermida et al. 2012). Thus, OSNs are hijacked, and their otherwise useful tools are misused to create chaos and spread false news, and to undermine intellectual convictions, ideological constants, and moral and social factors that could cause confusion within the community. This has become a threat to social security and social harmony as a result of the absolute freedom guaranteed by those sites (Mendoza, Poblete, and Castillo 2010);(Papadopoulos et al. 2016, Zhao et al. 2016, Ito et al. 2015).

On the other hand, the good quality content obtained from SBD has several significant impacts (Agichtein et al. 2008). The use of social media is an empowering force in the hands of the public and private sectors and can have a positive influence on a community's development. It is an important tool to spread (public health) awareness, ensure security, and improve social and economic practices. OSNs consolidate and strengthen relationships between the users by sharing factual information and exchanging views on a variety of topics. This gives individuals considerable experience in many domains, in addition to enabling them to acquire knowledge and skills. Furthermore, the extraction and examination of quality content can benefit several vital sectors. For example, high-quality social data leads to a better understanding of customer behaviour and keeps a company's audience updated with the latest developments which improve customers' experience and increases revenues (Shenoy and Prabhu 2016). Last but not least, the quality of data influences the decision-making process of business operators (Chengalur-Smith, Ballou, and Pazer 1999, Janssen, van der Voort, and Wahyudi 2017).

1.2.2 Unrelated and Ambiguous Data

The results of the MIT survey (Lavalle et al. 2011) demonstrated the importance of data analytics as a means of elevating a company's position. Respondents indicated that they use data analytics to drive future strategies and day-to-day operations. However, some companies may be under the misconception that because there is data, this in itself gives them fruitful results. In fact, the secret lies not in data collection, as decision makers may be swamped with more data, but in the acquisition of data that is relevant and meaningful (Schmidt, Galar, and Wang 2016). In particular, a company must first decide which information in the collected data is essential in terms of, for example, its strategic objectives. Then, data analysis makes the difference and achieves the anticipated outcomes. These endeavours are fortified by collecting and utilizing the massive amount of data generated by OSNs.

However, a major issue needs to be addressed to enhance the understanding of the contextual content, in order to maximise the benefits derived from the textual content of SBD. Due to the unstructured and ambiguous nature of social data, the reformatting of the data in a meaningful structure is a key challenge (Fu et al. 2012).

Although unstructured data is accessible and easily understood by a human, an adequate understanding of its contextual content using a machine is an arduous task due to its rapid propagation, heterogeneity, and ambiguity (Poria et al. 2015). As in linguistics, a single term can have different meanings in different contexts. Hence, it is essential that the textual content of social data be interpreted accurately in order to establish a clear and transparent interaction between the organisation and its customers through the means provided by the OSNs.

1.2.3 A Need for Domain-based Classification

As stated previously, users of OSNs are keen to establish strong relationships with others; they search for and connect with relevant content or users. Hence, to open dialogues between like-minded people so as to share opinions, and life experiences, the first step is to understand the users' domains of interest automatically. This is achieved through the analysis of users' content and determining all the domains that have been discussed among users. This allows the segmenting and searching of users according to their domains of interest (Michelson and Macskassy 2010), and this is motivated by its significance in a broad range of applications such as personalized recommendation systems (Silva et al. 2013), opinion analysis (Liu and Zhang 2012), expertise retrieval (Balog et al. 2012), and computational advertising (Yin et al. 2015). However, there are two main issues that have motivated the efforts to create domain-based classifications; first, it is not an easy task to analyse textual content due to the diversity of linguistics which makes it difficult to understand users' domains of knowledge; second, what is/are the appropriate technique(s) that should be incorporated to perform the classification task?

1.3 List of Concerns Need to be addressed by an Approach for SBD Analytics

This research is carried out to address several concerns that have emerged with the proliferation of SBD. An approach for SBD analytics is designed to address the following specific concerns:

- Veracity and Value of SBD which is discussed in Section 1.4.1
- An accurate Understanding of SBD which is discussed in Section 1.4.2
- Domain-based Classification of SBD which is discussed in Section 1.4.3

The following sub-sections address the above concerns in detail.

1.3.1 Veracity and Value of SBD

SBD analytics provides advanced technical capabilities to the process of analysing massive and extensive social data to achieve in-depth insights in an efficient and scalable manner. With the enormous increase in the volume and diversity of data that businesses are dealing with today, they find themselves at a crossroads, having to decide whether to ignore these data, or gradually start to adapt, understand and benefit from them (Katal, Wazid, and Goudar 2013, Schubmehl and Vesset 2014). Hence, to efficiently incorporate data analytics to benefit organisations, customers' data should be collected directly (i.e. internal operations collected from day-to-day transactional information systems) and indirectly (i.e. social data collected from OSN platforms). The data collected from OSNs should be examined to determine the high-quality content and to eliminate the poor-quality data from further data analysis. To obtain this objective, organisations should understand how to handle the propagation and the heterogeneity of BD. Thus, they should address the following two main concerns: (i) BD are huge in volume, of great variety, and spread rapidly; (ii) extracting valuable and accurate data is a key challenge. Hence, ***there is a need for an approach that can produce data analytics capable of handling BD features and effective in filtering out unsolicited data and inferring a value***. This approach should comprise an advanced technical solution able to capture huge amounts of generated data, scrutinise the collected data to eliminate unwanted data, measure the quality of the inferred data, and transform the amended data for further data analysis.

1.3.2 An Accurate Understanding of SBD

It is evident that unstructured data is produced exponentially. This necessitates further efforts to absorb such datasets and understand their contextual content. Textual content (a.k.a. natural language text) is considered the largest amongst all sources of information (Gupta and Lehal 2009). The wealth of free-form textual, social data has attracted researchers' attention and prompted them to find ways to discover knowledge hidden in the textual content. This problem has been addressed to some extent through the emergence of text mining technology, an extension of data mining, that is used to detect rules, patterns and trends from textual data such as tweets, HTML web pages, instant messages and emails (Feldman and Sanger 2007, He, Zha, and Li 2013). Natural language text is very ambiguous, and this is evident particularly when it comes to the continuous occurrences of the named entities. Hence, indicating and inferring key entities such as persons' names and professions, locations of cities and countries, products, companies, specialized terms etc. from text can significantly enhance several business processes and techniques such as knowledge base population, topics distillation, keyword search, and information integration (Shen, Wang, and Han 2015). Therefore, ***there is a need to develop and implement an approach to derive knowledge from SBD***. This approach will improve the overall comprehension of the processed textual datasets, and deliver knowledge in the form of unambiguous results by providing metadata which helps relevant data to be accurately interpreted and understood.

1.3.3 Domain-based Classification of SBD

In this fast-paced world, companies that are able to turn data into valuable opportunities are those that implement best practices and maintain prime competitive positions, while the ones that deal superficially with data will lag behind (Tole 2013, Labrinidis and Jagadish 2012, Schubmehl and Vesset 2014). In particular, classification of SBD enables companies to identify and cluster their customers and set specific values for each category based on its social impact (Orenga-Rogla and Chalmeta 2016). This, for example, helps companies to launch effective marketing campaigns addressing specific groups and designing programs to maintain customers' loyalty by fulfilling their desires and needs (Khobzi and Teimourpour 2015, Helal et al. 2016). In particular, this is achieved by analysing the overall posted content of the user, as well as each individual post, which is the more difficult task. Thus, ***it is essential to develop an approach that can perform classification tasks at the user level and the post level***. The factual grasp of the domains of interest extracted at the user level or post level enhances the customer-to-business engagement. This necessitates an accurate analysis of customer reviews and their opinions in order to strengthen brand loyalty, improve customer service, and increase an organisation's awareness of issues that need to be addressed

1.4 Research Objectives

The preceding sections introduced the issues facing the efforts intended to achieve the hoped-for added value from the SBD revolution. The motivations for this research are listed and explained in terms of three major concerns (i.e. veracity and value of SBD, an accurate understanding of SBD, and domain-based classification of SBD), demonstrating the need to develop approaches that can address vital SBD issues. In response to these concerns, the principal objective of this research is to ***develop approaches for social data analytics that can effectively measure the trustworthiness of contributors to SBD, derive domain knowledge and infer value from SBD***. This research objective is achieved via the following sub-objectives:

Sub-objective 1: Develop a framework to deduce the value and determine the veracity and credibility of SBD. It is envisaged that this framework will address the primary challenging features of the general problem of BD by implementing a technical solution to address the problem of handling the massive amount of data, and to facilitate data storage and analysis.

Sub-objective 2: Develop a systematic approach to extract knowledge captured from the textual content of SBD. This approach enhances the understanding of users' domains of interest.

Sub-objective 3: Develop an approach incorporating comprehensively advanced knowledge discovery and data classification techniques for domain-based detection. The proposed framework will be able to perform dual classification tasks at the user level and the post level.

Sub-objective 4: Develop and refine a prototype implementation of the design to verify and validate the effectiveness and efficiency of the proposed approaches and the associated frameworks.

It is anticipated that this thesis will make several contributions to both theory and practice, as discussed in the next section.

1.5 Thesis Contributions

The contributions made by this thesis to the field of SBD analytics and to application-based research in general are as follows:

Contribution 1 *- A comprehensive literature survey of state-of-the-art approaches.*

In consideration of the overall aim and objective of this thesis, in Chapter 2, a critical review is presented of the current state-of-the-art approaches relevant to the key areas of the research topic. The comprehensive and intensive review of the literature provides the necessary theoretical background to the research area and frames this research by positioning it against past research. Further, it demonstrates the rationale for conducting this research by identifying the research gaps and establishing research questions to address them.

This comprehensive survey of state-of-the-art approaches clusters the research issues according to three key research SBD areas:

- Lack of advanced domain-based trustworthiness approaches: Most of the current efforts apply generic-based credibility evaluation approaches for users and their content in OSNs. There is a lack of domain-based credibility techniques which examine the users' domains of interest, study their behaviour over time, incorporate the sentiment analysis of their posts' replies, and address the key features of BD.

- Lack of approaches to manage and extract high-level domains from the textual content of SBD: Most of the current topic distillation and modelling approaches rely on bag-of-words techniques. These endeavours have several limitations including: (i) inability to consider the semantic relationships of terms in the user's textual content; (ii) ineffectiveness in applying topic modelling techniques to short text messages such as tweets; and (iii) the high-level topics classifications that use these bag-of-words statistical techniques are inadequate and inferior.

- Lack of domain-based techniques for dual classification: The current efforts undertaken to understand the contextual content of short text messages, and classify them into appropriate domains of knowledge, are still immature; hence, extensive research in this area is vital due to ambiguity and brevity of short text messages such as tweets.

Contribution 2 *– An effective credibility framework for users of OSNs that addresses the main features of BD, and incorporates semantic analysis and the temporal factor.*

Chapter 4 discusses in detail the second contribution of this thesis. It presents CredSaT (***Cre****dibility incorporating* ***S****emantic* ***a****nalysis and* ***T****emporal factor*): a comprehensive SBD framework to measure users' credibility in domains of knowledge. The crawled datasets of user data and metadata are divided into several chunks where each chunk represents a specific period. A metric of key credibility attributes is incorporated to evaluate the user's credibility in each particular chunk, thus providing an overall credibility value. The mechanism used to calculate a user's value in each step takes into account the values of other users, thereby providing a normalisation approach for establishing a ranking list of credibility in each domain.

CredSaT consists of several related sub-contributions summarised as follows:

- A novel discriminating measurement for users in a set of knowledge domains is provided and demonstrated. Domains are extracted from the user's content using semantic analysis.
- A metric incorporating a number of attributes extracted from content/user analysis is consolidated and formulated to obtain the level of trustworthiness. A holistic trustworthiness approach is provided based on three main dimensions: (i) distinguishing OSNs' users in the set of their domains of knowledge; (ii) feature analysis of users and their contents; and (iii) time-aware trustworthiness evaluation.
- A distributed data processing solution is developed to facilitate data storage and trustworthiness evaluation.

- CredSaT is benchmarked against well-known baseline techniques over a curated labelled dataset. It outperforms other methods in all evaluation metrics. Additional experiments are conducted to evaluate CredSaT which validate the applicability and effectiveness of identifying highly domain-based trustworthy users. CredSaT further shows the capacity to capture spammers and other anomalous and untrustworthy users.

Contribution 3 *– A systematic approach to extract knowledge captured from the textual content of SBD.*

Another key contribution of this thesis is the development of a consolidated approach aimed to resolve the data ambiguity problem in the context of SBD. Through five integrated steps, the proposed approach has proven successful in enhancing semantic information extraction and enrichment of the textual data, thereby providing an adequate interpretation of the textual content of social data.

Domain ontology and semantic web tools facilitate the building of conceptual hierarchies and the process of populating the domain ontology with instances extracted from user tweets. Hence, an ontology-based approach as a means of extracting the semantics of textual data is integrated with different vocabularies and semantic repositories to enrich the semantic description of resources using an annotation component. In addition to ontology and vocabulary reuse, interlinking that includes the semantic relationship between similar entities stored in other datasets has been implemented in the developed framework.

Domain knowledge has been captured in ontologies which are then used to enrich the semantics of data with specific semantics conceptual representation of entities. The reuse of ontology and the interlinking of classes, entities and concepts with relevant entities from other repositories facilitates the interoperability of information. Therefore, the existing ontologies have been incorporated and reused.

The proposed approach is tested and evaluated with public datasets collected from Twitter and within the political domain. The results of experiments demonstrated that the developed approach outperforms another well-known semantic-based tool regarding quantity and accuracy of identified semantic entities and concepts. The findings indicate that by combining the proposed approach with other semantic repositories and APIs, the accuracy of concept identification is improved significantly. The resultant semantic data are processed and expressed as linked data and stored in RDF triples in the semantic-based repository database. The data are semantically represented under well-controlled vocabularies, useful taxonomical relationships, and with ontologies for inference of domain knowledge.

Contribution 4 *– An integrated framework incorporating domain knowledge discovery and machine-learning-based data classification techniques as a means of domain-based discovery, classification and prediction.*

The fourth contribution in this thesis fills a gap in the existing literature by presenting a consolidated framework for Twitter mining that aims to address the shortcomings of the current state-of-the-art approaches to topic distillation and domain discovery. As explained in Chapter 5, the proposed approach comprises two main analysis phases incorporating several semantic analysis tools and machine learning modules. In the first phase, the users' historical tweets are collected; their interests are examined over time, thereby providing a prediction of the users' interests, taking the temporal factor into consideration. In the second phase, the outcome of the previous analysis is used as a primary input to forecast the domain of future tweet content. Well-known machine learning classifiers are used for user classification. A comparison is conducted to benchmark the performance of the incorporated machine learning modules.

The main sub-contributions of Chapter 6 are as follows:

- A time-aware framework incorporating comprehensive knowledge discovery tools and well-known machine learning algorithms is proposed for domain-based discovery, which applies to the Twittersphere platform and is customisable to other OSNs.
- The proposed framework can perform classification tasks at the user level and tweet level.
- Unlike current statistical-based topic distillation techniques which retrieve search results without considering the temporal dimension, the proposed approach is better able to address the temporal factor; users' knowledge evolves, and their interests might be diverted elsewhere depending on their experience, work, study, or other factors.
- Unlike current unsupervised statistical approaches, the proposed approach incorporates supervised machine learning techniques to perform domain-based classification task for the already semantically-enriched temporally-segmented textual content.
- The conducted experiments using the Twitter platform as one of the dominant OSNs verify the effectiveness and applicability of our model as evident in the outstanding results of several performance evaluation

metrics.

The overall contribution of Chapter 6 provides the essential groundwork for a better understanding of user interest in several domains of knowledge. This is achieved by incorporating domain-based ontologies and semantic web analysis fortified with state-of-the-art machine learning techniques to achieve better familiarity with user interests. This facilitates the process of measuring user credibility in each domain of knowledge.

1.6 The significance of the Research

The BD revolution is relentlessly changing every part of society. Its horizon extends to comprise the digital trace originating every second from information-sensing Internet of Everything devices, RFID readers, huge metadata (e.g. trust, security, and privacy), the digitalization of business artifacts (e.g. files, documents, reports, and receipts), and last but not least, the data generated by social media. Therefore, understanding and analysing the semantics of BD is a goal of enterprises today. This provides an unprecedented scope for understanding our society and improving the way we live and conduct business.

This thesis contributes to the ongoing efforts to achieve the hoped-for added value from the BD revolution. In particular, due to the prominent role of OSNs in the lives of individuals and communities, the approaches developed in this research are primarily intended to address the various research issues arising from the propagation of SBD. The research impact in terms of its theoretical and practical significance is explained in the following sections.

1.6.1 Theoretical Significance

This research has theoretically constructed three main artifacts as follows:

- The development of a novel domain-based trustworthiness inference module for SBD contributes to the design theory of trust inference and evaluation methods. The mechanism followed in the developed approach incorporates several theoretically proven elements of Information Retrieval, Sentiment Analysis, Knowledge Engineering, etc.
- The development of an innovative approach for domain-based data extraction in SBD has proven effective for semantically annotating and enriching social data, thereby obtaining an accurate understanding of its textual content. This approach can be used as a guideline for other enriching textual data approaches to improve and enhance semantic analysis processes.
- The development of a new and advanced well-tested framework for Twitter mining contributes to the social data analysis design frameworks, and classification tools and techniques. The developed approach demonstrates its innovativeness through integrating semantic-based domain discovery tools and statistical-based machine learning algorithms to establish an effective groundwork for the related research theories and original technical methods.

1.6.2 Practical Significance

This research has practical significance and contributes to important business-related operations as follows:

- The continuous investment in the world of BD is widespread among several firms with dissimilar sizes, and operating in a variety of business sectors; 75 percent of the 400 companies surveyed by Gartner in 2015 reported that their BD investment had either begun or was anticipated to start in forthcoming years (Gartner 2015). Hence, organisations have had to become more decisive in regard to coping with the enormous influence of SBD in many aspects of business practices. This highlights the need to review the tools used to collect, transfer, store, and analyse such massive amounts of data.
- Another survey carried out by Gartner (Sallam et al. 2017) indicated that by 2020 business firms will continue to make significant investments in social and big data analytics, and will provide platforms for users to access curated and credible data collected from internal and external data sources. In addition, a Deloitte survey (Phillipps 2013) emphasized the importance of data analytics in the business domain; around seventy-five percent of all respondents believed that the continuous propagation of data will benefit their business strategies. Ninety-six percent of the respondents consider data analytics as an added value for their businesses in the coming three years.
- This thesis proposes practical frameworks that have widely significant implications for a variety of business-related applications such as the VoC/VoM, recommendation systems, the

discovery of domain-based influencers, and opinion mining through tracking and simulation. In particular, the accurate understanding of the domains of interest extracted at the user level or post level strengthens the engagement between businesses and their current and prospective customers. This contributes to an accurate analysis of indirect customer feedback that comprises social listening to customer reviews and opinions to improve brand loyalty, customer service, customer care interactions, etc.

- The developed approaches include techniques for inferring trustworthy social data, semantically enriching textual data, and domain-based classification and prediction. The practical impact of these techniques strengthens the ongoing efforts to develop technical solutions for social data analytics. Hence, organisations can utilise the proposed techniques and further develop more sophisticated systems to improve their internal business processes by incorporating information from a greater range of external data sources.

1.7 Thesis Structure

The structure of this thesis is outlined as follows:

Chapter 1 -- Introduction

This chapter provides an introduction to the issues that have emerged with the propagation of SBD. The chapter explains the motivation for this study and lists the concerns that need to be addressed. Hence, the problems confronting the era of SBD are identified, followed by the key objectives of this thesis and the methodology adopted to achieve them.

Chapter 2 -- Literature Review

Chapter 2 provides an in-depth analysis of the current state-of-the-art research in the areas of social trust, semantic analysis, and domain classification. The undertaken efforts in each research venue has been discussed. The chapter concludes with the identified research gaps that have motivated the research activities of this thesis.

Chapter 3-- Problem Definition

Chapter 3 presents an overview of the research problems addressed in this thesis, together with the key concepts and their definitions. The research issues are examined, and indicate the need to develop approaches and a framework for social data analytics that can measure the trustworthiness of users in SBD, derive domain knowledge, and infer value from SBD. The DSR methodology implemented in this thesis is described with the set of key activities undertaken throughout the research. Finally, the solution requirements to the research issues are explained.

Chapter 4 -- Domain-Based Analysis of Users' Trustworthiness in Social Big Data

This chapter presents an advanced domain-based credibility framework incorporating semantic analysis and the temporal factor to measure and rank the credibility of users in SBD. Through the theoretically proven techniques and the conducted experiments, the proposed framework has proven successful in fine-grained trustworthiness analysis of users and their domains of interest in OSNs using BD infrastructure.

Chapter 5 -- Semantic Data Extraction from Social Big Data

This chapter provides an applicable and effective framework for domain knowledge extraction by means of an ontology-based approach. Results of experiments indicate that the proposed mechanism can outperform a well-known semantic provider.

Chapter 6 -- Ontology-based Domain Discovery Incorporating Machine Learning

Chapter 6 presents an approach to mitigate ambiguity and provides domain classification to the textual content at the user and tweet levels. The approach incorporates external semantic web knowledge bases and machine learning modules. Its applicability and effectiveness are demonstrated in the chapter through experiments and benchmark comparisons.

Chapter 7 -- Recapitulation and Future Work

Chapter 7 revisits the research issues addressed in this study and concludes with the set of enhancements proposed to address the limitations of previous approaches and makes several recommendations for future research directions.

1.8 Conclusion

The abundance of SBD gives organisations the opportunity to maximise their use of a wealth of available information to increase their revenues. Hence, there is an urgent need to capture, load, store, process, analyse, transform, interpret, and visualise a diversity of social datasets to develop meaningful insights that are specific to an application's domain.

This chapter introduced the issues and trends related to the emergence of SBD. The necessity to tackle the constant growth of social data has motivated this research to develop approaches capable of mining enormous amounts of data in order to extract high quality, relevant, and meaningful information. In particular, this research is motivated to untangle certain associated issues abstracted as follows:

- Poor quality data: the rapid propagation of social data, in conjunction with the lack of a gatekeeper for the OSNs, raise a concern with regards to the quality of the data being generated from these platforms.

- Unrelated and ambiguous data: the unstructured and ambiguous nature of social data prevent the data from being reformatted in a meaningful structure. In particular, obtaining an adequate understanding of the contextual content of the social data using machine-based techniques is an onerous task due to the rapid propagation, heterogeneity, and ambiguity of social data.

- A need for domain-based classification: analysing the textual content of social data is not an easy task due to the linguistics diversity which makes it more difficult to understand users' domains of knowledge.

The key concerns and problems which comprise the scope of the proposed approaches are articulated. These concerns are reiterated as follows:

- Veracity and value of SBD: there is a need for a data analytics approach capable of handling the BD features and that effectively measures social data credibility, filter out unsolicited data and infer a value.

- Ann accurate understanding of SBD: there is a need to implement an approach to derive knowledge from SBD.

- Domain-based classification of SBD: it is essential to develop an approach that can perform classification tasks at the user level and the post level.

This thesis has made the following contributions:

- A comprehensive literature survey of state-of-the-art approaches.

- The development of an effective credibility framework for users of OSNs, addressing the main features of BD and incorporating semantic analysis and the temporal factor.

- The development of a systematic approach to extract knowledge captured from the textual content of SBD.

- The development of an integrated framework incorporating domain knowledge discovery and machine-learning-based data classification techniques in the quest for domain-based discovery, classification and prediction.

As discussed in Section 1.6, this thesis describes the key contributions made by this thesis to the body of scientific knowledge in terms of its theoretical and practical significance. This chapter concludes with an outline of the thesis structure.

Chapter 2 establishes the theoretical background of this thesis by reviewing the literature on state-of-the-art developments related to the research topic. In this review, various current approaches will be evaluated and recommendations will be made to bridge the research gap.

Chapter 2 Literature Review

2.1 Introduction

The dramatic increase in the social data impact - a testimony to our growing digital lifestyles - has taken on industries and activities ranging from marketing and advertising to intelligence gathering and political influence. Therefore, it is essential that BD be interpreted in order to bring new perspectives and to improve business practices, yet this revolution is still in its infancy. It is easy to assume that the social data revolution is about quantity, but social data is more than just volume. In fact, the implications of this revolution are far reaching as they involve building the data infrastructures required to effectively deal with the propagation of social data in order to infer the hoped-for added value. This has motivated the research communities to conduct in-depth investigations, provide solutions, and implement platforms enabling these datasets to benefit several applications.

This chapter presents a comprehensive survey of the state-of-the-art approaches drawn from academic works relevant to this thesis. The purpose of the review is to examine and evaluate the current approaches to Social Trust, Semantic Analysis, and Data Classification in the era of SBD. This chapter is structured as follows:

- In Section 2.2, a review is conducted of the literature pertaining to social trust comprising the generic and domain-based approaches, and concludes with an assessment of the current approaches used to determine trustworthiness.

- Section 2.3 presents an overview of the existing semantic analysis, machine-learning-based data classification, and topic discovery approaches for SBD, followed by an assessment of these techniques.

- An integrated overview of the existing approaches and techniques is presented in Section 2.4. This section provides a cohesive review of previous research efforts in the areas of trust, semantic analysis and machine learning in SBD.

2.2 SBD Incorporating Trust

In modern enterprises, social networks are used as part of the infrastructure for a number of emerging applications such as recommendation and reputation systems. In such applications, trust is one of the most important factors in the decision-making process. Sherchan et al. (Sherchan, Nepal, and Paris 2013) defined ‘trust’ as the measurement of confidence where a group of individuals or communities behave predictably. The significance of trust is evident in multiple disciplines such as computer science, sociology, and psychology. Trust evaluation in the social media environment is still immature; hence, extensive research is required in this area (Sherchan, Nepal, and Paris 2013). The current literature indicates that there have been ongoing efforts to improve the approaches used to measure, evaluate and quantify the trustworthiness values inferred from the users of OSNs and their content. These approaches can be divided into two main categories within the context of SBD: (i) solutions that address the problem of generic social trust; (ii) solutions that infer domain-driven social trust.

2.2.1 Generic-based Trustworthiness Approaches

Generic-based trust approaches in OSNs are those frameworks, techniques, and tools developed to calculate and infer the trustworthiness values of users and/or their content with no consideration being given to the domain(s) of interest which can be extracted from the user level or post level. The trustworthiness of social media data is now a crucial consideration. With such a vast volume of data being interchanged within the social media environment, data credibility is a vital issue, especially regarding personal data (Passant et al. 2009). Several approaches have been proposed for measuring trustworthiness in social media (Weng et al. 2010, Silva et al. 2013, Yeniterzi and Callan 2014, Kwak et al. 2010, Tsolmon and Lee 2014, Agarwal and Bin 2013, Podobnik et al. 2012b, Sikdar et al. 2013, Wu, Arenas, and Gomez 2017) (Podobnik et al. 2012a) (Jeong, Seol, and Lee 2014). Nepal et al. (Nepal, Paris, and Bouguettaya 2013) addressed the challenges of the trust in the web-based social media. The paper states the four main components that are involved in the trust evaluation: service consumers, service providers, services and content. As indicated in their paper, the key challenges in the social trust context are: (i) how to assign a trust value for a new entity (user/post), (ii) trust propagation issue in OSNs; (iii) trust for recommender systems; (iv) addressing the phenomenon of misleading information in the social web environment (wrong content, personal identity and location).

Graph-based Social Trust: Podobnik et al. (Podobnik et al. 2012b) proposed a model that calculates trust between friends in a network graph based on weights of the edges between user's connected friends in Facebook. Agarwal and Bin (Agarwal and Bin 2013) suggested a methodology for measuring the trustworthiness of a social media user by using a heterogeneous graph in which each actor in the Twitter domain was presented as a vertex type in the graph. The level of trustworthiness was measured using a regressive spread process. The paper, on the other hand, neglects to consider the importance of a weighting scheme and the time factor. Each edge category should be assessed at different credibility levels; hence, a weighting scheme should be used. Trustworthiness values differ over time; consequently, the temporal/time factor should be integrated. (Yu et al. 2013) proposed a trust propagation scheme to predict a consumer's trust value in the service provider in service-oriented social networks, taking into consideration the structural properties of social networks and exploiting the association between degree distribution and trust distribution for the purpose of performance optimization. TweetCred (Gupta et al. 2014) is a Support Vector Machine credibility ranking inference framework for Twitter users in real-time data motion. Abbasie et al. (Abbasi and Liu 2013) presented an algorithm called CredRank to cluster users of the social media based on their online behaviour to detect the coordinated users. (Weitzel, de Oliveira, and Quaresma 2013) described a new methodology for determining the reputation of Twitter posts based on weighted social interaction. Jeong et al. (Jeong, Seol, and Lee 2014) discussed the perspectives of followers of a specific followee in the Twitter domain. The paper identified and classified three types of followers based on their feelings (supportive, non-supportive and neutral). Naumann (Naumann 2013) discussed the relationship between the message aim or intention of the user and his/her level of trust in the company. The paper addressed the question regarding the effect of the message intention on the trust variable in the domain of B2B companies. Kopton et al. (Kopton et al. 2013) explored functional magnetic resonance imaging (fMRI) as a means of evaluating trust in order to obtain a better understanding of users' behaviour on the OSNs; the idea is to study the brain activity of users when they are engaged with social platforms. (Cha et al. 2010) considered a number of attributes -indegree (#followers), retweets, and mentions- to measure users' trustworthiness. Brown et al. (Brown and Feng 2011) adopted the k-shell algorithm to measure users' influence. The algorithm takes a graph of the followers/following relationship as input and evaluates the k-shell level which indicates the users' ranking. Arlei et al. (Silva et al. 2013) examined the influence of social media users and the significance of their contents in information dissemination data. Tsolmon and Lee (Tsolmon and Lee 2014) proposed a means for measuring the trustworthiness of Twitter users. Parameters of the Following-Ratio (#follower/#following) and Retweet-Ratio (total number of retweets of user/total number of tweets) are utilised to infer well-known users, using the HITS Algorithm mechanism. Cutillo et al. (Cutillo, Molva, and Strufe 2009) presented a technique to confirm the privacy of the OSNs' users using a new method to handle certain security and privacy issues.

Trust for recommendation systems: In (Massa and Bhattacharjee 2004), the authors showed a web of trust as an alternative to the standard way of ranking a user, i.e. standard recommendation systems. Further, Gupta et al. (Gupta et al. 2013) presented the "WTF: Who To Follow" service which is being used as a recommendation system for the Twitter social network. This service is used mainly as a recommendation driver and has a significant impact; by using it, numerous new connections have been created. Further work by (Gallege et al. 2014b) and (Sun et al. 2014) proposed trusted-based recommendation techniques.

Trust incorporating Sentiment Analysis: The use of sentiment analysis techniques to analyse the content of OSNs has significantly influenced several aspects of research. In the context of social trust, authors of (Alahmadi and Zeng 2015) propose a recommendation system framework incorporating implicit trust between users and their emotions. AlRubaian et al. (AlRubaian et al. 2015) presented a multi-stage credibility framework for the assessment of microbloggers' content. The development of sentiment-based trustworthiness approaches for OSNs is discussed further in (Zhang et al. 2015, Bae and Lee 2012, Kawabe et al. 2015).

Incorporating BD Technology: Apache™ software foundation has designed Hadoop® (Shvachko et al. 2010), a Java-based open-source platform, to process large-scale datasets using a physical cluster of machines. In fact, Hadoop is not one tool per se; it is an ecosystem, and it is an infrastructure for several commercial and open source platforms developed to process BD in all stages of data analysis. It is difficult to track all Hadoop's related technologies. Thus, in this section, only the key components of the Hadoop ecosystem are described.

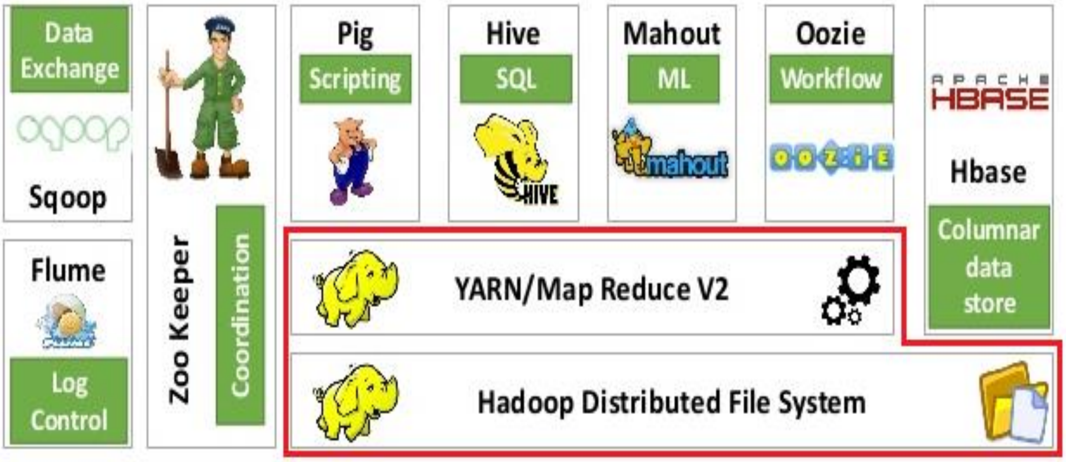


Figure 2-1: Hadoop ecosystem[1]

As depicted in Figure 2-1, Hadoop consists of several layers each of which is designed to perform certain tasks. These layers can be classified into two main categories: (i) *Hadoop core architectural blocks (surrounded by red)*: these are the

[1] Source: Apache Hadoop documentation.

major layers and components of the Hadoop infrastructure; (ii) *Hadoop supporting blocks*: comprise layers built at the top of the Hadoop core blocks to facilitate, monitor and manage data processing and storage. Table 2-1 lists each key component and its function.

Table 2-1: Hadoop Key Components

Cat.	Block Name	Usage
Hadoop Core Architectural Blocks	Hadoop Distributed File System (HDFS) (Shvachko et al. 2010)	A fault-tolerant storage unit that is the backbone data storage for a Hadoop cluster.
	MapReduce (Dean and Ghemawat 2008)	A tailor-made programming paradigm that is the processing core of Hadoop.
	YARN[2]	A large-scale distributed operating system designed to enhance data processing capabilities.
Hadoop Supporting Blocks (Management, Storage and Data Access)	Sqoop[3]	Facilitates the migration of data between the big data stores to the traditional databases/data warehouses.
	Flume[4]	Was designed to facilitate streaming and real-time processing by providing high reliable service to manage, collect and aggregate log data in large scale.
	Zookeeper[5]	Coordinates and manages tasks performed in the Hadoop distributed environment.
	Oozie[6]	Manages and schedules Apache Hadoop jobs.
	Mahoot[7]	A scalable framework built at the top of Hadoop to enable several machine learning tasks such as clustering, classification, etc.
	HBase™	Is the big data store. It provides a distributed and scalable data storing mechanism based on NoSQL(Not only SQL) notion. It is built on top of HDFS to offer fast data access.
	Hive (Thusoo et al. 2009)	Is an open source data model that enables users to build and execute SQL-like scripts.

[2] https://hadoop.apache.org/docs/r2.7.2/hadoop-yarn/hadoop-yarn-site/YARN.html
[3] http://sqoop.apache.org/
[4] http://flume.apache.org/
[5] http://zookeeper.apache.org/
[6] http://oozie.apache.org/
[7] http://mahout.apache.org/

Cat.	Block Name	Usage
	Pig Latin™	A module built at the top of Hadoop to provide high-level scripting language to facilitate data analysis.

Business Intelligence Incorporating Trust and Big Data Infrastructure: Business Intelligence applications are focused more on structured data; however, to understand and analyse the social trust, there is a need to collect data from various sources. Collective intelligence has spread to many different areas, with a particular focus on fields related to everyday life such as commerce, tourism, education and health, causing the size of the social Web to expand exponentially. SBD exhibits all the typical properties of big data: wide physical distribution, diversity of formats, non-standard data models, independently-managed and heterogeneous semantics. Labrinidis and Jagadish (Labrinidis and Jagadish 2012) summarised the challenges of BD as follows: 1. Data acquisition (infers useful data and discard irrelevant). 2. Building the right metadata for data description. 3. Data extraction and formation. 4. Data quality (value). 5. Automatic data analysis. 6. Coordination between traditional SQL with NoSQL methods.

The incorporation of BD technology to enhance data analysis tools is considered to be a hot topic, especially regarding the contents of social media because of its significance to data analytics. This has interestingly attracted researchers in industry and academia to leverage the BD techniques to benefit data analysis tools. The decision to incorporate BD technology (i.e. Hadoop/MapReduce) in trustworthiness social data analysis has been prompted by the huge size of social media content that needs an efficient and scalable technology to manage it so that the data volume issue is properly addressed. Moreover, recent literature has considered social networks as a form of BD in terms of volume (billions of social links), velocity (massive amount of generated content), and variety (videos, posts, mobile tweets, etc) (Paik et al. 2013). (Lim, Chen, and Chen 2013), (Cuzzocrea, Bellatreche, and Song 2013), (Shroff, Dey, and Agarwal 2013) and (Chen, Chiang, and Storey 2012) listed the main directions for BI over BD. Shroff et al. (Shroff, Dey, and Agarwal 2013) identified the importance of BD by highlighting the effect of its characteristics on the BI domain. BD in this context comprises the data derived from social networks which is unstructured and which BI tools are incapable of handling. In this context, authors of (Shroff, Dey, and Agarwal 2013) showed three use-cases where social contents dramatically affect business intelligence applications: Supply-Chain Disruptions, Voice of Customer and Competitive Intelligence. However, Trust and its impact in the socio-business analysis were not addressed by the paper. Authors of (Chen, Chiang, and Storey 2012) summed up the main areas related to BD and BI analysis; the paper lists the evolution of BI and Analytics, their application and the research opportunities. It implemented various research frameworks including BD analytics, text analytics, web analytics, network analytics and mobile analytics. Cuzzocrea et al. (Cuzzocrea, Bellatreche, and Song 2013) initiated the future research trends in the area of DW/OLAP and big data. The papers listed the main directions for the area of building and designing DW-OLAP over BD: 1. A methodology for designing OLAP that is capable of processing BD; 2. Efficient and complex paradigms to build OLAP's cubes over BD; and 3. Building semantically BD cubes. Lim et al. (Lim, Chen, and Chen 2013) listed the main directions for future research in the BI 2.0 in terms of BD, Text Analysis and Network analysis. Saha and Srivastava (Saha and Srivastava 2014) presented a summary to address the data veracity issue related to BD. Poor data quality has a major negative impact on the data analysis process, and the output will lack credibility and trustworthiness. The paper addressed the data quality issues and provided tools and solutions for various forms of data (relational, structured and semi-structured); however, the unstructured data types were not addressed. Moreover, hybrid approaches could be used that utilise ontology for data quality and trust inference purposes. The sentiment analysis of BD is now a hot topic. Khuc et al. (Khuc et al. 2012) proposed a methodology for sentiment analysis that incorporates BD technology (MapReduce/Hadoop) to process huge volumes of tweets. Although their solution addressed certain challenges, the issue of domain-based trust was omitted in their proposed approach; incorporating the notion of social trust will effectively increase the credibility of the sentiment extracted from tweets. To sum up, the research directions in the area of BD analytics include and are not limited to: incorporating BD technology (i.e. Map/Reduce, Hadoop) to benefit data analytics tools, developing methods to handle data in motion(real-time) for social data analytics and BI analysis, methodologies for designing OLAP tools innovatively to be capable to process SBD, and measures its credibility, and building semantic BD cubes. In addition, beginning with the characteristics of BD and sorting out issues related to these features will be the most efficient way to address BD and in addition will benefit the efforts of social data analytics and the expected outcomes of BD analysis.

2.2.2 Domain-based/Topic-Specific Trustworthiness Approaches

Adding a user-domain dimension when calculating trust in social media is an important step. This helps to enhance the understanding of users' interests. Hence, the notion of domain-based trust for the data extracted from the unstructured content (such as social media data) is significant. This is determined by calculating trustworthiness values which correspond to a particular user in a particular domain. The literature on trust in social media shows a lack of approaches for measuring domain-based trust. Several reviews have been carried out to highlight the importance of conducting a fine-grained trustworthiness analysis in the context of SBD (Momeni, Cardie, and Diakopoulos 2016, Sherchan, Nepal, and Paris 2013, Amalanathan and Anouncia 2016, Ruan and Durresi 2016). In particular, measuring the user's trustworthiness in each domain of knowledge is vital to acquiring a better understanding of users' behaviours in OSNs. The ontology represents the core of the domain where the knowledge is shared amongst different entities within the system that may include people or software agents (Chandrasekaran, Josephson, and Benjamins 1999). In this context, several efforts have been made to develop approaches for a fine-grain trustworthiness analysis (Embar et al. 2015a, Zhu, Su, and Kong 2015, Zhao et al. 2016, Lyu et al. 2015, Song, Li, and Zheng 2012, Abbasi and Liu 2013, Zhai et al. 2014) (Liu et al. 2013). An approach for microblogging ranking was proposed by Kuang et al. (Kuang et al. 2016). The authors incorporated three dimensions in their ranking technique (i.e. tweet popularity) the closeness between the tweet and the owner user, and the topics of interest. Zhao et al (Zhao et al. 2016) proposed a scalable trustworthiness inference module for Twitterers and their tweets that take into account the heterogeneous contextual properties. Another group of scholars have addressed the issue of influential users in OSNs (Brown and Feng 2011, Zhu, Su, and Kong 2015, Embar et al. 2015b, Zhou, Zhang, and Cheng 2014) (Pal et al. 2016). Authors of (Yeniterzi and Callan 2014) presented a method for discovering experts in topic-specific authority networks. They applied a modified version of the HITS Algorithm for more topic-specific network analysis. However, attributes such as (followers/following/friends counts, likes/favourites counts, etc.)

were not utilised to infer user reliability. Herzig et al. (Herzig, Mass, and Roitman 2014) presented an Author-Reader Influence (ARI) model that estimates a user content's attraction (i.e. content's uniqueness and relevance). In (Bozzon et al. 2013) the paper addressed the problem of selecting top-k expert users in the social group based on their knowledge about a given topic. In (Song, Li, and Zheng 2012), the authors built a model to discover popular topics by analysing users' relationships and their interests. Jiyeon and Sung-Hyon (Jang and Myaeng 2013) analysed the flow of information amongst users of social networks to discover "dedicators" who influence others by their ideas and specific topics. One of the top cited works in topic-based user ranking is Twitterrank (Weng et al. 2010). Authors of Twitterrank incorporate topic-sensitive PageRank to infer topic-specific influential users of Twitter.

2.2.3 Assessment of Approaches Incorporating Trust in SBD

The subsections above present a review of several key techniques and approaches implemented to define and formulate the trustworthiness of users and/or their content in OSNs. These approaches can be divided into two research streams. Firstly, there are the methods conducted to evaluate social trustworthiness in general. These approaches address the problem of social trust and credibility in the OSNs but do not take into consideration the users' domains of interest. Secondly, there are those methods that examine the textual content of users to infer their topic(s)/domain(s) of interest first, then determine their credibility in each topic/domain (Yu et al. 2013, Pal et al. 2016, Zhai et al. 2014, Zhao et al. 2016). These methods are advanced versions of the generic trust evaluation systems.

Two key techniques are used to construct the trustworthiness formula for both approaches, namely the feature-based techniques (Weerkamp and de Rijke 2012, Castillo, Mendoza, and Poblete 2011b, Duan et al. 2010, Gupta et al. 2014, Morris et al. 2012), and/or graph-based techniques (Yeniterzi and Callan 2014, Ravikumar et al. 2013, Abbasi and Liu 2013, Zhu, Su, and Kong 2015, Fiala 2012, Tsolmon and Lee 2014). The feature-based techniques measure the trustworthiness of users and their content by incorporating the list of key attributes which are associated with the metadata of the users and their content such as #followers, #friends, #likes/favourites #retweet/share etc. The graph-based approaches evaluate the trustworthiness of users and their content in OSNs by scrutinising their social connections, where the social trustworthiness values are propagated throughout the whole network of users. These techniques adopt graph propagation solutions such as PageRank, HITS Algorithm, etc.

Despite the considerable efforts conducted to resolve the social trust problem, there are still vital issues that need to be addressed to consolidate the proposed approaches. These include: tracking and monitoring users trustworthiness over time; improving the existing semantic analysis techniques in order to further enhance the contextual understanding of the users' textual content; further incorporation of the sentiment analysis methods in order to 'hear' the voices of the user's followers and their opinions of him/her. Last, but no less important, the implementation of the proposed approaches should address the key features of BD and provide technical solutions for handling the massive amount of data being generated steadily and incessantly from online social platforms.

2.3 SBD Incorporating Semantic Analysis, and Machine Learning for Data Classification and Topic Distillation

In the latter part of the 20th century, researchers in the field of Artificial Intelligence (AI) became active in the computational modelling of statistical analysis techniques and the defining of ontologies that would deliver automated reasoning capabilities. This section reviews the two main categories of these existing approaches: (i) techniques incorporating semantic analysis for domain discovery; (ii) machine learning statistical techniques for classification of data and the identification of topics.

2.3.1 SBD Incorporating Semantic Analysis

The Semantic Web (SW) was introduced by Berners Lee who envisaged the next web where data would be given semantic meanings via data annotation and manipulation in a machine-readable format (Berners-Lee and Hendler 2001). Ontology, based on Gruber's (Gruber 1995) definition, is the formal explicit specification of a shared conceptualization within a domain, as a form of concepts and relationships between these concepts which is used to describe the domain. By incorporating semantic analysis, semantic data can be inferred from social media data. Ontology has been widely applied in social media as a means of inferring semantic data in a broad range of applications. De Nart et al. (De Nart et al. 2016) proposed a content-based approach to extract the main topics from the tweets. This approach was an attempt to understand the research communities' activities and their emerging trends. Chianese et al. (Chianese, Marulli, and Piccialli 2016) proposed a data-driven and ontology-based approach to identify cultural heritage key performance indicators as expressed by social network users. This approach can be used in different domains but is only relevant to or ad-hoc in user domains. Michelson and Macskassy (Michelson and Macskassy 2010) used the DBpedia knowledge base to annotate entities in users' tweets, and extract the users' main interests by using the categories proposed in Wikipedia. Wikipedia as a knowledge base repository has been utilised for topic discovery in (Schonhofen 2006, Hassan, Karray, and Kamel 2012). (Carrasco et al. 2014) presented an ontology-based, multi-agent solution for the wild animal traffic problem in Brazil. (Iwanaga et al. 2011b) and (Ghahremanlou, Sherchan, and Thom 2014) both applied ontology to build applications in crisis situations. The former designed ontology for earthquake evacuation to help people find evacuation centres in an earthquake crisis based on data posted on Twitter. The latter showed a geo-tagger that processes unstructured content and infers locations with the help of existing ontologies. (Bontcheva and Rout 2012) conducted a survey that addressed research issues related to processing social media streams using semantic analysis. Some of the key questions which were the focus of this paper included: (i) How could Ontologies be utilised with Web of Data for semantically annotating social media contents? (ii) How could the annotation process discover hidden semantics in social media? (iii) How could trustworthiness of data be extracted from massive and noisy data? (iv) What are the techniques to model user identity in the digital world? (v) How could information retrieval techniques incorporate semantic analysis to retrieve highly relevant information? (Maalej, Mtibaa, and Gargouri 2014) built an ontology-based context-aware module for mobile social networks that helps mobile users to search social networks. Their approach includes: 1. knowledge extraction from social networks (implicit, explicit, (none) contextual data using API; 2. data cleansing; 3. knowledge modelling (knowledge of user's details and contextual information); 4. comparing user profiles and the contextual information; and 5. presenting retrieved data in mobile format. Further, the approach could incorporate trust to enhance the retrieved information by adding a trustworthiness layer to the information retrieval process. (Narayan et al. 2010) proposed an approach intended to explore events from a Twitter platform and enrich an ontology designed for that purpose. Coutinho et al. (Coutinho, Lang, and Mitschang 2013) discussed the IBM concept of a component system which has the following benefits: 1. It solves the problem of the ambiguity of concepts (homonym) in the social media content. 2. It solves the missing items and concepts issue by suggesting extra meaningful concepts as well as clustering concepts in taxonomies. The paper suggested statistical models for solving the concept ambiguity problem. The use of ontology, on the other hand, could solve the problem in a concrete manner; relating each concept to its corresponding meaning in the existing ontologies will extract the actual meaning and would solve the problem significantly. Context summarization was recently addressed in the social networks domain to better understand the content of SNs; Yi Chang et al. (Chang et al. 2013) suggested, using Twitter, a methodology to infer from a large context tree an accurate summary of multiple replies to a specific tweet which form the context tree itself where the root of the tree is the original tweet and the rest are tweets that reply to this original one. Building information retrieval systems on top of ontology is an interesting approach. Evrim et al. (Evrim and McLeod 2014) presented a search methodology that is based on the dynamic information retrieval dedicated to a particular domain to satisfy the user's request by using semantic information. Yuangang et al. (Yuangang et al. 2014) presented a mechanism to enrich the semantic web with semantic forms of JSON object format. The paper proposed an automatic extraction method to semantically model a schema-less data format such as JSON. However, the mechanism that is used to evaluate merged Ontologies is inefficient because it is requires manual processing; an automatic or semi-automatic means of validation would be a valuable improvement.

Ontology and SW technology employment in the BD context has been studied interestingly to benefit BD analysis. Optique (Calvanese et al. 2013) which is the next generation of Ontology-Based Data Access (OBDA), addresses BD characteristics and data access problem in particular. Moreover, Hoppe (Hoppe, Nicolle, and Roxin 2013) proposed an ontology-based approach for user profiling in the BD context. Reddy (Reddy 2013) suggested a future research project comprising distributed semantic data management. The project is divided into two main parts: 1. Design of an actor-based approach paradigm for the storage and execution RDF Data in a distributed environment utilising the MapReduce Framework. 2. Proposal of a pay-as-you-go approach for providing Semantic OWL data as a service in the cloud infrastructure; this includes data cleansing and ontologies construction and alignment using the Hadoop/MapReduce platform. The incorporation of Ontology and trust in the business domain has been addressed in the literature. Hussien et al. (Hussain, Chang, and Dillon 2006) presented a new paradigm in trust for e-business that is built on ontology to describe trust in the e-business environment. They built an ontological representation of trust between agents, products, and services. Trust for privacy concerns has also been evident in the literature; Cutillo et al. (Cutillo, Molva, and Strufe 2009) presented a method to ensure the privacy of OSN users using a new approach to handle certain security and privacy exposures.

In the sentiment analysis context, Cambria (Cambria 2013) stated the importance of concept-based sentiment analysis by semantic analysis of the social content, which can be done via the web ontologies. Zhang et al. (Zhang, Wang, and Huang 2013) suggested a sentiment-oriented method built on top of emotion Ontology for reasoning users' emotions and used YAGO Ontology to explain associative topics. Their approach extracts emotions regardless of the fact that the emotions extracted could be false and misleading. In (Kontopoulos et al. 2013), the authors developed an approach that utilises ontology for sentiment analysis. Their basic idea is to have a system that takes as an input a tweet(s) of a particular subject and produces a sentiment score for each feature of the subject. The former approaches, however, did not consider the notion of trust and its significance for data quality inference. Capturing external data for contextualising data analysis operations is a time-consuming and complex task but may bring large benefits to current BI environments (Manuel Pérez-Martínez et al. 2008). It is crucial that data analysts take into account the VoC and the VoM in their analysis as these are major external contexts. The VoC includes the customer opinions about products and services while VoM comprises all targeted market information that can affect the company business (Berlanga et al. 2014b). As in(Reidenbach 2009), both VoC and VoM are important to building a long-term competitive advantage. Some work has been done in opinion mining and sentiment analysis to extract and summarise sentiment data effectively. As a result, we can hear VoC and VoM in social media. In (Liu 2012a), the sentiment analysis directly deals with the content to identify the reputation of products or services. There are also many commercial applications for analysing social space data. However, the challenges are: (i) the reliability and quality of the external data sources; (ii) the nature of unstructured data; and (iii) the integration of traditional BI and social BI. BI requires corporate data together with trusted external data for reliable decision-making.

In term of the tools available for knowledge extraction, the following tools have been considered and reviewed. AlchemyAPI (Harris 2013) uses machine learning and natural language parsing technology for text-based content for named entities extraction and sentiment analysis. DBpedia Spotlight is a tool for automatically annotating mentions of DBpedia resources in the text. General Architecture for Text Engineering (GATE) framework is for the development and deployment of language processing technology (Cunningham et al. 2002). Social media data can be collected and stored in plain text and then loaded in GATE. Annotations are processed during text analysis in GATE in which ontological information is encoded. Even though GATE comes with a default information extraction system, ANNIE (Maynard et al. 2001), a set of rules needs to be established for specific ontology. The adoption of domain ontologies and controlled vocabularies means that knowledge can be reusable (Ashraf, Hussain, and Hussain 2012), which is one of the core contributions of ontology use. The reuse of ontology and interlinking it with other relevant entities facilitate the interoperability of information; therefore, where possible, ontologies are reused and used in the community to produce network effects. This was also highlighted in (Hepp 2007): "ontologies exhibit positive network effects, such that their perceived utility increases with the number of people who commit to them which comes with wider usage". Ontology Usage Analysis Framework (OUSAF) (Ashraf, Hussain, and Hussain 2012) empirically analyses the use of ontologies and ranks them based on their usage to promote adoption and uptake. In the interlinking and enrichment process, different vocabularies such as Friend-of-a-Friend (FOAF) (Brickley and Libby 2010), Dublin Core (DC) (Weibel 1998), Simple Knowledge Organization System (SKOS) (Miles and Bechhofer 2009), Semantically-Interlinked Online Communities (SIOC) (Breslin 2005) can be used to enrich the semantic description of resources using an annotation component. In addition to ontology and vocabulary reuse, interlinking includes the semantic relationship between similar entities stored in other datasets.

Table 2-2 summarises the existing efforts, showing: (i) the level of semantics analysis, (ii) whether it makes use of an ontology, and (iii) whether it can be applied to Online Social Networks (OSNs).

Table 2-2: Summary of reviewed papers for semantic analysis, domain-based classification and topic distillation

Approach/ Model / Authors	**Brief Description**	**Semantic Analysis**		**Use of Ontology**	**Applied in OSNs**
		Entity-Level	Domain Level		
WATES (Carrasco et al. 2014)	The ontology-based solution for wild animal traffic problem in Brazil.	Yes	No	Yes	Yes
Evacuation Ontology (Iwanaga et al. 2011b)	An ontology for earthquake-evacuation for a real-time solution that assists people to find evacuation centres.	Yes	No	Yes	Yes
OZCT (Ghahremanlou, Sherchan, and Thom 2014)	Identifying geographic events by referencing geolocation in tweets.	Yes	No	Yes	Yes
(Bontcheva and Rout 2012)	Addressing research issues related to processing social	Yes	No	Yes	Yes

Approach/ Model / Authors	**Brief Description**	**Semantic Analysis**		**Use of Ontology**	**Applied in OSNs**
		Entity-Level	Domain Level		
	media streams using semantic analysis.				
(Saif, He, and Alani 2011)	Sentiment analysis for Twitter.	Yes	No	No	Yes
TweetLDA (Quercia, Askham, and Crowcroft 2012)	A new supervised topic model for assigning "topics" to a collection of documents.	No	Yes	No	Yes
Twitterrank (Weng et al. 2010)	Aim to find topic influential Twitterers.	No	Yes	No	Yes
(Berlanga et al. 2014a), (Garcia-Moya et al. 2013), (Louati, El Haddad, and Pinson 2014)	New infrastructure for Social BI.	Yes	No	No	Yes
(Albanese 2013)	To access, retrieve and reuse semantic OLAP databases effectively and efficiently.	Yes	No	Yes	No
Epic (Jiang et al. 2014)	A capable, efficient and reliable system to handle data variety well.	No	No	No	Yes
SOLID (Cuesta, Martínez-Prieto, and Fernández 2013)	Answers Big Data requirements considering the data that is in-motion (real-time).	Yes	No	Yes	No

Approach/ Model / Authors	**Brief Description**	**Semantic Analysis**		**Use of Ontology**	**Applied in OSNs**
		Entity-Level	Domain Level		
Optique (Calvanese et al. 2013)	Address Big Data characteristics and data access problem in particular.	Yes	No	Yes	No
(Hoppe, Nicolle, and Roxin 2013)	Explore an Ontology-based approach for user profiling.	No	Yes	Yes	No
(Reddy 2013)	Distributed semantic data management over cloud-based infrastructure.	Yes	No	Yes	No

2.3.2 SBD Incorporating Machine Learning for Data Classification and Topic Distillation

Machine learning applications enable real-time predictions by leveraging high quality and well-proven learning algorithms. Based on the current dominant position and high impact on business in several use cases, according to Gartner's recent report on emerging technologies[8], incorporating machine learning, in particular, enhances the decision-making process and provides valuable insights on large-scale data. Topic distillation (a.k.a *topic discovery, or topic modelling, or latent topic modelling or statistical topic modelling*) is an automatic approach that applies statistical techniques to distil topics from a corpus of words in a set of documents

[8] http://www.gartner.com/document/3383817?ref=solrAll&refval=175496307&qid=34ddf525422cc71383ee22c858f2238a, Visited in 25/10/2016.

(Anthes 2010, Wang et al. 2009, Blei, Ng, and Jordan 2003a). The main reason for developing topic discovery techniques is to improve information retrieval tasks including the searching and structuring of a huge corpora of data, and indexing.

These statistical-based techniques have been used as another means of topic modelling and discovery in social data mining. Examples of the application of these statistics-based techniques are LDA (Latent Dirichlet Allocation) (Blei, Ng, and Jordan 2003b), Latent Semantic Analysis (LSA), and recently Fuzzy Latent Semantic Analysis (FLSA)(Karami et al. 2017). LDA is based on an unsupervised learning model used to identify topics from the distribution of words. In LSA, an early topic modelling method has been extended to pLSA (Hofmann 1999), which generates the semantic relationships based on a word-document co-occurrence matrix. FLSA assumes that the list of documents and the words within them can be fuzzy clustered where each cluster represents a certain topic. LDA and similar unsupervised techniques have been widely used in several modelling applications (Chen et al. 2016, Nichols 2014, Weng et al. 2010, Asharaf and Alessandro 2015, Quercia, Askham, and Crowcroft 2012, Onan, Korukoglu, and Bulut 2016). Vicient et al. (Vicient and Moreno 2015) presented a methodology for unsupervised topic discovery that involves linking social media hashtags to WordNet terms . In their approach, the authors of (Alam, Ryu, and Lee 2017) applied statistical techniques that are able to detect interpretable topics. The utilisation of such techniques in social data analysis approaches is also evident in other literature; Twitterrank (Weng et al. 2010) applied the LDA modelling technique to the overall content of each user in order to identify and classify users’ interests. Ito et al. (Ito et al. 2015) adopted LDA for topic discovery to validate the credibility of the content on Twitter. Xiao et al. (Xiao et al. 2013) proposed an approach for predicting users’ influence in the social data context.

They applied the LDA technique to determine the topic distribution of users.

2.3.3 Assessment of Approaches Incorporating Semantic Analysis, Machine Learning for Data Classification and Topic Distillation

The incorporation of Semantic Analysis in the era of SBD has become popular among several research communities (Iwanaga et al. 2011a, Carrasco et al. 2014, Reddy 2013, Bontcheva and Rout 2012) in their attempts to address the ambiguity of texts in unstructured data content and discover the domain of knowledge. Semantic analysis techniques are applied to identify, annotate, and enrich entities embodied in the social data content. Moreover, the unsupervised statistical topic distillation and discovery techniques have helped to make more sense of unstructured data and support several information retrieval tools and techniques. This section addresses issues related to the existing approaches for semantic analysis and machine-learning-based topic distillation.

There have been two main research avenues in which domains of interest have been investigated and inferred from the textual content of users in OSNs. The first avenue focuses on topic modelling and discovery in social data mining. Despite the popularity and the substantial importance of the statistics-based topic discovery approaches, they are unable to deal adequately with several issues: (i) the number of topics *K*, to be discovered, is set as a parameter in the experiment, and thus it is hard to identify the optimal *K* number which represents the adequate number of topics extracted from the document (Zhang, Cui, and Yoshida 2017); (ii) the topics extracted by these models do not take the time dimension into consideration. A document’s corpus evolves with time and, subsequently, so do its themes (Alghamdi and Alfalqi 2015); (iii) these models are considered as monolingual topic models, and thus do not differentiate idioms of the same language (Zoghbi, Vulic, and Moens 2016); (iv) these models are unable to infer topics from short texts such as tweets (Li et al. 2016).

Ontologies, semantic web and Linked Data have been used in conjunction with each other as another means of domain discovery. This is done by enriching textual data and extracting knowledge, thereby linking the textual data with a particular user domain. However, some techniques leverage the use of one or several knowledge bases to enrich the textual content but neglect another knowledge base. For instance, Michelson and Macskassy (Michelson and Macskassy 2010) used the DBpedia knowledge base. The incorporation of other knowledge base repositories such as Freebase, YAGO and OpenCyc will enhance their proposed approach. Furthermore, these endeavours incorporating semantic web technology to extract knowledge from the SBD should be supported by ontologies designed to capture the domains of knowledge.

Conducting SBD analytics requires implementing technical solutions capable of handling the massive amounts of propagated data. Hence, the aforementioned previous approaches lack an appropriate and applicable data infrastructure to address the key features of BD. BD technology infrastructure facilitates the implementation of these techniques and achieves the goal of the data analytics being conducted.

2.4 Critical Synthesis Review of the Current Approaches

These sections present an integrated review of the existing methods and approaches in the era of SBD incorporating trust, semantic analysis and machine learning.

- Lack of advanced domain-based trustworthiness approaches, which is discussed in subsection 2.4.1.
- Lack of an approach to manage and extract high-level domains from the textual content of SBD, which is discussed in subsection 2.4.2.
- Lack of domain-based techniques for dual classification, which is discussed in subsection 2.4.3.

2.4.1 Lack of an Advanced Domain-based Trustworthiness Approaches

A need for domain-based trustworthiness: Several researchers have applied generic-based credibility evaluation approaches for users and their content in OSNs (Cha et al. 2010, Silva et al. 2013, Jiang, Wang, and Wu 2014). However, they do not take the topic or subject factor into consideration; the classification has been computed in general. Users will have a certain reputation in one domain, but that does not always apply to any other domain. The users' credibility should be domain-driven. For example, evaluating users' trustworthiness in a specific domain has been driven by its implication in several applications such as personalized recommendation systems (Silva et al. 2013), opinion analysis (Liu and Zhang 2012), expertise retrieval (Balog et al. 2012), and computational advertising (Yin et al. 2015).

Lack of Incorporating Temporal Factor: Subsequent studies have focused on the users' topic(s) of interest or their domains of knowledge. However, no study has been conducted that examines users' behaviour over time (Gupta et al. 2013) (Weng et al. 2010, Abbasi and Liu 2013). The users' behaviours may change over time. It follows that their trustworthiness values vary over time; hence, the temporal factor should be considered. Moreover, Spammers' behaviours are inconsistent as they are not legitimate users although they pretend to be. Therefore, their "temporal patterns of tweeting may vary with frequency, volume, and distribution over time" (Yardi et al. 2009).

Lack of a sentiment analysis of conversations: In the context of social trust, frameworks have been developed to analyse the users' content, taking into consideration the overall feelings regarding what they have chosen to expose in their content (AlRubaian et al. 2015, Gallege et al. 2014a, Bae and Lee 2012). However, these efforts did not attempt a sentiment analysis of a post's replies by measuring the trustworthiness values. It is important to understand the sentiments of a user's followers as these reveal the followers' opinions of the user. Consequently, users who receive a high number positive replies should achieve a better reputation than users who receive a large number of negative replies to their content.

Lack of addressing key features of BD: BD technology for data storage and analysis offers advanced technical capabilities for the analysis of massive and extensive amounts of data to achieve comprehensive insights in an efficient and scalable manner. Manyika et al. (Manyika et al. 2011) listed some of the big data technologies such as Big Table, Cassandra (Open Source DBMS), Cloud Computing, Hadoop (Open Source framework for processing large sets of data) etc. Chen et al. (Chen et al. 2014) discussed the various open issues and challenges of BD and listed the key technologies of BD. The incorporation of BD technology to facilitate the trustworthiness measuring and inferring tools is unavoidable, especially regarding the nature of the contents of social media which are extensive. This has attracted researchers of social trust to apply BD techniques in their experiments (Lavbič et al. 2013, Herzig, Mass, and Roitman 2014). However, several key features of BD have not been addressed, such as volume (i.e. massive social data datasets), veracity (i.e. reputation of the data sources), and value (outcome product of the data analytics). Hence, the most efficient way to deal with BD and improve the analysis of trustworthiness, is to start with the features of BD and address their related issues. The outcomes of SBD analysis will in turn assist with the analysis of trustworthiness.

To sum up, measuring the user's trustworthiness in SBD is not a trivial task. This is due to volume and heterogeneous features of social data alongside the unstructured nature of data generated from the online social platforms. This makes it more difficult to evaluate users' trustworthiness and obstructs the process of obtaining an adequate understanding of the textual content of users, as will be discussed in the next section.

2.4.2 Lack of Managing and Extracting High-Level Domains from the Textual Content of SBD

Previous attempts to improve the understanding of the contextual content of social data is noteworthy. Most of the existing approaches to topic distillations rely on bag-of-words techniques such as LDA (Blei, Ng, and Jordan 2003a). However, despite the importance and popularity of these techniques for inferring the users' topics of interest, when it comes to the context of OSNs previous approaches have three main shortcomings: (i) they do not consider the semantic relationships of terms in the user's textual content; (ii) the topic modelling technique cannot be applied efficiently to short text messages such as tweets; (iii) the high-level topics classifications that use these bag-of-words statistical techniques are inadequate and inferior (Michelson and Macskassy 2010).

LDA extracts latent topics by presenting each topic as a distribution of words. However, this statistical mechanism does not consider the semantic relationships of terms in a document (Michelson and Macskassy 2010). Furthermore, the high-level topics classifications that use these bag-of-words statistical techniques are inadequate and inferior. Furthermore, using this technique is inappropriate for clustering and searching for users based on high-level topics (Michelson and Macskassy 2010). On the other hand, the utilisation of Semantic Web tools such as AlchemyAPI™ offers a comprehensive list of taxonomies divided into hierarchies where the high-level taxonomy represents the high-level domain and the deeper-level taxonomy provides a fine-grain domain analysis. For instance, "art and entertainment" is considered a high-level taxonomy in which "graphic design" is a deep-level taxonomy. LDA is unable to extract high-level topics such as "art and entertainment" from a corpus of posts or tweets unless this term exists in the corpus. Semantic analysis, conversely, extracts semantic concepts and infers high-level domains by using an ontology to analyse the semantic hierarchy of each topic, which is not possible using the LDA technique.

2.4.3 Lack of Domain-based Techniques for Dual Classification.

This section discusses the mechanisms and limitations of the current approaches to domain-based classification.

Domain-based classification (including both user and post levels): OSNs have spurred researchers to develop several methods for discovering the main interest(s) of OSN users. Due to ambiguity and shortness of posts such as tweets (Michelson and Macskassy 2010), these endeavours are still immature; Hence, extensive research in this area is vital (Shen, Wang, and Han 2015). For example, Twitter tools (Sherchan, Nepal, and Paris 2013, 2016) are used to explore user networks to obtain information about user interests and topics. These approaches extract only the keywords to obtain a summary of Twitter data. However, the use of keywords only cannot fully cover user domains and may generate misleading user information. Both user level and tweet level content should be examined, which involves the semantics of words and accurate disambiguation for social networks study. The accurate classification of the users' interest assists in providing an accurate understanding of short textual content of future tweets. This benefits several applications, the aim of which is to obtain correct domain-based trustworthiness of users and their content in OSNs.

Incorporation of domain ontology and semantic web, and machine learning: As indicated earlier, the high-level topics classifications that use the bag-of-words statistical techniques are inadequate, and the brevity and ambiguity of short texts make more difficult the process of topic modelling using these statistical models (Li et al. 2016). Besides, these methods do not consider the temporal factor. In other words, the users' knowledge evolves with time and their interests might be diverted elsewhere depending on their experiences, work, study, etc. Hence, it is important to scrutinise users' interests over time to infer intrinsic topics of interest to OSN users. Hence, there is a need for a comprehensive approach which leverages the external domain-based ontology and semantic web knowledge bases to help disambiguate the textual content at the user and tweet levels. This approach should include advanced and state-of-the-art machine learning techniques to perform domain-based classification at the user and tweet levels.

2.5 Conclusion

Currently, several research efforts are being made to handle and manage the large scale of SBD in the quest for added value. There have been some attempts to provide technical solutions to cope with the volume and speed of social data generation. This is by facilitating the process of data capturing, acquisition and storage. Other scholars have attempted to develop data analytics solutions to enhance the quality of the collected social data, to understand the users' topics of interest, to classify and categorise users into segments, and to acquire better insights, thus improving the decision-making process. Despite the significance of these endeavours and initiatives to improve the understanding and interpretation of the extracted social data, these current frameworks for SBD analysis only partially consider SBD features. Furthermore, previous efforts are conducted merely to address one or few issues of SBD. A comprehensive framework is required to resolve the issues of data quality, extract hidden knowledge, and infer the credibility of users and their OSN content by extracting the domains of interest at the user and the post levels. This consolidates the development of data classification techniques which leads to better anticipation of users' interest in their future published content.

In this chapter, an extensive survey of the existing literature is conducted. The discussion of several approaches for measuring the generic-based and domain-based trustworthiness in SBD is carried out followed by a review of the relevant literature pertaining to semantic analysis, machine learning and data classification within the context of SBD. Previous works related to these themes are critically reviewed, and demonstrate the considerable achievements that have been made in SBD analytics. However, the current approaches and techniques are still inadequate in terms of: (i) the lack of domain-based trustworthiness approaches; (ii) the lack of approaches for managing and extracting high-level domains from the textual content of SBD; (iii) the lack of domain-based approaches for dual classification of the textual content of SBD.

The next chapter establishes the research ground to address the deficiencies of previous researches depicted in this chapter. The related preliminary concepts are identified and defined, and the main research problems and the underlying research issues are explained. The methodology adopted for this study is explained and justified, followed by the solution requirements to develop the proposed approaches.

Chapter 3 Problem Definition

3.1 Introduction

Chapter 2 surveys the literature to provide the essential research background and reviews of the current state-of-the-art research in the areas of social trust, semantic analysis, and domain-based classification. The research gaps which motivated the research activities of this thesis are identified and depicted. These research gaps comprise three key research shortcomings. First, there is a lack of advanced domain-based trustworthiness approaches. Despite the significant efforts conducted to measure users' trustworthiness in OSNs, the current credibility mechanisms generally apply generic-based credibility techniques. The users credibility should be domain-driven. Second, there is no approach for effectively managing and extracting high-level domains from the textual content of SBD. The previous endeavours in regard to topics classification and modelling relied on bag-of-words techniques which show several significant deficiencies. There is an imperative need to incorporate semantic analytics techniques which can effectively provide fine-grain domain analysis. Third, there is a lack of domain-based techniques for dual classification. The attempts to discover the domain interest of users and their content are still immature. An accurate and effective classification of the users' domain of interest will enable a better contextual understanding of the short textual content of their future tweets.

This chapter presents an in-depth analysis of the aforementioned research problems and discusses the underlying research issues. Further, the chapter presents the research methodology adopted in this thesis to address the emerging research issues, followed by several requirements deemed necessary to solve the related problems.

3.2 Preliminary Concepts and Definitions

This section introduces several main concepts and the terminology used throughout this thesis.

Big Data

Definition: Big Data (BD) is the technical term used in reference to the vast quantity of heterogeneous datasets which are created and spread rapidly, and for which the conventional techniques used to process, analyse, retrieve, store and visualise such massive sets of data are now unsuitable and inadequate. This can be seen in many areas such as sensor-generated data, social media, uploading and downloading of digital media. BD has several V-features: Volume, Velocity, Variety, Veracity, Variability and Value. (Marz and Warren 2015) (Cukier 2010) (Beyer 2011), (Marz and Warren 2015), (Fan and Bifet 2013) (Kaisler et al. 2013).

Social Big Data

Definition: Social Big Data (SBD) and Big Social Data (BSD) are a combination of two terms – social media and Big Data – and are used interchangeably in reference to the massive amount of user-generated content, mainly in the form of unstructured data such as posts, photos, audios, videos etc. SBD comprises processes and methods that are designed to provide sensitive and relevant knowledge from social media data sources to any user or company. The distinctive features of social media data sources are their different formats and contents, their very large size, and the online or streamed generation of information (Bello-Orgaz, Jung, and Camacho

2016).

Online Social Networks

Definition: Online Social Networks (OSNs) or Social Network Sites (SNSs) are defined as the systems comprising certain tools, applications and platforms that enable the online social interactions of individuals and communities. These sites enable individuals and organisations to create public or semi-public profiles by establishing connections and providing the capacity to view and interact with others who have similar personal or career backgrounds. These web-based social networks include, but are not limited to, Facebook®, Twitter®, LiveBoon®, Orkut®, Pinterest®, Vine®, Tumblr®, Google Plus®, and Instagram® (Nepal, Paris, and Bouguettaya 2013, Amichai-Hamburger and Hayat 2017).

Domain of Knowledge

Definition: Domain of Knowledge is a particular area of an individual's work, expertise, or specialisation within the scope of subject-matter knowledge such as politics, sports, education etc. (Alexander and Judy 1988, Hjørland and Albrechtsen 1995b, Dinsmore 2017).

Social Trust

Definition: Social Trust or Social Credibility, in information science, is the measurement of confidence where a group of individuals or communities behave predictably. It can be described by offering reasonable grounds for being believed (Sherchan, Nepal, and Paris 2013, Castillo, Mendoza, and Poblete 2011b).

Domain-based User Trustworthiness

Definition: Domain-based Trustworthiness in social media, in the context of this thesis, refers to the credibility of users and their posted and shared content in a particular domain of knowledge.

Sentiment Analysis

Definition: Sentiments Analysis (a.k.a. opinion mining) is the process of recognising and quantifying the emotions inferred from textual content by means of statistical analysis, natural language processing, computational linguistics etc. (Liu 2012b).

Ontology Engineering

*Definition**:*** Tom Gruber attracted a great deal of interest from the computer science community by defining ontology as “an explicit specification of a conceptualisation" (Gruber 1993). Conceptualisation is the formulation of knowledge about entities. The specification is the representation of the conceptualisation in a concrete form (Stevens 2001). The specification will lead to commitment in semantic structure. In short, an ontology is the working model of entities. Notably, new software tools have been developed to facilitate ontology engineering. Ontology engineering involves the formulation of an exhaustive and rigorous conceptual schema within a given domain. Ontology captures the domain knowledge through the defined concrete concepts (representing a set of entities), constraints, and the relationship between concepts, to provide a formal representation in machine-understandable semantics. The purpose of ontology is to represent, share, and reuse existing domain knowledge.

Named Entity Recognition

*Definition**:*** Named Entity Recognition is a process of information extraction intended to examine the textual content and to classify and locate terms which belong to certain pre-defined categories such as countries, persons, organisations etc.

(Nadeau and Sekine 2007).

Semantic Annotation and Enrichment

Definition: Semantic Annotation is the process of identifying and capturing concepts within a text and assigning to them their semantic description based on the concepts defined in the domain ontology. The annotation is then enriched with a description of the concepts referring to the domain ontologies and using controlled vocabularies such as DC[1], SKOS[2], SIOC[3]. This allows each entity in the textual data to be specified with its semantic concept (Kiryakov et al. 2004, Oren et al. 2006).

Semantics Interlinking

Definition: Semantic Interlinking is the process of linking the descriptions of defined entities that exist in dissimilar datasets and vocabularies to extend the view of the entities representing the same real-world concepts/object in a certain domain (Ferraram, Nikolov, and Scharffe). This is leveraged through the semantic web technology introduced by Berners Lee who provided a new vision for the next web where data is given semantic meanings via data annotation and manipulation in a machine-readable format (Berners-Lee and Hendler 2001).

Machine Learning

Definition: Machine Learning (ML) is a branch of Artificial Intelligence(AI) which comprises several statistical techniques designed to enable applications to learn and make predictions without explicit programming (Samuel 1959).

Supervised Learning

[1] dublincore.org/

[2] http://www.w3.org/2004/02/skos/

[3] http://sioc-project.org/

Definition: Supervised Learning is a machine learning technique used to infer a function from a labelled trained dataset where the output is already known (Mohri, Rostamizadeh, and Talwalkar 2012).

Logistic Regression

Definition: Logistic Regression or logit regression is a predictive statistical analysis method used to conduct binary classification of a dataset comprising more independent variables which determine a certain outcome (Freedman 2009).

Support Vector Machine

Definition: Support Vector Machine (SVM) is commonly used for conducting binary classification tasks, in particular those involving contradictory matrix analysis (true-positive and false-negative) (Cortes and Vapnik 1995).

Decision Tree Classifier

Definition: A Decision Tree Classifier is a flow-chart-like structure, where each internal (or non-leaf) node denotes a test of an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node contains a class label. A decision tree expresses a recursive partition of the instance space (Lior 2014).

Cross-Validation

Definition: Cross-Validation or rotation estimation is a significant evaluation technique that evaluates predictive machine learning models by dividing the original training into partitions (or folds) and conducting a predictive analysis of each fold (Kohavi 1995).

Ground Truth

Definition: Ground Truth in machine learning refers to the labelled dataset (known to be correct) which comprises accurate data that has been observed or measured, and is used for training the intended classification module (Theiler et al. 1999).

3.3 Problems Definition

3.3.1 Problem Statement 1: Lack of Domain-based Credibility Approaches in the Era of SBD

Problem: Lack of domain-based credibility approaches in the era of SBD, in the context of this thesis, refers to the deficiency of implementing platforms to measure and evaluate the credibility of users in SBD considering their domain(s) of interest. These platforms should address the BD features, and facilitate data storage, data processing and data analysis.

Despite the diverse depictions of the BD problem, BD is usually described in terms of several **V**-features as depicted in Figure 3-1. These include, but are not limited to: ***Volume*** – refers to the vast increase in the data growth; ***Velocity*** – represents the accumulation of data in high speed and real-time from several data sources; ***Variety*** – involves fuzzy and heterogeneous types of data; **V**eracity – refers to the accuracy, correctness and trustworthiness of data; ***Variability*** – refers to variance in meaning(Fan and Bifet 2013); and ***Value*** – represents the outcome of BD analysis (i.e. new insights) (Demchenko et al. 2013). Hence, the impact of the volume of BD extends beyond business-related data to include political and governmental data, healthcare data, education-related data, and the data of many other sectors. The key challenge of BD analysis is the mining of enormous amounts of data in the quest for added value.

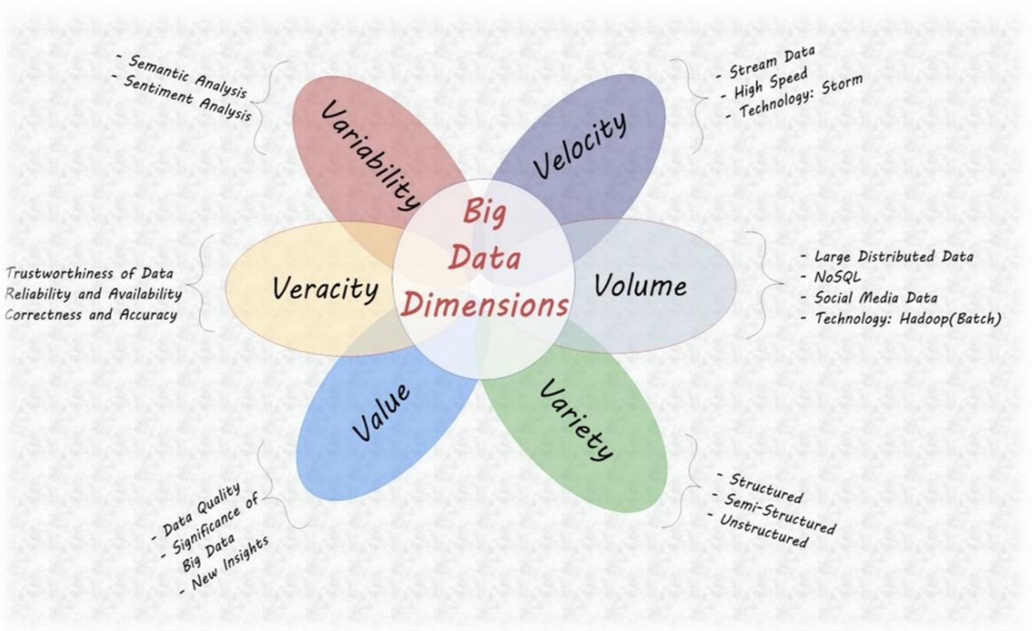


Figure 3-1: Big Data V-features

OSNs provides a momentum dense of social data which require a thorough scrutiny. These web social networks provide platforms for people to express their opinions and establish new avenues of social communication based on these virtual societies. OSNs offer effective mediums of communication through which legitimate users as well as spammers can publish their content. The spamming activities in social platforms has increased dramatically (Wang 2010). Spammers misuse the OSNs' features and tools, sending annoying messages to legitimate users, publishing contents that include malicious links, and hijacking popular topics (McCord and Chuah 2011). Spammers post contents on various topics, and they duplicate posts (Wang 2010). Further, to propagate their vicious activities, spammers abuse other OSN features such as hashtags, and mention other users and Link-shortening services (Miller et al. 2014). Hence, it is important to understand the users' behaviour because of the dramatic increase in the usage of online social platforms. For example, since there are over 310 million monthly active users of Twitter[4] , a significant question arises regarding the quality of the enormous amount of data that is being proliferated every minute by users of these computer-generated environments. This explains the importance of measuring the users' credibility and ascertaining the users' influence in a particular domain. Hence, the factual grasp of the users' domains of interest and an appropriate judgement of their emotions enhances the customer-to-business engagement. This necessitates an accurate analysis of customer reviews and their opinions to obtain a better understanding of their needs, thereby enhancing their customer service.

Several previous approaches have been attempts to address data quality and social credibility issues. However, most of these endeavours have tackled the credibility of OSN users with no consideration given to their domains of interest. Furthermore, the users' credibility over time has not been measured. Users' domains of interest evolve over time, as does their credibility. Hence, it is essential to have frameworks that can measure users' credibility temporally in all domains categorised under SBD.

3.3.2 Problem Statement 2: Ambiguous Understanding of the Textual Content of SBD

Problem: Ambiguous understanding of the textual content of SBD, in the context of this thesis, indicates the complication of obtaining an accurate understanding to the contextual meaning conveyed by the textual content posted by users of OSNs.

[4] https://about.twitter.com/company , Accessed 01 07. 2017

Because of the large volume of data and information generated by a multitude of sources, it is a huge challenge to manage and extract useful knowledge, especially given the different forms of data, streaming data and uncertainty of data. Hence, there are still challenges in this area of BD analytics research to capture, store, process, visualise, query, and manipulate datasets to derive meaningful information that is specific to an application's domain. In particular, the discovery and understanding of social data is a goal of enterprises today. The rapid increase in the amount of unstructured social data has highlighted the importance of unstructured social data as a means of acquiring deeper and more accurate insights into businesses and customers in order to achieve a competitive advantage in the current business environment (Feldman and Sanger 2007, Joa et al. 2012, Das and Kumar 2013).

Another noteworthy social issue which requires a thorough understanding of the social content involves the bad experiences that users of OSNs encounter through the great overlap between their personal space and those opinions that come from their general followers which leads to misinterpretation to what the users mean by their posts (Vitak 2012). OSNs have contributed to the spread of this phenomenon due to the open nature of these platforms that allow social content to be disseminated among large audiences, mostly outside the cultural contexts of the content generator. In comparison, traditional communication is characterised by direct face-to-face interaction. Hence, individuals who communicate via these limited virtual mediums misconstrue the factual meaning of the textual content. The social content published by the users of OSNs reflects their personality, their thoughts, their tendencies, and their nature. An accurate interpretation of the users' content will ensure that the intended meaning is being conveyed which includes the topics of interest embedded in its content.

3.3.3 Problem Statement 3: Lack of Domain-Based Approaches for Dual Classification in SBD

Problem: Lack of domain-based approaches for dual classification in SBD, in the context of this thesis, refers to the lack of effective platforms capable of providing domain-based classification to SBD. These platforms should be able to classify and predict the domain of knowledge at the user level and the content level (dual classification).

Given the constantly increasing amounts of social data, there arose the need to develop efficient tools for knowledge extraction. Hence, data analytics emerged as an approach for extracting knowledge from vast amounts of data. It is a modern technology that has strongly established itself in the information age. Data analysis techniques enable companies and organisations in various sectors to explore and focus on information that is the most important to their operations. Furthermore, these techniques lead to building future predictions and exploring behaviours and trends. Hence, several leading companies today use a methodical and systematic approach to explore knowledge derived from the massive amount of stored data, thus improving their practices and achieving a competitive advantage.

The immense and continuous propagation of the usage of OSN platforms, emphasises the significance of these virtual communities not only as a means of linking remote people, but also as avenues for knowledge exchange, skills improvements, expressions of opinion etc. Hence, it is important to discover the domain of the textual content generated by users of these platforms. This is likely to lead to better performance of a variety of business-related applications such as the VoC / the VoM, recommendation systems, and the discovery of domain-based influencers.

Consequently, OSNs have motivated researchers to develop several ways of discovering the main interests of their users. However, short text messages such as tweets can be ambiguous and confusing, and therefore not suited to the frequency-based topic-distillation techniques (i.e. LDA). Hence, it is essential to obtain an accurate understanding of the contextual meaning of the user's textual content in order to determine his/her domain of knowledge. This will help to indicate the domain of tweets that will be posted by the user in future.

3.4 Underlying Research Issues

3.4.1 Research Issue 1: Elicit Value and Attain Veracity of SBD through Domain-based Credibility Approach

The literature review has shown a lack of domain-based credibility approaches in the era of SBD. This section addresses the following primary research issue articulated based on the aforementioned problem:

Research Issue: elicit value and attain veracity of SBD through domain-based credibility approach. *This is through the development of a comprehensive framework that aims to infer value from SBD by measuring the domain-based trustworthiness of OSN users, addressing the main features of BD, and incorporating semantic analysis and the temporal factor.*

OSNs are a fertile platform through which users can express their opinions and share their views, experiences and knowledge of numerous topics. There is a lack of evaluation mechanisms that incorporate domain-based trustworthiness. In OSNs, discovering users' influence in a specific domain has been motivated by its significance in a broad range of applications such as personalized recommendation systems (Silva et al. 2013), opinion analysis (Liu and Zhang 2012), expertise retrieval (Balog et al. 2012), and computational advertising (Yin et al. 2015).

Domain of Knowledge is a particular area of an individual's work, expertise, or specialisation within the scope of subject-matter knowledge such as politics, sports, education, etc. (Hjørland and Albrechtsen 1995a). The Semantic Web provides a new vision for the next web where data is given semantic meanings via data annotation and manipulation in a machine-readable format (Berners-Lee and Hendler 2001). The incorporation of semantic analysis in OSNs, in particular, reduces the ambiguity of SBD by clarifying the actual context of the users' content. This mitigates the *variability* of BD (Emani, Cullot, and Nicolle 2015) (Hitzler and Janowicz 2013), distinguishes users' domains of interest, and deduces their actual sentiments.

Sentiment analysis (a.k.a. opinion mining) has become a core dimension of researchers' endeavours to create applications that leverage the massive increase of user-generated content (Kumar and Sebastian 2012). For example, User Generated Content (UGC) in OSNs has been examined to study the effective data extracted and applied to numerous applications (Zhang et al. 2015, Bae and Lee 2012, Kawabe et al. 2015). In the context of social credibility, several attempts have been carried out to measure the credibility of users and their content by leveraging the effective data distilled from their content. These attempts have not extended the sentiment analysis to the textual content of the inclusive conversations, which should include the attitudes derived from the posts' replies. The followers' replies to the user's content indicate the positive and negative opinions of the followers which is an important dimension to measure the user's credibility. Likewise, most of these efforts assimilated the sentiment analysis of the content regardless of its context. Hence, the semantic analysis should be included to improve the accuracy of the resultant sentiment finding. Moreover, the users' behaviours may change over time. It follows, therefore, that trustworthiness values vary over time; hence, the temporal factor should be considered.

The *veracity* of BD refers to the accuracy, correctness and trustworthiness of data. Demchenko et al. (Demchenko et al. 2013) presented multiple factors to ensure the *veracity* of BD. These factors include, but are not limited to, the following: (i) trustworthiness of data origin; (ii) reliability and security of data store; and (iii) data availability. The list established by Demchenko et al. could be enhanced by including a further two essential aspects: *correctness* and *consistency*. Although the origin and storage of data are critical, the trustworthiness of the source does not guarantee data correctness and consistency. Data cleansing and integration should be incorporated to ensure the *veracity* of data as well.

3.4.2 Research Issue 2: Extracting Knowledge and Defining Domain in SBD

The current approaches examined in the literature review do not adequately consider the semantic relationships of terms in the user's textual content, particularly in short text messages such as tweets. Moreover, these approaches are unable to obtain classifications of high-level topics as they utilise only the bag-of-words statistical techniques. Addressing the following research issue will be a major objective of this chapter:

Research Issue: Extracting knowledge and defining a domain in SBD. *This is through the development of an approach that aims to semantically analyse social content, thus enriching social data with semantic conceptual representation for domain-based discovery.*

The challenge of managing and extracting useful knowledge from social media data sources has attracted much attention from academia and industry. SBD is an important BD island; thus, social data analytics are intended to make sense of data and to obtain value from data. SBD provides a wealth of information that businesses, political governments, organisations, etc. can mine and analyse to exploit value in a variety of areas. However, there are still challenges in this area of SBD analytics research to capture, store, process, visualise, query, and manipulate dataset to derive meaningful information that is specific to an application's domains (Tole 2013, Labrinidis and Jagadish 2012).

The main obstacle to discovering the accurate domain(s) of the users is determining the contextual meaning of their textual content accurately. This is because of several linguistic features such as polysemy (the *same word has several meanings*), homonymy (*words have the same spelling and pronunciation, but have different meanings),* contronymy (the *same word has contradictory meanings*), etc. This linguistic diversity makes more difficult the process of determining the correct domain of interest at the user level and harder at the post level.

Most of the current endeavours to infer the topics of interest from textual content in OSNs use bag-of-words statistical techniques such as Latent Dirichlet Allocation (LDA). The problematic nature of these tools means that they cannot adequately infer high-level topics. Besides, the brevity and ambiguity of short texts such as tweets make it more difficult to process topic modelling using these statistical models. On the other hand, ontology, Linked Data, and a knowledge base can be utilised to identify, annotate, and enrich entities in tweets prior to conducting the semantic analysis. The semantic analysis extracts semantic entities and concepts and deduces high-level domains/topics by analysing the semantic hierarchy of each topic, leveraging an ontology.

3.4.3 Research Issue 3: Domain-Based Classification and Prediction of SBD

Previous works in the area of topic distillation, and discovery lack an appropriate and applicable technical solution that can handle the complex task of obtaining an accurate interpretation of the contextual social content. This is evident through the inadequacy of these endeavours in addressing the topics of microblogging short messages like tweets, and their inability to classify and predict the messages' actual and precise domains of interest at the user level. Hence, this chapter intends to address this problem by investigating the following underlying research issue:

Research Issue: Domain-based Classification and Prediction of SBD, *in the context of a domain discovery framework incorporating machine learning, is defined as domain-based discovery in OSNs at the user and tweet levels incorporating comprehensive knowledge discovery tools and well-known machine learning algorithms.*

Businesses have benefited from the prevalence of OSNs as these enable them to establish interactive-based dialogues with their customers (Chen, Chiang, and Storey 2012). These dialogues which appear in companies' social pages permit customers to express their views freely and without restrictions on products and services provided by those companies, using comments or replies to either praise these products or services, or to point out their shortcomings. This indeed provides an opportunity for the business firms to study and respond to these opinions, thereby enhancing its customer service, which greatly extends their customer knowledge, customer acquisition, and customer retention (Sashi 2012, Nitzan and Libai 2011). This objective can be achieved through an accurate interpretation of the users' social content, and automatically classifying their topics of interest into correct categories (Khobzi and Teimourpour 2015).

The existing approaches to topic extraction, modelling and classification rely on statistical bag-of-words techniques such as LDA (Blei, Ng, and Jordan 2003a). However, despite the importance and popularity of these techniques for inferring the users' topics of interest, when it comes to the use of Twitter, these approaches have significant shortcomings: (1) the number of topics to be discovered is set as a parameter in the experiment; thus, it is hard to identify the optimal number which represents the adequate number of topics extracted from the document (Zhang, Cui, and Yoshida 2017), (2) the topics extracted by these models do not contemplate the temporal aspects. A document's corpus evolves through time and subsequently so do its themes (Alghamdi and Alfalqi 2015), (3) these models are considered as monolingual topic models, and therefore do not differentiate idioms of the same language (Zoghbi, Vulic, and Moens 2016); and (4) these models are unable to infer high-level topics particularly from short text such as tweets (Li et al. 2016).

An accurate domain-based categorisation of a tweet's textual content is the foremost challenge. This is evident due to various linguistic traits[5] which hamper and make more difficult the efforts to resolve this research problem. These traits include: amphibology, polysemy, homonymy, contronymy, heterosemy, and many other interrelated linguistics features. This diversity in linguistics makes the process of determining the correct domain of interest from the short textual content of the tweet more difficult. Hence, it is essential to obtain an accurate understanding of the semantics of the overall tweets' textual content in order to determine the user's domain of knowledge. This will assist in determining the topic/domain of the short textual content of a tweet.

3.5 Research Methodology

3.5.1 Overview of Research Methods

A comprehensive understanding of a research problem is necessary in order to develop adequate and appropriate solutions to resolve the problem. Hence, it is essential to adopt a scientific methodology to acquire knowledge which will be used to design solutions across several problem domains. The research methods in information systems can be commonly categorised under two main themes: (i) social or natural science research, and (ii) design science research. Social science research aims to address certain interrelated social issues, concepts, ideas etc. to describe human behaviour (Simon 1996) or to understand reality (March and Smith 1995). This is done by adhering to several experimentally created and naturalistic scientific settings using various methodologies (for example, quantitative and qualitative

[5] https://plato.stanford.edu/archives/spr2016/entries/ambiguity/

research). (Krathwohl 1993, Sieber 2012, Wheeldon and Ahlberg 2011). The design science research was created as a standard paradigm for research conducted in the Information Systems (IS) field to provide guidelines for researchers in IS, enabling them to construct artifacts to address and resolve a specific research problem (von Alan et al. 2004). In the same context, Gregor et al. (Gregor and Jones 2007) and Venable (Venable 2013) contributed to the establishment and formulation of design theory. The former presented the structural components that are needed to communicate a design theory, including both core components and additional components. Venable criticised the still-arguable issues about design theory and provided a simplified formulation.

3.5.2 Choice of Design Science Research Paradigm

It is evident that research based on both social science and design science is intuitively significant to improve knowledge and to elicit the views of subjects being investigated. This thesis follows the Design Science Research Methodology (DSRM) due to its applicability to comprehend the nature of the research carried out in this thesis. In particular, the objective of this thesis is to develop systems for implementing social data analytics intended to measure the trustworthiness of SBD users, derive domain knowledge, and infer value from SBD. Hence, this research presents scientific and technical solutions by developing systems able to adequately address the identified research issues. Therefore, this research is aligned with the purpose of the DSR paradigm which is commonly adopted to create utilities and artifacts that serve stakeholders (March and Smith 1995, Weber 2010). Further, this thesis encompasses several IT-related theories, techniques, and systems' implementation which are the key focus of the DSR approach (Benbasat and Zmud

2003, Weber 2010, Alter 2008, von Alan et al. 2004). Last but not least, this research involves the collection of a large amount of social data, which would be impossible to achieve through population sampling and other natural research approaches.

Amongst several methodologies proposed and positioned under the DSR umbrella, Peffers et al. (Peffers et al. 2007) presented their methodology via a comprehensive list of activities for any research conducted in the IS discipline.

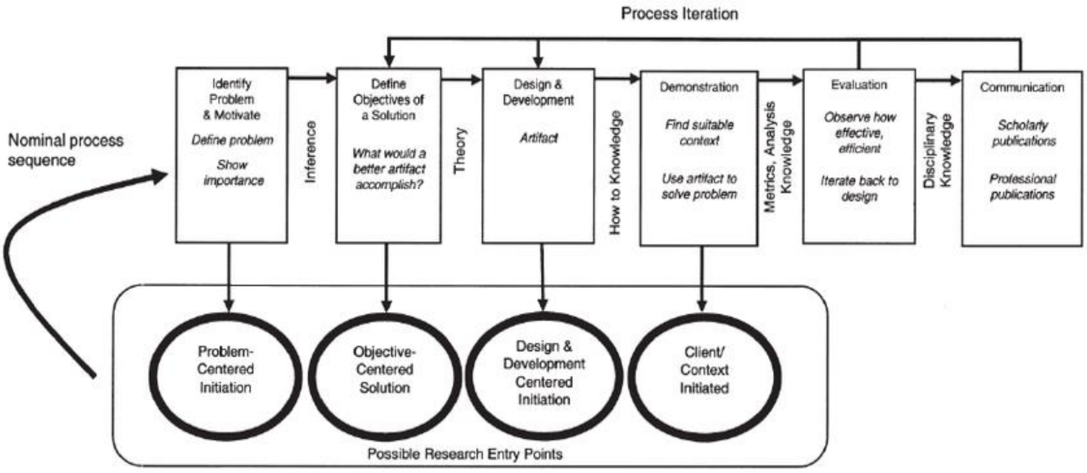

Process Iteration
Nominal process sequence
Identify Problem & Motivate
Define problem
Show importance
Inference
Define Objectives of a Solution
What would a better artifact accomplish?
Theory
Design & Development
Artifact
How to Knowledge
Demonstration
Find suitable context
Use artifact to solve problem
Metrics, Analysis Knowledge
Evaluation
Observe how effective, efficient
Iterate back to design
Disciplinary Knowledge
Communication
Scholarly publications
Professional publications
Problem-Centered Initiation
Objective-Centered Solution
Design & Development Centered Initiation
Client/ Context Initiated
Possible Research Entry Points

Figure 3-2: Design Science Research Methodology Process Model (Peffers et al. 2007)

Figure 3-2 depicts a commonly accepted framework model proposed by Peffers et al. for the production and presentation of design science research. The list of activities within this framework and carried out in this thesis include:

Activity 1 - Problem identification and motivation: “define the specific research problem and justify the value of a solution” (Hevner and Chatterjee 2010). After the problem has been deconstructed, it will be better understood, and this will assist in the development of a rigorous solution.

This research is carried out to precisely identify and examine certain problems that have emerged with the overwhelming amassing of SBD. This is done by conducting a literature review of current state-of-the-art approaches which revealed the following shortcomings:

- Lack of advanced domain-based trustworthiness approaches.
- Lack of approaches for managing and extracting high-level domains from the textual content of SBD.
- Lack of domain-based techniques for dual classification of the social content at the user level and post level.

Activity 2 - Define the objectives for a solution: "infer the objectives of a solution from the problem definition and knowledge of what is possible and feasible" (Hevner and Chatterjee 2010). This is done by explicitly defining the quantitative or qualitative nature of the objective.

This thesis aims to develop approaches for social data analytics intended to derive knowledge and infer value from SBD. This overall research objective is segmented into several sub-objectives as follows:

Sub-objective 1: Develop a framework to deduce the value and determine the veracity and credibility of SBD. It is envisioned that this framework will address the major challenging features that constitute the general problem of BD. This is done by implementing a technical solution to resolve the problem of handling the massive amount of data, and to facilitate data storage and analysis.

Sub-objective 2: Develop a systematic approach to extract knowledge captured from the textual content of SBD. This approach enhances the understanding of users' domains of interest.

Sub-objective 3: Develop an approach incorporating comprehensively advanced knowledge discovery and data classification techniques as a means of domain-based detection. The proposed framework can perform dual classification tasks at the user level and the post level.

Sub-objective 4: Develop and refine a prototype implementation of the design to verify and validate the effectiveness and efficiency of the proposed approaches and their related frameworks.

Activity 3 – Design and Development: this is the core part of the research; it results in the creation of artifacts and determines the architecture and functionality of these artifacts. "The objective of the design-science research is to develop technology-based solutions to important and relevant business problems" (von Alan et al. 2004).

This thesis reveals the development of several effective artifacts that address the research problems identified in the first activity. Various cutting-edge technical solutions were applied to all the research activities. Chapters 4-6 explain these technical solutions which are summarised as follows:

Chapter 4 presents a framework called CredSaT (***Cred****ibility incorporating* ***S****emantic* ***a****nalysis and* ***T****emporal factor*): a comprehensive SBD framework intended to measure users' credibility is based on their domains of knowledge.

Chapter 5 presents a framework developed to address the lack of approaches for managing and extracting high-level domains from the textual content of SBD.

Chapter 6 presents a consolidated framework leveraging former knowledge obtained from an analysis of the user's historical content to effectively classify and predict domains of interest at the user level and the post level.

Activity 4 – Demonstration: the artifact is used to solve a problem in order to demonstrate its efficacy. "Demonstration illustrates the use of the artifact to solve one or several problem instances and is considered as an early evaluation activity" (Prat, Comyn-Wattiau, and Akoka 2014).

The implementation of artifacts designed in this thesis has been successful in addressing the problems and issues identified in this research. This was done by developing prototypes as a Proof of Concept (POC). POC verifies the feasibility of the proposed methods and concepts to ensure their validity for building a potential real-world application.

Activity 5 – Evaluation: "observe and measure how well the artifact supports a solution to the problem" (von Alan et al. 2004). Evaluation is a crucial phase (von Alan et al. 2004) in design science as it assures the rigour of the research through the provided feedback on the implemented artifacts (Venable, Pries-Heje, and Baskerville 2016). This emphasises the importance of the evaluation phase to substantiate the design's efficacy and its effectiveness. In this context, Venable et al. (Venable, Pries-Heje, and Baskerville 2012) proposed a comprehensive evaluation framework as an evaluation guideline strategy for DSR researches.

The approaches developed in this research have been intensively verified and validated through several experiments and case studies as follows:

Chapter 4 shows the effectiveness of CredSaT by benchmarking it against other state-of-the-art baseline models. The reported performance of CredSaT indicates the highly trustworthy, domain-based influencers. The capability of CredSaT to infer anomalous users is confirmed.

Chapter 5 presents several evaluation methods using the developed approach in the political domain incorporating public data collected from Twitter. The work has produced optimistic results which establishes a foundation for predicting and classifying users' domain of knowledge.

Chapter 6 discusses experiments conducted to evaluate the developed framework. These experiments validate the applicability and effectiveness of our approach to acquiring a better understanding of Twitter content at the user and tweet levels. This is evident through the notable performance of the machine learning experiments conducted at both the user and tweet levels.

Activity 6 – Communication: the artifacts and their importance should be presented to appropriate audiences, including technical personnel. This is to demonstrate its novelty, importance, rigor and utility. Several scholarly research publications have emerged from this thesis in various research avenues.

These include the following papers published in conference proceedings and peer-reviewed journals throughout the course of this research:

- B. Abu-Salih, P. Wongthongtham, K. Y. Chan. 2018. "Twitter Mining for

Ontology-based Domain Discovery Incorporating Machine Learning", *Journal of Knowledge Management (JKM)*. Vol. 22 Issue: 5, pp.949-981, https://doi.org/10.1108/JKM-11-2016-0489.

- B. Abu-Salih, P. Wongthongtham, K. Y. Chan, Z. Dengya. 2018. "CredSaT: Credibility Ranking of Users in Big Social Data incorporating Semantic Analysis and Temporal Factor", *Journal of Information Science (JIS)*, https://doi.org/10.1177/0165551518790424 .
- P. Wongthongtham, and B. Abu-Salih, "Ontology-based Approach for Identifying the Credibility Domain in Social Big Data", *Journal of Organizational Computing and Electronic Commerce (JOCEC), Inpress - Accepted Mar 2018*.
- P. Wongthongtham, K. Y. Chan, V. Potdar. 2018. B. Abu-Salih, S. Giakwad, J. Pratima, "State-of-the-Art Ontology Annotation for Personalised Teaching and Learning and Prospects for Smart Learning Recommender Based on Multiple Intelligence and Fuzzy Ontology". *International Journal of Fuzzy Systems (IJFS)*. Vol. 20 Issue: 4, pp. 1357-1372, https://doi.org/10.1007/s40815-018-0467-6.
- B. Abu-Salih, P. Wongthongtham, Z. Dengya, SH.Alqrainy. 2015. "An Approach for Time-Aware Domain-Based Analysis of Users' Trustworthiness in Big Social Data", *Services Transactions on Big Data (STBD)*, Vol. 2 Issue: 1, pp. 41-56, https://doi.org/10.29268/stbd.2015.2.1.4.
- B. Abu-Salih, P. Wongthongtham, S.-M.-R. Beheshti, and Z. Dengya, "A Preliminary Approach to Domain-based Evaluation of Users' Trustworthiness in Online Social Networks," in IEEE International Congress on Big Data (BigData Congress-2015), New York, USA, 2015.

- B. Abu-Salih, P. Wongthongtham, S.-M.-R. Beheshti, and B. Zajabbari, "Towards A Methodology for Social Business Intelligence in the era of Big Social Data incorporating Trust and Semantic Analysis" in Second International Conference on Advanced Data and Information Engineering (DaEng-2015), ed. Bali, Indonesia: Springer, 2015.
- P. Wongthongtham, B. Abu-Salih, "Ontology and trust-based data warehouse in new generation of business intelligence: State-of-the-art, challenges, and opportunities", in IEEE 13th International Conference on Industrial Informatics (INDIN) 2015: 476-483.
- R. Nabipourshiri, B. Abu-Salih, P. Wongthongtham, "Tree-based Classification to Users' Trustworthiness in OSNs", in 10th International Conference on Computer and Automation Engineering (ICCAE 2018) 2018.
- J. Kaur, P. Wongthongtham, B. Abu-Salih, S. Fathy, "Analysis of Scientific Production of IoE Big Data Research", in the 32nd IEEE International Conference on Advanced Information Networking and Applications Workshops (WAINA), Krakow, Poland, 2018, pp. 715-720. doi: 10.1109/WAINA.2018.00173.

3.6 Solution Requirements

3.6.1 Solution Requirements for Domain-based Credibility Approach.

In Section 3.4.1 the research issue is established that requires a solution to measure the domain-based credibility in OSNs by incorporating semantic analysis, sentiment analysis and temporal factor. A comprehensive framework for the era of

SBD intended to examine the content of users in OSNs and infer their credibility is required. This solution includes four main building blocks that enable the required framework to address the research issue:

- Semantic analysis: the factual grasp of the users' domains of knowledge leads to a better understanding of their interest(s). In addition, discovering users' influence in a specific domain has been motivated by its significance in several applications. Incorporating semantic analysis will reduce the ambiguity of the data, thereby decreasing the variability of BD (Hitzler and Janowicz 2013). Hence, the textual message of tweets need to be to analysed and enriched in order to provide the semantics of textual data and obtain the taxonomy of each message.
- Sentiment Analysis: understanding users' emotions and their impressions, including their positive or negative opinions, is vital if one is to determine the followers' effect on the user's social content. Hence, it is essential that the conversations conducted through the OSNs be examined in order to identify, extract and study the attitudes, emotions, opinions, and the subjective impressions of followers. This can be achieved by incorporating APIs utilising NLP techniques, statistics, or machine learning methods.
- Temporal Factor: the time factor should be considered when examining users behaviours in OSNs platforms. Hence, it is important to study user behaviours over time and reflect these using their domain-based credibility values. Further, the behaviour of social spammers is inconsistent in term of their usage of social platforms as their "temporal patterns of tweeting may vary with frequency, volume, and distribution over time"(Yardi et al. 2009). Hence, it is vital to scrutinise user's social content temporally so as

to find a degree of truthfulness in their behaviours.

Framework verification and validation: the underlying mechanism for measuring the domain-based trustworthiness of users and their content in SBD is a crucial tool that helps business firms to listen to the voice of the right customer. Hence, it is crucial to ensure the effectiveness and applicability of the developed approach in order to study the users' behaviours and infer their credibility. This is done by applying the appropriate evaluation metrics and benchmark comparisons to demonstrate the capacity and feasibility of the proposed framework to resolve related real-life problems.

3.6.2 Solution Requirements for Extracting Knowledge and Defining Domain in SBD

Section 3.4.2 addresses the need for a framework designated for semantic data extraction and domain knowledge inference of SBD content. The first requirement for creating such a framework is to build the domain ontology which depicts the domain of interest and gives the meaning of terms in the vocabulary. This ontology will be incorporated to conduct the process of data annotation and enrichment to capture the domain knowledge. Constructing a domain ontology is a basic step and necessary to acquiring a thorough understanding of the domain of discourse. This process requires identifying the formally named and defined classes, types, attributes, and mutual relationship of the entities and concepts that indicate a specific domain of study.

The requirement above is consolidated with the utilisation and reuse of other datasets, semantic repositories and light-weight ontologies to conduct the process of semantically interlinking similar entities. This helps to provide additional metadata for the concepts defined in the ontology which extends the knowledge captured for each annotated entity inferred from the social content. Textual data is identified according to its relevant concepts, properties, and relations captured in an ontology. For examples, Twitter messages posted on a public account are considered as a source. Each word in each message is tokenised. If the tokens or words match the class and properties in an ontology, then they are populated. There is a list of common words which are used to find occurrences of general terms, such as a pronoun etc. There is also a set of rules to identify matched annotations to be populated as instances of certain classes and properties in the ontology. The ontology population is stored in knowledge base tuples which can then be queried or used as valuable information for customers, managers, analytics, or applied to a statistical model to be used in the decision-making process. Once the designated domain of knowledge has been annotated and populated through the ontology instantiations, it can be used to reduce the ambiguity of SBD by providing an accurate interpretation of the context of the users' content. This helps to decrease the variability of BD, and better manage and extract high-level users' domains of knowledge.

The framework developed to extract semantics and infer knowledge form SBD should be implemented in practice and its ability to address the underlying research problem should be validated. Hence, a prototype system as a proof-of-concept is required to evaluate the developed framework in order to demonstrate its effectiveness in accomplishing the intended task and its feasibility in providing a better understanding of the user's domain of interest.

3.6.3 Solution Requirements for Domain-based Classification of SBD

A solution to domain-based classification of the textual content in SBD requires extracting the semantics of the textual data from which meaningful information is derived. As in linguistics, a single term can have different meanings in different contexts. To obtain the right information, we need the right source. In this research, ontology is utilised to obtain the semantics of textual data. This is done by reusing and extending the currently available ontologies that are stored in an ontology repository. By means of an ontology, entities in the extracted textual data are linked with their corresponding concepts. Useful knowledge captured in the ontology can be inferred and used in the analysis process. The conceptualisation of ontology is the formation of knowledge which includes classes (or concepts), attributes (or properties) and relationships (or relations between classes and class members). The specification is the representation of the knowledge in a substantial form. The structure of ontologies depends on the type of knowledge and, importantly, the usage of ontology. Ontology is used in this research to enrich textual data semantics to understand the meaning expressed through the text. This enables the next requirement to be fulfilled, which is the automatic classification and prediction of the domain of interest inferred from the textual content at the user and tweet levels. Hence, consolidated machine learning techniques are incorporated.

Machine Learning is specifically intended for the processing and analysis of data for understanding and building an analytical model capable of learning and making predictions and classifications without explicit programming. The key to the success of these techniques is their ability to understand data without having to provide clear instructions and rules, leaving room for learning and reasoning according to the issue being addressed. In this context, several machine learning modules should be utilised to perform the intended tasks of prediction and classification.

The framework incorporating semantic analysis techniques and machine learning algorithms for domain-based discovery should be evaluated by means of practical experiments carried out to verify the effectiveness and applicability of the framework reinforced by several performance evaluation metrics to show its capacity to perform classification tasks at the user level and tweet level.

3.7 Conclusion

This chapter presents an overview of the research problems addressed in this thesis. The key concepts relevant to this thesis are presented together with their definitions. The research issues are discussed, indicating the necessity to develop approaches for social data analytics that will effectively measure the trustworthiness of SBD users, derive domain knowledge, and infer value from SBD. The DSR methodology adopted in this thesis is described as are the key activities undertaken throughout the research. Finally, the solutions required to address the research issues are described.

The next chapter presents a detailed overview of the framework designed to address and resolve the research problem and related issues presented in Sections 3.3.1 and 3.4.1 respectively. The set of solution requirements in Section 3.6.1 are included in the development of the proposed system. Moreover, Chapter 4 illustrates the effectiveness of the developed system by making several benchmark comparisons and reporting on the rigorously conducted experiments.

Chapter 4 Domain-Based Analysis of Users’ Trustworthiness in Social Big Data

4.1 Introduction

Deep insights into BD require a better understanding of the massive amount of data being generated every second, necessitating the leveraging of new data analysis techniques and the continuous improvement of existing practices. Section 2.1 provides an integrated review of the existing methods and approaches in the context of trustworthiness in SBD. This review indicates the need to develop a framework that will enable a better comprehension of the domain-based trustworthiness of users and their content in OSNs. Assessing the user’s credibility within SBD context is an arduous task due to the massive size and diverse features of SBD in addition to the unstructured nature of social data content.

The preceding chapter presents an in-depth analysis of the related research problem and the underlying issues. In particular, Section 3.3.1 addresses the deficiency of implementing platforms to measure and evaluate the credibility of users in SBD while taking into consideration their domain(s) of interest. These platforms should address the BD features, and facilitate data storage, data processing and data analysis. The research issue elaborated in Section 3.4.1 indicates the necessity to develop a platform to elicit value and determine the veracity of SBD through a domain-based credibility approach. The requirements for this solution platform are presented in Section 3.6.1. These requirements comprise the need for incorporating semantics analytics, sentiment analytics, and the temporal factor.

This chapter presents an overview of an advanced approach designed to address the domain-based credibility in SBD. A detailed description is given of the overall mechanism of the developed approach and includes a metric of the key attributes used to measure the trustworthiness of domain-based users. The last section of this chapter discusses several experiments conducted to demonstrate the applicability of the proposed approach.

4.2 CredSaT System Architecture

This section presents an overview of the framework solution provided to address lack of domain-based credibility approaches. This framework is called CredSaT (***Cre****dibility incorporating* ***S****emantic* ***a****nalysis and* ***T****emporal factor*): a comprehensive SBD framework intended to measure users' credibility is based on their domains of knowledge.

Figure 4-1 illustrates the CredSaT framework. This framework adopts the BD value chain presented by Hu et al. (Han et al. 2014) which covers the life cycle of BD. This chain comprises four main stages: data generation, data acquisition, data storage, and data analysis. Historical and newly-generated tweets with their metadata are collected and pre-processed to ensure dataset cleansing and consistency. Data storage provides distributed BD infrastructure to facilitate data analysis. The data analysis phase is the focal area of this approach. Collected data from the distributed environment are processed in two main analysis stages: (1) Semantic Analysis; (2) Credibility Analysis. Next sub-sections presents a detailed depiction of these stages.

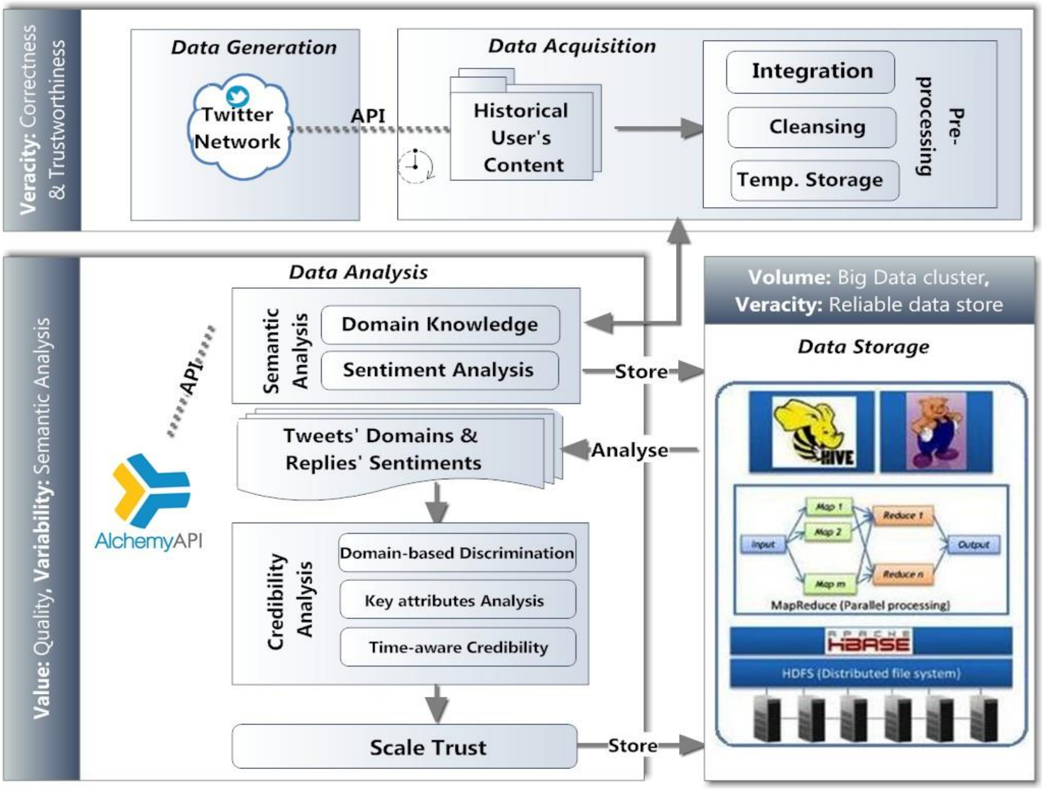


Figure 4-1: CredSaT Framework[1]

4.3 Data Generation and Acquisition

4.3.1 Data Generation

Twitter micro-blogging has been extensively utilised to research several domains of interest. Twitter has been chosen in this research due to the following reasons: (i) Twitter platform has been studied broadly in the research communities (Chen, Madhavan, and Vorvoreanu 2013); (ii) It facilitates retrieving public tweets through providing APIs; (iii) the Twitter messages' "max 140 characters" feature enables data analysis and prototype implementation for a proof of concept purpose. Although this research focuses on the data generated from Twitter, the proposed approach applies to all other social media platforms.

[1] Source of Big Data Infrastructure Design: http://www.bourntec.com/big-data/

4.3.2 Data Acquisition

As previously mentioned, the origin of the data is crucial to ensure the *veracity* of BD. The trustworthiness and certainty of the available data will help to produce reliable and consistent results during the analysis phase. Twitter data access mechanisms have been harnessed in this study for data collection purposes. Users' information and their tweets and all related metadata were crawled using *TwitterAPI* (Makice 2009). A PHP script was implemented to crawl users' content and their metadata using the *User_timeline* API method. This API allows access and retrieve the collection of tweets posted by a certain user_id associated with each API request. This approach is used rather than a keyword search API due to the reasons as follows. Keyword-based search API has certain limitations listed in (Chen, Madhavan, and Vorvoreanu 2013), i.e. Twitter index provides only tweets posted within 6-9 days thus it is hard to acquire historical Twitter dataset before this time span. Further, Search API retrieves results based on the relevance to the query caused by uncompleted results. This implies missing tweets and users in the search results. Using user's timeline approach, on the other hand, retrieves up to 3,200 of the recent users' tweets. Last but not least the purpose of this thesis is to measure the users' trustworthiness hence user-driven tweets collection is the suitable approach. Further, *acTwitterConversation*[2] API was used to retrieve all public conversations related to the tweets being fetched using Twitter API.

Data acquisition is carried out using a PHP script triggered by running a cron job which selects a new user_id and starts collecting historical user information, tweets, replies and the related metadata. The list of Twitterers' user_ids used in the

[2] https://github.com/farmisen/acTwitterConversation

data acquisition phase is extracted from a Twitter graph dataset crawled by Akcora et al. (Akcora et al. 2014). This graph is chosen since it includes the list of users who had less than 5,000 friends in 2013. This threshold was established by Akcora et al. to discover bots, spammers and robot accounts. This threshold is used to measure their credibility as well. This assists in finding domain influencers from a dataset of general users whose domains of knowledge are not explicitly known. Further, the developed framework is capable of identifying anomalous users as illustrated in the evaluation section.

Users' public tweets were crawled during the period from 20/02/2015 to 15/12/2015. The retrieved dataset includes (i) all public metadata for tweets, such as tweet_id, tweet_text, created_at, user_id, retweet_count, favorite_count, replies_count etc.; (ii) public metadata for users, such as user_id, screen_name, name, location, followers_count, friends_count, favourites_count, etc; (iii) all tweet conversations with the related metadata such as replier_username, reply_content, data_of_reply, etc.

Figure 4-2 shows the number of crawled tweets posted between 2006 and 2015. As can be seen in Figure 4-2 there has been a dramatic increase in the number of tweets since 2006. This indicates the great extent to which people are engaging with social media. These social platforms enable them to publish their content, taking advantage of the open environment and fewer restrictions.

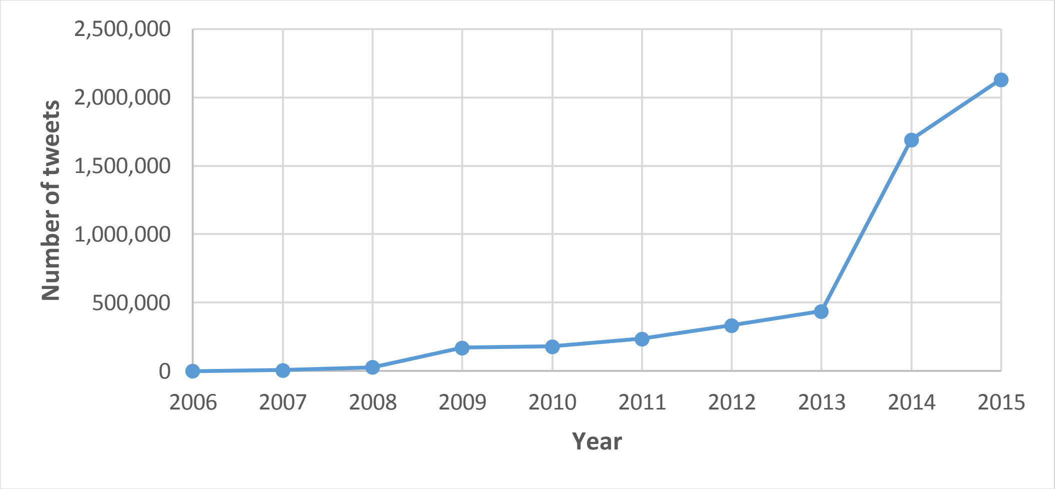


Figure 4-2: Number of Tweets collected per year

4.3.3 Pre-processing phase

To ensure the *veracity* of BD, the accuracy, correctness and trustworthiness of data should be ascertained. Although the data's origin and storage are critical to ensuring the *veracity* of BD, the trustworthiness of the source does not guarantee data correctness and consistency. Data cleansing and integration should also be used to guarantee the *veracity* of data. Further improvements to the quality of the collected data will be discussed later in the analysis phase. To address the data veracity regarding data correctness, the raw extracted tweets were subjected to a pre-processing phase. This phase includes the following steps:

4.3.3.1 Data integration and temporary storage

The raw extracted tweets from TwitterAPI are in JSON format. *AcTwitterConversation* fetches conversation tweets as an array. Data integration will provide a better-structured data format for the next analysis phase. This has been achieved by reformatting and unifying the raw tweets (JSON) and replies (ARRAY) to fit with the relational database model, the design of which is based on the metadata of the tweet, reply and the user. Then, the reformatted data is stored in a temporary location (i.e. MySql database).

4.3.3.2 Data Cleansing

Data at this stage may include many errors, meaningless data, irrelevant data, redundant data, etc. Thus, data is cleansed to remove noisy data and ensure data consistency. The following sequential steps were taken for the data cleansing process (i) all redundant content are eliminated (i.e. the same dataset crawled more than once) such as tweets, replies or users' data and their metadata; (ii) users who posted fewer than 50 tweets were excluded. This particular threshold was experimentally established because the aim of this research is to discover domain-based influential users; thus, it is assumed that those users post a relatively large number of tweets in their domain(s) of interest; (iii) media *URLs* are eliminated such as photos uploaded to Twitter, or media uploaded to one of the popular media sharing websites listed in (Saravanakumar and SuganthaLakshmi 2012) such as Instagram, Flickr, YouTube, and Pinterest. This is because these have no actual text that can be extracted for further analysis. Moreover, *URLs* directing to Facebook websites are eliminated due to the restrictions that apply to the public access of their content.

In the implementation stage of this phase, a set of PHP scripts and MySQL procedures are developed to process the raw data to achieve the above objectives. The pre-processing phase has been implemented and deployed onto the NeCTAR research cloud[3].

[3] https://nectar.org.au/research-cloud/

4.3.4 Data Storage

Data storage is the third phase of the BD lifecycle. *Volume* is an essential dimension to be considered when describing BD. It refers to the vast increase in the data growth where proper tools and techniques are required to manage such huge blocks of data. The data stored in this chain provides distributed and parallel data processing infrastructure based on the *Hadoop*[4] platform for Big Data. Hadoop is a distributed computing platform for data processing. It is an open source project developed by Apache™ to provide scalable, reliable and fault tolerant framework for Big Data. The BD infrastructure at the School of Information Systems, Curtin University, is utilised for data storage. This is a 6-node BD cluster, each with 64 GB RAM, 2 TB Storage, and 8 Core Processors. The temporal-temporary data dumps its contents to this distributed environment after the data integration process. Several *Hive*[5] tables are designed and implemented in this distributed environment. The data dumps are stored in Hive tables to facilitate data access and manipulation through the SQL-like format.

4.4 Domain-based Trustworthiness Analysis

The data analysis phase is the production stage of the BD value chain, and it is the focal area of the proposed approach. The credibility of users, incorporating semantic dimension and the temporal factor, is established as explained in the following sub-sections.

[4] https://hadoop.apache.org/
[5] https://hive.apache.org/

4.4.1 Domain Extraction and Sentiment analysis

Deep insights into BD require new data analysis techniques and the continuous improvement of existing practices. *Variability* (Fan and Bifet 2013) is an important BD dimension. Variability refers to variance in meaning. Incorporating semantic analysis will reduce the ambiguity of the BD. This mitigates its variability (Emani, Cullot, and Nicolle 2015, Hitzler and Janowicz 2013), distinguishes users' domains of interest, and infers their genuine sentiments.

In this context, AlchemyAPI[6] is used as a domain knowledge inference tool to infer the content's taxonomies. AlchemyAPI is a powerful tool and outperforms other entities' recognition and semantic mapping tools such as DBpedia Spotlight7, Extractiv8, OpenCalais9 and Zemanta10 (Rizzo and Troncy 2011b). In addition, in March 2015 IBM has acquired Alchemy for IBM's development of next-generation cognitive computing applications (IBM 2015). AlchemyAPI analyses the given text or URL and categorises the content of the text or webpage according to three domains (taxonomies) with the corresponding *scores* and *confident* values. *Scores* are calculated using AlchemyAPI, range from "0" to "1", and convey the correct degree of an assigned Taxonomy/Domain to the processed text or webpage. *Confidence* is a flag associated with each response, indicating whether AlchemyAPI is confident with the output. AlchemyAPI is used further to identify the overall positive or negative sentiment of the provided content. Table 4-1 shows an example of incorporating this API to extract taxonomies and the sentiment of the provided tweet. As illustrated in

[6] AlchemyAPI is now acquired and supported by IBM, a prestigious software company in 2015, and renamed as Natural Language Understanding (https://www.ibm.com/watson/services/natural-language-understanding/)

[7] http://dbpedia.org/spotlight/

[8] http://wiki.extractiv.com/w/page/29179775/Entity-Extraction

[9] www.opencalais.com

[10] www.zemanta.com

Table 4-1, the content of the tweet is analysed by AlchemyAPI using two main modules: Taxonomy Inference and Sentiment Analysis. The scores are provided for each module to represent the relevancy of the retrieved taxonomy and sentiment to the provided tweets. The taxonomy inference module is used in this research in the domain discovery process, while sentiment analysis is used to discover the sentiments of tweets' replies as discussed later.

Table 4-1: An example of incorporating AlchemyAPI for taxonomies inference and sentiment analysis

Tweet	**"Achieved a new personal record with @RunKeeper: Longest duration in a week #FitnessAlerts"**			
Taxonomy Inference			**Sentiment Analysis**	
Taxonomy	**Score**	**Confidence**	**Sentiment**	**Score**
/sports	0.658138	yes	positive	0.742995
/health and fitness	0.476813	yes		
/sports/running and jogging	0.335338	no		

A tweet's content has one or two main components: *text* and *URL*. Due to the limitation of a tweet's length, a normal or legitimate Twitterer attaches with her tweet a URL to a particular webpage, photo, or video to help her followers obtain further information on the tweet's topic. Twitter scans URLs against a list of potentially harmful websites, then URLs are shortened using *t.co service* to maximise the use of the tweet's length. Anomalous users such as spammers abuse this feature by hijacking trends, using unsolicited mentions, etc., to attach misleading URLs to their tweets. Thus, it is important to study the tweet's domain and the comprised URL's domain to obtain a better understanding of the user's domain(s) of knowledge, which is then used to measure user domain-based credibility.

AlchemyAPI is used to analyse and infer taxonomies of each user's tweet and the website content of the associated URL rather than analysing the user's timeline as one block. This is to provide a fine-grained analysis of tweet data. AlchemyAPI may not able to infer a domain for any particular tweet or URL when the tweet is very short, or the content is unclear or nonsensical or written in a language other than English. Likewise, if the URL is invalid, corrupted or contains non-English content, domains cannot be inferred. Currently, English language contents are the only contents supported by AlchemyAPI in their taxonomy inference technique. Hence, a tweet and its metadata are removed from the dataset if the tweet was written in another language.

For the *veracity* aspect, and as an advanced cleansing approach, the quality of crawled data is improved by selecting tweets and URLs if the domains inferred from embedded content have acquired a score above "0.4" and confident value of "yes" as thresholds. Hence, domains in which the score is below "0.4" or confident value of "no" are skipped. This threshold is selected after noting that the retrieved domains are closely related to the tweets' context when the score is above "0.4", and confident value suggests "yes". This rule is intended to increase the quality and correctness of the retrieved domains, thus satisfying the *veracity* aspect of BD. This will improve the credibility values which will be discussed later. The "minimum 50 tweets" threshold is applied again to the new dataset.

AlchemyAPI is utilised further in this chapter to derive the sentiment of a given reply whether it is positive, negative or natural with the corresponding sentiment score. Consequently, all of a tweet's set of replies are crawled and the sentiments of these replies are incorporated in the analysis to enhance the credibility process as discussed later in this chapter.

Table 4-2: Total count of users, tweets and replies before and after cleansing phase

	Before Cleansing	**After Advanced Cleansing**	**Eliminated (%)**
Total # Users	9,772	7,401	24.3%
Total # Tweets	5,220,478	2,810,362	46.2%
Total # Replies	2,010,992	1,443,932	28.2%

Table 4-2 shows the total number of users, tweets and replies before and after the advanced cleansing process. It is worth noting the importance of data cleansing to purify the raw dataset and enhance its quality. Although the selection criteria for the OSNs' users in this research are quite restricted, the number of eliminated content highlights some significant issues as follows: (i) the quality of contents posted in the social media should be critically studied before conducting further analysis; (ii) it is important in the cleansing phase to ensure the data *veracity* in the BD context; (iii) part of the eliminated tweets were written in a non-English language; however, since these tweets may contain valuable content, sophisticated semantic analysis tools are required to address multilingual contents.

Figure 4-3 displays the list of all domains and corresponding numbers of users, tweets and URLs extracted from the dataset after the cleansing process. It is evident that "arts and entertainment" are the area of most interest to tweeters. For example, 4,550 users tweeted about 'art and entertainment' domain, and the total number of tweets in this category is 471,883. On the other hand, "real estate" seems to be the area of less interest to users; 2,596 and 9,163 are the total numbers of users and tweets respectively. As depicted in Figure 4-3, the number of users in each domain is relatively high. This is because AlchemyAPI inferred a wide range of domains from users’ content. For example, 32% of the users posted at least one tweet or URL for each domain. This issue will be discussed further in the credibility analysis phase.

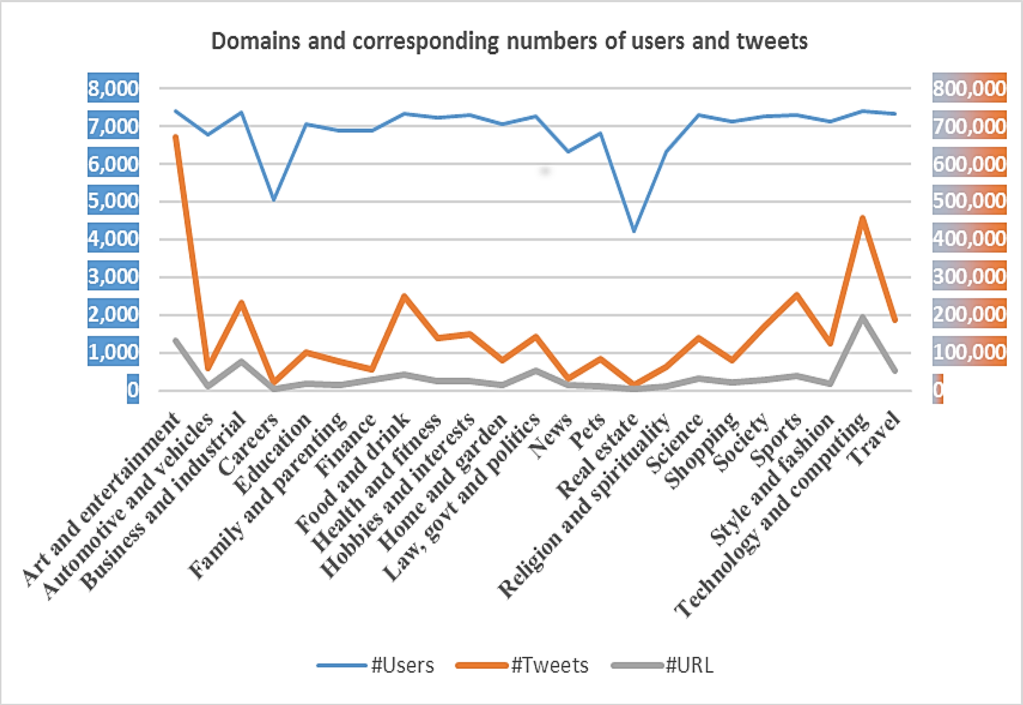


Figure 4-3: Total Numbers of users, tweets and URLs in each taxonomy

4.4.2 Time-aware domain-based user’s credibility analysis

The key challenge for BD analysis is the mining of enormous amounts of data in the quest for added value. *Value* of BD (Kaisler et al. 2013) measures the quality and significance of data with new insights. Acquiring substantial and valuable information from data in big data scale is a vital task. Researchers are trying to capture the *value* of BD in dissimilar contexts. In OSNs, understanding the users’ behaviour is significant due to the dramatic increase in the usage of online social platforms. This indicates the importance of measuring the users’ trustworthiness, thereby discovering users' influence in a particular domain. In this chapter, a domain-based analysis of users’ credibility is suggested to provide a comprehensive, scalable framework. This is achieved by analysing the collection of a user's tweets to measure the initial user's credibility value based on the user's historical data. This is done through the time-aware, domain-based user credibility ranking approach. Table 4-3 describes the notations used in this chapter.

Table 4-3: Notations

Symbol	Description
$\boldsymbol{U}$	Set of the collected user ids in the dataset, $\boldsymbol{|U| = m}$
$\boldsymbol{D}$	Set of domains of knowledge, $\boldsymbol{D = \{d_1, d_2, .., d_n\}, |D| = n = 23}$ (By AlchemyAPI)
$\boldsymbol{Twt\_Sim}$	$\boldsymbol{m \times 1}$ vector of Tweet Similarity Penalty
$\boldsymbol{URL\_Sim}$	$\boldsymbol{m \times 1}$ matrix of URL Similarity Penalty
$\boldsymbol{Sc'}$	$\boldsymbol{m \times n}$ matrix of normalized domain based user score
$\boldsymbol{DF}$	$\boldsymbol{m \times 1}$ vector of domain frequency
$\boldsymbol{IDF'}$	$\boldsymbol{m \times 1}$ vector of normalized inverse domain frequency
$\boldsymbol{W'}$	$\boldsymbol{m \times n}$ matrix of normalized users weight
$\boldsymbol{R'}$	$\boldsymbol{m \times n}$ matrix of normalized domain-based user's retweets
$\boldsymbol{L'}$	$\boldsymbol{m \times n}$ matrix of normalized domain-based user's likes
$\boldsymbol{P'}$	$\boldsymbol{m \times n}$ matrix of normalized domain-based user's replies
$\boldsymbol{SP}$	$\boldsymbol{m \times n}$ matrix of domain-based user positive sentiment replies
$\boldsymbol{SN}$	$\boldsymbol{m \times n}$ matrix of domain-based user negative sentiment replies
$\boldsymbol{S'}$	$\boldsymbol{m \times n}$ matrix of normalized domain-based user sentiments replies
$\boldsymbol{FOL}$	$\boldsymbol{m \times 1}$ vector of total count of users' followers
$\boldsymbol{FRD}$	$\boldsymbol{m \times 1}$ vector of total count of user's friends (followees)
$\boldsymbol{FF\_R'}$	$\boldsymbol{m \times 1}$ vector of normalised user followers-friends ratio
$\boldsymbol{C}$	$\boldsymbol{m \times n}$ matrix of domain-based user credibility
$\boldsymbol{C^T}$	$\boldsymbol{m \times n}$ matrix of domain-based user credibility for the time period $\boldsymbol{T}$
$\boldsymbol{TC}$	$\boldsymbol{m \times n}$ matrix of time-aware domain-based user credibility
$\boldsymbol{TC'}$	$\boldsymbol{m \times n}$ matrix of normalized time-aware domain-based user credibility
$\boldsymbol{I}$	Credibility window of time periods

4.4.2.1 Methodology

One of the main objectives of CredSaT is to discover influential domain-based users from the list of users whose domain(s) of knowledge is tacit, incorporating the temporal factor. Hence, the outcome should be a ranked list of users with a corresponding credibility value for each specific domain. To achieve this purpose, the dataset of a user's data and metadata is divided into several chunks where each chunk represents a specific time period. A metric of credibility measurements is used to evaluate the user's trustworthiness in each particular chuck, thus providing overall credibility values. The mechanism used to calculate a user's value in each step takes into account other users' values, thereby providing a normalisation approach for building the relative ranking list of credibility in each domain. Hence, each particular key value obtained from the user's data and metadata should be measured against other users' values. In other words, each of the key attributes is normalised in each domain by dividing the value of the user's attribute by the maximum value achieved by all users in that domain. The following subsections explain the metric used to measure the domain-based credibility of users in SBD. A list of the technical terminologies included in this chapter is provided to elucidate the proposed approach. For demonstration purposes, the "*Technology and Computing*" domain is selected to illustrate the proposed approach.

4.4.2.2 Distinguishing domain-based OSNs users

The analysis of a user's content to discover his/her main domains of interest is an essential start to the process of measuring the user's credibility. In OSNs, a user ***u*** achieves a higher weight value in a certain domain(s) of knowledge if ***u*** shows a strong interest in these domains through the posted tweets and attached URLs. This weight should be higher than those of other users who posted content in a broad range of domains. This is because no user could be conversant with all domains of knowledge (Gentner and Stevens 1983). Therefore, the theoretical notion of Term Frequency-Inverse Document Frequency (TF-IDF) has been used to distinguish domain-based users of OSNs from others (Abu-Salih, Wongthongtham, and Zhu 2015).

Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF is considered as a core component embodied onto the vector space model (VSM)(Salton, Wong, and Yang 1975) which is one of the classical approaches to Information Retrieval statistical models. “The intuition was that a query term which occurs in many documents is not a good discriminator”(Robertson 2004). This implies that a term which occurs in many documents decreases its weight in general as this term does not show the particular document of interest to the user (Ramos 2003). TF-IDF measures the importance or significance of a term to a certain document exists in a corpus of documents. It comprises standard notions which formulate its structure; Term Frequency (TF): is used to compute the number of times that a term appears in a document. TF expresses the importance of the term in the document; Document Frequency (DF): is a statistical measure to evaluate the importance of a term to a document in a corpus of texts (Rajaraman and Ullman 2011). Inverse Document Frequency (IDF) is a discriminating measure for a term in the text collection. It was proposed in 1972 (Sparck Jones 1972) as a cornerstone of term weighting, and a core component of TF_IDF. It is used as a discriminating measure to infer the term’s importance in a certain document(s) (Robertson and Sparck-Jones 1976). TF_IDF combines the definitions of TF (the importance of each index term in the document) and IDF(the importance of the index term in the text collection), to produce a composite weight for each term in each document. It assigns to a word t a weight in document d that is:

- Highest when t occurs many times within a few number of documents.
- Lower when the term t occurs fewer times in a document d, or occurs in many documents.
- Lowest when the term t exists all documents.

In the context of this research, this heuristic aspect can be incorporated into a model to evaluate the trustworthiness of users. Consequently, it is argued that a user who posts in all domains has a low trustworthiness value in general. This argument is justified based on the following facts: (i) No one person is an expert in all domains (Gentner and Stevens 1983); (ii) A user who posts in all domains does not declare to other users which domain(s) s/he is interested in. A user shows to other users which domain s/he is interested in by posting a wide range of contents in that particular domain; (iii) There is the possibility that this user is a spammer due to the behaviour of spammers posting tweets about multiple topics (Wang 2010). This could end up by tweets being posted in all domains which do not reflect a legitimate user's behaviour. In other words, a user $\boldsymbol{u}$ whose posts in general are discussing a particular domain(s), $\boldsymbol{u}$ gets a higher distinguishing value in this domain(s) and overcomes other users who usually post in a broad range of domains. This involves studying the content of users' tweets and their embedded URLs to obtain a thorough understanding of their domain(s) of knowledge as it will be elaborated through this section. To this end, a domain-based user's content score matrix is proposed:

Domain-based User's Content Scores Matrix (*Sc*): $Sc_{u,d}$ refers to the refinement summing of the corresponding scores achieved by AlchemyAPI for all tweets' texts ($Sc_{u,d}^{Twt}$), and the refinement summing of scores retrieved from URLs' webpage content ($Sc_{u,d}^{URL}$) posted by a user u where a domain d was inferred. It can be calculated as follows:

$$\boldsymbol{Sc_{u,d}} = (\boldsymbol{Twt\_Sim_u} \times \boldsymbol{Sc_{u,d}^{Twt}} + \boldsymbol{URL\_Sim_u} \times \boldsymbol{Sc_{u,d}^{URL}}) \text{ , for each domain d} \quad (4.1)$$

where $\boldsymbol{Sc_{u,d}^{Twt}}$ is computed by adding all scores retrieved from AlchemyAPI of tweets' texts posted by user $\boldsymbol{u}$ in domain $\boldsymbol{d}$, $\boldsymbol{Sc_{u,d}^{URL}}$ is calculated by accumulating scores for all websites' content of the URLs embedded in users $\boldsymbol{u}$'s tweets in domain **d,** $\boldsymbol{Twt\_Sim_u}$ is the *Tweets Similarity Penalty* factor, it can be defined as follows:

User's Tweets Similarity Penalty ($\boldsymbol{Twt\_Sim}$***):*** where $\boldsymbol{Twt\_Sim_u}$ represents the count of unique keywords (#***DistinctWords***) in the overall user's tweets to the total number of keywords in the user's tweet (#***Words***). $\boldsymbol{Twt\_Sim_u}$ can be calculated as:

$$\boldsymbol{Twt\_Sim_u} = \frac{\boldsymbol{\#DistinctWords_u}}{\boldsymbol{\#Words_u}} \quad (4.2)$$

$\boldsymbol{URL\_Sim_u}$ in Eq. (4.2) is the *URL Similarity Penalty* factor, and is defined as follows:

User's URL Similarity Penalty($\boldsymbol{URL\_Sim}$)***:*** where $\boldsymbol{URL\_Sim_u}$ represents the percentage of non-redundant URLs (***#DistinctURLs***) with non-redundant hosts of URLs (***#DistinctURLsHosts)*** to the total number of URLs (***#URLs***) posted by user $\boldsymbol{u}$**;** it is computed as follows**:**

$$\boldsymbol{URL\_Sim_u} = \boldsymbol{0.5}\left(\frac{\boldsymbol{\#DistinctURLs_u + \#DistinctURLsHosts_u}}{\boldsymbol{\#URLs_u}}\right) \quad (4.3)$$

$\boldsymbol{\#DistinctURLs_u}$, $\boldsymbol{\#DistinctURLsHosts_u}$, and $\boldsymbol{\#URLs_u}$ could have the same value; i.e. the user might add a unique *url* for each time the user attaches a *url* to a tweet. Thus, “0.5” is added to normalize $\boldsymbol{URL\_Sim_u}$ value. Table 4-4 shows a list of synthetic tweets to illustrate the idea of similarities penalty. As illustrated in Table 4-4, only the highlighted words are counted in calculating the similarity of tweet texts $\boldsymbol{Twt\_Sim_x}$**.** The $\boldsymbol{Twt\_Sim_x}$ is computed after eliminating the stopwords and URLs from the tweets text. $\boldsymbol{URL\_Sim_x}$ is calculated by extracting all URLs, and finding the non-redundant URLs and hosts.

Table 4-4: Synthetic Tweets to illustrate the use of URL and Tweets’ texts similarities.

List of Tweets	***Tweet1:*** “This website is amazing and useful: http://www.example.com/subdirectory1/index.html” ***Tweet2:*** “Check this website for a recent update: http://www.example.com/index.html” ***Tweet3:*** “Check this website for an update: http://www.example.com/subdirectory2/index.html”						
Words→Count	‘website’ → 3 , ‘amazing’→1, ‘useful’→1, ‘check’→ 2, ‘recent’→1, ‘update’→2						
Distinct Words	‘website’, ‘amazing’, ‘useful’, ‘check’, ‘recent’, ‘update’						
DistinctURLs	‘http:// www.example.com /subdirectory1/index.html’ ‘http://www.example.com/index.html’ ‘http://www.example.com/subdirectory2/index.html’						
DistinctURLs Hosts	www.example.com						
#Distinct Words	6	***#Words***	10	$\boldsymbol{Twt\_Sim_x}$	0.6	***#Distinct URLs***	3
#DistinctURLs Hosts	1	***#URLs***	3	$\boldsymbol{URL\_Sim_x}$	0.666	***#Distinct URLs Hosts***	1

Assumption*:* *URLs similarity penalty* and *Tweets similarity penalty* are proposed to address the similarities of embedded URLs and texts of the users’ tweets. Generally, legitimate (normal) users who are knowledgeable or influencers in a certain domain(s) do not post the same tweet(s) repeatedly or embed in their tweets in the same URL(s). Users who post the same content are likely to be engaging in the anomalous behaviour, and they might be spammers, or campaign promoters, etc. In this research, the intention is to distinguish between the knowledgeable domain-based users and users engaged in the anomalous behaviour. Therefore, the aforementioned user categories should be assigned less weight than the normal user category; hence, $\boldsymbol{Twt\_Sim_u}$ and $\boldsymbol{URL\_Sim_u}$ are proposed as penalty factors applied to the scores given to the text of the user's tweet or content of embedded URLs' websites.

AlchemyAPI infers a maximum of three different taxonomies for each text or webpage; however, the corresponding scores are considered as an important factor. Taxonomy with a score of value '1' should acquire a higher weight than taxonomy with a score of value '0.4'. Thus, $\boldsymbol{Sc_{u,d}}$ is proposed which accumulates for user $\boldsymbol{u}$ the refinement resultant scores of domain $\boldsymbol{d}$ from the list of historical user's tweets and the websites' content of the attached URLs.

Table 4-5 shows the highest five $\boldsymbol{Sc}$ scores in the "Technology and Computing" domain for the list of users in the dataset. For each user, the table shows the total number of cleansed tweets***(#Twts)***, number of tweets where the "Technology and computing" domain was inferred***(#DomainTweets)***. The user's content scores achieved only from the tweets analysis $\boldsymbol{Sc_u^{Twt}}$, and the user's scores achieved only from the URLs analysis ($\boldsymbol{Sc_u^{URL}}$).

Table 4-5: Domain-Based User Score $\boldsymbol{Sc_{u,d}}$

Technology and Computing												
Twitterer	***#Twts***	***#Domain Tweets***	$\boldsymbol{Sc_u^{Twt}}$	$\boldsymbol{Sc_u^{URL}}$	***#URLs***	***#Distinct URLs***	***#Distinct URLsHosts***	$\boldsymbol{URL\_Sim_u}$	***#Words***	***#Distinct Words***	$\boldsymbol{Twt\_Sim_u}$	$\boldsymbol{Sc_u}$
PaaSDev	1932	1541	1074.5	1168.8	2432	1397	595	0.4095	10,615	4005	0.3773	884.1
misterron	**2307**	**1588**	**1438.2**	**882.9**	**3381**	**1348**	**18**	**0.2020**	**9,987**	**3146**	**0.315**	**631.4**
eBasiony	2314	1254	798.7	968.4	3922	2251	10	0.2882	1,1970	4286	0.3581	565.2

Rap_Payne	1619	1159	834.8	727.6	2282	1368	121	0.3262	10,615	4072	0.3836	557.6
RealWireFeed	1397	885	580.5	677.3	1934	1376	5	0.3570	10,442	4029	0.3858	465.8

Table 4-5 further shows for each user the count of embedded URLs (***#URLs***), count of non-redundant embedded URLs(***#DistinctURLs***), count of non-redundant sites' hosts of the embedded URLs(***#DistinctURLsHosts***), Total count of words in users' tweets(***#Words),*** Total count of unique words in user's tweets (#***DistinctWords***), the computed URL penalty factor ($\boldsymbol{URL\_Sim_u}$), and the tweet's text penalty factor($\boldsymbol{Twt\_Sim_u}$). For example, @***misterron*** has reattached the same full URL (including the path with the query string) in **2.5** tweets (i.e. **3381/1348**), and the number of hosts (domains) to which these links are pointing equals "**18"**. The *url similarity penalty* calculated for this user $\boldsymbol{URL\_Sim}_{\textbf{misterron}} = \mathbf{0.5}\left(\frac{\mathbf{18+1,348}}{\mathbf{3,381}}\right) = \mathbf{0.202}$. Furthermore, after aggregating the overall tweets' content of the user @***misterron***, it turns out that every word was repeated around three times, thus the *tweet similarity penalty* for this user $\boldsymbol{Twt\_Sim}_{\textbf{misterron}} = \left(\frac{\mathbf{3,146}}{\mathbf{9,987}}\right) = \mathbf{0.315}$.

Although the figures presented in Table 4-5 imply a strong interest for those users in the "Computer and Technology" domain, it is important to obtain an overall understanding of the domain(s) in which each user is interested; hence, domain frequency ($\boldsymbol{DF_u}$) is incorporated which calculates the number of domains the user $\boldsymbol{u}$ has tweeted about. To distinguish users among the list of their domains of interest, the inverse domain frequency set **(**$\boldsymbol{IDF}$**)** is implemented as follows:

$$\boldsymbol{IDF_u = log(\frac{n}{DF_u})} \qquad (4.4)$$

Where $\boldsymbol{n}$ = the number of domains, $\boldsymbol{DF_u}$ = the domain frequency for user $\boldsymbol{u}$.

$\boldsymbol{IDF_u}$ is proposed to assign the user a high weight if her content and embedded URLs are indicating only a few domains. On the other hand, $\boldsymbol{IDF_u}$ assigns a low weight to a user if the content of her tweets and URLs are indicating a large range of domain(s).

The last step of this phase is used to calculate the weight for each user in each domain by combining the following factors: (i) domain-based user's content scores $\boldsymbol{Sc}$ (users' interest in each domain); (ii) the normalized inverse domain frequency $\boldsymbol{IDF}$ (distinguish users amongst domains of interest) as follows:

$$\boldsymbol{W_{u,d}} = \boldsymbol{Sc_{u,d}} \times \boldsymbol{IDF_u}\,, \qquad where\ \boldsymbol{Sc_{u,d}} > \boldsymbol{\rho} \qquad \text{,for each domain } \boldsymbol{d} \qquad (4.5)$$

where $\boldsymbol{W_{u,d}}$ represents the weight of each user $\boldsymbol{u}$ obtained in the domain $\boldsymbol{d}$, $\boldsymbol{\rho}$ is a threshold value provided as a fine-tuning parameter representing the minimum total scores for each user in each domain ($\boldsymbol{\rho}$ is set to '10' for demonstration purposes). The imposition of this threshold is intended to provide more accurate and reasonable results. In particular, having this threshold means that the small $\boldsymbol{Sc}$ scores achieved by each user in each domain will be disregarded. This is because small user's scores in each domain could end up decreasing the overall discriminating value of this user in all domains. These small scores should be considered due to the following: (i) incorrect domain assignment may occur to a tweet in the domain analysis phase that assigns a user's tweet to an unrelated domain(s); (ii) users may deviate from their domain of expert to discuss general, unrelated or trending topics. Hence, the fine-tuning parameter $\boldsymbol{\rho}$ is proposed as a thresholding value when counting the total scores for each user in each domain to provide more precise and reasonable results.

$W_{u,d}$ assigns to a user u a weight in domain d that is: (i) highest when a user

u has a Large number of tweets within a tiny number of domains; (ii) lower when a user u has fewer tweets in a particular domain(s) or has tweets in a wide range of domains; (iii) lowest when a user u tweets in all domains since this user does not declare which domain (s)he is interested in.

Table 4-6: Highest weight values of users in Technology and Computing Domain ($\mathbf{W_{u,d}}$)

Technology and Computing									
Twitterer	**#Total Tweets**	**#Domain Tweets**	DF_u	IDF_u	$URL\_ Sim_u$	$Twt\_ Sim_u$	Sc_u	W_u	W'_u
TheBFF	710	670	4	0.76	0.273	0.379	353.0	268.3	1.000
kennethvs	498	441	2	1.061	0.372	0.302	207.2	219.8	0.819
Morgancomputers	339	296	1	1.362	0.254	0.281	141.2	192.3	0.717
PaaSDev	1,932	1,541	14	0.216	0.410	0.377	884.1	191	0.712
chuckfreeze	1,735	930	3	0.885	0.315	0.231	212.6	188.1	0.701

Table 4-6 presents a list of users who achieved the highest weight values in the "Technology and Computing" domain. Entries of $\boldsymbol{W}$ are normalized into the range of [0, 1] by dividing each entry by the maximum weight values of the corresponding domain. For example, all users' weight values in the domain d are normalized as follows:

$$\boldsymbol{W'}_{\boldsymbol{u,d}} = \frac{\boldsymbol{W_{u,d}}}{\boldsymbol{max(W_{*d})}}, \quad \text{for each domain } d \tag{4.6}$$

where $\boldsymbol{W'}_{\boldsymbol{u,d}}$ is the normalized weight of user u in domain d, $\mathbf{max}(\boldsymbol{W}_{*\boldsymbol{d}})$ represents the maximum weight value of all collected users in domain d. As illustrated in Table 4-6, *@TheBFF* has achieved the highest weight value in the "Technology and Computing" domain compared with other users in the crawled dataset. This occurred because this user had the following advantages: (i) his domains of interest are relatively few (4 out of 23); (ii) 94% of the total tweets discussed "Technology and Computing" topics; (iii) the score he achieved in this domain is quite high (353.082).

The weight assigned to each user in each domain is important to address the interests of OSNs' users; however, this is insufficient to rank users based merely on their domains of interest. This dimension will be addressed in the next section through the analysis of users' metadata to obtain a comprehensive insight into their behaviour based on their interactions with other users in the OSN platform.

4.4.2.3 Feature-based user ranking

It is important to have an understanding of the interactions-based attributes of OSN users, as this is a significant factor when discovering socially reliable, domain-based users. This involves studying the followers' interest in the users' content, their positive or negative opinions, etc. In this section, a metric incorporating several key attributes are used to build the feature-based ranking model.

As mentioned previously, AlchemyAPI infers a maximum of three taxonomies for each processed text (i.e. tweet's text or URL's website content). The tweets' metadata (such as #likes, #Retweet, #Replies, etc.) does not indicate the particular domain in which the follower has valued the tweet. Hence, the user's scores produced by AlchemyAPI for each domain are used to provide a weighting distribution mechanism for all metadata items in the inferred domains; this mechanism is termed in this research by *domain-base relativeness factor*. More details will be provided under each feature in the following subsections.

Domain-based user retweet matrix ($\boldsymbol{R}$), where $\boldsymbol{R_{u,d}}$ represents the frequency of retweets for user' content in each domain $\boldsymbol{d}$.

The *domain-base relativeness factor* is used to calculate $\boldsymbol{R_u}$ based on the $\boldsymbol{u}$'s score obtained for each domain $\boldsymbol{d}$. In particular, total count of retweets "*retweet_count*" is distributed among the $\boldsymbol{u}$'s domain(s) based on her score for each one. For example, suppose the domain-base scores spreading for a tweet $(\boldsymbol{t_x})$ posted by user $\boldsymbol{u}$ is (1, 0.5, and 0.5) in ("Sports", "Arts and Entertainment", and "Education") domains respectively, and the total retweets of $\boldsymbol{u}$**'**s tweet **= 10,** then the distribution number of retweets for user $\boldsymbol{u}$ is **(**$\boldsymbol{R_{u,sports}}$ **= 5,** $\boldsymbol{R_{u,arts}}$ **= 2.5,** $\boldsymbol{R_{u,education}}$ **= 2.5).** $\boldsymbol{R}$ is normalized as follows:

$$\boldsymbol{R'_{u,d}} = \frac{\boldsymbol{R_{u,d}}}{\boldsymbol{max(R_{*d})}} \quad \text{, for each domain } d \qquad (4.7)$$

Where $\mathbf{max}(\boldsymbol{R_{*d}})$ is the maximum count of retweets obtained for all users' content in domain $\boldsymbol{d}$.

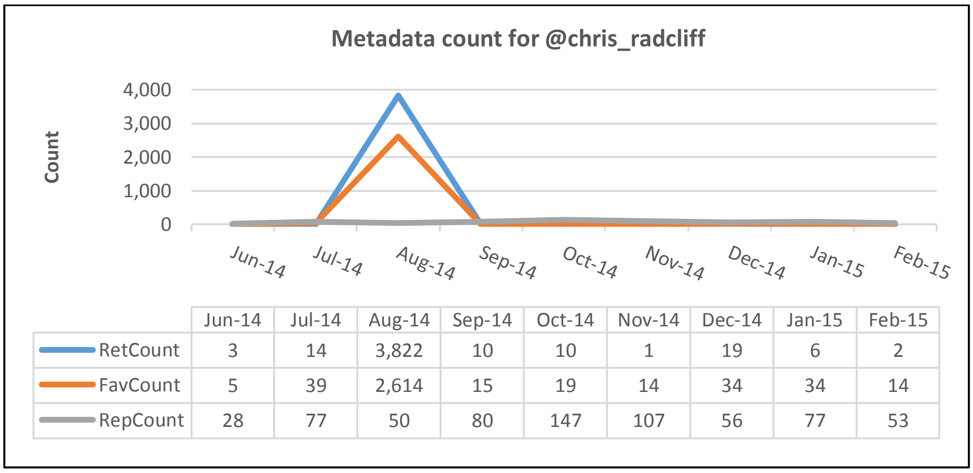
It is evident that the crawled dataset for any user might contain one or more of the following categories: original tweets, retweets or replies to other tweets. The content of retweets has been retained and used for domain discovery purposes. When a user retweets a certain tweet $\boldsymbol{t_y}$ then this user supports the context of $\boldsymbol{t_y}$ despite $\boldsymbol{t_y}$ being originated by someone else. However, all retweets with the associated metadata have been eliminated and are not counted for credibility purposes. This is because the metadata such as (retweet_count, favorite_count, and replies_count) which are associated with this tweet's category indicate the original tweet and cannot be used to support the credibility of the reTwitterer.

Table 4-7: Domain-based User Retweets $R_{u,d}$

Technology and Computing					
Twitterer	***#Total Tweets***	***#Domain Tweets***	W'_u	R_u	R'_u
chris_radcliff	**768**	**148**	**0.074**	**3831**	**1**
nfreader	542	206	0.174	962	0.251
nukeador	165	44	0.076	627	0.164
IvorCrotty	1841	398	0.032	604	0.158
LocalJoost	609	249	0.180	398	0.104

The Twitterer ***@chris_radcliff,*** shown in Table 4-7, achieved the highest percentage of domain-based retweets although this user acquired a relatively low weight in the "Tech. and Comp." domain ($\boldsymbol{W'_{chris\_radcliff}} = \mathbf{0.074}$). Figure 4-4 depicts the total count of retweets, favourites, and replies obtained for ***@chris_radcliff***'s content each month. It is evident that the total count of retweets for this users' content reached a peak in Aug-2014; this is due to one of his tweets[11] posted that month which has been retweeted a relatively high number of times (***3813*** *retweets*), and the total retweets count for the user content in Aug-2014 (3,822). However, the average retweets count of this user's content in other months equals "**8.125**" retweets. Tracing retweet counts by time is important to measure the consistent interest in a user's content temporally, and this applies to all other metadata attributes. This accentuates the importance of measuring the credibility of users incorporating the temporal factor. This dimension will be addressed later in this chapter.

[11] The tweet can be viewed through this link: https://twitter.com/chris_radcliff/status/504400669571178496



	Jun-14	Jul-14	Aug-14	Sep-14	Oct-14	Nov-14	Dec-14	Jan-15	Feb-15
RetCount	3	14	3,822	10	10	1	19	6	2
FavCount	5	39	2,614	15	19	14	34	34	14
RepCount	28	77	50	80	147	107	56	77	53

Figure 4-4: Metadata count over time for @chris_radclif

Domain-based user likes matrix (***L***), where $\boldsymbol{L_{u,d}}$ represents the percentage of likes/Favourites count for the users' content in each domain ***d***. ***L*** is normalized as follows:

$$L'_{u,d} = \frac{\mathbf{L_{u,d}}}{\mathbf{max(L_{*d})}} \qquad \text{,for each domain } d \qquad (4.8)$$

Where $\mathbf{max}(\boldsymbol{L_{*d}})$ is the maximum percentage of likes/Favourites obtained for all users' content in domain ***d***. "*fav_count*" metadata value is distributed based on the *domain-base relativeness factor* mechanism.

Table 4-8: Domain-based User Likes $\boldsymbol{L_{u,d}}$

Technology and Computing					
Twitterer	*#Total Tweets*	*#Domain Tweets*	$\boldsymbol{W'_u}$	$\boldsymbol{L_u}$	$\boldsymbol{L'_u}$
chris_radcliff	768	148	0.074	2615.6	1
tigga7d6	2560	1696	0.339	1274.1	0.251
nfreader	542	206	0.174	816.8	0.166
scout2i	1626	1005	0.409	659.2	0.163
SpnMaisieDaisy	1836	212	0.028	585.9	0.104

Table 4-8 illustrates the top five values in ***L*** for the "Tech. and Comp."

domain. @***chris_radcliff*** has achieved the highest value due to the popularity of aforementioned tweet which posted in Aug-2014 (***2,614*** *Total Likes as illustrated in* Figure 4-4). Despites these figures, the high numbers of domain-based retweets, or likes in a certain domain, do not necessary indicate an influential user in that domain and vice versa. For example, a celebrity might post a tweet about a certain trending topic which is not particularly related to his/her main area of interest(s). It stands to reason that this user will receive a high number of retweets, replies, or likes due to her popularity. This emphasises the importance of acquiring a thorough understanding of the user's data and metadata, thereby providing a correct inference of the users' domains of knowledge.

Domain-based user replies matrix **(*P*)**, where $\boldsymbol{P_{u,d}}$ embodies the count of replies to the users' content in each domain $\boldsymbol{d}$. ***P*** is normalized as follows:

$$\mathbf{P'_{u,d}} = \frac{\mathbf{P_{u,d}}}{\mathbf{max(P_{*d})}} \text{ ,for each domain } d \qquad (4.9)$$

Where $\mathbf{max}(\boldsymbol{P_{*d}})$ is the maximum percentage of replies obtained for all users' contents in domain $\boldsymbol{d}$. "*replies_count*" metadata is distributed based on *domain-base relativeness factor* mechanism**.** Still**,** the domains associated with the content of tweets' replies can be analysed to extract the actual domain(s) of each reply. This will be addressed in the future research to enhance the entries of ***P***. Table 4-9 shows the list of highest domain-based replies values in ***P***.

Table 4-9: Domain-based User's Content Replies $\mathbf{P_{u,d}}$

Technology and Computing					
Twitterer	***#Total Tweets***	***#Domain Tweets***	$\boldsymbol{W'_u}$	$\boldsymbol{P_u}$	$\boldsymbol{P'_u}$
tigga7d6	2560	1696	0.339	1908	1

Technology and Computing					
grahamgilbert	1040	432	0.220	992	0.52
Xantiriad	2298	577	0.084	992	0.52
Aurynn	2222	558	0.117	985	0.516
markdrew	2005	731	0.072	917	0.481

Although Table 4-9 shows that the top five users achieved the highest number of replies in the “Tech. and Comp” domain, the sentiments expressed in these replies should be considered to obtain a better understanding of the repliers’ opinions about users’ content. In OSNs, sentiment analysis has been utilised in several aspects of research. In the context of social trust, frameworks have been developed to analyse the trustworthiness of users’ content taking into consideration the overall feelings towards what users expose in their content. However, these efforts did not analyse the sentiment in a post’s replies in evaluating the trustworthiness of users and their content. The following are the features proposed to address the analysis of replies regarding sentiment.

Domain-based user positive sentiment replies matrix $(\boldsymbol{SP})$, where $\boldsymbol{SP}_{\boldsymbol{u},\boldsymbol{d}}$ refers to the sum of the positive scores of all replies to a user $\boldsymbol{u}$ in domain $\boldsymbol{d}$. Positive scores are achieved from AlchemyAPI with values greater than “0” and less than or equal to “1”. The higher the positive score, the greater is positive attitude the repliers have to the users’ content.

Domain-based user negative sentiment replies matrix $(\boldsymbol{SN})$, where $\boldsymbol{SN}_{\boldsymbol{u},\boldsymbol{d}}$ represents the sum of the negative scores of all replies to a user $\boldsymbol{u}$ in domain $\boldsymbol{d}$. Negative scores are those values greater than or equal to “-1” and less than “0”. The lower the negative score, the greater is the negative attitude the repliers have to the users' content.

Domain-based user sentiments reply matrix ($\boldsymbol{S}$), where $\boldsymbol{S_{u,d}}$ embodies the difference between the positive and negative sentiments of all replies to user $\boldsymbol{u}$ in the domain $\boldsymbol{d}$. $\boldsymbol{S}$ is normalized as follows:

$$\boldsymbol{S'_{u,d}} = \frac{\boldsymbol{S_{u,d} - min(S_{*d})}}{\boldsymbol{max(S_{*d}) - min(S_{*d})}} \quad \text{,for each domain } \mathbf{d} \qquad (4.10)$$

where $\boldsymbol{S_{u,d} = SP_{u,d} - |SN_{u,d}|}$

$\boldsymbol{S_{u,d}}$ embodies the difference between the positive scores and the negative scores for all replies to user $\boldsymbol{u}$ in domain $\boldsymbol{d}$. $\mathbf{max}(\boldsymbol{S_{*d}})$ represents the maximum differences between the positive and negative replies to all collected users in domain $\boldsymbol{d}$. $\mathbf{min}(\boldsymbol{S_{*d}})$ represents the minimum difference between the positive and negative replies to all collected users in domain $\boldsymbol{d}$. It is evident that the list of replies could include responses from the tweet's owner as a part of the conversation. All replies posted by the tweet's owner are eliminated from the conversation and are not included in the above equations. This is to provide accurate sentiments results which reflect the actual positive or negative opinions of the tweet expressed by its followers. The entries of $\boldsymbol{SP}$ and $\boldsymbol{SN}$ are computed using the domain-base relativeness factor mechanism. For example, suppose replies_count for the tweet $(\boldsymbol{t_x})$ of the example mentioned before is equal to **10**, and the sum of the positive and negative replies for $\boldsymbol{t_x}$ are (**15, -10**) respectively, then the dispersal of the positive scores amongst the extracted domains will be **(**$\boldsymbol{SP_{u,sports}}$ **= 7.5,** $\boldsymbol{SP_{u,arts}}$ **= 3.75,** $\boldsymbol{SP_{u,education}}$ **= 3.75),** and the dispersal of the negative scores is **(**$\boldsymbol{SN_{u,sports}}$ **= -5,** $\boldsymbol{SN_{u,arts}}$ **= -2.5,** $\boldsymbol{SN_{u,education}}$ **= -2.5).** Table 4-10 shows the top-5 $\boldsymbol{S_u}$ scores for the list of users in the dataset. It is worth noting that some users received strongly positive sentiments toward their content despite the fact that their domain-based number of tweets is relatively low. This is evident because followers establish their opinion of the user's content by considering the quality rather the quantity of his/her content. This involves creating new, unique, valuable and domain-related content, which is received well by her audience. Furthermore, none of the top five users listed in Table 4-9 are mentioned in Table 4-10. This implies that if a user **u** received a relatively high number of replies, this does not necessarily reflect a positive attitude toward her content. Therefore, studying the sentiment in the content's replies is a significant way of ascertaining the users' actual feelings. The correlation between all entries of $\boldsymbol{S}$ and $\boldsymbol{P}$ will be provided later in this chapter.

Table 4-10: Domain-based user sentiments replies $\mathbf{S_{u,d}}$

Technology and Computing							
Twitterer	*#Total Tweets*	*#Domain Tweets*	$W'_{u,d}$	SP_u	SN_u	S_u	S'_u
scout2i	1626	1005	0.33	75.198	-13.434	61.764	1
agardnahh	815	520	0.386	67.483	-9.570	57.913	0.988
CodrutTurcanu	2251	1100	0.383	60.068	-7.580	52.488	0.971
johnjwall	1695	229	0.285	70.107	-21.318	48.789	0.96
MLanghans410	840	632	0.276	63.303	-16.022	47.281	0.955

The last dimension from the list of user's key attributes is the relationship between the number of followers and friends of each user. Twitter applies certain rules to ban the following aggressive behaviour; Twitter defines the aggressive following as "indiscriminately following hundreds of accounts just to garner attention" (Twitter). Twitter limits the total number of users that a user can follow to 2,000 users. Any addition to this number requires an addendum to the list of followers first; hence, the follower-following relationship remains balanced. The dramatic increase of friends that a user u follows compared to the steadiness in the number of followers is considered to be suspicious behaviour, and such a user is most likely to be a spammer (Twitter 2009, Wang 2010). This relation has been incorporated in the literature to measure the credibility of the OSNs' users; Wang (Wang 2010) used this relation to provide a reputation measurement for the user. This measurement tool is improved in this research as follows:

User Followers-Friends Relation matrix **($\boldsymbol{FF\_R}$),** where $\boldsymbol{FF\_R_u}$ refers to the difference between the number of followers and friends that user $\boldsymbol{u}$ obtains to the *age* of user's profile. $\boldsymbol{FF\_R_u}$ is calculated as follows:

$$\boldsymbol{FF\_R_u} = \begin{cases} \frac{FOL_u - FRD_u}{Age_u}, & if\ FOL_u - FRD_u \neq 0 \\ \frac{1}{Age_u}, & if\ FOL_u - FRD_u = 0 \end{cases} \qquad (4.11)$$

Where $\boldsymbol{FOL_u}$ is the number of $\boldsymbol{u}$'s followers, $\boldsymbol{FRD_u}$ is the number of $\boldsymbol{u}$'s friends, and $\boldsymbol{Age_u}$ is the age of $\boldsymbol{u}$'s profile in years. The variance between the numbers of followers and friends could be due to the profile's age. Users who obtained a dramatic positive difference between number of followers and friends during a relatively short period have an advantage over those who have achieved the same difference albeit over a long period of time. $\boldsymbol{FF\_R_u}$ is normalised as follows:

$$\boldsymbol{FF\_R'_u} = \frac{\boldsymbol{FF\_R_u - min(FF\_R)}}{\boldsymbol{max(FF\_R) - min(FF\_R)}} \qquad (4.12)$$

Where $\mathbf{max}(\boldsymbol{FOL})$ is the maximum *Followers-Friends Ratio* value of all users in the network, $\mathbf{min}(\boldsymbol{FRD})$ is the minimum *Followers-Friends Ratio* value of all users in the network. Table 4-11 shows the list of users who achieved the highest $\boldsymbol{FF\_R'_u}$ values. It is evident that $\boldsymbol{FF\_R'_u}$ key attribute is not a good measurement to rank the domain-based users per se; users with high $\boldsymbol{FF\_R'_u}$ might obtain a general reputable position, and they are highly unlikely to be spammers. However, it is sometimes difficult to convey the main topic(s) of interest to those users with high $\boldsymbol{FF\_R'_u}$ value by studying the relatively few number of user tweets such as in the ***@kyrii*** case**.**

Table 4-11: Twitter Followers - Friends Ratio, and #Tweets in Technology and Computing domain

Twitterer	#Total Tweets	#Domain Tweets	$W'_{u,d}$	FOL_u	FRD_u	Age_u	$FF\_R'_u$
michaelfrisby	433	64	0.049	4150	29	7	1
roseandgrey	293	45	0.034	4686	733	7	0.972
brettdetar	535	54	0.039	4037	121	7	0.966
captdirectory	140	75	0.043	4501	660	7	0.953
kyriii	122	48	0.121	4852	119	9	0.927

4.4.2.4 Correlation between the key attributes

This section provides the correlation between the six key attributes in “Technology and Computing” domain. These key attributes are used for the credibility evaluation. For each key attribute, a ranking list of users has been constructed, and the *Pearson Correlation Coefficient* is computed between each ranking list as depicted in Table 4-12. The correlation between $\boldsymbol{R'}$ and $\boldsymbol{L'}$ is the highest, this is evident because it is likely that “fav/like” activity is associated with “retweet” activity. It is apparent that the correlation between the ranking list of Follower-Friends Ratio $\boldsymbol{R\_FF\_R'}$ and other key attributes is the weakest; this implies that obtaining a high number of followers and a low number of friends does not necessarily achieve high values in the number of domain-based tweets, retweets, replies , etc. and vice versa. For example, the top five users indicated in Table 11 obtained the highest $\boldsymbol{FF\_R'}$ values; however, they did not achieve the same position in other matrices that have already been calculated for the main key attributes of the "Technology and computing" domain. Despite the fact that one or more of those users might obtain dominant positions in other domains of knowledge, their high $\boldsymbol{FF\_R'}$ values alone should not be considered to indicate their influence. In other words, all associated users' data and metadata should be studied comprehensively to determine domain-based influencers.

Table 4-12: Correlation between the ranking lists of all key attributes in the Technology and Computing Domain

$\boldsymbol{vs}$	$\boldsymbol{R\_W'}$	$\boldsymbol{R\_R'}$	$\boldsymbol{R\_L'}$	$\boldsymbol{R\_P'}$	$\boldsymbol{R\_S'}$	$\boldsymbol{R\_FF\_R'}$
$\boldsymbol{R\_W'}$	1	-	-	-	-	-
$\boldsymbol{R\_R'}$	0.5167	1	-	-	-	-
$\boldsymbol{R\_L'}$	0.5070	**0.7651**	1	-	-	-
$\boldsymbol{R\_P'}$	0.3919	0.5931	0.7087	1	-	-
$\boldsymbol{R\_S'}$	0.3427	0.4082	0.4398	0.4409	1	-
$\boldsymbol{R\_FF\_R'}$	0.0969	0.2979	0.2570	0.2022	0.1262	1

4.4.2.5 Monitoring user's credibility over time

So far, the list of features used to measure the credibility of OSNs users is presented. It is evident that the individual preliminary results of each key attribute cannot be used to judge the credibility of the user as such; to ascertain credibility, all available data and metadata should be analysed thoroughly to produce an accurate measurement of the trustworthiness of users. An initial holistic domain-based user credibility formula incorporating all key attributes is proposed as follows:

$$\boldsymbol{C_{u,d} = \alpha * FF\_R'_u + \beta * W'_{u,d} + \gamma * R'_{u,d} + \delta * L'_{u,d} + \theta * P'_{u,d} + \vartheta * S'_{u,d}} \qquad (4.13)$$

Where $\boldsymbol{C_{u,d}}$ represents the user $\boldsymbol{u}$'s credibility in domain $\boldsymbol{d}$, while (**α**, **β**, **γ**,

$\boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{\vartheta}$) are introduced to adjust the significance of each key attribute ($where\ \boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\delta} + \boldsymbol{\theta} + \boldsymbol{\vartheta} = \mathbf{1}$). Although $\boldsymbol{C_{u,d}}$ provides a broad view of the user's trustworthiness in each domain of knowledge, the temporal factor is necessary to observe the user behavior over time thus consolidate the proposed approach. For example, the total number of domains inferred from the user's content could increase or decrease by time. Because this is subject to change, it is worth monitoring the users' interest(s) temporally and indicating their credibility values accordingly. The temporal factor is significant due to the following observations: (i) At time t a user u is likely to be more trustworthy than a user v whose vivacity is low, considering both users hold the same trustworthiness values at time $t - 1$. (ii) Similarly, if a user u has shown a dramatic decrease over time in one or more of ($R_{u,d}, L_{u,d}, P_{u,d}$ $and\ FF\_R_u$) ratios, this implies a reduction in the u's trustworthiness value and vice versa.(iii) Spammers' behaviours are unsteady as they are not legitimate users although they pretend to be. Hence, their "temporal patterns of tweeting may vary with frequency, volume, and distribution over time"(Yardi et al. 2009). The temporal dimension is addressed in this research through; (i) dividing all tweets with all related metadata into chunks, where each chunk includes user's tweets and their metadata of a particular period; (ii) calculate the domain-based trust based on the steps provided in the aforementioned trustworthiness mechanism. The only feature that is common in these chunks is the Twitter Follower-Friends FF_R_u Ratio. This is because one snapshot for FF_R_u has been obtained which represents the follower to friends relationship at the time of crawling Twitter. The temporal factor is assimilated as follows:

$$\boldsymbol{TC_{u,d}} = \frac{\sum_{k=1}^{I} \boldsymbol{w(k)} \times \boldsymbol{C_{u,d}^{k}}}{\sum_{k=1}^{I} \boldsymbol{w(k)}} \qquad (4.14)$$

where $\boldsymbol{TC_{u,d}}$ is the new time-aware domain-based credibility of user $\boldsymbol{u}$ in domain $\boldsymbol{d}$, $\boldsymbol{C^k}$ is the domain-based credibility matrix which is calculated for the time period $\boldsymbol{k}$, and $\boldsymbol{w(k)}$ is a weighting function introduced to provide a weighting mechanism for each credibility value of each time period. $\boldsymbol{I}$ is a *credibility window* defined as follows:

Credibility Window (***I***): is the number of the Twitter datasets corresponding to several recent and sequential time periods.

Figure 4-5 depicts the proposed idea. For example, if ***I*** **= 6,** this indicates six timely sequential snapshots of the users' data and metadata (such as the last six years, months, weeks, etc.) which are incorporated to measure the credibility of users in each period, and the overall credibility $\boldsymbol{TC}$. The users' credibility values in all time periods are indexed sequentially starting from the oldest time period.

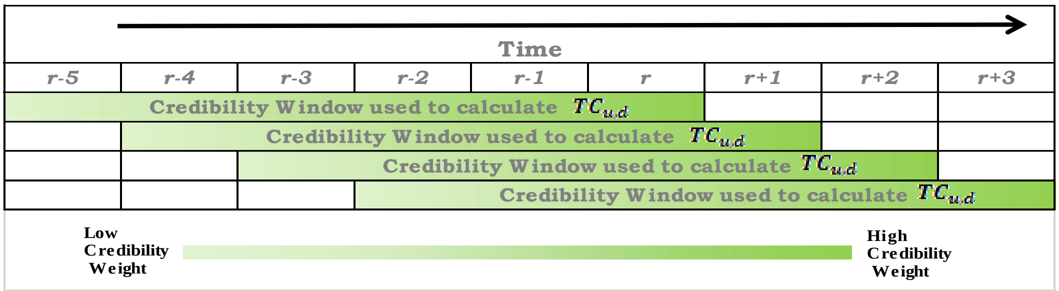

Time
r-5
r-4
r-3
r-2
r-1
r
r+1
r+2
r+3
Credibility Window used to calculate $TC_{u,d}$
Credibility Window used to calculate $TC_{u,d}$
Credibility Window used to calculate $TC_{u,d}$
Credibility Window used to calculate $TC_{u,d}$
Low Credibility Weight
High Credibility Weight

Figure 4-5: Credibility Window

The threshold (***I***) is used to facilitate the credibility analysis by focusing on recent users' content in the past (***I***) time periods. Furthermore, it is more efficient to measure the credibility of users based on their current and recent behaviour. This is logical since the user's interest(s) could change, and their knowledge evolves. Hence, the user's older content and metadata should be considered as a legacy chunk of data and therefore should not be incorporated in the credibility analysis.

4.4.2.6 Scale credibility values

The last step in this approach is to use a scale as a measurement system to interpret the numeric values resulting from the evaluation approach and convert them to a meaningful presentation. Thus, one of the most popular scale systems which use a 7-level trustworthiness scale (Hussain, Chang, and Dillon 2006) is customised and used in this research. This trustworthiness measure helps to rate trust by numerically quantifying the trust values and qualifying the trust levels numerically. Table 4-13 shows the seven levels of trustworthiness determined by this method.

Table 4-13: Seven levels of trustworthiness (Hussain, Chang, and Dillon 2006)

Trustworthiness Level	**Semantics (Linguistic Definitions)**	**Trustworthiness Value (User defined)**	**Visual Representation**
Level -1	New User	$TC'_{u,d} = ""$	Not displayed
Level 0	Very Untrustworthy User	$TC'_{u,d} = 0$	Not displayed
Level 1	Untrustworthy User	$0 < TC'_{u,d} \leq 1$	From ★
Level 2	Partially Trustworthy User	$1 < TC'_{u,d} \leq 2$	From ★ to ★★
Level 3	Largely Trustworthy User	$2 < TC'_{u,d} \leq 3$	From ★★ to★ ★★
Level 4	Trustworthy User	$3 < TC'_{u,d} \leq 4$	From ★★★ to ★★★★
Level 5	Very Trustworthy User	$4 < TC'_{u,d} \leq 5$	From ★★★★ to★★★★★

Entries of $\boldsymbol{TC}$ are scaled to values between “0” and “5” as follows:

$$\boldsymbol{TC'_{u,d}} = \frac{\left(\boldsymbol{TC_{u,d} - min\left(TC_{*,d}\right)}\right)*\boldsymbol{5}}{\boldsymbol{max\left(TC_{*,d}\right) - min\left(TC_{*,d}\right)}} \qquad (4.15)$$

The next sections demonstrate the implementation of the time-aware credibility mechanism and provide an evaluation metric for the proposed credibility approach.

4.5 Experimental Results

To evaluate the effectiveness of CredSaT, several experiments are conducted as follows; (i) the developed approach is benchmarked against other state of the art baseline models; (ii) the performance of CredSaT is reported to indicate the highly trustworthy, domain-based influencers; (iii) the capability of CredSaT to infer anomalous users is presented.

4.5.1 Subsets selection and experiments settings

To evaluate the credibility of users incorporating the temporal factor, the cleansed dataset is divided into six chunks starting at Nov-2014 and ending in Apr-2015, i.e. $\boldsymbol{I} = \mathbf{6}$**,** see Eq. (4.14), where each chunk is comprised of the data and metadata of each particular month. These chunks embody the chronologically sequential snapshots indicating the recent user's activity amongst the crawled dataset. Table 4-14 shows the total count of *users*, *tweets* and their *replies* for the determined time. The number of users shown in Table 4-14 (i.e. 6,066) represents the total distinct number of users who posted tweets in one or more of the determined months. In other words, 6,066 users out of 7,401[12] have been active in the six months. The remaining users posted their tweets before that, although they have been inactive on Twitter recently. This signifies the importance of studying users' content temporally.

[12] Refer to Table 4-2

Table 4-14: Total monthly count of users, tweets and replies

Month	Nov-14	Dec-14	Jan-15	Feb-15	Mar-15	Apr-15	Total
#Users	4,531	4,596	4,718	4,690	4,388	4,309	6,066
#Tweets	119,847	123,304	145,768	147,145	144,529	137,567	818,160
#Replies	55,949	58,956	76,561	73,867	70,135	61,352	396,820

The aforementioned set of equations (4.1) to (4.14) have been implemented for the datasets of each selected month. The value of $\boldsymbol{\rho}$ indicated in Eq. (4.5) is set to "2" experimentally as it represents the monthly threshold value. Significant adjustments introduced in Eq. (4.13), $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\theta}$ and $\boldsymbol{\vartheta},$ are set to (0.2, 0.2, 0.2, 0.1, 0.1 and 0.2) respectively. These values are set as they have proven successful to provide high credibility results after conducting several experiments incorporating different weight values. Function $\boldsymbol{w}(\boldsymbol{k})$ indicated in Eq. (4.13) is defined as $\boldsymbol{w}(\boldsymbol{k}) = \boldsymbol{k},$ this implies that $\boldsymbol{C}_{\boldsymbol{u},\boldsymbol{d}}^{\boldsymbol{k}}$will be assigned a value weight equal to the associated $\boldsymbol{k}$ value. Hence, the highest $\boldsymbol{k}$ value is assigned to the most recent dataset; conversely, the older the dataset, the lower is the assigned $\boldsymbol{k}$ value. The time-aware, domain-based user's credibility matrix $\boldsymbol{TC}$ is calculated for all AlchemyAPI's 23 domains of knowledge for every month of the credibility window (I), where I = 6. The time-aware, domain-based normalized credibility matrix $\boldsymbol{TC}'$ is calculated. This matrix includes a ranked list of users in each particular domain.

The top users in each domain comprise the trustworthy and very trustworthy users in that domain. Those users embody the influential users in each domain of knowledge. Although the domain of knowledge for some influencers is explicitly indicated in their Twitter's bio, the domain of interest is a tacit knowledge for other domain-based, highly trustworthy users. The proposed approach determines those users, assigns them trustworthiness values, and places them at various levels of trustworthiness.

4.5.2 Monthly domain-based trustworthy users

The six key attributes used to calculate the users' credibility are computed in each domain for each selected month. For example, Table 4-15 shows the monthly top five credible users in the "*Technology and Computing*" domain, and the highest five values of each key attribute which are used to measure $\boldsymbol{C}$**.** Values of $\boldsymbol{FF'_R}$ matrix are found in Table 4-11. Values of $\boldsymbol{FF'_R}$ are domain- and time-independent because the number of followers and friends have been captured once and they do not reflect any particular domain or period. The regular updating of the $\boldsymbol{FF'_R}$ matrix will be addressed in the future work.

Figures of Table 4-15 highlights the following issues: (i) there is a noticeable unsteadiness in the Twitterers' value for each key attribute in each month. For example, @***chuckfreeze*** achieved the highest weight ($\boldsymbol{W'}$) amongst other users in Feb-2015, and attained the second position in Nov-2014; however, the domain based weight positions of this user were (6th, 11st ,118th, and 12th) in Dec-14, Jan-15, Mar-15 and Apr-2015 respectively. (ii) It is evident that users achieved high values in some keys and low values in other keys. In other words, users might have obtained high domain-based weight due to their interest in one or few domains; however, their metadata exposed a deficiency in the count of domain-based replies, sentiment ratio, likes, and retweets. For example, all the key attributes extracted from @***chuckfreeze*** data were almost zero or null; hence, it is anticipated that this user would obtain a low domain-based trustworthiness value. This accentuates, again, the importance of monitoring user behaviour over time, and this is reflected in their credibility. On the other hand, low values for some key attributes should not mean that high values for other key attributes should be ignored. For example, ***@TexMitchell, @Delsant, @geekz1,*** etc,. presented a dedicated domain-based utilization of Twitter for some months although their content had not attracted much attention. To sum up, all key attributes analysed in this research should be considered to provide an accurate measurement of the user’s credibility in each domain.

The top five trustworthy users listed at the bottom of Table 4-15 show a noticeable domain-based interest for each month obtained by acquiring a high value for one or more of the key attributes. For example, ***@wolf_gregor*** attained a dominant position for almost every month. This is due to his continuous interest in the “Comp. and Tech.” domain, and the positive sentimental attitude to his content.

Table 4-15: Monthly top-5 credible users in “Technology and Computing” domain

	Nov-14		**Dec-14**		**Jan-15**		**Feb-15**		**Mar-15**		**Apr-15**	
	Twitterer	***Val***	***Twitterer***	***Val***	***Twitterer***	***Val***	***Twitterer***	***Val***	***Twitterer***	***Val***	***Twitterer***	***Val***
W′	misterron	1.000	misterron	1.000	bernardumali	1.000	**chuckfreeze**	**1.000**	Zeroplane	1.000	**Delsant**	**1.000**
	chuckfreeze	**0.689**	**TexMitchell**	**0.916**	FAWKSIE1	0.774	RickMartinezTX	0.783	Leaskh	0.501	Guerrerotook	0.982
	TexMitchell	**0.635**	anchorsouth	0.732	Kalpers	0.722	martinscerri	0.736	RealWireFeed	0.447	kennethvs	0.857
	Mapio	0.512	whichwdc	0.684	martinscerri	0.707	RealWireFeed	0.718	geekz1	0.382	the_arnon	0.790
	jehb	0.510	jehb	0.677	nathanholloway	0.698	bernardumali	0.651	misterron	0.358	johnjwall	0.630
R′	SpnMaisieDaisy	1	arieldiaz	1	nukeador	1	LocalJoost	1	rexguo	1	spbivona	1
	rasputnik	0.99	zxombie	0.966	LocalJoost	0.485	macguitar	0.69	edithyeung	0.512	afigman	0.658
	wolf_gregor	**0.621**	SchwartzTV	0.898	Lmotsh	0.234	edithyeung	0.48	whichwdc	0.419	IvorCrotty	0.339
	jehb	0.303	hazelmist	0.797	Zxombie	0.211	**wolf_gregor**	**0.37**	jehb	0.419	neuecc	0.208
	barrett	0.292	keyle	0.695	TimKrajcar	0.176	JustinCampPhoto	0.37	jkc137	0.419	DJTRASE	0.181
L′	SpnMaisieDaisy	1	hazelmist	1	nukeador	1	macguitar	1	edithyeung	1	aevanko	1
	rasputnik	0.441	iamWALP	0.968	LocalJoost	0.731	LocalJoost	0.95	rexguo	0.963	afigman	0.867
	wolf_gregor	**0.399**	arieldiaz	0.823	SpnMaisieDaisy	0.69	JustinCampPhoto	0.882	LauraORourke	0.835	CodrutTurcanu	0.682
	iamWALP	0.283	benjaminedgar	0.79	edithyeung	0.472	zpao	0.639	rzonmrcury	0.817	cbroyles	0.564
	jehb	0.193	draahwl	0.645	Zxombie	0.426	edithyeung	0.563	lennarz	0.55	neuecc	0.441
P′	mykola	1	markdrew	1	h0bbel	1	LocalJoost	1	Xantiriad	1	aevanko	1
	mrbill	0.818	ade	0.702	commadelimited	0.572	jtrs73	0.436	LauraORourke	0.901	trdibo23	0.74

	Nov-14		Dec-14		Jan-15		Feb-15		Mar-15		Apr-15	
	markdrew	0.676	GnTrobby1051	0.667	Dshafik	0.399	Perfume_Girl	0.326	rzonmrcury	0.721	Xantiriad	0.597
	trdibo23	0.581	h0bbel	0.655	Peeja	0.356	jukesie	0.257	commadelimited	0.601	chrisrisner	0.514
	developit	0.561	mrbill	0.625	Ade	0.333	h0bbel	0.232	pwSociety	0.511	Elle4DDubOnlyxx	0.497
S'	xeraa	1	ade	1	samillingworth	1	jimhanas	1	LauraORourke	1	CodrutTurcanu	1
	mrbill	0.963	daylemajor	0.947	commadelimited	0.994	grahamgilbert	0.926	andreaLG	0.68	aevanko	0.989
	johnjohnston	0.933	johnjohnston	0.849	hailpixel	0.775	JeremyKendall	0.924	**wolf_gregor**	**0.631**	JeremyKendall	0.639
	wolf_gregor	**0.854**	AlvinNg	0.849	BrianPurkiss	0.614	bkraft	0.901	samillingworth	0.562	agardnahh	0.625
	steveavery	0.755	macguitar	0.837	PDCExeter	0.607	LocalJoost	0.9	scout2i	0.561	FinessIHS	0.614
C	**wolf_gregor**	**0.6398**	arieldiaz	0.6018	nukeador	0.5162	edithyeung	0.5298	edithyeung	0.5678	afigman	0.4674
	SpnMaisieDaisy	0.5945	hazelmist	0.5594	commadelimited	0.487	JeremyKendall	0.4765	rexguo	0.5168	spbivona	0.4289
	jehb	0.393	edithyeung	0.502	samillingworth	0.4496	**wolf_gregor**	**0.4459**	**wolf_gregor**	**0.5009**	johnjwall	0.3809
	xeraa	0.389	markdrew	0.5015	edithyeung	0.4357	jimhanas	0.3711	commadelimited	0.4169	kennethvs	0.3785
	edithyeung	0.3571	JeremyKendall	0.4969	scout2i	0.3618	SpnMaisieDaisy	0.3709	scout2i	0.4085	**wolf_gregor**	**0.3694**

4.5.3 Scale credibility values

Figure 4-6 depicts the monthly credibility distribution of four domains: “Computing and Technology”, “Sports”, “Education”, and “Arts and Entertainment”. Further experimental results of other domains are illustrated in Figure A-1. As depicted in Figure 4-6, most of the users are domain-based and are in a category of “*Very untrustworthy*” or “*Untrustworthy*”. This is because of three main facts: (i) As mentioned in the *Data Acquisition* section, a user has selected if the number of followers is less than 5000. Thus, it is likely to find anomalous or untrustworthy users in the dataset. (ii) User content is divided into several months; some users show a fluctuation in their usage of online social platforms; (iii) a fine-grained model is built in this approach to measure users’ credibility; thus, users should show a continuous interest in a certain domain(s), and they should attract a good amount of attention from other users to acquire a decent level of domain-based credibility.

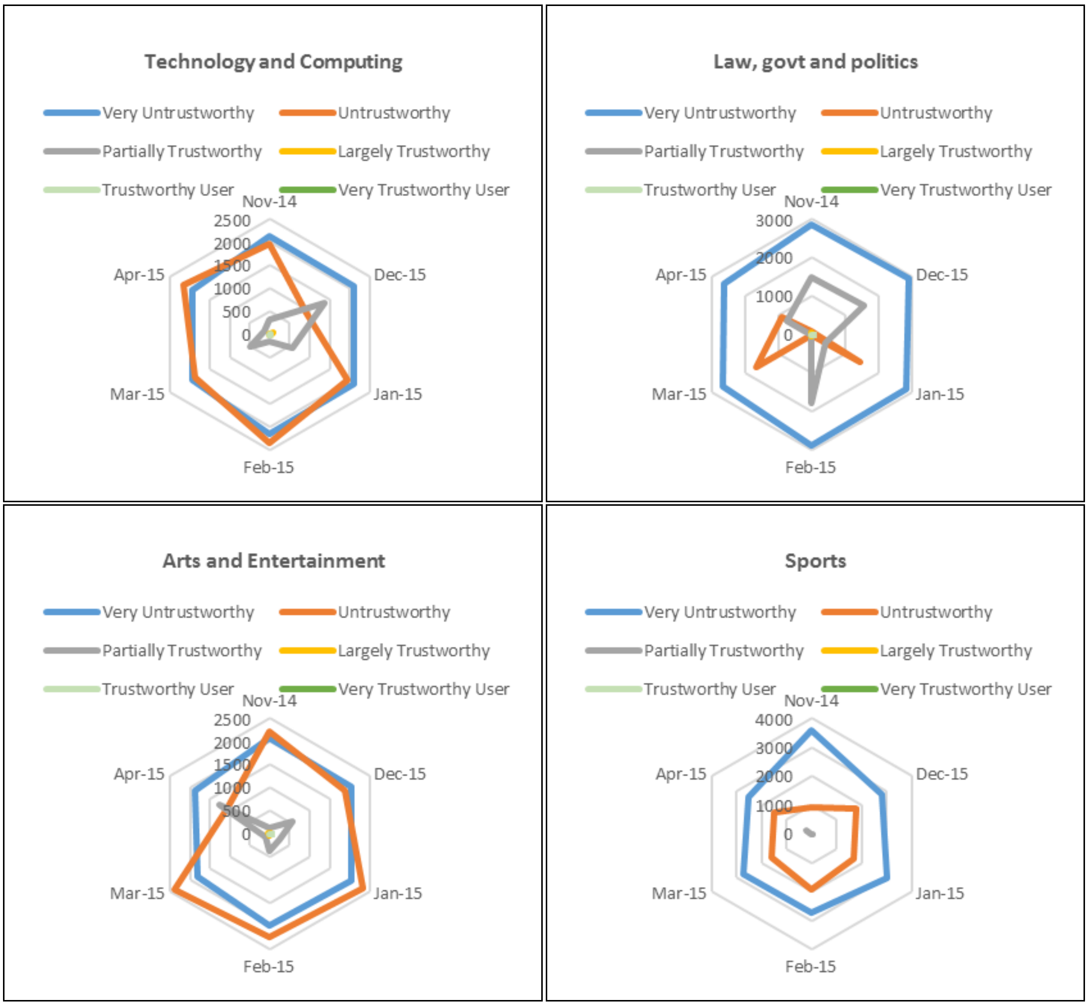


Figure 4-6: Monthly users' trustworthiness levels in four selected domains

4.5.4 Discovery of domain-based influencers - Baseline Comparison

A benchmark comparison against a set of evaluation techniques over a curated labelled dataset is conducted. This dataset contains four domains ("*Computing and Technology*", "*Sports*", "*Law, Gov, and Politics*", and "*Arts and Entertainment*"), and a set of "20" selected influential users in each domain. The list of influential users is selected by carefully examining their tweets, and collected metadata (bio information, #followers, #friends, etc.), thus choosing the list of users who have shown a noticeable and capacious interest in the selected domains consolidated with the figures captured from their metadata. The list of methods incorporated in the conducted comparison includes:

- ***Twitterrank***(Weng et al. 2010): aims to find topic-based influential Twitterers incorporating LDA statistical model for topic distillation, and topic-sensitive PageRank for credibility propagation. As topics of users are identified in TwitterRank based on the words' distribution of their tweets, the high-level topics classifications are inadequate and inferior(Michelson and Macskassy 2010). Therefore, the topics identified by Twitterrank(*namely LDA*) may not match the high-level domains which are identified incorporating ontology and semantic analysis facilitated by AlchemyAPI. To establish a common ground to conduct the comparison, a python implementation of Twitterank[13] is adopted and customised, and several trials of Twitterrank are reported over the collected dataset to find closely matching topics to the four domains of knowledge, and to infer the top influential users of each topic accordingly.
- ***High In-degree****:* measures the influence of Twitter by studying the number of followers. This feature is incorporated by several service providers(Weng et al. 2010).
- ***High domain-based key attributes****:* this method extracts five lists indicating the key attributes explained in this research and as summarised in Eq. (4.13). Each list comprises the set of users obtained the highest domain based values in each corresponding key attribute.

Evaluation Metric: The performance of finding domain-based influencers of

[13] http://bit.ly/TwitterRankPython

each method is measured based on the obtained *Precision, Recall, F-score and nDCG.* Let $\boldsymbol{HC_d}$ presents the set of influencers of domain $\boldsymbol{d}$ as indicated in the curated dataset, $\boldsymbol{HR_d^Q}$ embodies the top $\boldsymbol{Q}$ users of domain $\boldsymbol{d}$ retrieved by each incorporated method. The evaluation metrics can be calculated as the following: *Precision*^{1}: it measures the ratio between the numbers of correct retrieved domain-based influential users to the number of top-$\boldsymbol{Q}$ returned users by the method. *Precision* is assigned a number "^{1}" as this metric will be utilised later for a different purpose in a different experiment (see section 4.5.6). *Recall:* indicates the ratio between the numbers of correct retrieved domain-based influential users to the actual number of domain-based influential users identified in the curated dataset. *F-score*: is used to provide the trade-off between *Precision* and *Recall. Normalized Discounted Cumulative Gain (NDGC)*: measures the performance of the model incorporating graded relevance metric. The later metric is adopted in this experiment to provide a fine grain evaluation analysis. This is through assessing the retrieved user in each method by a scale of four relevance degrees; *highly influential, influential, somehow influential, not an influential.* These metrics can be defined as follows:

$$Precision_{d_Q}^{1} = \frac{\left|HC_d \cap HR_d^Q\right|}{|HC_d|} \tag{4.16}$$

$$Recall_{d_Q} = \frac{\left|HC_d \cap HR_d^Q\right|}{\left|HR_d^Q\right|} \tag{4.17}$$

$$F - score_{d_Q} = \frac{2 * Precision_d^1 * Recall_d}{Precision_d^1 + Recall_d} \tag{4.18}$$

$$nDCG_{d_Q} = \frac{DCG_{d_Q}}{(ideal)DCG_{d_Q}}, \text{ where } DCG_{d_Q} = \sum_{i=1}^{Q} \frac{2^{rel_{d_i}-1}}{\log_2(i+1)} \tag{4.19}$$

The conducted experiment retrieves the top 150 influencers in each domain

$\boldsymbol{d}$ for each model (i.e. Q=150). $Precision_d^1$, $Recall_d$, $Fscore_d$ and $nDCG_d$are calculated for each domain, and the average is computed for each metric in all domains. Table 4-16 shows the performance of each model.

Table 4-16: Evaluation of domain-based Influential Retrieval

Baseline Model	**Precision[1]**	**Recall**	**F-Measure**	**nDCG**
CredSaT	**0.95**	**0.75**	**0.85**	**0.93**
TwitterRank	0.73	0.65	0.69	0.86
High In-degree	0.83	0.35	0.49	0.75
High W'	0.76	0.59	0.66	0.86
High P'	0.51	0.31	0.39	0.36
High L'	0.83	0.4	0.54	0.75
High R'	0.90	0.62	0.77	0.68
High S'	0.85	0.66	0.74	0.69

The figures in Table 4-16 indicate that the CredSaT model outperforms other methods in all metrics. It is intuitive that CredSaT overshadows Twitterrank task of inferring influential users. This is because the mechanism followed by CredSaT considers several focal dimensions which are neglected by Twitterank such as mentoring users' credibility over time, sentiments analysis of the tweets replies, etc. CredSaT as a comprehensive framework performs better than harnessing each key attributes separately to measure the user's influence. For example, although the weight assigned to each user in each domain is important to address the interests of OSNs' user, this is insufficient to rank users based merely on their domains of interest. Likewise, obtaining a high number of followers does not imply an influence in all domains of knowledge; yet, these number of followers might be attained due to the importance of the user in a certain domain(s). Hence, it is essential to possess an understanding of the user's interests in all domains which include the interactions-based attributes of users in OSNs. This involves analysing the user's content, studying the overall followers' interest in the user's content, followers' sentiments toward the user, etc.

4.5.5 Highly domain-based trustworthy users

Figure 4-7 shows a closer look at the CredSaT's top five trustworthy users in each selected domain of the crawled dataset. The results shown in these charts are broadly acceptable. Further details of the calculated metadata obtained through this experiment can be found in Table A-1, Table A-2, Table A-3, and Table A-4. In the "*Computing and Technology*" domain ***@edithyeung, @wolf_gregor, @johnjwall, @commadelimited*** *and* ***@JeremyKendall*** attained the highest positions. ***@edithyeung***, for example, obtained a domain-based "*Very Trustworthy*" level. This is because ***@edithyeung*** shows a continuing interest in IT aspects in most of the posted tweets and links on Twitter. Moreover, a recent visit[14] to the user profile exhibited more than 300% increase in the number of followers since this metadata was crawled during the dataset acquisition phase. This is supported by the high number of positive replies, retweets, and favourites. This also applies to the other top four users in the "Computing and Technology" domain.

@SpnMaisieDaisy obtained a "very trustworthy" level in "Art and Entertainment". This user often tweeted about movies and TV series, and the metadata shows that other users pay particular attention to his "Art and Entertainment"-related tweets. ***@SpnMaisieDaisy*** has maintained his leading position in almost every month, which indicates his continuous interest in this

[14] https://twitter.com/edithyeung, Visited in 30/04/2017

domain. In the “sports” domain, ***@nwipreps,*** presents a platform to distribute tweets about many kinds of sports. This user keeps the followers updated on all sports-related news. With the highest values in the number of likes and retweets in the sports domain, ***@nwipreps*** deserves to be placed in this position. In the “Law, govt, and politics” domain, the top five users, in general, tweeted about topics related to Law, government or politics. For example, **@englishvoice** is the official Twitter account for the English Democrats, the nationalist political party in England. It is reasonable to expect that their Twitter accounts would achieve a five-star ranking because this account is dedicated to discussing political topics, which is supported by their followers. Apart from his interest in politics-related news, ***@IvorCrotty*** indicates in his bio that he is the head of a social media extension “***@rt_com***” for “Russia Today”: a Russian government-funded television network. Thus, @***IvorCrotty*** has maintained his dominant position in the “Law, govt, and politics” domain.

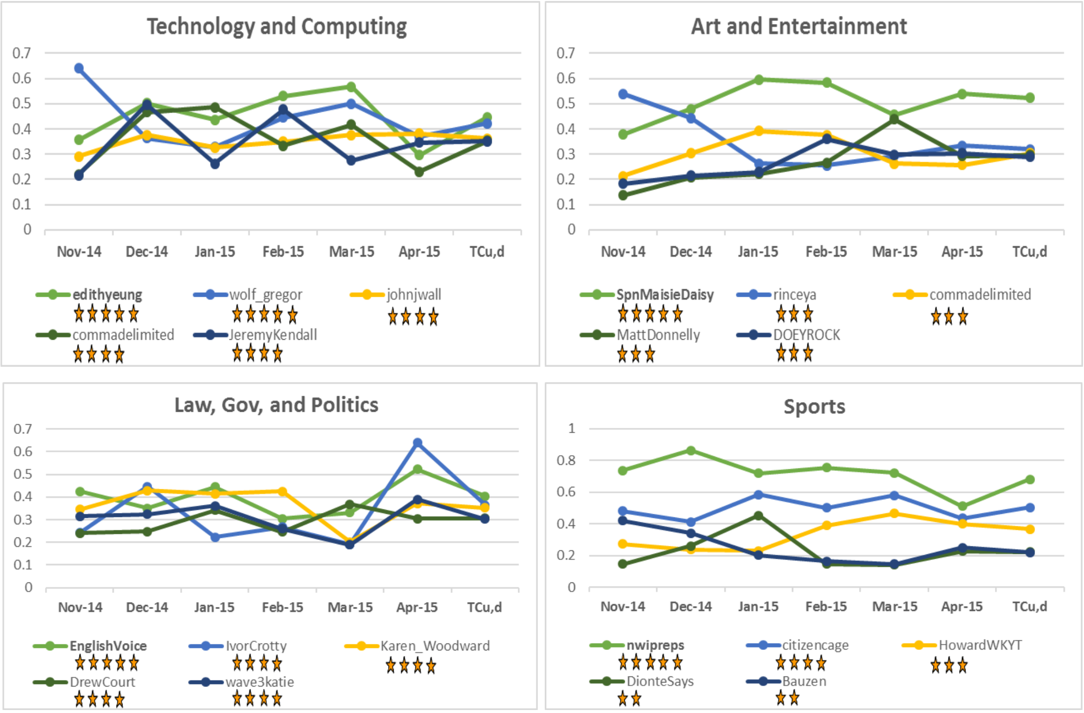


Figure 4-7: Highest $TC_{u,d}$ values in four selected domains

4.5.6 Discovery of Anomalous users

In addition to all the levels of trustworthy users, $\boldsymbol{TC_{u,d}}$ comprises a wide range of domain-based “*Untrustworthy*” and “*Very Untrustworthy*” users; thus, it is highly likely that this indicates that spammers, or other illegitimate user categories are amongst them. To capture these categories more easily, two criteria to narrow this research are applied:

- Selecting a set of users who have been placed at a “Very Untrustworthy” level (i.e $\boldsymbol{TC_{u,*} = 0}$) in ALL 23 domains, and achieved the lowest values in Tweets Similarity Penalty ($\boldsymbol{Twt\_Sim}$)(i.e strong similarity of tweets).
- Selecting a set of users who have been placed at a “Very Untrustworthy” level (i.e $\boldsymbol{TC_{u,*} = 0}$*)* in ALL 23 domains and achieved the lowest values in URL Similarity Penalty ($\boldsymbol{URL\_Sim}$)(i.e. strong similarity of URLs).

The results of the above criteria are compared with a set of retrieved users based on the following criteria:

- *Low In-degree*: selects users who obtained the least number of followers.
- *Anomaly detection toolkit of Graphlab*™[15]*:* This machine learning based module indicates the data items/points which are different from other data items. It assigns an anomaly score of value between "0" to "∞", where the higher the score, the more likely the data item is anomalous. All users with the following features were passed to this toolkit; #***DistinctWords***, #***Words***, $\boldsymbol{Twt\_Sim}$, and the $\boldsymbol{TC}$ values of all 23 domains. The users who achieved the highest score; i.e. detected anomalies, are used in this benchmark.

The examination process was conducted manually by reading all the crawled tweets of each user in each criterion's set, and labeling each user with one of two main categories (i) *Normal* users: are those legitimate users whose tweeting behavior is normal; (ii) *Anomalous* users: are those who utilise the Twitter platform for scamming, spamming, and other anomalous activities. The precision evaluation metric is computed as follows:

$$Precision^2 = \frac{Number\ of\ perceived\ anomalous\ users}{Number\ of\ retrieved\ users} \tag{4.20}$$

Figure 4-8 presents the retrieval precision of the top-K at 10, 20, 30, 40, 50 and the Average Precision. As depicted in this figure, the experiments conducted on the retrieved users of each criterion verify the effectiveness of the developed approach to discover anomalous users. For example, the first ten users retrieved by enquiring users who obtained zero credibility value in all domains, along with their

[15] https://turi.com/products/create

tweets' similarities are the highest (i.e. lowest *Twt_Sim* values), were all exhibiting anomalous behavior. However, only one user of the first 10 retrieved users, whose in-degree features are the lowest, was anomalous. Although the anomalous users discovered using the criterion "*Very Untrustworthy with Low Twt_Sim*" are relatively similar to those detected using "*Graphlab-anomaly_detection*" module, the average precision accumulated using the proposed approach is promising for building anomalies detection frameworks consolidated with the features proposed in this chapter. This will be investigated further in the future work.

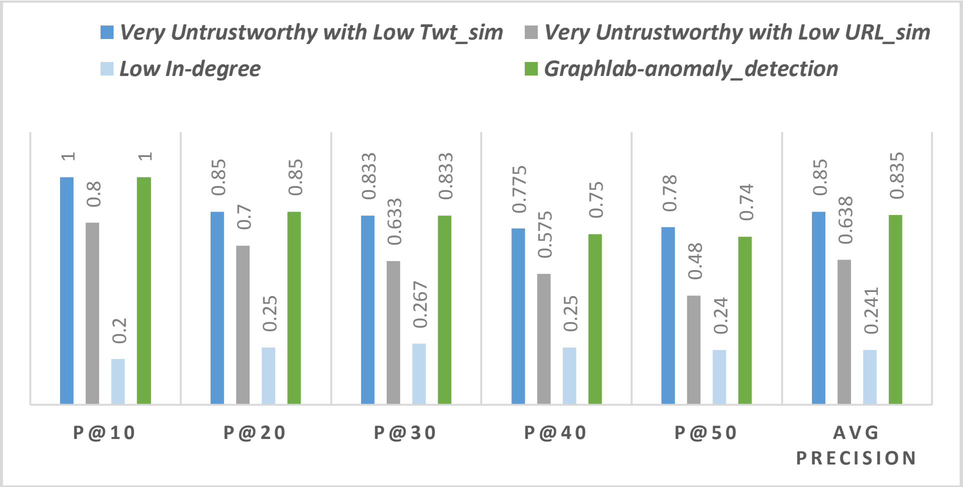


Figure 4-8: Evaluation of Anomalous Retrieval (*precesion*[2])

4.6 Conclusion

This chapter presents an approach for measuring time-aware domain-based users' trustworthiness in OSNs. CredSaT (*Credibility incorporating Semantic analysis and Temporal factor*), a domain-based credibility framework incorporating semantic analysis and the temporal factor, has been developed to measure and rank the credibility of users in SBD. The BD value chain is included in CredSaT to cover the life cycle of BD. CredSaT addresses four main BD features: *Veracity* through data trustworthiness, data certainty and a reliable data store; *Volume* through BD storage cluster; *Variability* by incorporating semantic analysis; and *Value* by creating a comprehensive framework to measure the credibility of users in SBD.

The main contributions of this chapter are as follows:

- CredSaT is proposed as an effective credibility framework for users of OSNs, addressing the main features of BD and incorporating semantic analysis and the temporal factor.
- A novel metric incorporating fine-grained key attributes is utilised to create the feature-based ranking model.
- The temporal factor is used to study the users' behaviour over time and reflect this behaviour using their domain-based credibility values.

The experiments conducted to evaluate this approach validate the applicability and effectiveness of determining domain-based highly trustworthy users, as well as capturing untrustworthy users; for example, when querying the list of anomalous users whose credibility is very low in all domains and the similarity of their tweets is very high, the results indicate an average precision value of "0.85".

AlchemyAPI™ has been used in this study as the sole semantics provider. Although its usage has shown optimistic results, the utilisation of AlchemAPI™ for analysing the domain of short text messages (such as tweets) is inadequate due to the brevity of these texts which increases the difficulty of obtaining an accurate understanding of their contextual content. Hence, it is necessary to conduct deeper investigation in the field of web semantic and ontology engineering in order to tackle the problem of semantic analysis of messages such as tweets in a more sophisticated manner. An accurate understanding of the contextual content of short messages such as tweets will assist domain-based classification to be made at the tweet level which facilitates further domain-based credibility approaches for short text messages. The next chapter resolves this issue by developing a framework incorporating domain ontology and semantic web technology to extract semantics of textual data and define the domain of data. Domain knowledge is captured in ontologies which are then used to enrich the semantics of tweets provided with a specific semantic conceptual representation of entities that appear in the tweets.

Chapter 5 Semantic Data Extraction from Social Big Data

5.1 Introduction

Social media plays a major role in our societies, reshaping the media arena, changing the rules of the game to a large extent, and providing ample space for expression and mutual dialogue. These forces have helped create platforms for the formation of public space or the public domain. The new media have also weakened the role of the so-called gatekeeper and the role of the major traditional media (newspapers, radio, television, news outlets) in prioritising public opinion (Hermida et al. 2012). It has the active ability to set and shape the nature of discourse and "Manufacturing Consent (Herman and Chomsky 2010)". Social media have been used excessively and irresponsibly by a certain slice of the community; their use has resulted in a state of political polarisation, ideological alignment, spreading lies, rumours and sedition, and promoting extremism, racism and terrorism. The literature review emphasises the pivotal significance of understanding the contextual content of social data. This leads to a close acquaintance with users' beliefs and their attitudes. Hence, improving the current technology tools and approaches in order to better understand the user's social content is inevitable.

The previous chapter proposes a framework for measuring users' credibility in OSNs. This is through aggregating a set of factors that contribute to the user's credibility and performing experiments on Twitter data using different domains. The mechanism followed in the previous chapter integrates semantics generated from a sole semantics provider (i.e. AlchemyAPI™), and experiments have been undertaken to measure the credibility at the user level. This leaves a room for a key challenge:

how can we discover the domains of short text messages such as tweets?

Sections 3.3.2 and 3.4.2 discuss this research problem and underlying issues. In particular, the current techniques examined in the literature review do not adequately take into account the semantic relationships of terms in the user's textual content, particularly in short text messages such as tweets. For example, in the political domain, the term 'Labor', that is extracted from tweets, would be represented under the concept 'Political Party' in the "Politics" domain but would be a different concept in another domain such as the "Work" domain. This is a significant problem due to the brevity of tweets which prevents the machine from obtaining an accurate understanding of their textual content.

This chapter attempts to resolve this research problem by presenting an ontology-based approach to extract the semantics of textual data and define the domain of data. The specific research questions that are addressed in this chapter are:

- How could Ontology, Linked Data, and a Knowledge Base be utilised to identify, annotate, and enrich entities in tweets for semantic analysis in Twitter?
- What are the components of the system applies a particular domain, i.e. Political domain?
- How can the ontology-based approach incorporated with AlchemyAPI enhance semantic information extraction?
- How can the findings of this project be used in practice and serve as a foundation for future expansion?

This chapter presents a systematic framework to extract knowledge captured from the textual content of SBD. Through five integrated steps, the proposed approach has proven to be successful in improving semantic information extraction and enrichment of the textual data, thereby providing an adequate interpretation of the textual content of social data. The overall system architecture is depicted in Section 5.2, and its semantic analysis components are discussed in Section 5.3. Case studies and the metric of system evaluation are described in sections 5.4 and 5.5 respectively. Finally, this chapter concludes with a summary of the developed framework, its innovativeness and applicability.

5.2 System Architecture

An overview of the framework proposed to address the lack of managing and extracting high-level domains from the textual content of SBD is presented in this section. The aim of the developed framework is to semantically analyse tweets to enrich data with a specific semantic conceptual representation of entities. Essentially, the proposed system has five main processes shown in Figure 5-1 as follows:

- Pre-processing data, which is given in Section 5.2.1.
- Domain Knowledge Inference, which is given in Section 5.2.2.
- Annotation and Enrichment, which is given in Section 5.2.3.
- Interlinking, which is given in Section 5.2.4.
- Semantic Repository, which is given in Section 5.2.5.

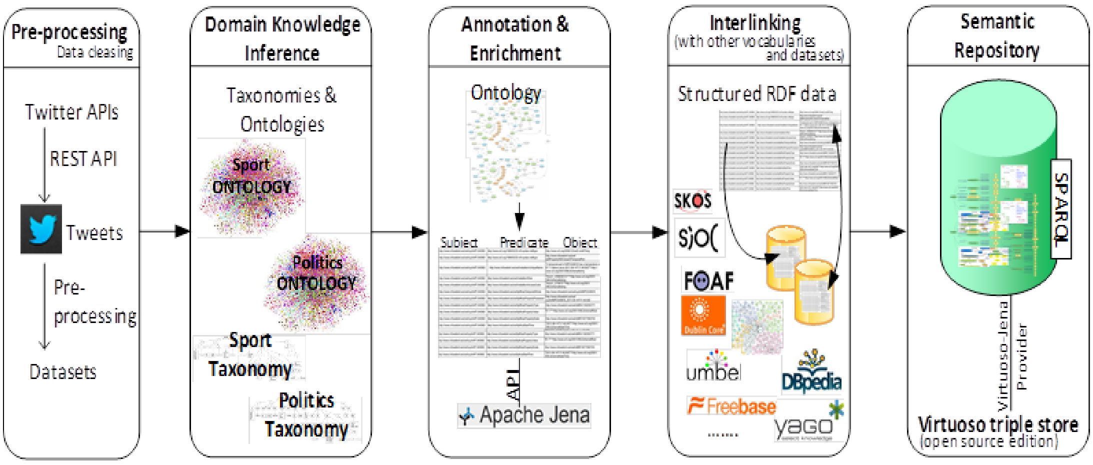


Figure 5-1: System Architecture

5.2.1 Pre-processing data (data cleansing)

This framework incorporates one of Twitter APIs named REST APIs to collect public archived tweets. The collected tweets are processed using standard data cleansing and pre-processing approaches to ensure data quality based on the following filtration criteria:

i. Remove Twitter handles "@" to get only the Twitterers' usernames.

ii. Remove the following to get only content: Twitter hashtags "#", URLs and hyperlinks, Punctuations, and Emoji.

iii. Correct and unify the encoding format as some tweets include some complex characters such as â, €™, œ, ¦, â€, etc. thus all tweets are decoded with the UTF-8 standard format to transform such symbols to the understandable data output.

There are other comprehensive metrics used in pre-processing Twitter data particularly for sentiment analysis, for example, handling negation and duplicate removal for retweets which are important for sentiment classification (Arias, Arratia, and Xuriguera 2014). However, Twitter textual data are considered hence those metrics are not necessary. Those internet slang, e.g. lol or acronyms and typos are collected and processed within text mining approach. The ‘lol’ slang is not relevant to the task of what this approach is trying to accomplish; however, some acronyms and typos can be relevant and may not be detected which will be alleviated in future work.

5.2.2 Domain Knowledge Inference

In the domain knowledge inference process, the domain knowledge is captured in domain ontologies is identified and used in the enrichment of the semantics of the tweets. In each tweet that users post, the semantics and the domains of the tweets can be extracted; the extracted domain knowledge is then used to enrich the tweets.

The inference process consists of two stages, i.e. start-up stage and learning stage. The start-up stage is a set-up stage that uses AlchemyAPI to identify domain ontologies. Figure 5-2 shows the domain knowledge inference process during the start-up stage.

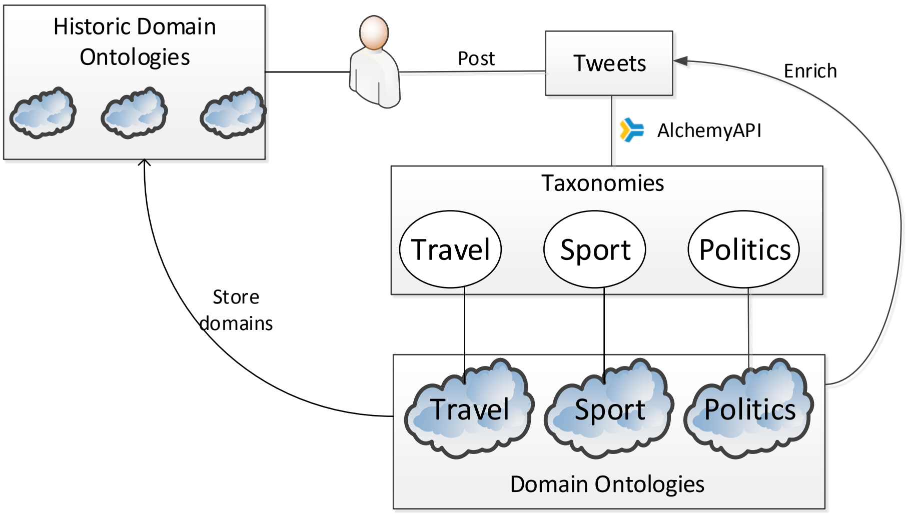
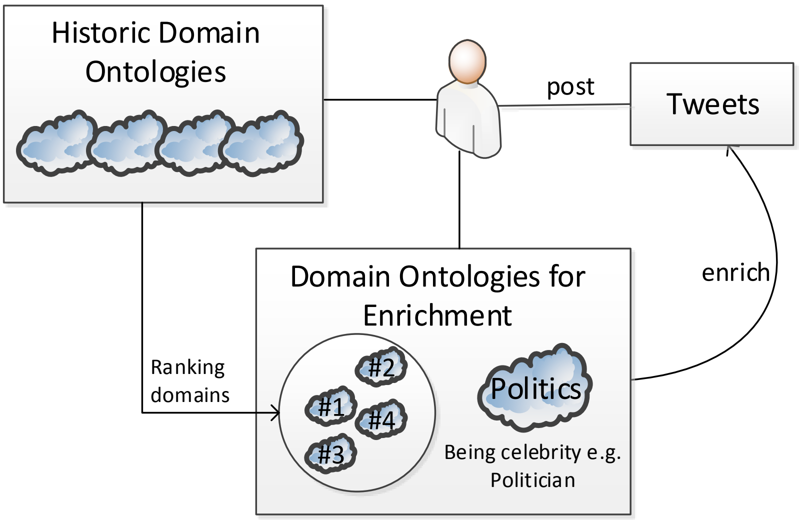


Figure 5-2: Domain knowledge inference process during the start-up

As shown in the figure, it starts when a user post tweets. Each textual tweet data is processed by AlchemyAPI to obtain its taxonomy. AlchemyAPI infers a maximum of three different taxonomies for each text or webpage. The reason of choosing this number is not justified by AlchemyAPI; however, we find it reasonable because it can be irrational to infer more than three topics from short texts such as tweets. Each domain ontology is identified based on the taxonomy and is then used in the enrichment process. For example, AlchemyAPI identifies three (3) taxonomies for a tweet, i.e. Travel, Sport, and Politics, so three (3) domain ontologies of Travel, Sport, and Politics are assigned as domain knowledge. The three ontologies will be used in the enrichment process and are stored as historical domain ontologies for the particular user who posted the particular tweet.

Once users have a list of historical domains, it will move into next stage, i.e. learning stage where machine learning is utilised. Machine learning ranked the historic domains and based on the ranking it provides the ability to select the particular domain ontologies for the enrichment process. Domain ontologies are ranked in an orderly number of tweets posted most. Rule-based learning method is applied here. For example, if a user has been posted most tweets about the sport, the sports domain ontology is firstly used for tweet enrichment. In a case of the user being a celebrity, domain(s) will be obvious, and the particular domain ontologies can be applied in the enrichment process. Figure 5-3 shows the domain knowledge inference process during the learning stage.



Historic Domain Ontologies
post
Tweets
enrich
Domain Ontologies for Enrichment
Ranking domains
#2
#1
#4
#3
Politics
Being celebrity e.g. Politician

Figure 5-3: Domain knowledge inference process during the learning stage

5.2.3 Annotation and Enrichment

For the annotation and enrichment process, the textual data of tweets is semantically annotated with the concepts in the domain ontologies; the annotation is then enriched with a description of the concepts referring to the domain ontologies and using controlled vocabularies, e.g. Dublin Core (DC[1]), Simple Knowledge Organization System (SKOS[2]), Semantically-Interlinked Online Communities (SIOC[3]). This allows each entity in the textual data to be specified with its semantic concept. The particular concepts can be further expanded into other related concepts and other entities instantiated by the concepts. The consolidation of this semantic information provides a detailed view of the entity captured in domain ontologies. The

[1] dublincore.org/
[2] http://www.w3.org/2004/02/skos/
[3] http://sioc-project.org/

domain ontologies are manipulated in this research using Apache Jena API. Jena, which is a Java framework for building semantic web applications, provides functionalities of creating, read, and modify triples (subject – predicate – object) in ontologies.

5.2.4 Interlinking

For the interlinking process, entities are interlinked with similar entities defined in other datasets to provide an extended view of the entities represented by the concepts (Ferraram, Nikolov, and Scharffe). The focus is on equivalence links specifying URIs (Universal Resource Identifiers) that refer to the same resource or entity. Ontology Web Language (OWL) provides support for equivalence links between ontology components and data. The resources and entities are linked through the 'owl#sameAs' relation; this implies that the subject URI and object URI resources are the same. Hence, the data can be explored in further detail. In the interlinking process, different vocabularies i.e. Upper Mapping and Binding Exchange Layer (UMBEL[4]), Freebase[5] – a community-curated database of well-known people, places, and things, YAGO[6] – a high quality knowledge base, Friend-of-a-Friend (FOAF[7]), Dublin Core (DC[8]), Simple Knowledge Organization System (SKOS[9]), Semantically-Interlinked Online Communities (SIOC[10]), and DBpedia[11] knowledge base, are used to link and enrich the semantic description of resources annotated.

[4] http://umbel.org/
[5] http://www.freebase.com/
[6] https://github.com/yago-naga/yago3/
[7] http://www.foaf-project.org/
[8] dublincore.org/
[9] http://www.w3.org/2004/02/skos/
[10] http://sioc-project.org/
[11] wiki.dbpedia.org/

5.2.5 Semantic Repository

A semantic repository represents a knowledge base which continues and updates the semantically rich annotated structured data. Ontology formalises the conceptualised knowledge in a particular domain and provides explicit semantics by splitting concepts, their attributes, and their relationships with the instances. In the repository, there are terminological data that define concepts (classes), attributes (data properties), relationships (object properties), and axioms (constraints) as well as data that enumerates the instances (individuals). This enables different services to support such as concept-based search, entailment to retrieve implied knowledge, instance-related information retrieval, etc. By using the semantic repository, query expansion can be performed for entity disambiguation, and semantic description of entities are retrieved.

In the repository, the structured data are stored as the RDF graph[12](a standard data model for data interchange on the World Wide Web) for persistence. Virtuoso (open source edition) triple store is used to store the RDF triples, ontologies, schemas, and expose it using SPARQL endpoint. SPARQL endpoint enables applications, users, software agents, and the like to access the knowledge base by posing SPARQL queries.

5.3 Semantic Analysis Components

5.3.1 Politics Ontology

The BBC[13] produces a plethora of rich and diverse content about things that matter to the BBC's audiences ranged from athletes, politicians, artists, etc. (2015a).

[12] https://www.w3.org/RDF/

[13] http://www.bbc.com/

The BBC uses domain Ontologies to describe the world and content the BBC creates and to manage and share data within the Linked Data platform. Linked Data provides an opportunity to connect the content through various topics. Among the nine domain ontologies that the BBC has developed and used, Politics Ontology is an ontology which describes a model for politics, specifically regarding local government and elections (Carrasco et al. 2014). This was originally designed to cope with UK (England and Northern Ireland) Local, and European Elections in May 2014. The focus of the project is on Australian Politics. Hence, the domain-specific Politics Ontology for Australian Politics is developed by extending the BBC Politics Ontology. The ontology in Australian Politics is specified having Australian politicians and Australian political parties. The concepts, instances, and relations are used in the annotation process. At this stage, concept Politician has 53 instances, i.e. 53 Australian politicians and concept Political Party has four instances, i.e. four main Australian political parties as shown in Appendix. The politics ontology is being incrementally extended over time. Figure 5-4 shows the BBC Politics Ontology; Figure 5-5 shows the extended version of the BBC Politics Ontology using OntoGraf for visualisation of the relationships in ontologies.

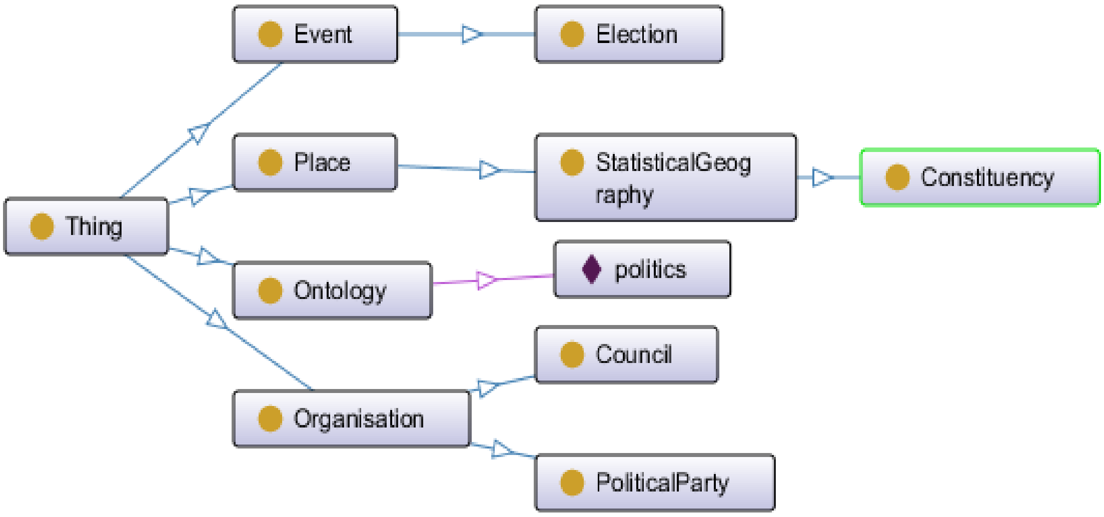
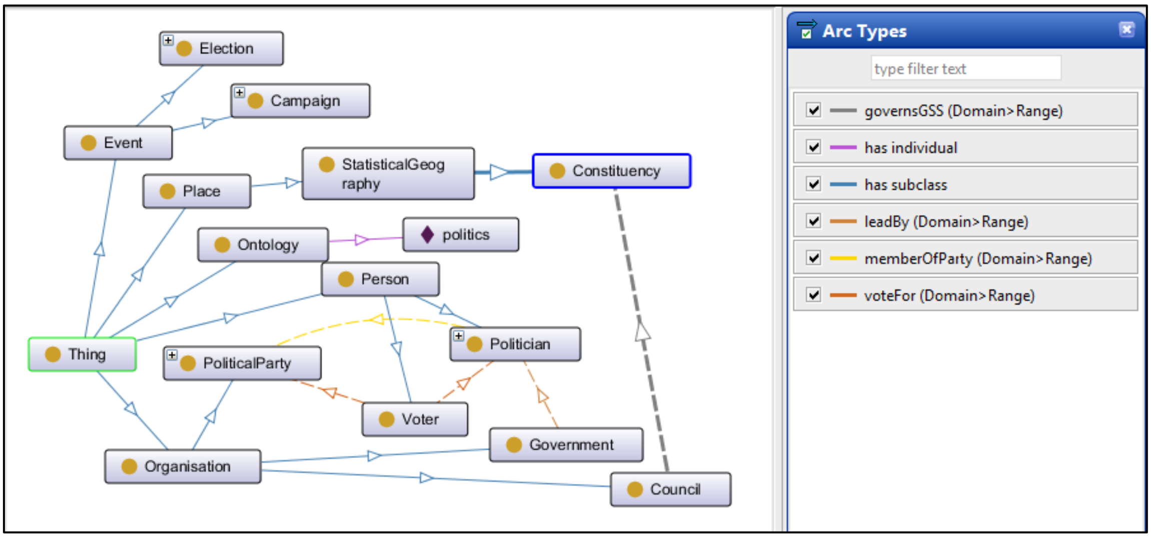


Figure 5-4: BBC politics ontology



☑ governsGSS (Domain>Range)
☑ has individual
☑ has subclass
☑ leadBy (Domain>Range)
☑ memberOfParty (Domain>Range)
☑ voteFor (Domain>Range)

Figure 5-5: BBC politics ontology extension

To ensure the extended version of Politics Ontology is consistent which is important as part of an ontology's development and testing, the Ontology needs to undergo reasoning process. No reliable conclusion can be deduced otherwise. The extended version of Politics Ontology has been reasoned to check its logical consistency using FaCT++(Tsarkov and Horrocks 2006), HermiT(HermiT 2016), Pellet(Sirin et al. 2007), Pellet (Incremental), RacerPro(Haarslev et al. 2012) and TrOWL(Thomas, Pan, and Ren 2010) reasoners. The reasoners checked the class, object/data property hierarchies, the class/object property assertions, and whether there were the same individuals contained within the ontology. Consistency verification through a reasoner includes consistency checking, concept satisfiability, classification, and realisation which are all standard inference services conventionally provided by a reasoner. The extended version of the Politics Ontology does not contain any contradictory facts.

5.3.2 Text Mining Tool

Text mining techniques have been applied for name entity recognition, text classification, terminology extraction, and relationship extraction (Cohen and Hersh 2005). To convert unstructured textual data from large-scale collections to a semi-structured or structured data filtering based on the need, natural language processing algorithms are used (Bello-Orgaz, Jung, and Camacho 2016). However, this can be difficult because the same word can mean different things depending on context. Ontologies can help to automate human understanding of the concepts and the relationships between concepts. Ontologies allow for achieving a certain level of filtering accuracy. Hence in this research, text mining tool is harnessed together with domain-specific ontologies for better accuracy of concept identification.

There are several text mining tools like open APIs that can extract entities and map the entities with concepts for online textual data. (Rizzo and Troncy 2011a) evaluate five popular entity extraction tools on a dataset of news articles i.e. AlchemyAPI, Zemanta, OpenCalais, DBpedia Spotlight, and Extractiv. (Saif, He, and Alani 2012) chose to evaluate the first three of the five entity extraction tools on tweets. The results from their experiments in both types of research are in the same line which showed that AlchemyAPI performed best for entity extraction and semantic concept mapping. In addition, in March 2015 IBM has acquired Alchemy for IBM's development of next-generation cognitive computing applications (2015b). Hence, AlchemyAPI has been used and evaluated, the best tool of all, in this project. Evaluation of other tools can be done in the future work.

5.3.3 Politics Twitter data

Twitter REST API is incorporated to collect public archived tweets. For the work and experiments, the collected tweet data are run through AlchemyAPI and tweets are selected for the dataset based on the set thresholds as follows which are defined by AlchemyAPI:

- Having confidence score above 0.4 AND
- Not having confidence response data status as no (not confidence).

AlchemyAPI provides a confidence score for the detected category ranged from 0.0 to 1.0 where higher is better (Turian 2013). The confidence score and response data convey the likelihood of the identified category is correct.

Table 5-1 shows dataset sources, the number of collected tweets and number of selected tweets, and period of collection. The number of collected tweets are those collected tweets during a period however only number of tweets are selected for experiments based on thresholds above mentioned. To get politics data, politicians are the main source and journalists' tweets are considered as an addition source. The two datasets contain politics data; the difference is that one from politicians' view and the other from journalists' view. Both datasets are chosen for experiments of this research.

Table 5-1: Details of two datasets

	Sources	**No. of collected tweets**	**No. of selected tweets**	**Period of collection**
Politics dataset	Twitter accounts of two Australian politicians.	4,122 (1,954 and 2,168)	3,653	25th Jan 2011 - 26th March 2015

	Sources	No. of collected tweets	No. of selected tweets	Period of collection
Politics Influence dataset	Twitter accounts of two Australian journalists.	3,479 (3,207 and 272)	210	5th Oct 2010 - 20th May 2015

5.4 Case Studies of Politics Twitter Data

AlchemyAPI can identify people, companies, organisations, cities, geographic features, and other types of entities from the textual data content in the general classification. It supports Linked Data and employs natural language processing technology to analyse the data and extract the semantic richness embedded within (Turian 2013). It is a comprehensive tool however it can only categorise the most general classification due to the lack of domain-specific knowledge. For a specific domain, AlchemyAPI will need ontologies to categorise content based on ontology concepts, instances, and relationships. Hence, the ontology-based approach proposed in this paper will be of benefit regarding extending the existing AlchemyAPI.

Tweet: "Launched Jennifer Kanis for Melbourne Campaign today. Outcomes instead of ineffective self indulgent commentary. Vote Labor in Melbourne."

AlchemyAPI entity extraction and concept mapping results:

ENTITY: Jennifer Kanis; TYPE of ENTITY: Person

ENTITY: Melbourne Campaign; TYPE of ENTITY: Organization

ENTITY: Melbourne; TYPE of ENTITY: City

AlchemyAPI taxonomy results:

/travel/tourist destinations/australia and new zealand

/society/work/unions

Figure 5-6 Output from AlchemyAPI for entity extraction, concepts mapping, and taxonomy

As can be seen from Figure 5-6, Alchemy fails to capture the keywords '*vote'* and '*labour'* as entities due to no specific domain knowledge. As a result, the taxonomy classifications of travel and society are inadequate. However, if politics ontology is applied as specific domain knowledge, the keywords 'Vote' and 'Labor' are annotated with its type respectively as relation 'voteFor' and concept 'Political Party'. By annotating two more entities of Labor and Vote and specifying particular entity Jennifer Kanis as Politician as shown in Figure 5-7, the political domain is counted as the domain of this tweet in addition to the travel and society domains. The more data that are annotated, the more entities are extracted in which the domain of tweet is clearer.

Jennifer Kanis - CONCEPT: Politician

Labor - CONCEPT: Political Party

Vote – Relation: voteFor

Figure 5-7: Politics Ontology Annotation

In addition, based on the concepts being referred to, the entities can be inferred to the knowledge captured in the Politics ontology. Figure 5-8 shows entities ‘Jennifer Kanis’, ‘Labor’, and ‘Vote’ being respectively inferred to concepts ’Politician’ and ‘Political Party’ and relation ‘voteFor’. Figure 5-8 also shows concepts being related to other concepts forming the domain of knowledge. The knowledge is captured in the politics ontology describes the semantics of concepts.

Table 5-2 shows the modelling notations that appear in Figure 5-8.

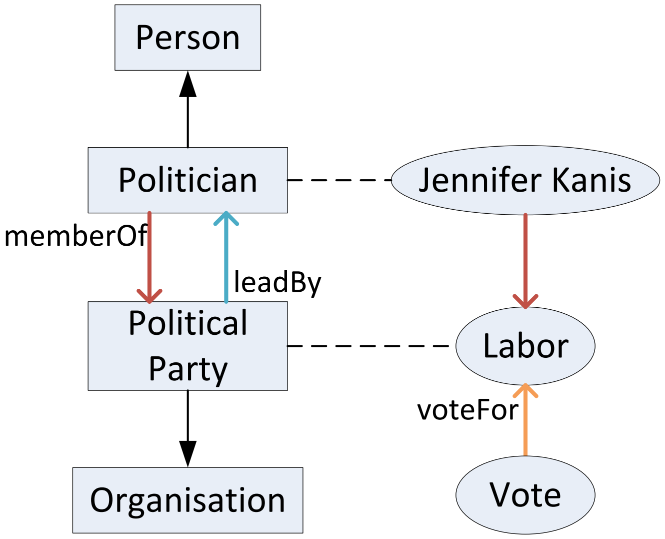


Figure 5-8: Knowledge captured in politics ontology

Table 5-2: Ontology modelling notations

Notations	**Semantics**
▭	Concept / Ontology class
⬭	Instance / Individual
⟶	Association semantical relationship (different colours represent different relationships)
➞	Generalisation / Taxonomical / Hierarchical relationship
– – -	Instance / Individual relationship

Hence, by integrating the results of the AlchemyAPI and the politics ontology annotation, the following information can be inferred from the particular tweet:

i. Jennifer Kanis is a Politician; Politician is a Person.

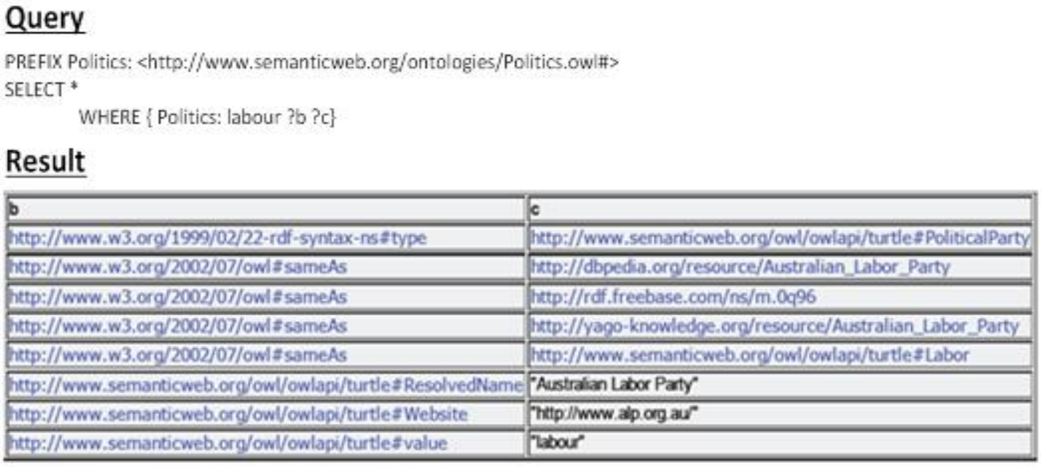
ii. Labor is a Political Party; Political Party is an Organisation.

iii. Jennifer Kanis is a member of Labor.

iv. Vote for Labor.

v. Melbourne is a city.

Figure 5-9 indicates the query and subsequent result to retrieve all information of Labor party. As can be seen, it shows entity ‘Labour’ enriched with its type of political party, website, and official name. The entity can also interlink with controlled vocabularies. Here, the entity ‘Labor party’ is interlinked with vocabularies from DBpedia, freebase, yago, and semanticweb.



Query

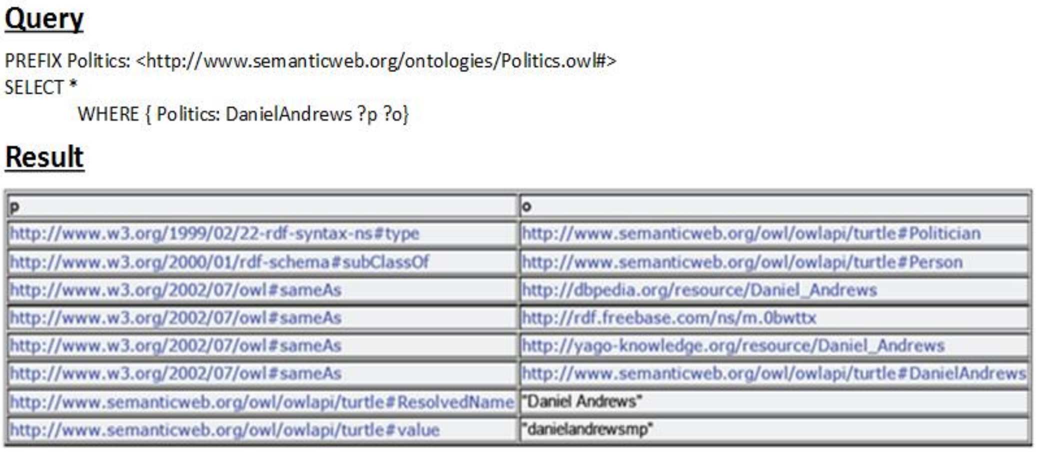
```
PREFIX Politics: <http://www.semanticweb.org/ontologies/Politics.owl#>
SELECT *
        WHERE { Politics: labour ?b ?c}
```

Result

b	c
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.semanticweb.org/owl/owlapi/turtle#PoliticalParty
http://www.w3.org/2002/07/owl#sameAs	http://dbpedia.org/resource/Australian_Labor_Party
http://www.w3.org/2002/07/owl#sameAs	http://rdf.freebase.com/ns/m.0q96
http://www.w3.org/2002/07/owl#sameAs	http://yago-knowledge.org/resource/Australian_Labor_Party
http://www.w3.org/2002/07/owl#sameAs	http://www.semanticweb.org/owl/owlapi/turtle#Labor
http://www.semanticweb.org/owl/owlapi/turtle#ResolvedName	"Australian Labor Party"
http://www.semanticweb.org/owl/owlapi/turtle#Website	"http://www.alp.org.au/"
http://www.semanticweb.org/owl/owlapi/turtle#value	"labour"

Figure 5-9: Enrichment and interlinking of Labor party

Figure 5-10 provides the query that retrieves all information of Politician Daniel Andrews. As can be seen, it shows the enrichment and interlinking of the entity with its name, its type of Politician, and its subclass of Person. The entity is also interlinked with vocabularies from DBpedia, freebase, yago, and semanticweb.



Query

```
PREFIX Politics: <http://www.semanticweb.org/ontologies/Politics.owl#>
SELECT *
        WHERE { Politics: DanielAndrews ?p ?o}
```

Result

p	o
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.semanticweb.org/owl/owlapi/turtle#Politician
http://www.w3.org/2000/01/rdf-schema#subClassOf	http://www.semanticweb.org/owl/owlapi/turtle#Person
http://www.w3.org/2002/07/owl#sameAs	http://dbpedia.org/resource/Daniel_Andrews
http://www.w3.org/2002/07/owl#sameAs	http://rdf.freebase.com/ns/m.0bwttx
http://www.w3.org/2002/07/owl#sameAs	http://yago-knowledge.org/resource/Daniel_Andrews
http://www.w3.org/2002/07/owl#sameAs	http://www.semanticweb.org/owl/owlapi/turtle#DanielAndrews
http://www.semanticweb.org/owl/owlapi/turtle#ResolvedName	"Daniel Andrews"
http://www.semanticweb.org/owl/owlapi/turtle#value	"danielandrewsmp"

Figure 5-10: Enrichment and interlinking of Politician Daniel Andrews

Another example is shown in Figure 5-11.

Tweet: “Thoughts and prayers with Karen Overington's family today. Karen was true Labor, a true friend and will be truly missed by all of us.”

AlchemyAPI entity extraction and concept mapping results:

ENTITY: Karen Overington; TYPE of ENTITY: Politician

AlchemyAPI taxonomy results:

/society/work/unions

/family and parenting

Figure 5-11: Output from AlchemyAPI for entity extraction, concepts mapping, and taxonomy classification of a tweet

In the tweet shown in Figure 5-11 above, AlchemyAPI captures only the entity ‘Karen’ Overington as a politician. The entity and keywords of ‘true friends’, ‘prayers’, ‘thoughts’, ‘family’, and ‘labour’ are used to classify the tweet under the taxonomy of society and family and parenting which is inadequate. Hence, if politics ontology is applied, the keyword ‘labour’ is annotated as an entity under the concept of the political party. This results in classifying Political domain as an additional domain of tweet.

The politics dataset has been used to evaluate the AlchemyAPI. AlchemyAPI classifies the politics dataset into various domains as shown in Figure 5-12. For two different users, it shows that most tweets are in the travel domain though it is supposed to be in political domain due to the politics dataset. In comparison to results from AlchemyAPI associated with politics ontology as shown in Figure 5-13, it classifies the same dataset into the proper domain, i.e. the political domain. This shows significant improvement when associated with specific domain knowledge of politics being captured in politics ontology.

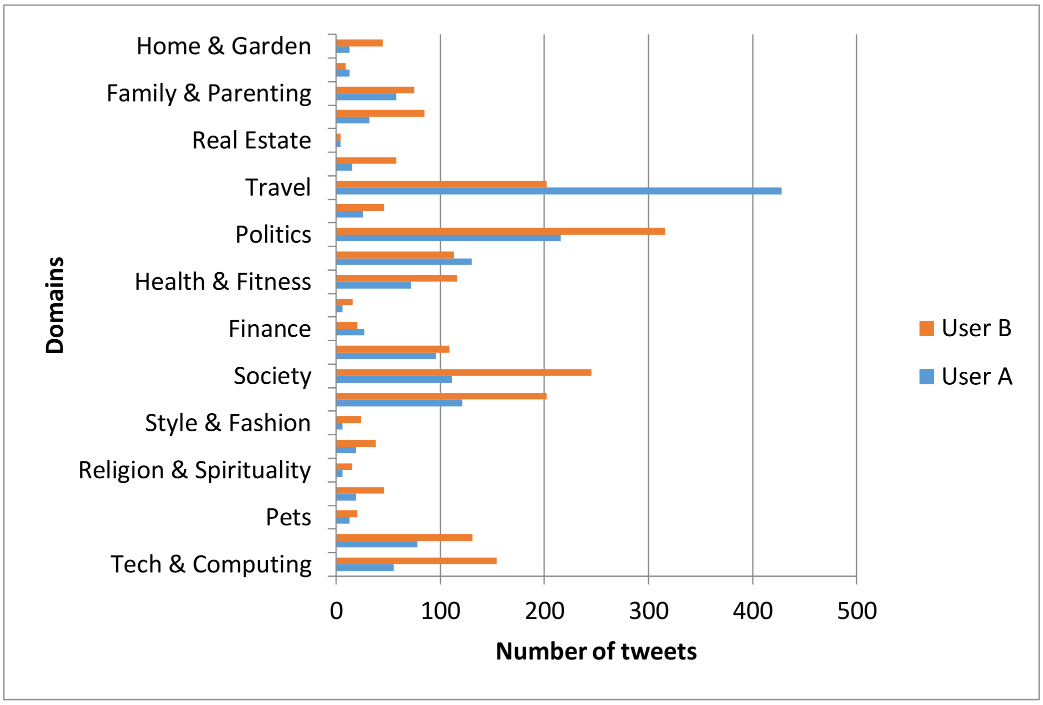


Figure 5-12: Results from Alchemy showing some tweets in various domains from the politics dataset

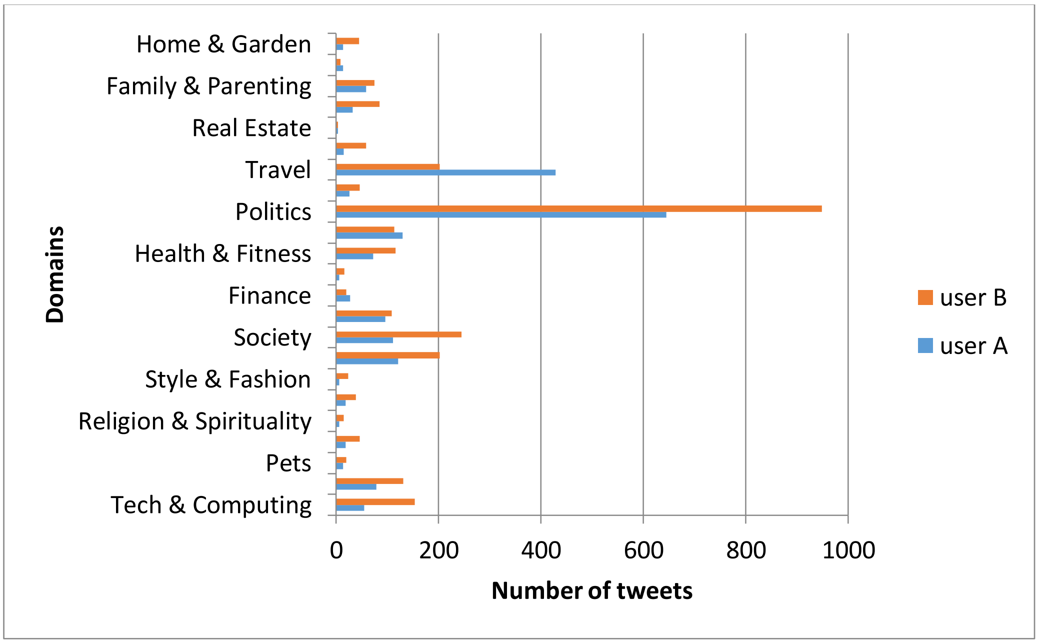


Figure 5-13: Results from Alchemy associated with politics ontology showing a number of tweets in various domains from the politics dataset

Once the domain can be correctly defined from user's tweets using the proposed ontology-based approach, users' influence in particular domains can be discovered, and domain-based trustworthiness can also be evaluated. In next section, the developed framework is evaluated for domain classification and entity annotation.

5.5 System Evaluation

In this section, the semantic information extraction at the domain level and the entity level is evaluated. The performance of AlchemyAPI alone is compared with the performance of AlchemyAPI when it is associated with the developed ontology-based approach.

5.5.1 Datasets

For evaluation purpose, 473 tweets are chosen from the selected politics dataset and chosen 209 tweets from the selected politics-influenced dataset. This subset is selected after conducting the pre-processing (cleansing) on the complete dataset as indicated in Section 5.2.1. Datasets are divided for evaluation purposes into four categories:

Category #1: Tweets that are classified by AlchemyAPI as political domain and the politics ontology annotate them

Category #2: Tweets that are classified by AlchemyAPI as NON-political domain however the politics ontology annotate them

Category #3: Tweets that are classified by AlchemyAPI as a political domain but the politics ontology does NOT annotate them

Category #4: Tweets that are classified by AlchemyAPI as NON-political domain and the politics ontology does NOT annotate them.

5.5.2 Evaluators

Three evaluators are incorporated to evaluate the concept extraction and domain identification outputs generated by AlchemyAPI alone compared with AlchemyAPI associated with the proposed ontology-based approach. One of the evaluators is considered as a domain expert in politics, i.e. this person is currently involved in politics and has worked in the area for more than five years. The other two evaluators are academics and considered as non-domain experts who have a general interest in the political domain.

5.5.3 Results and Discussion

The assessment of the outputs is based on the following criteria:

i. the correctness of the extracted politics entities;

ii. the correctness of inferring the extracted politics entities with its concept; and

iii. the correctness of political domain classified in tweets.

5.5.3.1 Politics dataset

This section discusses the evaluation results from the politics dataset. Table 5-3 shows the number of *correct* extracted politics entities. The results show that for tweets that are classified by AlchemyAPI as politics tweets, the politics ontology can annotate 98 more politics entities from just 41 entities from AlchemyAPI. The number of politics entities increases to 139 entities when combining the AlchemyAPI result with the politics ontology result; that is, the number of entities is almost tripled. For the non-politics tweets classified by AlchemyAPI, the politics ontology can annotate 161 more politics entities from just

62 entities from AlchemyAPI. The number of politics entities increases to 223 entities when combining the AlchemyAPI result with the politics ontology result, i.e. almost four times more entities.

Table 5-3: Number of correct extracted politics entities

Categories of dataset	**AlchemyAPI**	**Politics Ontology**	**AlchemyAPI and Politics Ontology**
Alchemy Politics tweet being annotated by Politics ontology	41	98	139
Alchemy NON-Politics tweet being annotated by Politics ontology	62	161	223
Alchemy Politics tweet NOT being annotated by Politics ontology	0	0	0
Alchemy NON-Politics tweet NOT being annotated by Politics ontology	0	0	0
Total	103	259	362
Percentage of *correct* extracted entities (sample size of 473)	22%	55%	77%

Table 5-4 shows the number of *incorrect* extracted politics entities in the four categories as explained in Section 5.5.1 for datasets. The results show some flaws in AlchemyAPI which can be overcome by incorporating it with specific domain knowledge captured in politics ontology.

Table 5-4: Number of incorrect extracted politics entities

Categories of dataset	AlchemyAPI
category #1: Alchemy Politics tweet being annotated by Politics ontology	35
category #2: Alchemy NON-Politics tweet being annotated by Politics ontology	35
category #3: Alchemy Politics tweet NOT being annotated by Politics ontology	8
category #4: Alchemy NON-Politics tweet NOT being annotated by Politics ontology	8
Total	86

In total, AlchemyAPI alone extracts 103 politics entities, failing to extract 259 politics entities which the politics ontology annotates as entities. Hence, by incorporating the politics ontology with AlchemyAPI, more political entities are extracted, totalling 362 entities rather than just 103 entities.

The pie chart shows all distinct 681 entities resulting from AlchemyAPI as seen in Figure 5. Number of TRUE refers to the number of political entities that are correctly annotated with its metadata. Number of FALSE refers to the number of political entities that are incorrectly annotated. Number of NULL refers to the number of non-politics entities.

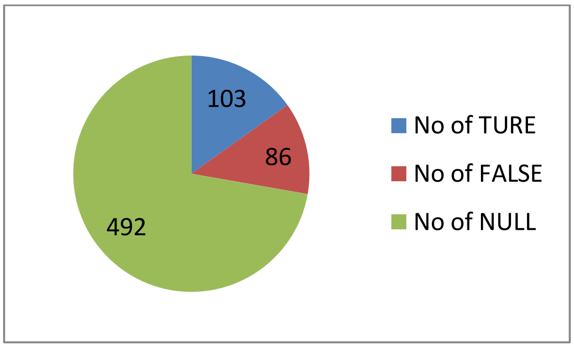


Figure 5 14: All distinct entities resulting from AlchemyAPI

The results show that AlchemyAPI identifies more entities in other domains outside the political domain in the politics dataset.

5.5.3.2 Politics-influenced dataset

This section discusses evaluation results from politics influence dataset. Table 5-5 shows that AlchemyAPI alone correctly extracts 44 politics entities, incorrectly extracts 15 politics entities, and fails to extract 59 politics entities which the politics ontology annotates as entities. By incorporating the politics ontology with AlchemyAPI, more political entities are extracted, totalling 103 entities which are over twice the number of entities extracted by AlchemyAPI alone.

Table 5-5: Politics entity extraction in AlchemyAPI from the politics-influenced dataset

	Correct extracted politics entities	**Incorrect extracted politics entities**	**Missing politics entities**	**Total number of retrieved entities**	**Total number of politics entities**
AlchemyAPI	44	15	59	59	103

5.5.3.3 Precision, recall, and F-measure

In this section, precision, recall, and F-measure are evaluated from AlchemyAPI results for both datasets. Precision is the fraction of retrieved entities that are politics-related as shown in equation (5.1) while recall is the fraction of political entities that are retrieved as shown in equation (5.2). Another metric known as the F-measure, which is the weighted harmonic mean of precision and recall, is used as shown in equation (5.3).

Precision = Number of Politics Entities Retrieved / Total Number of Retrieved Entities (5.1)

Recall= Number of Politics Entities Retrieved / Total Number of Politics Entities

(5.2)

$$\textbf{F-measure} = 2 \times \frac{precision \times recall}{precision + recall} \qquad (5.3)$$

Figure 5-14 shows a comparison of politics data and politics-influenced data on precision, recall, and F-measure. From the figure, it can be observed that AlchemyAPI performs better in data from various domains (politics influence dataset) rather than domain-specific data (politics dataset).

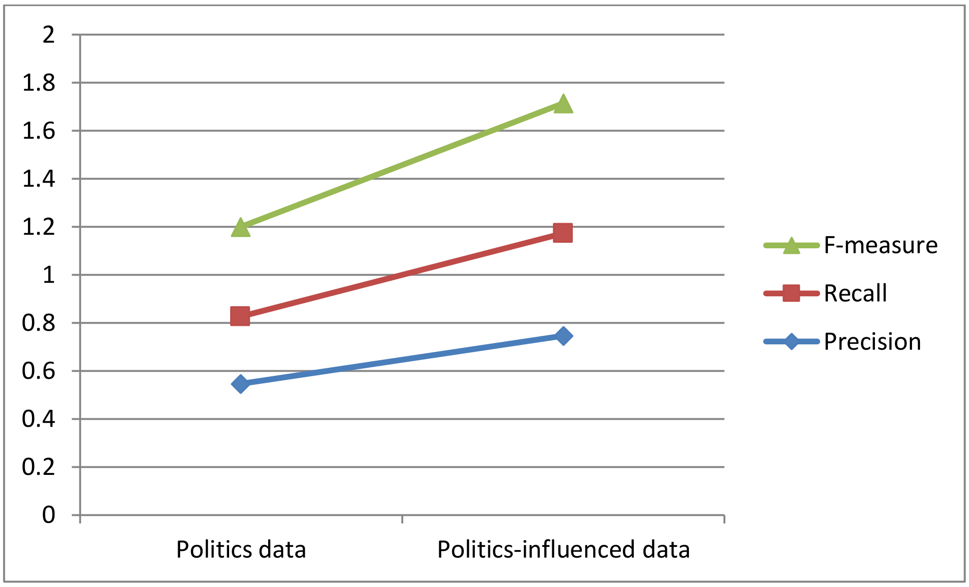


Figure 5-14: Comparison of politics data and politics influence data on precision, recall, and F-measure

Regarding precision, it shows that AlchemyAPI can retrieve more entities that are politics-related in a politics-influenced dataset than in a politics dataset. This is because fewer incorrect politics entities are retrieved from a politics-influenced dataset.

Regarding recall, it shows that AlchemyAPI can retrieve more politics entities in a politics-influenced dataset than in a politics dataset. This is because AlchemyAPI should have identified more politics entities in the politics dataset, but failed to do so.

5.5.3.4 Political domain classification

In this section, the correctness of the political domain classified in tweets is shown. The evaluators validate each tweet in the datasets and determine whether it is a politics-related post. Table 5-6 shows the percentage of tweets being classified as politics-related.

Table 5-6: Percentage of tweets being classified as politics-related

Categories of dataset	Politics dataset	Politics-influenced dataset
Alchemy Politics tweet being annotated by Politics ontology	99%	98%
Alchemy NON-Politics tweet being annotated by Politics ontology	98%	97%
Alchemy Politics tweet NOT being annotated by Politics ontology	27%	47%
Alchemy NON-Politics tweet NOT being annotated by Politics ontology	12%	32%

It can be observed from the results that almost all tweets that the politics ontology annotates are politics tweets. The politics ontology annotates less than 50% of the politics tweets, but more of the politics-influenced dataset than the politics dataset. This indicates that domain-specific ontology performs better in a domain-specific dataset.

5.6 Conclusion

This chapter presents an ontology-based approach as a means of extracting the semantics of textual data, thus inferring users' domains of interest. Most of the existing approaches to infer domains of interest of online social platform users rely on several statistics-based bag-of-words techniques. These approaches are inadequate for inferring high-level topics. Furthermore, these techniques are unable to infer the semantic relationships of terms in the textual content of users. More importantly, these techniques are inferior in terms of extracting topics from short text messages such as tweets. However, the approach proposed in this chapter attempts to resolve this issue by capturing domain knowledge through ontologies which are then used to enrich the semantics of data with specific semantics conceptualisation of entities of entities.

In brief, the key contributions of this chapter are as follows:

- Ontology, Linked Data, and a Knowledge Base have been utilised to identify, annotate, and enrich entities in tweets for semantic analysis in Twitter.
- An ontology-based approach that incorporates AlchemyAPI has proven to be successful and enhances semantic information extraction.
- The findings of this project are used in practice and serve as the foundation for future expansions to this work.

The developed system described in this chapter comprises five steps: pre-processing (through data cleansing and integration), domain knowledge inference (through semantics analytics and domain ontologies), annotation and enrichment (through semantically annotating textual text of tweets with the concepts in the domain ontologies), interlinking (entities are interlinked with similar entities defined in other datasets to provide an extended view of the entities represented by the concepts), and semantic repository (which is the knowledge base repository that continues and updates the semantically rich annotated structured data).

Experiments are conducted using the approach in the political domain incorporating public data collected from Twitter. These experiments include evaluating the performance of the system through assessing it by three evaluators (an expert in the politics domain, and two academics who have an interest in the politics domain). The assessment of the outputs is based on: (i) the correctness of the extracted politics entities; (ii) the correctness of inferring the extracted politics entities with its concept; and (iii) the correctness of political domain classification in tweets. An evaluation metric is used, comprising well-known evaluation measurements (i.e. precision, recall, and F-measure). The work has produced optimistic results which establishes a cornerstone for predicting and classifying users' domain of knowledge as will be discussed in detail in the next chapter. In other words, this chapter presents a mechanism for adequately understanding the domain knowledge inferred from textual content. The next chapter extends this work by applying further semantics analytics techniques (i.e. WordNet®). Further, the next chapter describes several domain-based classification and prediction models for SBD which can perform dual classification at the user level and the post level. This is done by scrutinising users' interest in the political domain, taking into account the temporal factor.

Chapter 6 Ontology-based Domain Discovery Incorporating Machine Learning

6.1 Introduction

People express their thoughts, feelings, activities, and plans etc. via Online Social Networks (OSNs). Often, their posts link to the product(s), service(s), event(s), society, or person(s) etc., and people in OSNs intuitively tend to seek and connect with others who are like-minded. This homophily results in building homogenous personal networks based on behaviours, interests, and feelings etc.(McPherson, Smith-Lovin, and Cook 2001). In particular, OSNs are a medium for content makers to express and share their thoughts, beliefs, and domains of interest. This gives individuals access to a wider audience which positively affects their social rank and provides other benefits, such as gaining political support (Rainie and Wellman 2012). This is consolidated with the rapid increase in unstructured social data which has highlighted its importance as a means of acquiring deeper and more accurate insights into businesses and customers. Therefore, the cornerstone of building users' online social profiles is an accurate understanding and classification of their domains of interest.

In this context, companies incorporate advanced social data analytics when designing effective marketing strategies and seeking to leverage the interactive feature of OSNs. Thus, to create the required interaction with their customers, companies use many modern means of communication to attract customers and visitors to their online social platforms. Consequently, it is necessary for companies to analyse their customers' social content and allocate the customers to appropriate categories based on their topics of interest, in order to deliver the right message to the right category. This can be obtained by collecting tweets of those customers who post their content publically. This allows companies to access and retrieve these tweets facilitated by using Twitter REST APIs. Given this ultimate goal, appropriate technical solutions should be adopted that have the capacity to infer the meaning of social content at the user level and post level.

In the previous chapter, an ontology-based approach is developed to semantically analyse the social data at two levels, i.e. the entity level and the domain level, in order to extract the semantics of textual data and define the domain of data. This approach has been successful in supporting and boosting the output of semantics analytic providers (i.e. AlchemyAPI). However, there is a need to extend this work to provide a platform for automatically classifying and predicting the domain of interest at the user level and post level. Sections 3.3.3 and 3.4.3 explain this problem and the related research issues. In particular, the existing approaches to topic extraction, modelling and classification rely on statistical bag-of-words techniques such as Latent Dirichlet Allocation(LDA) (Blei, Ng, and Jordan 2003a). These techniques have shown several limitations including: (i) the number of designated topics is fixed and should be known before the analysis (Zhang, Cui, and Yoshida 2017); (ii) the topics mined by these models do not take the temporal aspects into account (Alghamdi and Alfalqi 2015), (iii) these models are considered as monolingual topic models, and therefore do not differentiate idioms in the same language (Zoghbi, Vulic, and Moens 2016); and (iv) these models are unable to infer high-level topics particularly from short texts such as tweets (Li et al. 2016).

This chapter addresses this research issue and extends the work of the previous chapter. This is by means of leveraging the external semantic web knowledge bases and machine learning modules to reduce the ambiguity in the textual content and to classify and predict the domain of interest at the user and tweet levels as will be illustrated in the next sections.

This chapter begins with an outline of the key components of the system architecture of the proposed framework. This system incorporates the set of features extracted and selected from the social data. Then, an overview is presented of several well-known machine learning algorithms. Finally, the system evaluation and testing are described in detail, indicating several benchmark comparisons.

6.2 System Architecture

Figure 6-1 shows the architecture of the proposed framework which adopts a BD infrastructure. This framework comprises three main components, namely: (1) data collection and acquisition, (2) features extraction, and (3) the prediction module. The big data infrastructure at the School of Information Systems, Curtin University, is used as a distributed environment to facilitate data storage and analysis. This facility has a 6-node cluster, each with 2 TB Storage, 8 Core Processors, and 64 GB RAM.

The information flow through the proposed framework can be described in steps. As shown in Figure 6-1, steps 1 to 4 represent the processes required to achieve the predicted likelihood value of the user's interest in the political domain. This is the first outcome value (Politics Likelihood / User Level) indicated by the red dotted line. Steps 5 to 9 follow and predict the political domain-based likelihood value of a newly-posted user tweet. This is the second outcome value (Politics Likelihood / Tweet Level) indicated by the red dotted line. In the proposed framework, the user posts public content to the Twitter network, which facilitates data collection through the available application programming interfaces (APIs). The user's content is collected in two phases, namely, historical user's content and new user's content. The user's historical content represents the recent and former tweets which are collected in the first phase. The user's new content refers to their future tweets which will be collected during the second phase.

The collected historical tweets are pre-processed and passed to either the tweet features or user features extraction module. A list of user features is extracted and fed into a machine learning module to predict the political domain likelihood value, where the domain likelihood indicates the user's interest in the political domain. This domain likelihood is harnessed further and is added as another feature to the list of features extracted from the new user tweet after pre-processing during the second phase. The newly combined list of tweet features is fed into the machine learning module to predict the political domain likelihood of the newly posted tweet. The following subsections explain the mechanism of each component of the proposed framework.

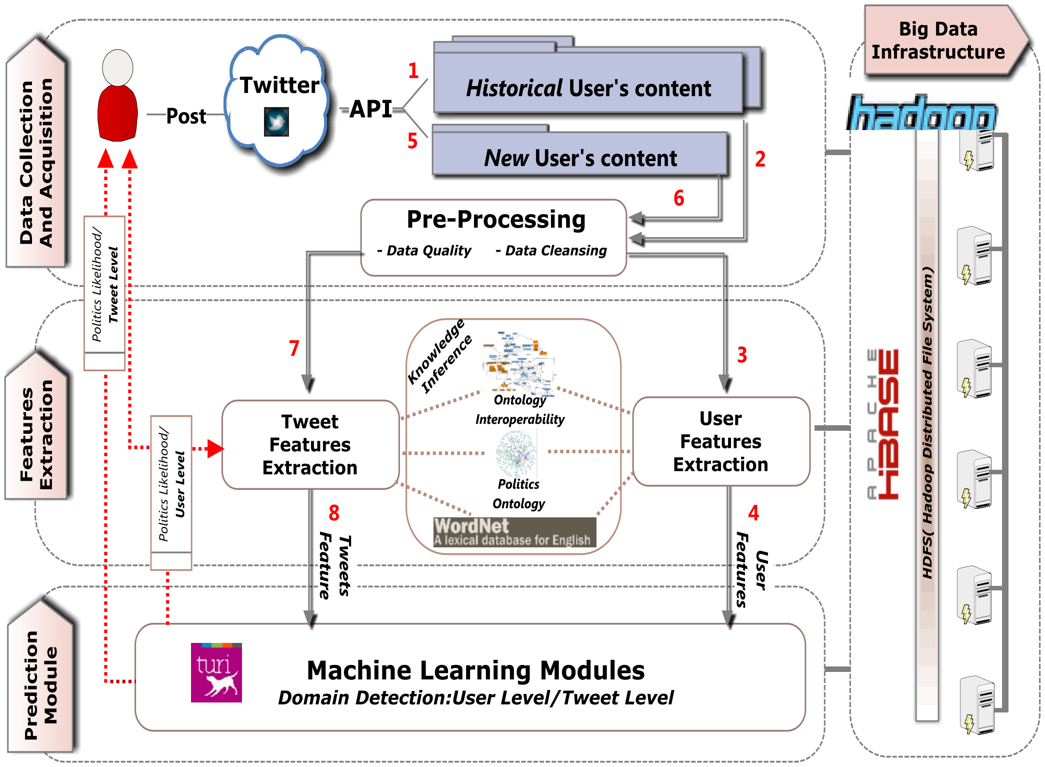


Figure 6-1: Semantic Analysis – Machine Learning: System Architecture

6.3 Data Collection and Acquisition

6.3.1 Data Generation and Selection

As indicated in the chapters above, Twitter has provided a rich dataset of over 500 million tweets daily which is around 200 billion tweets a year (Sayce 2016). Twitter mining is an emerging research field falling under the umbrella of data mining and machine learning. Twitter™ is the chosen subject of this research because: (1) Twitter is a fertile medium for researchers in diverse disciplines, leveraging the vast volume of content; (2) Twitter facilitates data collection by providing easy access APIs to The Twitter-sphere; (3) due to the economy and the ambiguity and brevity of a tweet’s content, it is challenging to determine the accurate domain(s) to which the user’s tweet is referring.

For proof of concept, this study is limited to an on/off domain classification to the content of OSNs. Hence, the political domain has been selected for the following reasons: (1) Twitter has been intensively incorporated as an important arena by politicians to express and defend their policies, to practice electoral propaganda and to communicate with their supporters (Shapiro and Hemphill 2017), (2) Twitter has raised considerable controversy regarding its usage as a platform to attack political opponents (Van Kessel and Castelein 2016), (3) Twitter is characterised by its growing social base to include broad political, social groups leveraged by ease of use, free access, and deregulated nature (Halberstam and Knight 2016), (4) the amount of the political discourse in social content is overwhelming; over one-third of OSNS's users believe that they are worn-out by the quantity of the political content they encounter (Duggan 2016). Such an abundance of data facilitates data aggregation and improves the outcome of the data analysis. For future work, this study aims to develop a multi-domain-based classification, leveraged by domain ontologies, semantic technologies and linked open data. Hence, besides the political domain, an analysis of other domains of interest may be further investigated in the future.

The dataset used for this study has been collected using Twitter's "User_timeline[1]" API method. This mechanism allows access to and retrieval of public users' content and metadata. The collection of the users' content was accomplished in two stages: (1) by collecting historical user content (up to "3,200" most recent tweets[2]). This dataset will be used to predict the user's interest in the political domain in general, and (2) by collecting the new content of those users

[1] https://dev.twitter.com/rest/reference/get/statuses/user_timeline.

[2] This threshold is set by Twitter™ as the maximum number of recent tweets the twitter API is allowed to retrieve.

whose historical tweets were obtained in the first phase. This is used to predict the political domain likelihood value of the new tweet. As will be described later, the dataset of the first stage is used to predict the user's interest in politics at the user level, i.e. to establish an understanding of the user's interest in the political domain based on the user's past content. The political domain likelihood value of the new user's tweet is predicted based on the analysis of its content, other than the politics interest likelihood value predicted at the user level.

6.3.2 Pre-processing Data

The veracity of data refers to the certainty, faultlessness and truthfulness of data (Demchenko et al. 2013). Although reliability, availably and security of data's nascence and storage are significant, these factors do not guarantee data correctness and consistency. Appropriate data cleansing and integration techniques should be incorporated to ensure certainty of data. The data collected for the user's content, and historical and new tweets, are pre-processed by data quality enhancement and data cleansing techniques which are discussed below:

- **Data cleansing** of user content is conducted by using the following techniques: (1) all redundant content (i.e. same dataset crawled more than once) such as tweets or user data is eliminated with their metadata; (2) removing stop words; (3) removing URLs; (4) decoding all HTML entities to their applicable characters; (5) eliminating all HTML tags such as <p>, <a>, etc.; (6) removing punctuation marks, correcting encoding format, etc.

- **In data quality enhancement,** the list of Twitter handles (a.k.a. Twitter user/screen name such as @example), which are indicated in the user's tweets, is collected and replaced with the user's corresponding names. This is achieved through the Application Programming Interface (API) of Twitter's "lookup[3]". These handles are normally neglected or deleted when mining user's tweets. However, these handles are important because they are used by Twitter users to mention other Twitter users in their tweets, replies or re-tweets. Hence, it is essential to identify and ascertain the actual names of those users. This assists in the process of domain extraction. For example, a user shows an interest in the political domain if she/he commonly indicates handles linked to politicians or political parties, in addition to publishing other politics-related content.

6.4 Feature Selection and Extraction

The pre-processed dataset is passed to the features extraction modules. For the new users, the features of their content (historical tweets) are extracted in the "User Features Extraction" module. As for the new tweets of the already existing users, features are extracted in the "Tweet Features Extraction" module.

This study aims to establish a fundamental ground for efficiently detecting the domain of interest of Twitter users, which will significantly contribute to a better understanding of the domain(s) of future users' tweets. As a proof of concept, the proposed system is validated by an application on the Political domain, where the proposed system attempts to detect whether the domain of a tweet is or is not politics-related. This validation is based primarily on former knowledge about a user's

[3] https://dev.twitter.com/rest/reference/get/statuses/lookup.

political interests obtained by analysing the user's historical content. To do so, the following politics-domain knowledge inference approach is designed to extract the semantics of a user's tweets, thereby uncovering the user's domain of interest.

6.4.1 Political domain Knowledge Inference

In the feature extraction module, domain knowledge inference is the main process used to extract user and tweet features from pre-processed datasets. For proof of concept, the study focuses on the political domain, using politics ontology, WordNet, and ontology interoperability to infer politics knowledge.

6.4.1.1 Politics Ontology and WordNet®

The political domain refers to the knowledge captured in politics ontology along with its knowledge base. BBC defines politics ontology as "an ontology which describes a model for politics, specifically regarding local government and elections". The BBC Politics ontology conceptualises a politics model especially for the UK government and elections. It was originally designed to cope with UK local government and European Elections in May 2014. This study applies the BBC Politics ontology to Australian politics by further extending politics concepts. BBC Politics ontology and its extension ontology harnessed in this chapter are depicted in Figure 5-4 and Figure 5-5 respectively. Furthermore, this study uses WordNet[4], which is a lexical dictionary used to construct relations between terms of synonymies. Synonyms (or synsets) are a set of interrelated terms or phrases which indicate the same semantic concept, such as the words "elections, public opinion poll, opinion poll, and ballot". All the synsets of the political concepts captured in politics ontology

[4] https://wordnet.princeton.edu/

depicted in Figure 5-5 are examined, and only the synonyms applicable to the political context are captured.

6.4.1.2 Ontology Interoperability

The interlinking with other relevant entities defined in other datasets supports interoperability. The approach taken in this study addresses information interoperability by focusing on the equivalence links that direct the URI to refer to the same resource or entity. The politics ontology supports the equivalence links between the ontology components and the tweet data. The resources and entities are linked through the owl#sameAs relation. This implies that the subject URI and object URI resources are the same, and hence the data can be further explored.

In the interlinking process, AlchemyAPI™ is incorporated as a one-stop shop, leveraging access to a wide variety of linked data resources[5] through providing easy access APIs. These resources include but are not limited to: different vocabularies such as Upper Mapping and Binding Exchange Layer (UMBEL), Freebase (which is a community-curated database for well-known people, places, and things), YAGO high-quality knowledge base, and DBpedia knowledge base, etc. These resources are used to help extend the knowledge base of the politics ontology by identifying (non-)Australian politicians and (non-)Australian political parties from users' tweets. For example, at this stage, "99,812" instances of "2009" politicians, and "48,704" instances of "59" political parties are captured in the politics ontology.

6.4.2 User Level Features

The political interest of users is primarily measured by two main proposed

[5] http://www.alchemyapi.com/products/alchemylanguage/linked-data.

factors: continuity and knowledgeability. Continuity refers to the frequent interest of a user in a certain domain. In other words, the user demonstrates an interest in the political domain by tweeting or retweeting content in this domain over a relatively long period. Continuity is measured by counting the number of political entities identified from the user' tweets in each period (such as every month, quarter, etc.). Knowledgeability (or Speciality) refers to the user's close acquaintance with the political domain and also refers to the user's dedicated pursuit of the political domain through a commitment such as work or study. Knowledgeability is measured by accumulating the distinct number of political entities annotated from the user's tweet, and the user's profile description. Table 6-1 shows the list of features used to classify whether the user's interest is **on-topic** or **off-topic**. **On-topic** refers to when the user demonstrates a continuous interest in the political domain. **Off-topic** users are those whose Twitter content shows their non-interest in the political domain.

Table 6-1: A List of User's Features

No	Features	Description
1	no_tweets, x_1	The total count of users' historical collected tweets up to 3,200 tweets.
2	unq_pol_entities, x_2	Total count of distinct/unique political entities extracted from all user's tweets
2	pol_entities_pre_QW_YYYY, x_3	Count of political entities annotated from the tweets posted before quarter 'W' of the year 'YYYY'
3	pol_entities_QW_YYYY, x_4	Count of political entities annotated from the tweets posted in quarter 'W' of the year 'YYYY'
4	pol_entities_QX_YYYY, x_5	Count of political entities annotated from the tweets posted in quarter 'X' of the year 'YYYY'
5	pol_entities_QY_YYYY, x_6	Count of political entities annotated from the tweets posted in quarter 'Y' of the year 'YYYY'
6	pol_entities_QZ_YYYY, x_7	Count of political entities annotated from the tweets posted in quarter 'Z' of the year 'YYYY'
7	profile_pol_entities, x_8	Count of political entities annotated from user's profile description
9	verified(Authentication Status), x_9	Authentication flag used for accounts of public interest (for example, politicians)

The features x_2 to x_8 as depicted in Table 6-1 are selected to primarily focus on users' ongoing interest in and knowledge about the political domain by extracting the political entities from their tweets and by leveraging the knowledge-inference tools explained in the previous section. In particular, features x_2 to x_8 are proposed to address the political knowledgeability of users. Moreover, features x_3 to x_7 address the continuing interest of users in the political domain. Features x_1 and x_9 are added to support the aforementioned features and will be discussed later in this research.

Unq_pol_entities (x_2), listed in Table 6-1, refers to the number of distinct political entities extracted from the history of a user's tweets. Profile_pol_entities (x_8) represents the number of all political concepts that are identified in the users' profile description on their Twitter accounts. The former feature represents the diversity of the political concepts embodied in the users' tweets, and the latter feature, x_8, is used to examine the explicit indication of the user's interest in the political domain, particularly if the user works in this domain. This is usually clearly indicated in their profile description.

The list of all political entities is counted periodically. The political entities extracted from the user content for each period is used to scrutinise political interest temporally rather than scrutinising the tweets as a whole. Therefore, the collected historical tweets are divided into five groups, x_3 to x_7. Four groups, x_4 to x_7, indicate the four sequential and recent quarters (W, X, Y and Z), where 'Z' is the most recent quarter and one group, x_3, indicates the rest of the tweets posted before the 'W' quarter. This mechanism is proposed because the user's interest(s) may change, and their knowledge may evolve. Hence, it is more efficient to examine the user's domain(s) of interest based on current and recent behaviours from the four-time groups. Furthermore, some users only show a particular interest in the political domain when popular political events are taking place. For example, a users' involvement in conversations during election campaigns does not necessarily indicate an interest in the political domain generally, as the election is a trending topic only, on which users with dissimilar interests share their thoughts, and/or anticipations about the potential candidates.

The remaining two features listed in Table 6-1 are the no_tweets and verified features. The no_tweets, x_1, relates to the number of collected historical tweets. This feature is important as a means of addressing the ratio between the number of political concepts accumulated for features x_2 to x_8 and the total number of tweets. For example, two users might archive the same number of distinct political concepts, although the number of tweets differs for each user. The verified feature, x_9, is the authenticated flag (i.e. blue verified badge ✓). Twitter may set this flag to '1' for users of public interest. Twitter currently offers this feature to help users find influential and high quality accounts in several domains.[6]

6.4.3 Tweet Level Features

In the previous section, the user's historical collected tweets were studied to obtain an accurate understanding of that user's interest in the political domain. A list of features extracted from the content at the user level is formulated and will be used to predict the user's political interest (likelihood). On this backdrop, the likelihood of the user's interest in the political domain would be the main driver facilitating an understanding of the domain of the users' future tweets. Table 6-2 summarises the list of features selected to predict the political likelihood at the tweet level.

[6] https://blog.twitter.com/2016/announcing-an-application-process-for-verified-accounts-0.

Table 6-2: A List of Tweet Features

No	Features	Description
10	political_entities, x_{10}	Count of political entities extracted from the tweet
11	words_count, x_{11}	Count of tweet's words
12	political_perc, x_{12}	Computed as $\frac{x_{10}}{x_{11}}$
13	pol_entities_recent_quarter, x_{13}	Count of political entities annotated from the user's tweets posted in the most recent quarter
14	user_pol_likelihood, x_{14}	Political likelihood value

As shown in Table 6-2, political_entities (x_{10}) represents the number of political entities annotated from the tweet using the knowledge above discovery tools. Words_count (x_{11}) is the number of remaining words in the tweet after the cleansing process. Political_perc (x_{12}) represents the ratio between the number of political entities annotated in the tweet to the total words used. Despite its brevity, a tweet might discuss more than one topic; thus, x_{12} is proposed as an indicator of the weight of the political domain in the tweet. Pol_entities_recent_quarter (x_{13}) represents the number of political entities from all tweets posted during the most recent quarter. This feature is included because it represents the user's most recent political (non-)interest. User_pol_likelihood (x_{14}) is the predicted value obtained from user analysis which signifies a user's general interest in the political domain.

Features x_{13}, and x_{14} are proposed to indicate the recent political interest of the user. These features assist in further understanding the actual context of the newly posted tweet, given their typically short length and ambiguity. Hence, users that have been predicted to be interested in the political domain will likely post politics-related content in future posts. This will be discussed and demonstrated further in the experiments section (Section 6.6).

6.5 Machine Learning Module for Classification

This section provides an overview of well-known machine learning classification algorithms. Based on the user and tweet features, $\bar{x} = [x_1, x_2, \ldots\ldots x_{14}]$, a machine learning module determines the likelihood of whether or not a user/tweet is in the political domain, namely y, where the following commonly used implicit or explicit classifiers including logistic regression (Hosmer Jr, Lemeshow, and Sturdivant 2013), decision tree (Quinlan 1993) (Ho 1995) (Friedman 2001), and support vector machine (Boser, Guyon, and Vapnik 1992), are used for user based classification, and logistic regression is used for tweet-based classification. For demonstration purposes, this overview will consider the domain-based classification at the user level. Logistic regression (Al-Tahrawi 2015, Yen et al. 2011), decision tree (Sharef et al. 2015), and support vector machine (Altınel, Can Ganiz, and Diri 2015, Dong et al. 2016) in particular have been used for text categorisations. Also these approaches are more narrow and computationally simpler than recently developed machine learning approaches such as the deep learning or deep networks approaches.

Development of a novel classifier is not the main research focus of this study. Hence, the study attempt to implement a computationally simple but effective approach. Five commonly used approaches are used, namely: logistic regression, support vector machine, top-down inducing based decision tree, random forest-based decision tree, and gradient-boosting-based decision tree.

6.5.1 Logistic Classifier

Logistic regression is commonly used for conducting binary classification tasks (Hosmer Jr, Lemeshow, and Sturdivant 2013). In logistic regression, the likelihood of whether the user is in the political domain is determined by a logistic function consisting of a linear summation of $\mathrm{x}_1$to x_9. The logistic function is given as:

$$f^{LR}(\bar{x}) = P(y=1|\bar{x}) = \frac{1}{1+\exp\left(-\left(b_0 + \sum_{i=1}^{14} b_i \cdot x_i\right)\right)} \tag{6.1}$$

In the study, $\mathrm{b}_0, \mathrm{b}_1,$ to , b_{14} are the logistic coefficients, which are determined by maximising the likelihood that the user is definitely in the political domain. Unlike linear regression which has normally distributed residuals, ordinary least square regression cannot be applied to determine the logistic coefficients. Hence, to determine "$\mathrm{b}_0, \mathrm{b}_1,$ to, b_{14}", here and elsewhere, Newton's method[7] is used. Newton's method begins with tentative logistic coefficients, and it adjusts the coefficients slightly to see whether they can be improved. It repeats this iterative process until the process converges. A user is classified in the political domain when the value of $f^{LR}(\bar{x})$ in (1) is large than 0.5. Otherwise, the user is classified as being in the non-political domain.

6.5.2 Support Vector Machine

Support vector machine (SVM) is commonly used for conducting binary classification tasks (Boser, Guyon, and Vapnik 1992) particularly involving with the confusion matrix analysis (true-positive and false negative). SVM is relatively new and was designed for applications involving text categorisation and recognition (see

[7] Newton Methods is known as an algorithm designed to find the roots of a function.

for example (Altınel, Can Ganiz, and Diri 2015, Dong et al. 2016)).

In SVM, the user is classified as either being in the politics or the non-political domains, based on the following formulation:

$$f^{SVM}(\bar{x}) = \operatorname{sgn}(D(\bar{x})) \tag{6.2}$$

where $D(\bar{x}) = \sum_{i=1}^{14} w_i \varphi(x_i) + b$; (6.3)

and $\operatorname{sgn}(D(\bar{x})) = \begin{cases} 0 \text{ if } D(\bar{x}) < 0 \\ 1 \text{ if } D(\bar{x}) \geq 0 \end{cases}$ (6.4)

φ is the transform function which is correlated to the kernel function and w_i with $i = 1,2,\text{to}14$ and b represents the SVM parameters. The five common kernel functions are; linear function, homogeneous polynomial, inhomogeneous polynomial, gaussian radial basis function and hyperbolic tangent. The kernel function is generally determined by a trial and error method. After the kernel function has been determined, w_i and b are reformulated as a quadratic programming problem, which is solved by the gradient descent algorithm. When the value of $f^{SVM}(\bar{x})$ in (2) is equal to 1, the user is classified as being in the political domain. Otherwise, the user is classified as being in the non-political domain.

6.5.3 Decision Tree Classifier

A Decision tree is a classifier which can express a recursive partition of the instance space. A decision tree is a flow-chart-like structure, where each internal (or non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The highest node in the tree is the root node. Figure 6-2 illustrates how a decision tree is used to determine whether a user is in the political domain. The study considers a simple decision tree with four features, x_2 , x_3, x_7 , and x_8. The red branches of the decision tree indicate that the user is in the political domain; this occurs if any of following three conditions are met: (1) if x_2 is larger than 100 and x_3 is larger than 50, (2) if x_2 is less than 100, x_7 is larger than 80 and x_8 is larger than 50, or (3) if x_2 is less than 100, x_7 is less than 80 and x_8 is larger than 100.

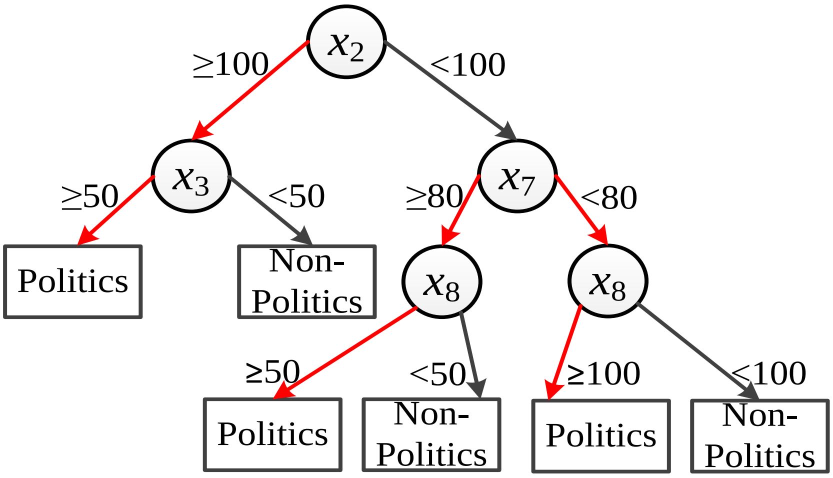


Figure 6-2: Example of a Decision Tree for the Political domain

Compared with logistic regression and SVM, decision trees are very intuitive and easy to interpret. In addition, empirical results have demonstrated that decision trees outperform SVM and logistic regression on 11 benchmark problems, regarding ten classification metrics (Caruana and Niculescu-Mizil 2006). Three commonly-used approaches, namely top-down inducing C4.5 (Quinlan 1993), random forest (Ho 1995), and gradient boosting (Friedman 2001) are used to develop decision trees for determining whether a user is in the political domain. In top-down inducing C4.5, the decision tree is constructed from top to bottom, based on a divide-and-conquer mechanism. The top-down inducing C4.5 trains the samples based on the splitting measures. After the selection of an appropriate split, which results in a minimum classification error, each node further subdivides the training samples into smaller subsets of samples, until the split gains satisfy the splitting measure. In a random forest, multiple trees are generated based on randomly selected subspaces of features. The trees generalise their classification in complementary ways, and their combined classification attempts to improve every single tree. In gradient boosting, a base decision classifier is expanded by adding additional branches to the base of the tree. The expansion continues until no further improvement can be obtained by adding branch.

6.6 System Evaluation

In previous sections, a system framework is proposed to detect the domain-based interest of users/tweets by incorporating machine learning. This section evaluates the effectiveness of the proposed system framework.

6.6.1 Datasets Collection and Ground Truth

To evaluate the proposed system framework, a list of Australian Twitter users and their public content was collected and pre-processed as discussed in section 6.3. The tentative list of users who are potentially interested in the political domain was selected from the following sources: (1) a list of Members of Parliament and Senators indicated on the official website of the Parliament of Australia.[8] (2) members and subscribers of three politics-based Australian Twitter lists,[9] and (3) miscellaneous

[8] http://www.aph.gov.au/.

[9] https://twitter.com/latikambourke/lists/australian-journalists/subscribers;
https://twitter.com/lizziepops/lists/politics/members ;
https://twitter.com/smh/lists/federal-politicians.

sources.[10] Due to the lack of online sources indicating those users interested in politics in OSNs, the lists above are selected because it is assumed that these people are interested in the political domain as is evident later in this study.

Users who are assumed to have little or no interest in the political domain were tentatively selected from two collected datasets: (1) members of various Australian Twitter lists established to discuss sports, information technology, and other non-political domains; (2) a list of Australian users who achieved the highest trustworthiness values in all domains except "news, government and politics", extracted from list of users collected and analysed as discussed in Chapter 5. The tentative selection criterion was established based on the user's profile description, choosing users who indicate a non-politics interest. The collected dataset thus comprises beside users who are interested in Politics domain, another category of users who have little or no interest in the political domain. This is to provide heterogeneous dataset which comprises manifold opinions and views and not politically polarized.

The collected and cleansed tweets of each user is then carefully examined to obtain an accurate understanding of the user's domain of interest, thereby establishing a truth dataset for developing and validating the proposed system framework at the user level. In this dataset, users are labelled and assigned to two categories: (1) **on-topic** users who show a particular interest in the political domain and (2) **off-topic** users who demonstrate no or minimal interest in the political domain. Table 6-3 shows a tentative list of collected users, and the actual number of users selected for

[10] http://earleyedition.com/2009/04/22/australias-top-100-journalists-and-news-media-people-on-twitter; Wikipedia: Australian political journalists : https://en.wikipedia.org/wiki/Category:Australian_political_journalists.

the ground truth, based on an examination of all tweets.

Table 6-3: Ground Truth - User level

	#Collected users (tentative list)	**Ground Truth**
on-Topic	310	227
off-Topic	350	283

The collected users of the ground truth dataset indicated in Table 6-3 are analysed with their historical tweets to develop the prediction model. This is used to predict the likelihood of users in the political domain.

The next phase involves conducting experiments at the user level to predict the politics classification of the new users' tweets. Therefore, another dataset is collected which contains new tweets posted by already-examined users. The new tweets are examined, and a subset of the tweets is selected to construct the ground truth for conducting experiments at the tweet level. The selection was based on four criteria; (1) tweets indicating a **political domain**, and posted by **politics** users; (2) tweets indicating a **political domain**, and posted by **non-politics** users; (3) tweets indicating a **non-political domain**, and posted by **politics** users; (4) tweets indicating a **non-political domain**, and posted by **non-politics** users. These four criteria are chosen to support the prediction model which will be constructed at the tweet level. Table 6-4 shows the total number of tweets collected based on the four selection criteria.

Table 6-4: Ground Truth - Tweets Level

	Politics users	**Non-politics users**
Politics tweets (on-topic)	150	125
Non-politics tweets (off-topic)	105	100

The proposed system framework is implemented in the Turi Graphlab Create™ which is used for these experiments using the Python programming environment(Low et al. 2014). Turi Graphlab Create is used as it is scalable and can, therefore, accommodate relatively huge datasets. The proposed system framework is used to conduct the experiment at the user level with the nine features (x_1 to x_{14}) illustrated in Table 6-1 and the five classifiers discussed in section 6.5, logistic regression (LR), support vector machine (SVM), top-down inducing based decision tree (TD-DT), random forest-based decision tree (RF-DT) and gradient-boosting-based decision tree (GB-DT). Turi Graphlab Create is also used to conduct experiments at the tweet level with the features listed in Table 6-2. 10-fold cross validation[11] is used on the datasets to evaluate the generalisation capability of the proposed system framework which is embedded with the five classifiers.

At the user level analysis, and as depicted in Figure 6-3, the proposed system framework can be used to determine (classify) whether or not a user is interested in politics. The circled ones are classified as the politics-interested users, and the non-circled ones are the users who are not interested in politics. Four scenarios are

[11] 10-fold cross validation: a process of randomly splitting datasets into 10 partitioned subsets. One subset is kept as the validation data for testing the model, and the remaining 9 subset are used to train the data. The cross-validation process is iterated 10 times where each subset is used once as a validation data.

illustrated by the classification as:

1. True-positives (TP), which indicate the number of actual politics users that are classified correctly as politics users;
2. False-positives (FR) which indicate the number of non-politics users that are classified incorrectly as politics users;
3. False-negative (FN) which indicate the actual politics users that are classified incorrectly as non-politics users; and
4. True-negative (TN) which indicate the non-politics users that are classified correctly as non-politics users.

These four scenarios can also be shown in the confusion matrix (Table 6-5) which depicts the performance of the prediction. The model illustrated in Figure 6-3 is also applicable to the tweets classification which is the second analysis phase of the proposed approach.

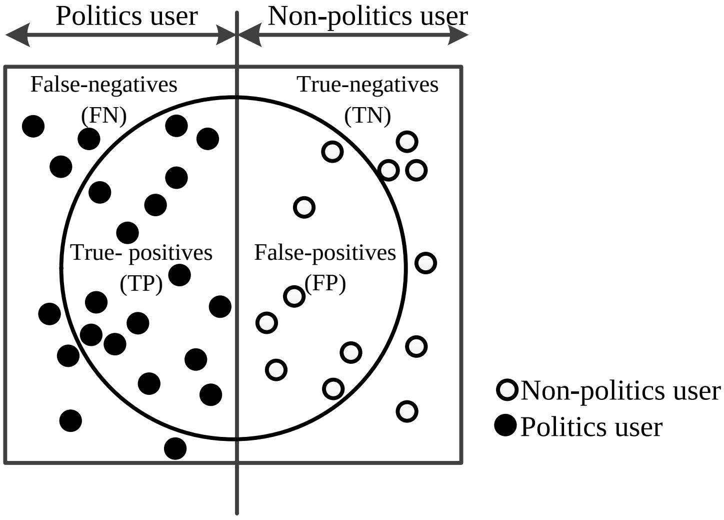


Figure 6-3: Classification of Politics/Non-politics Users

Table 6-5: Confusion Matrix

		Prediction	
		on-topic	**off-topic**
True	**on-topic**	TP	FN
	off-topic	FP	TN

In Graphlab Create ™, the confusion matrix is often a table used to provide further details on the true and false predictions. This table comprises three columns:

1. Target_label: the classification label of the ground truth. It represents the **on-topic** and **off-topic** label in this study;
2. Predicted_label: the classifier prediction label; and
3. Count: the number of times the predicted_label matches the target_label.

The evaluation has been performed by using the following metrics to evaluate the classification performance in predicting whether or not the user/tweet is in the political domain.

Accuracy indicates the correctness of the incorporated classifier in making the correct prediction. This is essentially the ratio between the correct predictions (i.e. $\mathrm{TP}+\mathrm{TN}$) and the total predictions ($\mathrm{FN}+\mathrm{TP}+\mathrm{FP}+\mathrm{TN}$). This is computed as:

$$Accuracy = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{FN}+\mathrm{TP}+\mathrm{FP}+\mathrm{TN}} \tag{6.5}$$

Log-loss (logarithmic loss) is a fine-grained classification evaluation metric. This value is computed by the negative of the accumulation of the log probability of each sample, normalised by the number of samples:

$$\textit{Log-Loss} = -\frac{1}{n}\sum_{i\epsilon 1,..N}(y_i \log(P_i) + (1 - y_i)\log(1 - P_i)) \qquad (6.6)$$

Where y_i is the i-th target value, and P_i is the i-th predicted probability. This metric is used because the likelihood probability is addressed to predict the **on-topic** or **off-topic** likelihood of the user or tweet.

Precision, Recall and F-score are metrics commonly used to evaluate classification performance. Precision, Recall and F-score are shown in (7), (8) and (9) respectively.

$$\Pr ecision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad (6.7)$$

$$\mathrm{Re}\,call = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad (6.8)$$

$$F\text{ - }score = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (6.9)$$

Precision indicates the ratio between the number of actual politics users/tweets that are classified correctly and the total number of correct and incorrect classifications of politics users/tweets. Recall indicates the ratio between the number of actual politics users/tweets that are classified correctly and the total number of actual politics users/tweets. Hence, high precision indicates that the classifier is capable of generating substantially more relevant predictions for actual politics users/tweets than the irrelevant ones. High recall indicates that the classifier is capable of generating most of the relevant predictions for actual politics users/tweets. Precision with a value of '1' indicates that every prediction is the actual politics user/tweet, but it does not mean that all the actual politics users/tweets are retrieved; while a recall score with a value of '1' indicates that all predictions are actual politics users/tweets, but it does not indicate the number of non-politics predictions that are retrieved. Hence, the F-score is used to provide the trade-off between precision and recall.

6.6.2 Domain Detection – User Level

The features above in Table 6-1 are analysed for each user where tweets are divided temporally into five groups to address the temporal dimension. The second and third columns in Table 6-1 show the feature values concerning the on-topic samples and off-topic samples respectively, where **on-topic samples** represent the list of users interested in the political domain and **off-topic samples** show the users who did not have an interest in the political domain. For the on-topic samples, the i-th feature is denoted as x_i on-topic. For the off-topic samples, the i-th feature is denoted as x_i off-topic. ARD (Absolution Relative Difference) in (10) is used to indicate the relative difference between the **on-topic samples** and the **off-topic samples.**

$$\text{ARD} = 100 \times \text{Abs}\left(\frac{x_i^{\text{on\_topic}} - x_i^{\text{off\_topic}}}{x_i^{\text{on\_topic}} + x_i^{\text{off\_topic}}}\right) \qquad (6.10)$$

The higher the ARD value, the higher the impact of the corresponding feature used to discriminate **on-topic** and **off-topic** users. For example, an ARD of x_8 equal to 100 indicates that x_8 is highly significant in identifying the (non-)interested users in the political domain by examining their profile description. This evidence will be discussed later.

Table 6-6: Dataset Statistics – User Level

	on-topic samples	**off-topic samples**	**ARD**
Total #Users	227	283	
Total #Tweets, x_1	499,475	611,014	10.044
Total #unq_pol_entities, x_2	14,818	2,833	67.9
Total #pol_entities_pre_Q3_2015, x_3	110,128	8,770	85.248
Total #pol_entities_Q3_2015, x_4	18,492	869	91.023
Total #pol_entities_Q4_2015, x_5	14,842	522	93.205
Total #pol_entities_Q1_2016, x_6	21,562	601	94.577
Total #pol_entities_Q2_2016, x_7	39,712	1,218	94.048
Total #profile_pol_entities, x_8	237	0	100
Total #Verified, x_9	167	94	27.969

As depicted in Table 6-6, the political entities detected in features x_2 to x_8 for **on-topic** users are much greater than the entities detected for the **off-topic** users. This is because **on-topic** users have shown extensive interest in the political domain through their content on Twitter.

To evaluate the effectiveness of the proposed system framework embedded with the five classifiers (LR, SVM, TD-DT, RF-DT and GB-DT) 10-fold cross-validations were used. In the cross-validations, the total observations (i.e. 510 users) are randomly split into two datasets, namely the training dataset (which is 80% of the total sample) and the validation dataset (which is 20% of the total sample). Table 6-7 illustrates the main settings and parameters used to train each of the five classifiers in the proposed system framework.

Table 6-7: Classifiers Settings

Classifier	Main settings	Parameters
LR	Hyperparameters- L1 penalty	0
	Hyperparameters-L2 penalty	0.01
	Solver	Newton-Raphson
	Solver iterations	9
SVM	Solver	L-BFGS[12]
	A predefined number of iterations	10
	Hyperparameters Mis-classification penalty	1
TD-DT	Number of trees	1
	Max tree depth	6
RF-DT	Number of trees	10
	Max tree depth	6
GB-DT	Number of trees	10
	Max tree depth	6

Table 6-8 depicts the confusion table used to quantify the performance of each classifier. It can be seen that the LR performs better in the classification task of this study; of the 107 samples used to validate each algorithm, only two samples were incorrectly classified by LR. However, all other classifiers, TD-DT and RF-DT algorithms for example, wrongly classified more samples in the prediction validations. Nevertheless, the classification performance of the incorporated algorithms is acceptable. These methods can perform effectively regarding this domain classification problem.

[12] L-BFGS: is a Limited memory of Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization algorithm. This is a robust solver for datasets with many coefficients

Table 6-8: Confusion Table

Target_ label	Predicted_ label	LR	SVM	TD-DT	RF-DT	GB-DT
on-topic	on-topic	59	58	41	57	48
off-topic	off-topic	46	46	52	45	45
on-topic	off-topic	2	3	2	4	3
off-topic	on-topic	0	0	3	1	1

Table 6-9 shows the evaluation performance metrics of each classifier, where the means and variances for the 10 fold cross-validations are given. The metric means that the non-bracketed values and the metric variances are the bracketed values. It can be seen from Table 6-9 that LR achieves the better metric means for the five classification metrics among all of the methods where Accuracy, Precision, Recall and F1_score are “the larger-the-better” and Log_loss is “the smaller-the-better”. The metric variances generated by LR are the smallest. Therefore, LR can yield the best and most robust classification when compared to the other four methods.

Despite the classifier’s convergence on the four metrics (i.e. Accuracy, Precision, Recall, and F1-score), LR is better than the other four methods, particularly regarding log_loss. This indicates that the predicted likelihoods of the validation dataset using FR closely match with the assigned labels. TD-DT, on the other hand, is the poorest method when compared with the others.

Table 6-9: Performance Comparison of Five Classifiers to Detect User Political Interest

	Accuracy	Log_loss	Precision	Recall	F1_score
LR	0.9824 (0.0002653)	0.0406	1.0000	0.9672	0.9833

	Accuracy	Log_loss	Precision	Recall	F1_score
SVM	0.9784 (0.003417916)	0.5781	1.0000	0.9508	0.9748
TD-DT	0.9157 (0.033453)	0.4816	0.9318	0.9535	0.9425
GB-DT	0.9255 (0.032357)	0.1321	0.9831	0.9508	0.9667
RF-DT	0.9490 (0.009473)	0.2321	0.9828	0.9344	0.9580

Note: Accuracy, Precision, Recall and F1_score are "the larger-the-better". Log_loss is "the smaller-the-better". Further details to the training and validation results of cross-validation are shown in Table A-5

Table 6-10 shows the highest estimated coefficient values calculated for each feature using LR. It shows that "profile_pol_entities, x_8" is the highest estimated coefficient. This is consistent with the results illustrated in Table 6-8, where x_8 has the highest impact when compared with the other features. This is due to the importance of this feature in distinguishing the user's interest in the political domain. In particular, users whose profile descriptions include politics-related entities such as a parliament member, political journalist, etc., are likely to suggest the political domain in their tweets.

Table 6-10: Highest Positive Coefficients- User Level

Feature	Value
profile_pol_entities, x_8	8.601
verified, x_9	2.162
unq_pol_entities, x_2	0.144
pol_entities_Q4_2015, x_5	0.02

In addition, the t-test (Box, Hunter, and Hunter 2005) was used to evaluate the significance of the hypothesis that the accuracy means obtained by the best method LR are higher than those obtained by the other methods (SVM, TD-DT, RF-DT and GB-DT). The t-values between LR and the other methods are shown in Table 6-11. Based on the t-distribution table, if the t-value is higher than 1.699, the significance is 95% confidence, which means that the accuracies obtained by the LR are higher than those obtained by the other methods with a 95% confidence level. The t-value can be determined by:

$$t\text{-value} = \frac{\mu_2 - \mu_1}{\sqrt{\left(\sigma_1^2 + \sigma_2^2\right)/N}}\ , \qquad (6.11)$$

where μ_1 is the mean accuracy obtained by the LR and μ_2 is for the other methods, σ_1^2 is the accuracy variance obtained by the LR and σ_2^2 is for the other compared methods. Is equal to 10 as this is 10-fold cross-validation. In general, the results indicate that there is no significant difference between LR and the other tested methods, although the LR can generally obtain better accuracies.

Table 6-11: T-values Between LR and the Other Tested Methods

	LR and SVM	**LR and TD-DR**	**LR and RF-DR**	**LR and GB-DR**
T-values	0.20842	1.1487	1.0703	0.99622

Therefore, the decision trees obtained by TD-DT can be interpreted and explained to executives of the user domain, as the accuracies obtained by TD-DT are similar to those obtained by LR. Figure 6-4 illustrates the resultant decision tree of the TD-DT classifier generated by Graphlab Create. It is evident that a feature is selected as a root node in TD-DT if this feature achieves the lowest classification error among the other features by applying the same dataset. The values associated with each leaf node in Figure 6-4 represent the “margins” which are a form of prediction showing the distance of samples from the decision boundary. The greater the distance, the higher the confidence in the classifier's prediction that the user is interested in the political domain. These margins can be converted to likelihood values (predictions) by applying the sigmoid function to the margins.

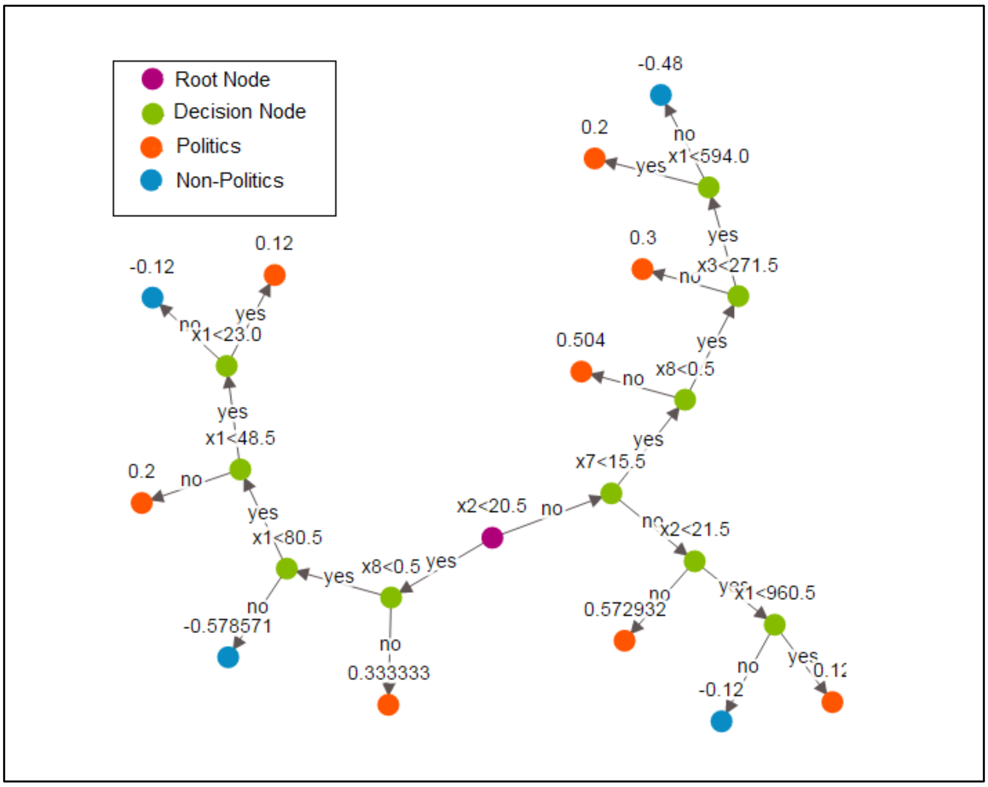


Figure 6-4: Decision Tree created by TD-DT

As depicted in Figure 6-4, feature x_2 (fuchsia node) has been selected by the classifier as the root node at which to split the tree. To evaluate this tree, the root node should be the first to start with, and follow the correct path through the decision nodes (green nodes) until the leaf node (red/blue node) is approached which indicates whether the user is interested in politics (red node) or not (blue node). For example, consider the two observations provided in Table 6-12; one indicates a user interested in politics (**@SenatorWacka**) and one does not show an interest in this domain (**@LabGallerie**). This Table shows the margins and the associated predictions for each sample. To apply the tree represented generated for **@SenatorWacka,** we start with the root node. "$x_2 < 20.5$" is **no** because $\mathbf{x_2}$ SenatorWacka = 42, "$x_7 < 15.5$" is **no** because $\mathbf{x_7}$ SenatorWacka = 20, "$x_2 < 20.5$" is **no** because $\mathbf{x_2}$ SenatorWacka = 42. This leads to a red leaf with the value of "0.572932" which represents a user who is interested in the political domain. The application of the same tree on **@LabGallerie** leads to a blue leaf with a value of -0.578571, which indicates a non-politics user. This is evident in both users, whose classification labels match with the resulting predictions.

Table 6-12: Margins and Predictions of Two Samples

Twitter_ID	x1	x2	x3	x4	x5	x6	x7	x8	x9	Label (1:Politics, 0:non-Politics	Margins	Predictions
@SenatorWacka	880	42	468	12	3	1	20	1	1	1	0.572932	0.639440
@LabGallerie	1498	4	19	1	2	7	3	0	0	0	-0.578571	0.359261

6.6.2.1 A Comparison with LDA and SLA

As discussed, LDA and SLA are statistically well-known models used for several topic modelling applications. In this section, an experiment is conducted to benchmark the applicability of the proposed model at the user level against these two models, to identify a user's main topic of interest. Gensim's python implementation (Rehurek and Sojka 2010) of LDA and SLA is used. The collected historical tweets of two politicians' accounts (i.e. @sarahinthesen8 and @stephenjonesALP) have been fed to the three models: LDA, SLA and the developed model incorporating a Politics Knowledge Inference. The experimental settings for LDA and SLA are set to one topic modelling, and the extracted terms indicate the 25 most contributed terms to this topic. In this approach, the top 25 frequently annotated entities from the user's tweets are extracted. Table 6-13 and Table 6-14 show the top 25 terms/entities extracted using the three approaches for @sarahinthesen8 and @stephenjonesALP respectively.

Table 6-13: Top entities/terms Extracted using LDA, SLA and the Developed Approach For @sarahinthesen8

LDA	**LSA**	**Politics Knowledge Inference**	
		Entity	SubType
refuge	refuge	Government of Australia	Organization
young	young	Australian Greens	Political Party
sarah	sarah	Member of Parliament	Politician
hanson	hanson	Elections	Event
nauru	nauru	Australian Labor Party	Political Party
children	children	Parliament	Organization
detent	detent	Liberal Party of Australia	Political Party
govt	govt	Malcolm Turnbull	Politician
australia	australia	Peter Dutton	Politician
green	green	Tony Abbott	Politician
abbott	abbott	Politics	Ontology
today	today	Sarah Hanson-Young	Politician
asylum	asylum	Electorate	Voter
manu	manu	Council	Organization
aust	aust	Politician	Person
people	people	inequality	Political_Slogan
senate	senate	Coalition	Political_Slogan
seeker	seeker	Joe Hockey	Politician
abuse	abuse	George Brandis	Politician
news	news	Liberal National Party of Queensland	Political Party
minister	time	welfare	Political_Slogan
time	minister	Barnaby Joyce	Politician
dutton	dutton	Nick McKim	Politician
turnbull	turnbull	Kristina Keneally	Politician
australian	australian	Simon Birmingham	Politician

Table 6-14: Top Entities/terms Extracted Using LDA, SLA and the Developed Approach For @ stephenjonesALP

LDA	LSA	Politics Knowledge Inference	
		Entity	SubType
illawarra	illawarra	Member of parliament	Politician
qt	qt	Elections	Event
today	today	Parliament	Organisation
great	great	Australian Labor Party	Political Party
mp	mp	Government of Australia	Organisation
stephen	stephen	Liberal Party of Australia	Political Party
good	good	Coalition	Political Slogan
post	post	Tony Abbott	Politician
school	school	Council	Organisation
abbott	abbott	Anthony Albanese	Politician
jone	jone	Politics	Ontology
photo	photo	Julia Gillard	Politician
auspol	auspol	Electorate	Voter
parliament	day	Greg Combet	Politician
day	jame	Sharon Bird	Politician
jame	parliament	Joe Hockey	Politician
big	big	Mark Butler	Politician
support	support	Malcolm Turnbull	Politician
nbn	nbn	Kate Ellis	Politician
house	house	Barack Obama	Politician
facebook	facebook	Joel Fitzgibbon	Politician
time	time	Jamie Briggs	Politician
fb	fb	Australian Greens	Political Party
australia	australia	Steven Ciobo	Politician
purser	purser	Greg Hunt	Politician

The list of the top contributed terms identified using 1-topic modelling for each user incorporating LDA and SLA illustrates the inadequacy of these approaches in identifying a high-level topic. Further, some of the inferred contributed terms using LDA and SLA are jumbled terms and don't make sense such as the last indicated terms in Table 6-13. On the other hand, with the top 25 entities annotated for both users using the developed approach, the high-level topic (i.e. politics) is highly noticeable. In the developed proposed system framework, each entity is linked with a specific class in the ontology. The knowledge obtained for each entity can be enriched to facilitate the overall semantic interlinking which leads to a better understanding of the domain of knowledge. Interlinking and enrichment are not applicable to LDA and SLA. Furthermore, all the top entities annotated using the developed proposed system framework indicate politics entities, although some of the most frequently occurring terms extracted using LDA and SLA are political entities. In a nutshell, the outcome of this experiment shows the applicability and effectiveness of the developed proposed framework.

6.6.3 Domain Detection – Tweet Level

Table 6-15 shows the statistics of the dataset used for this experiment at the tweet level. The new tweets are collected from the list of users indicated in the previous section. These tweets represent the new tweets posted after quarter 2, 2016. Hence, the tweet-level experiments are conducted on the set of tweets which have not been included in the user's historical tweets as discussed in the previous section.

The features shown in Table 6-2 are formulated for each tweet. **On-topic samples** in Table 6-1 represent the list of tweets labelled as politics tweets. **Off-topic samples** show the list of tweets labelled as non-political tweets. The ARD value is calculated for each feature. Table 6-15 shows the statistics calculated for the ground truth which is used to classify tweets according to a particular domain. It is evident that the calculated ARD for the two mean values of x_5 is the smallest value due to the noticeable convergence of $\boldsymbol{x_5}$ in both categories. This is because a user who has been classified as belonging to the political domain does not necessarily post all of his/her future tweets in this domain. Likewise, a user who has been classified as a non-politically interested user may show an interest in this domain in future tweets. Nevertheless, x_5 is most likely to distinguish the ambiguous political entities annotated from the textual content of a tweet, thereby helping to accurately ascertain the tweet's domain. This will be discussed further in this section.

Table 6-15: Dataset Statistics – Tweet Level

	on-topic samples	**off-topic samples**	**ARD**
Total #Tweets	255	225	
Total #political_entities(x_1)	880	71	85.068
Total #words_count(x_2)	3,762	2,391	22.282
Average political_perc, (x_3)	0.249	0.033	76.596
Total #pol_entities_recent_quarter(x_4)	65,049	37,248	27.177
Average user_pol_likelihood, $\mathbf{x_5}$	0.638	0.563	6.245

Due to the ability of the LR to detect the domain of interest at the user level, LR is further used to classify tweets in this phase with the same set of parameters listed in Table 6-7. To validate the efficiency of the proposed approach, 10-fold cross validation is performed where the 480 samples are randomly split into a training dataset (80%) and a validation dataset (20%). To further validate the effectiveness of the proposed approach, another experiment was conducted which excluded user_pol_likelihood, x_5 from the feature sets. This is to measure the significance of this feature to predict a tweets domain. Table 6-16 shows the confusion table used to quantify the performance of the LR classifier in each experiment, where Exp.1 refers to the first experiment conducted incorporating all features listed in Table 6-2. Exp.2 refers to the second experiment conducted on the same dataset excluding x_5.

Table 6-16: Confusion Table –Tweet Level

Target label	Predicted label	Exp.1	Exp.2
on-topic	on-topic	58	56
off-topic	off-topic	42	39
on-topic	off-topic	0	3
off-topic	on-topic	0	2

As depicted in the confusion matrix in Table 6-16, Exp.1 achieved better results than Exp. 2; incorporating all features including the past user's political prediction (x_5) leads to zero incorrect classifications. However, eliminating x_5 from the list of features results in five out of 100 incorrect classifications. This is confirmed by the comparison of the performance results of the two experiments illustrated in Table 6-17.

Table 6-17: Performance Comparison of Two Experiments – Tweet Level

	Accuracy	Log_loss	Precision	Recall	F1_score
Exp.1	1	0.01	1	1	1
Exp.2	0.95	0.072	0.949	1	0.957

Despite the convergence in each metric listed in Table 6-17, the predicted likelihoods of the validation dataset incorporating all features closely match the assigned labels.

Table 6-18: Highest Positive Coefficients- Tweet Level

Exp.1		Exp.2	
Feature	Value	Feature	Value
political_perc, x_3	24.86	political_perc, x_3	25.126
user_pol_likelihood, x_5	12.095	political_entities, , x_1	1.823

Exp.1		Exp.2	
political_entities, x_1	5.409	words_count, x_2	0.825
words_count, x_2	0.623	pol_entities_recent_quarter, x_4	0.009

Table 6-18 shows the highest estimated coefficient values calculated for each tweets feature in each of the conducted experiments. It is evident that political_perc (x_3**)** obtained the highest coefficient value in Exp.1 and Exp.2. This is due to the impact of the tweets political weight, indicating the tweets domain. This feature is supported by considering the number of political entities (x_1) and the total number of words in the tweet, x_2. User_pol_likelihood (x_5**)** obtained the second highest estimated coefficient after conducting Exp.2. This is due to the significance of incorporating former knowledge about the user's political interest in the process of predicting the domain of their future tweets.

Table 6-19 elucidates further the significance of incorporating x_5. Table 6-19 shows two real tweets of the ground truth dataset; one is labelled "politics" and the other is labelled "non-politics", posted by two Twitter users (i.e. @tamaleaver, non-politics user, and @peterjblack, politics user). The list of features included in Table 6-2 is calculated for each tweet. As depicted in Table 6-19, features x_1, x_2, and x_3 obtained the same values for each tweet. This exacerbates the process of obtaining the correct domain by considering only the number of political entities and counting the words in each tweet. It is evident that features x_4 and x_5 are important for identifying the tweets domain due to their significance for the classification task.

Assumption: It is argued that the annotated political entity of a tweet posted by a user who has already been predicted to be interested in the political domain, and who has included a relatively large number of political entities annotated in their tweets, is likely to indicate an actual political concept. Likewise, a user who has not shown an interest in politics in the past is not likely to indicate politics-related content in future tweets. This helps to eliminate the ambiguity for those entities which might have dissimilar meanings in several contexts. Moreover, this applies to all domains of knowledge.

Therefore, despite obtaining one political entity (Labour/Labor) for each tweet in Table 6-19, these tweets convey two different messages which are unrelated regarding context.

Table 6-19: Features Extracted from Two Tweets Posted by Two Users (Politics and Non-politics)

Twitterer	Tweet	x_1	x_2	x_3	x_4	x_5	Label
@tamaleaver	“Researching microcelebrity: Methods, access and labour, Jonathan Mavroudis”.	1	7	1/7	4	0.019	non-politics
@peterjblack	“Labor could support 'self-executing' same-sex marriage plebiscite”.	1	7	1/7	927	0.98	politics

6.7 Conclusion

This chapter presents an approach intended to provide in-depth insights into users’ domains of interest inferred from their pervasive propagation of tweets. This is achieved through a systematic approach that begins by addressing the volume feature of SBD, incorporating data generation and acquisition techniques, and then inferring the added value obtained from the data analysis. It is anticipated that this will contribute to an advanced domain-based trustworthiness approach that can filter out unsolicited tweets and increase the value of content. To achieve this objective, this research presents a consolidated framework leveraging former knowledge obtained from an analysis of the user's historical content. In this context, the political domain is used to determine the user's interest in this domain. Hence, an effective approach to classify Twitter users and their new updates is proposed according to two main categories: (i) **on-topic**: a user or tweet is classified under the political domain; (ii) **off-topic**: a user or tweet is classified under the non-political domain.

The major contributions of this chapter are as follows:

- A time-aware framework incorporating comprehensive knowledge discovery tools and well-known machine learning algorithms is proposed for domain-based discovery, which applies to the Twittersphere platform and customisable to other OSNs.
- Unlike current statistics-based topic distillation techniques which retrieve search results but neglect the temporal dimension, the proposed approach is better able to address the temporal factor; users' knowledge evolves and their interest might be diverted elsewhere depending on their experience, work, study, or other factors.
- Unlike current unsupervised statistical approaches, the proposed approach incorporates supervised machine learning techniques to perform domain-based classification task for the already semantically-enriched temporally-segmented textual content.
- The conducted experiments using the Twitter platform as one of the dominant OSNs verify the effectiveness and applicability of our model as evident in the outstanding results of several performance evaluation metrics.

In particular, this chapter introduces a framework comprising three main components, namely: (1) data collection and acquisition, (2) features extraction, and (3) the prediction module. The proposed framework comprises two analysis phases. In the first phase, the users' historical tweets are collected; their interest is examined over time, thereby providing a prediction of the users' interest, while taking the temporal factor into consideration. In the second phase, the prediction likelihood values obtained in the first phase have been utilised to predict the domain of the users' future tweets. Users' classification is achieved through the use of well-known machine learning classifiers.

The framework is evaluated, and its feasibility is demonstrated by means of three significant experiments: (i) the evaluation performance of the incorporated five classifiers (i.e. LR, SVM, TD-DT, GB-TD, and RF-DT) have proven successful in the domain-based classification task at the user level. This is strongly indicated by the results obtained by means of four evaluation metrics (i.e. Accuracy, Precision, Recall, and F1_score); (ii) the developed approach is benchmarked against well-known bag-of-words statistical approaches (i.e. LDA and SLA). The outcome of this experiment shows the applicability and effectiveness of the proposed framework in identifying high-level topics, domain-based semantic interlinking and enrichment task, and the notable performance of the retrieved top annotated entities compared to LDA and LSA models; (iii) the discovery andunderstanding of the user domains of interest facilitate the tests performed at the user's post level. This is evident through the promising results of two experiments conducted at the tweet level which show high accuracy in predicting the domain inferred from a short text message (i.e. tweet).

The next chapter revisits the research issues and problems addressed in this thesis. The set of improvements made to redress the limitations of the frameworks proposed in this thesis are introduced followed by recommendations and suggestions for future research directions.

Chapter 7 Recapitulation and Future Work

7.1 Research Overview

Over the recent decade, thanks to the communicative technologies produced by the scientific revolution, people have been able to exchange knowledge and experiences at a speed that has overcome the constraints of time and space. This can lead to the dissolution of cultural differences between groups and individuals. These new information and communication technologies have opened new horizons and brought profound changes to various aspects of human, cultural, intellectual and social life. They have also had a significant impact on all forms of human communication, opening the way to the realisation of the "Global Village" referred to by Marshall McLuhan (McLuhan 1994).

These technologies leveraged by the widespread use and the prevalence of the Internet have not only produced enormous technical, economic, educational and social benefits for humanity, but have contributed to the emergence of new forms of media which consist of different norms, orientations and means. OSNs are one of the most significant outcomes of the communication revolution. They have firmly established themselves and have significantly helped to transform the user from a recipient of information, as in traditional media, into an information producer and contributor. OSNs are public in nature and easy to access, enabling anyone to establish a platform through which to communicate with a wide audience of people. This has contributed favourably to raising community awareness, bringing views together, and to the exchange of thoughts, information and knowledge. However, the utilisation of these platforms has unfortunately led to the propagation of incorrect, misleading and deceptive information. Namely, instead of being used to promote informed social practices by disseminating accurate information and facilitating communication between people, and accelerating it to fight corruption, fanaticism and lies, OSNs platforms have been hijacked and misused to spread misinformation, rumours, and poor data. This causes great damage to values and ideals and even disrupts security and stability.

In light of the above, it is essential to create effective and applicable technical and analytical solutions capable of thoroughly examining the social content and distinguishing the credible users and content from the untrustworthy users and their content. In this thesis, several approaches for data analytics aimed at domain-based classification and trustworthiness of users and their SBD content are presented. The previous chapters have presented these approaches in detail, in addition to describing the various experiments conducted to demonstrate their applicability to the research issues.

The next section revisits the research issues that are central to this thesis. The last section makes several suggestions for areas of improvement and the avenues available for future research.

7.2 Recapitulation of the Thesis

The rapid growth of enterprise requirements in conjunction with an increase in the volume of modern data repositories, and the nature of the data that can be stored, have made traditional statistical methods inadequate in meeting all data analysis requirements. This has necessitated the development of advanced data analytics to extract useful knowledge from such vast volumes of data.

The main objective of this theses is to develop approaches for data analytics in order to derive knowledge and infer value from SBD. Data sources have increased from transactional data sources and limited external data sources to many other data sources such as data coming from a global environment in the form of news, etc. and from VoM and VoC in the form of OSNs, web blogs etc. This thesis presents solutions to address certain research issues summarised as follows:

1. Propose a framework to infer the value and determine the veracity and domain-based credibility of SBD.

   The widespread use of SBD has pointed the research community in several significant directions. In particular, the notion of social trust has attracted a great deal of attention from information processors/computer scientists and information consumers/formal organisations. This is evident in various applications such as recommendation systems, viral marketing and expertise retrieval. One of the main reasons for determining the value of SBD is to provide frameworks and methodologies by means of which the credibility of OSNs users can be evaluated. These approaches should be scalable to accommodate large-scale social data. Hence, there is a need to have a thorough understanding of social trust in order to improve and expand the analysis process and infer the credibility of social big data. Given the environment's exposed settings and the fewer limitations imposed on OSNs, the medium allows legitimate and genuine users as well as spammers and other untrustworthy users to publish and spread their content. Hence, it is vital to measure users' trustworthiness in numerous domains and thereby define domain-based influences and filter out untrustworthy users. Most of the current methods applied to evaluate the credibility of OSN users are generic-based approaches. The developed framework aims to address the current lack of a mechanism for the evaluation of domain-based trustworthiness.

2. Propose an approach to extract domain knowledge captured from the textual content of SBD.

The challenge of managing and extracting useful knowledge from social media data sources has attracted much attention from academia and industry. One of the major challenges of OSNs analysis is to be able to better understand the domain of knowledge in which the user is interested. This problem is exacerbated by: (1) inconsistent user behaviour (a user's interests can evolve and change over time), and (2) the brevity and economy of tweet content. Hence, understanding users' domain(s) of interest is a significant step in addressing their domain-based trustworthiness by acquiring an accurate understanding of their content temporally in OSNs. The developed framework proposes an ontology-based approach to extract semantics of textual data and define the domain of data. Semantic analysis, through the utilisation of existing Ontologies and Linked Data, enables knowledge to be elicited from social data, thereby enriching its textual content to deliver semantics and link each message with a particular domain. This process of semantic extraction enables the intended meaning of the textual content to be understood, which leads to inferring the actual interests of OSN users.

3. Propose an approach incorporating knowledge discovery and data classification techniques for the purpose of domain-based detection and classification.

In light of the general perception of advanced data analytics, companies compete to implement unconventional social data analytics in order to establish effective marketing strategies, taking advantage of the interactive feature of OSNs. Thus, to create the desired interaction with their customers, companies use many modern forms of communication to attract customers and visitors to their online social platforms. Consequently, it is necessary for companies to analyse the customers' social content and allocate the customers to appropriate categories, so that the right message is delivered to the right consumer category. If companies succeed in building effective clusters of customers and then determine the basic criteria for each cluster in making their buying decisions, companies will be able to take clear actions to implement them. Twitter is designed to track public figures and news and provide a platform for users to follow their friends and associates. The "maximum 140 characters" feature has made Twitter particularly important and widespread; however, this feature constricts the size of each user's published content which is needed to conduct an adequate analysis.

The most well-known approaches for inferring users' topics of interest (such as LDA-related techniques) fail to address several key issues, namely: (1) the inability to consider the semantic relationships of the terms in the user's textual content; (2) the inadequacy of their application to a topic modelling technique using short text messages such as tweets; and (3) the high-level topic classifications that use these bag-of-words statistical techniques are inadequate and inferior. The proposed approach provides an effective system framework to address these issues by taking into consideration the sporadic quality of tweets, and resolving the problematic issue of obtaining a factual understanding of the contextual meaning of a user's social content, thus placing the user's content in its appropriate category.

4. Evaluation of the proposed approaches and their frameworks.

This thesis aims to develop approaches for data analytics to facilitate the process of obtaining an accurate understanding of the textual context of the users' content and discovering the users' interest, both of which allow their domain-based trustworthiness to be determined. The implementation of the proposed approaches involves the development of prototype systems incorporating several technical solutions. These systems have been deployed and experimentally evaluated through several performance measures, benchmark comparisons, and case studies.

7.3 Future Work

This thesis is a report on work in progress as it is an ongoing project the purpose of which is to develop a methodology for Social Business Intelligence (SBI) that incorporates Semantic analysis and Trust notions to enrich textual data and consider data trustworthiness respectively (Abu-Salih, Wongthongtham, Beheshti, et al. 2015, Abu-Salih, Wongthongtham, Zhu, et al. 2015, Abu Salih et al. 2015, Wongthongtham and Abu-Salih 2015). The approaches developed in this thesis have produced optimistic results. However, there are certain limitations that need to be addressed and possible enhancements to be elucidated and marked as future work. Future work will focus on: (i) enhancing the proposed approaches of semantic extraction, domain-discovery and domain-based credibility and classification; (ii) devising a methodology integrating semantic analysis and trust notions for SBI. The following sub-sections explain in more detail the future research directions.

7.3.1 Future Work on Enhancements to the Proposed Approaches

7.3.1.1 Enhancements to the Domain-based Credibility Approach

CredSaT is an ongoing project. For future work, several enhancements will be implemented to consolidate the proposed approach:

- CredSaT will be improved to handle two additional BD features: Variety through the importation of more data sources; and Velocity through the addition of a new module to measure the credibility of the new content in real time (i.e. assign a credibility value to a new user’s tweet).
- AlchemyAPI has been used in this framework as the sole semantics provider. Although this service provider is supported by IBM, a prestigious software company, the resultant semantics could be further enhanced by utilising an ontology-based approach. In particular, domain knowledge is

captured in ontologies which are then used to enrich the semantics of tweets provided with a specific semantic conceptual representation of entities that appear in the tweets.

- A new graph-based model will be created to propagate the users' credibility throughout the entire network. Hence, an enhanced version of Twitterrank (Weng et al. 2010) is anticipated that takes into consideration the semantics of the textual content and the temporal factor.
- An anomaly detection approach will be developed that incorporates machine learning and an advanced list of features.

7.3.1.2 Enhancements to the Approach for Semantic Data Extraction from SBD

The developed approach for semantic data extraction from SBD has produced optimistic results. It shows a capacity to infer semantically annotated concepts and entities from the textual content leveraged by light-weight ontologies and other semantic repositories. However, the mechanism followed in this approach can be enhanced further:

- AlchemyAPI has been harnessed and evaluated in the proposed approach as the semantic provider, which has proven to be superior to the other semantic analysis tools. However, in future work, other tools can be utilised and evaluated.
- Twitter user_timeline REST API method has been used to collect public tweets from Twittersphere, which is mainly used to retrieve the historical tweets posted by a certain Twitterer_id. In future work, experiments involving other Twitter APIs will be conducted in order to examine and

analyse the textual content related to hashtags and topics mapped, and the latest news and world events and trends.

- Comprehensive ontologies are being continuously updated by applying machine learning technologies, i.e. driving data to obtain the domain knowledge (a reversal of the proposed approach).
- Twitter has been the sole social network on which the experiments and case studies are performed. Hence, a future goal is to analyse other social media networks such as Facebook, LinkedIn, Weblogs etc.

7.3.1.3 Enhancements to the Approach for Domain Discovery Incorporating Machine Learning

Through experiments conducted using the Twitter platform as one of the dominant OSNs, the proposed approach for domain discovery incorporating machine learning techniques has established an important foundation for a better understanding of user interest in several domains of knowledge. This is achieved by incorporating domain-based ontologies, semantic web analysis and machine learning classification techniques to gain better familiarity with user interests. This facilitates the process of measuring user credibility in each domain of knowledge. The following are the possible enhancements and research directions to be addressed in the anticipated future work:

- Besides politics, a domain-based analysis of several domains of knowledge will be conducted to acquire a more comprehensive insight into each domain. This is to facilitate the development of several domain-based ontologies leveraged by semantic web technologies and Linked Open Data.

- Machine learning will be utilised to achieve the abovementioned research objectives through multi-classification applications, to predict the likelihood of user interest in several domains of knowledge.
- Mobile cloud computing technologies have good potential for the future of social media. Mobile devices such as iPhones, tablets, laptops, smartphones etc can be connected to the Internet. The Internet of Things can capture all social media data. The centre cloud using the machine learning algorithm, such as deep learning, can be harnessed to analyse people’s needs and behaviours inferred from social data stored in the cloud.
- The current machine learning approaches assume that uncertainty and incompleteness do not significantly affect the accuracy of the Twitter classification. In fact, data uncertainty and incompleteness may exist. In the future, we will formulate the data uncertainty and incompleteness as fuzzy numbers which can be used to address imprecise, uncertain and vague data. Based on the fuzzy numbers, a fuzzy-based machine learning algorithm will be developed to estimate the effect of data uncertainty and incompleteness.

7.3.2 Future Work on a Methodology for Social Business Intelligence

In a competitive environment, one of the main challenges facing organizations over the past few years is how to understand data and discover hidden values embedded within in order to deliver timely, accurate, and advanced information and knowledge for decision-making. Identifying the best customers and their behaviour (such as what they buy, when they buy, the frequency of their shopping and how much they spend during each visit), reviewing product profitability over time, sales history of each store over time, understanding geographic regions and many other factors are the key objectives of all businesses, and can be achieved only by using effective tools for data collection, data transfer, storage and analysis. The purpose of business intelligence is to support better decision-making (Power 2007). Business Intelligence (BI) refers to the set of software and hardware solutions (data warehousing, data mining, OLAP, etc.) which add value to enterprises by providing new insights about data using sophisticated analysis tools (Chaudhuri, Dayal, and Narasayya 2011b). Moreover, the goal of BI is to translate collected historical data to useful business knowledge to support decision-makers.

There are different types of data, ranging from structured data in relational databases to unstructured data in file systems and semi-structured data that is neither raw nor strictly typed as in conventional database systems. Structured data is usually produced by the day-to-day operational activities of a business. However, most businesses also produce unstructured or semi-structured data that need to be discovered, i.e. those data produced by communication between business and customer such as customer feedback, contracts, complaint emails or transcripts of telephone conversations which are in semi-structured or unstructured formats. Moreover, the widespread increase of several OSNs has given businesses the opportunity to study customer views and market data on very large scales and for very large populations (De Choudhury et al. 2010). As a result, analysts today are able to conduct in-depth analysis of external business data such as customer blog postings (Gruhl et al. 2004), Internet chain-letter data (Liben-Nowell and Kleiberg 2008), social tagging (Anagnostopoulos, Kumar, and Mahdian 2008), Facebook news feed (Sun et al. 2009) and many other data sources. Hence, this demand for real-time business intelligence and the popularity of OSNs has created a need for SBI. SBI aims to reveal the fundamental factors derived from social perspectives that determine an organisation's performance.

The importance of trust in the social media context comes from its importance for market analysis, listening to the VoC and for sentiment analysis to feed BI applications (Berlanga et al. 2014a). Therefore, a methodology is required to infer the trustworthiness of unstructured data from different sources such as social media networks, news agencies, and web logs, and to store a collection of trustworthy unstructured data using the existing data warehousing solutions. In a related context, semantically enriching textual data and imposing structure on the unstructured nature of SBD when transferring these to data warehouses is another significant task. Discovering the semantics of social data will enhance the quality and accuracy of data stored in data warehouses which will dramatically affect the decision-making process as well as the quality of extracted reports. The application of semantic web technology resolves the issue of the ambiguity of data and provides metadata which helps related data to be understood and interpreted accurately. Meanwhile, ontology is utilised to define and collect semantically-related concepts and relations between concepts (Ahmed and Gerhard 2010). This could be done in particular by using existing ontologies (or new ones) which facilitate the extraction of data semantics. Although a few approaches have incorporated semantic web technology and trustworthy social data to feed BI platforms, there is still a gap in the literature regarding methodologies that incorporate semantic analysis and trust in regard to big SBI. Therefore, the aim of future research is to provide a methodology for social trust and semantic analysis of SBI. Hence, trust and semantic analysis notions will be applied in SBI to confirm the credibility of information, to determine the reputation of the sources, and to define the legitimate contributors with a degree of trustworthy information, the sources, and the users, thereby providing new insights that will benefit the BI domain.

7.4 Conclusion

This chapter provides a recapitulation of the overall approaches developed in this thesis. The chapter includes the main research directions that are proposed in order to extend and improve the outcomes of this thesis. The technical solutions presented in this thesis have been accepted by the research community evidenced in the scholarly research publications in several international refereed and peer-reviewed conferences and journals. Appendix B presents a list of publications derived from this thesis.

Bibliography

"World Wide Web Consortium, Internet Live Stats." Accessed 19 July 2017 http://www.internetlivestats.com/one-second/.

"BBC Ontologies." http://www.bbc.co.uk/ontologies. Accessed 19 May 2017.

"IBM Acquires AlchemyAPI, Enhancing Watson's Deep Learning Capabilities." https://www-03.ibm.com/press/us/en/pressrelease/46205.wss . Accessed 18 May 2015.

"Free Social Media Analytics Tools." Simply Measured. Accessed 16 September 2015. http://simplymeasured.com/free-social-media-tools/.

Abbasi, Mohammad-Ali, Shamanth Kumar, Jose Augusto Andrade Filho, and Huan Liu. 2012. "Lessons learned in using social media for disaster relief-ASU crisis response game." International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction.

Abbasi, Mohammad-Ali, and Huan Liu. 2013. "Measuring User Credibility in Social Media." In *Social Computing, Behavioral-Cultural Modeling and Prediction*, edited by ArielM Greenberg, WilliamG Kennedy and NathanD Bos, 441-448. Springer Berlin Heidelberg.

Abu-Salih, Bilal, Pornpit Wongthongtham, Seyed-Mehdi-Reza Beheshti, and Behrang Zajabbari. 2015. "Towards A Methodology for Social Business Intelligence in the era of Big Social Data incorporating Trust and Semantic Analysis." Second International Conference on Advanced Data and Information Engineering (DaEng-2015), Bali, Indonesia.

Abu-Salih, Bilal, P. Wongthongtham, and C. Y. Kit. 2018. "Twitter mining for ontology-based domain discovery incorporating machine learning." *Journal of Knowledge Management* 22 (5):949-981. doi: 10.1108/Jkm-11-2016-0489.

Abu-Salih, Bilal, Pornpit Wongthongtham, and Dengya Zhu. 2015. "A Preliminary Approach to Domain-Based Evaluation of Users' Trustworthiness in Online Social Networks." 2015 IEEE International Congress on Big Data.

Abu-Salih, Bilal, Pornpit Wongthongtham, Dengya Zhu, and Shihadeh Alqrainy. 2015. "An Approach For Time-Aware Domain-Based Analysis of Users' Trustworthiness In Big Social Data." *Services Transactions on Big Data (STBD)* 2 (1):16.

Abu-Salih, Bilal, Pornpit Wongthongtham, Seyed-Mehdi-Reza Beheshti, and Dengya Zhu. 2015. "A Preliminary Approach to Domain-Based Evaluation of Users' Trustworthiness in Online Social Networks." Big Data (BigData Congress), 2015 IEEE International Congress on.

Abu-Salih, Bilal, Pornpit Wongthongtham, Kit Yan Chan, Dengya Zhu. 2018. "CredSaT: Credibility Ranking of Users in Big Social Data incorporating Semantic Analysis and Temporal Factor", Journal of Information Science (JIS), https://doi.org/10.1177/0165551518790424.

Agarwal, M., and Zhou Bin. 2013. "Detecting Malicious Activities Using Backward Propagation of Trustworthiness over Heterogeneous Social Graph." Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, 17-20 Nov. 2013.

Agichtein, Eugene, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. "Finding high-quality content in social media." Proceedings of the 2008 international conference on web search and data mining.

Ahmed, Zeeshan, and Detlef Gerhard. 2010. "Role of ontology in semantic web development." *arXiv preprint arXiv:1008.1723*.

Akbari, Elham, Ahmad Naderi, Robert-Jan Simons, and Albert Pilot. 2016. "Student engagement and foreign language learning through online social networks." *Asian-Pacific Journal of Second and Foreign Language Education* 1 (1):4.

Akcora, Cuneyt Gurcan, Barbara Carminati, Elena Ferrari, and Murat Kantarcioglu. 2014. "Detecting anomalies in social network data consumption." *Social Network Analysis and Mining* 4 (1):1-16.

Al-Tahrawi, M. . 2015. "Arabic text categorization using logistic regression." *International Journal of Intelligent Systems and Applications* 7 (6):71-78.

Alahmadi, D. H., and X. J. Zeng. 2015. "ISTS: Implicit social trust and sentiment based approach to recommender systems." *Expert Systems with Applications* 42 (22):8840-8849. doi: 10.1016/j.eswa.2015.07.036.

Alam, M. H., W. J. Ryu, and S. Lee. 2017. "Hashtag-based topic evolution in social media." *World Wide Web-Internet and Web Information Systems* 20 (6):1527-1549. doi: 10.1007/s11280-017-0451-3.

Albanese, Massimiliano. 2013. "Measuring Trust in Big Data." In *Algorithms and Architectures for Parallel Processing*, edited by Rocco Aversa, Joanna Kołodziej, Jun Zhang, Flora Amato and Giancarlo Fortino, 241-248. Springer International Publishing.

Alexander, D. E. 2014. "Social media in disaster risk reduction and crisis management." *Sci Eng Ethics* 20 (3):717-33. doi: 10.1007/s11948-013-9502-z.

Alexander, P. A., and J. E. Judy. 1988. "The Interaction of Domain-Specific and Strategic Knowledge in Academic-Performance." *Review of Educational Research* 58 (4):375-404. doi: 10.3102/00346543058004375.

Alghamdi, Rubayyi, and Khalid Alfalqi. 2015. "A Survey of Topic Modeling in Text Mining." *International Journal of Advanced Computer Science and Applications* 6(1). doi: 10.14569/IJACSA.2015.060121

AlRubaian, Majed, Muhammad Al-Qurishi, Mabrook Al-Rakhami, Sk Md Mizanur Rahman, and Atif Alamri. 2015. "A Multistage Credibility Analysis Model for Microblogs." Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015.

Alter, S. 2008. "Defining information systems as work systems: implications for the IS field." *European Journal of Information Systems* 17 (5):448-469. doi: 10.1057/ejis.2008.37.

Althoff, Tim, Pranav Jindal, and Jure Leskovec. 2017. "Online actions with offline impact: How online social networks influence online and offline user behavior." Proceedings of the Tenth ACM International Conference on Web Search and Data Mining.

Altınel, Berna, Murat Can Ganiz, and Banu Diri. 2015. "A corpus-based semantic kernel for text classification by using meaning values of terms." *Engineering Applications of Artificial Intelligence* 43:54-66. doi: http://dx.doi.org/10.1016/j.engappai.2015.03.015.

Amalanathan, Anthoniraj, and S Margret Anouncia. 2016. "A review on user influence ranking factors in social networks." *International Journal of Web Based Communities* 12 (1):74-83.

Amichai-Hamburger, Yair, and Tsahi Hayat. 2017. "Social Networking." In *The International Encyclopedia of Media Effects*. John Wiley & Sons, Inc.

Anagnostopoulos, A., R. Kumar, and M. Mahdian. 2008. "Influence and correlation in social networks." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM .

Anthes, G. 2010. "Topic Models Vs. Unstructured Data." *Communications of the Acm* 53 (12):16-18. doi: Doi 10.1145/1859204.1859210.

Arias, Marta, Argimiro Arratia, and Ramon Xuriguera. 2014. "Forecasting with twitter data." *ACM Trans. Intell. Syst. Technol.* 5 (1):1-24.

Asharaf, S, and Zonin Alessandro. 2015. "Generating and visualizing topic hierarchies from microblogs: An iterative latent dirichlet allocation approach." Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on.

Ashraf, Jamshaid, Omar Khadeer Hussain, and Farookh Khadeer Hussain. 2012. "A framework for measuring ontology usage on the web." *The Computer Journal* 56(9), 1083-1101.

Bae, Y., and H. Lee. 2012. "Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers." *Journal of the American Society for Information Science and Technology* 63 (12):2521-2535. doi: 10.1002/asi.22768.

Balog, K., Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. 2012. "Expertise Retrieval." *Foundations and Trends in Information Retrieval* 6 (2-3):127-256. doi: 10.1561/1500000024.

Bello-Orgaz, Gema, Jason J. Jung, and David Camacho. 2016. "Social big data: Recent achievements and new challenges." *Information Fusion* 28:45-59. doi: http://dx.doi.org/10.1016/j.inffus.2015.08.005.

Benbasat, I., and R. W. Zmud. 2003. "The identity crisis within the is discipline: Defining and communicating the discipline's core properties." *Mis Quarterly* 27 (2):183-194.

Berlanga, Rafael, María José Aramburu, Dolores M. Llidó, and Lisette García-Moya. 2014a. "Towards a Semantic Data Infrastructure for Social Business Intelligence." *New Trends in Databases and Information Systems*:319-327.

Berlanga, Rafael, MaríaJosé Aramburu, DoloresM Llidó, and Lisette García-Moya. 2014b. "Towards a Semantic Data Infrastructure for Social Business Intelligence." In *New Trends in Databases and Information Systems*, edited by Barbara Catania, Tania Cerquitelli, Silvia Chiusano, Giovanna Guerrini, Mirko Kämpf, Alfons Kemper, Boris Novikov, Themis Palpanas, Jaroslav Pokorný and Athena Vakali, 319-327. Springer International Publishing.

Berners-Lee, T., and J. Hendler. 2001. "Publishing on the semantic web." *Nature* 410 (6832):1023-4. doi: 10.1038/35074206.

Beyer, Mark. 2011. "Gartner says solving 'big data'challenge involves more than just managing volumes of data." *http://www.gartner.com/it/page.jsp?id=1731916*.

Bhattacharya, Parantapa, Muhammad Bilal Zafar, Niloy Ganguly, Saptarshi Ghosh, and Krishna P Gummadi. 2014. "Inferring user interests in the twitter social network." Proceedings of the 8th ACM Conference on Recommender systems.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003a. "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3 (4-5):993-1022. doi: 10.1162/jmlr.2003.3.4-5.993.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003b. "Latent dirichlet allocation." *Journal of machine Learning research* 3 (Jan):993-1022.

Bontcheva, Kalina, and Dominic Rout. 2012. "Making sense of social media streams through semantics: a survey." *Semantic Web* 5(5), 373-403.

Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik. 1992. "A training algorithm for optimal margin classifiers." Proceedings of the fifth annual workshop on Computational learning theory.

Box, George EP, J Stuart Hunter, and William Gordon Hunter. 2005. *Statistics for experimenters: design, innovation, and discovery*. Vol. 2: Wiley-Interscience New York.

Bozzon, Alessandro, Marco Brambilla, Stefano Ceri, Matteo Silvestri, and Giuliano Vesci. 2013. "Choosing the right crowd: expert finding in social networks." Proceedings of the 16th International Conference on Extending Database Technology.

Breslin, John G., et al. . 2005. *Towards semantically-interlinked online communities, The Semantic Web: Research and Applications*. Springer Berlin Heidelberg.

Brickley, Dan, and Miller Libby. 2010. FOAF vocabulary specification 0.98. In *Namespace document 9*.

Brown, Phil E, and Junlan Feng. 2011. "Measuring user influence on twitter using modified k-shell decomposition." Fifth International AAAI Conference on Weblogs and Social Media.

Calvanese, Diego, Martin Giese, Peter Haase, Ian Horrocks, Thomas Hubauer, Y Ioannidis, Ernesto Jiménez-Ruiz, Evgeny Kharlamov, Herald Kllapi, and J Klüwer. 2013. "Optique: OBDA Solution for Big Data." In *The Semantic Web: ESWC 2013 Satellite Events*, 293-295. Springer.

Cambria, Erik. 2013. "An Introduction to Concept-Level Sentiment Analysis." In *Advances in Soft Computing and Its Applications*, edited by Félix Castro, Alexander Gelbukh and Miguel González, 478-483. Springer Berlin Heidelberg.

Carrasco, Rafael da Silva, Alcione de Paiva Oliveira, Jugurta Lisboa Filho, and Alexandra Moreira. 2014. "Ontology supported system for searching evidence of wild animals trafficking in social network posts." *Revista Brasileira de Computação Aplicada* 6 (1):16-31.

Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. "An empirical comparison of supervised learning algorithms." Proceedings of the 23rd international conference on Machine learning.

Castillo, C., M. Mendoza, and B. Poblete. 2011a. "Information credibility on twitter." In *Proceedings of the 20th International Conference on World Wide Web*. Hyderabad, USA: ACM.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. 2011b. "Information credibility on twitter." Proceedings of the 20th international conference on World wide web.

Cha, Meeyoung, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. 2010. "Measuring User Influence in Twitter: The Million Follower Fallacy." *ICWSM* 10:10-17.

Chamorro-Premuzic, T. 2014. "How the web distorts reality and impairs our judgement skills." *The Guardian*. 13 May 2014.

Chandrasekaran, Balakrishnan, John R Josephson, and V Richard Benjamins. 1999. "What are ontologies, and why do we need them?" *IEEE Intelligent systems* 14 (1):20-26.

Chang, W. L., A. N. Diaz, and P. C. K. Hung. 2015. "Estimating trust value: A social network perspective." *Information Systems Frontiers* 17 (6):1381-1400. doi: 10.1007/s10796-014-9519-0.

Chang, Yi, Xuanhui Wang, Qiaozhu Mei, and Yan Liu. 2013. "Towards Twitter context summarization with user influence models." Proceedings of the sixth ACM international conference on Web search and data mining, Rome, Italy.

Chaudhuri, S., U. Dayal, and V. Narasayya. 2011a. "An Overview of Business Intelligence Technology." *Communications of the Acm* 54 (8):88-98. doi: 10.1145/1978542.1978562.

Chaudhuri, S., U. Dayal, and V. Narasayya. 2011b. An overview of business intelligence technology.

Chen, H. C., R. H. L. Chiang, and V. C. Storey. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact." *Mis Quarterly* 36 (4):1165-1188.

Chen, Min, Shiwen Mao, Yin Zhang, and VictorC M. Leung. 2014. "Open Issues and Outlook." In *Big Data*, 81-89. Springer International Publishing.

Chen, Xin, Krishna Madhavan, and Mihaela Vorvoreanu. 2013. "A Web-Based Tool for Collaborative Social Media Data Analysis." Cloud and Green Computing (CGC), 2013 Third International Conference on.

Chen, Ye, Bei Yu, Xuewei Zhang, and Yihan Yu. 2016. "Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals." Proceedings of the Sixth International Conference on Learning Analytics & Knowledge.

Chengalur-Smith, I. N., D. P. Ballou, and H. L. Pazer. 1999. "The impact of data quality information on decision making: An exploratory analysis." *Ieee Transactions on Knowledge and Data Engineering* 11 (6):853-864. doi: 10.1109/69.824597.

Chianese, Angelo, Fiammetta Marulli, and Francesco Piccialli. 2016. "Cultural Heritage and Social Pulse: A Semantic Approach for CH Sensitivity Discovery in Social Media Data." 2016 IEEE Tenth International Conference on Semantic Computing (ICSC).

Cohen, A.M., and W.R. Hersh. 2005. "A survey of current work in biomedical text mining." *Brief. Bioinform.* 6 (1):57–71.

Colace, F., M. De Santo, L. Greco, V. Moscato, and A. Picariello. 2015. "A collaborative user-centered framework for recommending items in Online

Social Networks." *Computers in Human Behavior* 51:694-704. doi: 10.1016/j.chb.2014.12.011.

Cortes, C., and V. Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3):273-297. doi: 10.1007/Bf00994018.

Coutinho, Fabio Cardoso, Alexander Lang, and Bernhard Mitschang. 2013. "Making Social Media Analysis more efficient through Taxonomy Supported Concept Suggestion." In 15th GI-Symposiun Database Systems for Business, Technology and Web .

Cuesta, CarlosE, MiguelA Martínez-Prieto, and JavierD Fernández. 2013. "Towards an Architecture for Managing Big Semantic Data in Real-Time." In *Software Architecture*, edited by Khalil Drira, 45-53. Springer Berlin Heidelberg.

Cukier, Kenneth. 2010. *Data, data everywhere: A special report on managing information*: Economist Newspaper, Volume 394, Issue 8671.

Cunningham, Hamish , Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications." the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, US.

Cutillo, L. A., R. Molva, and T. Strufe. 2009. "Safebook: A Privacy-Preserving Online Social Network Leveraging on Real-Life Trust." *Ieee Communications Magazine* 47 (12):94-101. doi: Doi 10.1109/Mcom.2009.5350374.

Cuzzocrea, Alfredo, Ladjel Bellatreche, and Il-Yeol Song. 2013. "Data warehousing and OLAP over big data: current challenges and future research directions." Proceedings of the sixteenth international workshop on Data warehousing and OLAP, San Francisco, California, USA.

Das, TK, and P Mohan Kumar. 2013. "Big data analytics: A framework for unstructured data analysis." *International Journal of Engineering Science & Technology* 5 (1):153.

De Choudhury, M., Y. Lin, H. Sundaram, K. Candan, L. Xie, and A. Kelliher. 2010. "How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media?" The Fourth International AAAI Conference on Weblogs and Social Media, Washington DC, USA, 23-26, May.

De Nart, Dario, Dante Degl'Innocenti, Marco Basaldella, Maristella Agosti, and Carlo Tasso. 2016. "A Content-Based Approach to Social Network Analysis: A Case Study on Research Communities." In *Digital Libraries on the Move: 11th Italian Research Conference on Digital Libraries, IRCDL 2015, Bolzano, Italy, January 29-30, 2015, Revised Selected Papers,* edited by Diego Calvanese, Dario De Nart and Carlo Tasso, 142-154. Cham: Springer International Publishing.

Dean, J., and S. Ghemawat. 2008. "Mapreduce: Simplified data processing on large clusters." *Communications of the Acm* 51 (1):107-113. doi: Doi 10.1145/1327452.1327492.

Demchenko, Yuri, Paola Grosso, Cees De Laat, and Peter Membrey. 2013. "Addressing big data issues in scientific data infrastructure." Collaboration Technologies and Systems (CTS), 2013 International Conference on.

Dinsmore, Daniel L. 2017. *Strategic processing in education*: Taylor & Francis.

Dong, L., N. Feng, P. J. Quan, G. P. Kong, X. Y. Chen, and Q. N. Zhang. 2016. "Optimal kernel choice for domain adaption learning." *Engineering Applications of Artificial Intelligence* 51:163-170. doi: 10.1016/j.engappai.2016.01.022.

Duan, Yajuan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. 2010. "An empirical study on learning to rank of tweets." Proceedings of the 23rd International Conference on Computational Linguistics.

Duggan, Maeve. 2016. "The Political Environment on Social Media." Pew Research Center: Internet, Science & Tech Accessed 15/09/2017. http://www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/.

Dumbill, Edd. 2012. *Planning for big data*:" O'Reilly Media, Inc.

Emani, C. K., N. Cullot, and C. Nicolle. 2015. "Understandable Big Data: A survey." *Computer Science Review* 17:70-81. doi: 10.1016/j.cosrev.2015.05.002.

Embar, Varun R, Indrajit Bhattacharya, Vinayaka Pandit, and Roman Vaculin. 2015a. "Online topic-based social influence analysis for the wimbledon championships." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Embar, Varun R., Indrajit Bhattacharya, Vinayaka Pandit, and Roman Vaculin. 2015b. "Online Topic-based Social Influence Analysis for the Wimbledon Championships." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia.

Emrouznejad, Ali. 2016. *Big Data Optimization: Recent Developments and Challenges*. Edited by Springer. Vol. 18, *Studies in big data*.

Evrim, V., and D. McLeod. 2014. "Context-based information analysis for the Web environment." *Knowledge and Information Systems* 38 (1):109-140. doi: 10.1007/s10115-012-0493-x.

Fan, Wei, and Albert Bifet. 2013. "Mining big data." *ACM SIGKDD Explorations Newsletter* 14 (2):1. doi: 10.1145/2481244.2481246.

Feldman, Ronen, and James Sanger. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*: Cambridge university press.

Ferraram, Alfio, Andriy Nikolov, and François Scharffe. 2013. "Data linking for the semantic web." *Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications* 169:326.

Fiala, D. 2012. "Time-aware PageRank for bibliographic networks." *Journal of Informetrics* 6 (3):370-388. doi: 10.1016/j.joi.2012.02.002.

Freedman, David A. 2009. *Statistical models: theory and practice*: cambridge university press.

Friedman, Jerome H. 2001. "Greedy function approximation: a gradient boosting machine." *Annals of statistics*:1189-1232.

Fu, Yanwei, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. 2012. "Attribute learning for understanding unstructured social activity." European Conference on Computer Vision. (pp. 530-543). Springer, Berlin, Heidelberg.

Gallege, Lahiru S, Dimuthu U Gamage, James H Hill, and Rajeev R Raje. 2014a. "Towards trust-based recommender systems for online software services." Proceedings of the 9th Annual Cyber and Information Security Research Conference.

Gallege, Lahiru S., Dimuthu U. Gamage, James H. Hill, and Rajeev R. Raje. 2014b. "Towards trust-based recommender systems for online software services." Proceedings of the 9th Annual Cyber and Information Security Research Conference, Oak Ridge, Tennessee.

Gantz, John, and David Reinsel. 2010. "The digital universe decade-are you ready." *External publication of IDC (Analyse the Future) information and data*:1-16.

Gantz, John, and David Reinsel. 2012. "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east." *IDC iView: IDC Analyze the future* 2007 (2012):1-16.

Garcia-Moya, L., S. Kudama, M. J. Aramburu, and R. Berlanga. 2013. "Storing and analysing voice of the market data in the corporate data warehouse." *Information Systems Frontiers* 15 (3):331-349. doi: 10.1007/s10796-012-9400-y.

Gartner. 2015. "Gartner Survey Shows More Than 75 Percent of Companies Are Investing or Planning to Invest in Big Data in the Next Two Years." Accessed 11/06/2017.

Gentner, Dedre, and Albert L Stevens. 1983. " Mental models." Hillsdale, N.J: L. Erlbaum Associates.

Ghahremanlou, Lida, Wanita Sherchan, and James A Thom. 2014. "Geotagging Twitter Messages in Crisis Management." *The Computer Journal*. 58.9: 1937-1954.

Grajales III, Francisco Jose, Samuel Sheps, Kendall Ho, Helen Novak-Lauscher, and Gunther Eysenbach. 2014. "Social media: a review and tutorial of applications in medicine and health care." *Journal of medical Internet research* 16 (2):e13.

Gregor, S., and D. Jones. 2007. "The anatomy of a design theory." *Journal of the Association for Information Systems* 8 (5):312-335.

Griffin, Abbie, and John R. Hauser. 1993. "The Voice of the Customer." *Marketing Science* 12 (1):1-27. doi: doi:10.1287/mksc.12.1.1.

Gruber, T. R. 1993. "A translation approach to portable ontology specification." Knowledge Acquisition 5(2): 199-220

Gruber, Thomas R. 1995. "Toward principles for the design of ontologies used for knowledge sharing?" *International journal of human-computer studies* 43 (5):907-928.

Gruhl, D., R. Guha, D. Liben-Nowell, and A. Tomkins. 2004. "Information diffusion through blogspace." the 13th International World Wide Web Conference(WWW’04), New York, USA.

Guille, A., H. Hacid, C. Favre, and D. A. Zighed. 2013. "Information Diffusion in Online Social Networks: A Survey." *Sigmod Record* 42 (2):17-28.

Gupta, Aditi, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. "Tweetcred: Real-time credibility assessment of content on twitter." International Conference on Social Informatics (pp. 228-243).

Gupta, Pankaj, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. 2013. "WTF: the who to follow service at Twitter." Proceedings of the 22nd international conference on World Wide Web, Rio de Janeiro, Brazil.

Gupta, Vishal, and Gurpreet S Lehal. 2009. "A survey of text mining techniques and applications." *Journal of emerging technologies in web intelligence* 1 (1):60-76.

Haarslev, V., K. Hidde, R. Moller, and M. Wessel. 2012. "The RacerPro knowledge representation and reasoning system." *Semantic Web* 3 (3):267-277. doi: 10.3233/Sw-2011-0032.

Hajli, M. N. 2014. "A study of the impact of social media on consumers." *International Journal of Market Research* 56 (3):387-404. doi: 10.2501/Ijmr-2014-025.

Halberstam, Y., and B. Knight. 2016. "Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter." *Journal of Public Economics* 143:73-88. doi: 10.1016/j.jpubeco.2016.08.011.

Han, Hu, Wen Yonggang, Chua Tat-Seng, and Li Xuelong. 2014. "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial." *IEEE Access* 2:652-687. doi: 10.1109/access.2014.2332453.

Harris, Derrick. 2013. "The Gigaom guide to deep learning: Who's doing it, and why it matters." Gigaom Research. November 1 http://gigaom.com/2013/11/01/the-gigaom-guide-to-deep-learning-whos-doing-it-and-why-it-matters/.

Hassan, M. M., F. Karray, and M. S. Kamel. 2012. "Automatic Document Topic Identification using Wikipedia Hierarchical Ontology." Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on, 2-5 July 2012.

He, W., S. H. Zha, and L. Li. 2013. "Social media competitive analysis and text mining: A case study in the pizza industry." *International Journal of Information Management* 33 (3):464-472. doi: 10.1016/j.ijinfomgt.2013.01.001.

Helal, N. A., R. M. Ismail, N. L. Badr, and M. G. M. Mostafa. 2016. "A novel social network mining approach for customer segmentation and viral marketing." *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery* 6 (5):177-189. doi: 10.1002/widm.1183.

Hepp, Martin. 2007. "Possible ontologies: How reality constrains the development of relevant ontologies." *Internet Computing,IEEE* 11 (1):90-96.

Herman, Edward S, and Noam Chomsky. 2010. "Manufacturing consent: The political economy of the mass media." Random House.

Hermida, A., F. Fletcher, D. Korell, and D. Logan. 2012. "SHARE, LIKE, RECOMMEND Decoding the social media news consumer." *Journalism Studies* 13 (5-6):815-824. doi: 10.1080/1461670x.2012.664430.

HermiT, OWL. 2016. "Reasoner." *URL: http://www. hermit-reasoner. com.*

Herzig, Jonathan, Yosi Mass, and Haggai Roitman. 2014. "An author-reader influence model for detecting topic-based influencers in social media." Proceedings of the 25th ACM conference on Hypertext and social media.

Hevner, Alan, and Samir Chatterjee. 2010. *Design research in information systems: theory and practice*. Vol. 22: Springer Science & Business Media.

Hitzler, P., and K. Janowicz. 2013. "Linked Data, Big Data, and the 4th Paradigm." *Semantic Web* 4 (3):233-235. doi: 10.3233/Sw-130117.

Hjørland, Birger, and Hanne Albrechtsen. 1995a. "Toward a new horizon in information science: Domain-analysis." *Journal of the American Society for Information Science* 46 (6):400-425. doi: 10.1002/(sici)1097-4571(199507)46:6<400::aid-asi2>3.0.co;2-y.

Hjørland, Birger, and Hanne Albrechtsen. 1995b. "Toward a new horizon in information science: Domain-analysis." *Journal of the Association for Information Science and Technology* 46 (6):400-425.

Ho, Tin Kam. 1995. "Random decision forests." Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on.

Hofmann, Thomas. 1999. "Probabilistic latent semantic indexing." Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

Hoppe, Anett, C. Nicolle, and A. Roxin. 2013. "Automatic ontology-based user profile learning from heterogeneous web resources in a big data context." *Proceedings of the VLDB Endowment* 6 (12):1428-1433. doi: 10.14778/2536274.2536330.

Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. Vol. 398: John Wiley & Sons.

Hudy, Anissa C. 2015. "Turning the big data crush into an advantage." *Information Management Journal* 49 (1):38-41.

Hussain, F. K., E. Chang, and T. S. Dillon. 2006. "Trust Ontology for Service-Oriented Environment." IEEE International Conference on Computer Systems and Applications.

IBM. 2015. "IBM Acquires AlchemyAPI, Enhancing Watson's Deep Learning Capabilities." Accessed 16/10/2015. http://www-03.ibm.com/press/us/en/pressrelease/46205.wss.

Immonen, A., P. Paakkonen, and E. Ovaska. 2015. "Evaluating the Quality of Social Media Data in Big Data Architecture." *Ieee Access* 3:2028-2043. doi: 10.1109/Access.2015.2490723.

Ito, Jun, Jing Song, Hiroyuki Toda, Yoshimasa Koike, and Satoshi Oyama. 2015. "Assessment of tweet credibility with LDA features." Proceedings of the 24th International Conference on World Wide Web.

Iwanaga, I. S. M., Nguyen The-Minh, T. Kawamura, H. Nakagawa, Y. Tahara, and A. Ohsuga. 2011a. "Building an earthquake evacuation ontology from twitter." 2011 IEEE International Conference on Granular Computing (GrC), 8-10 Nov. 2011.

Iwanaga, I. S. M., Nguyen The-Minh, T. Kawamura, H. Nakagawa, Y. Tahara, and A. Ohsuga. 2011b. "Building an earthquake evacuation ontology from twitter." Granular Computing (GrC), 2011 IEEE International Conference on, 8-10 Nov. 2011.

Jang, Jiyeon, and Sung-Hyon Myaeng. 2013. "Discovering Dedicators with Topic-Based Semantic Social Networks." International Aaai Conference On Weblogs And Social Media.

Janssen, M., H. van der Voort, and A. Wahyudi. 2017. "Factors influencing big data decision-making quality." *Journal of Business Research* 70:338-345. doi: 10.1016/j.jbusres.2016.08.007.

Jeong, Kwang-Yong, Jae-Wook Seol, and KyungSoon Lee. 2014. "Follower Classification Based on User Behavior for Issue Clusters." In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, edited by Tutut Herawan, Mustafa Mat Deris and Jemal Abawajy, 143-150. Springer Singapore.

Jiang, Dawei, Gang Chen, Beng Chin Ooi, Kian Lee Tan, and Sai Wu. 2014. "epiC: an Extensible and Scalable System for Processing Big Data." *PVLDB*:541–552.

Jiang, Wenjun, Guojun Wang, and Jie Wu. 2014. "Generating trusted graphs for trust evaluation in online social networks." *Future generation computer systems* 31:48-58.

Jin, L., Y. Chen, T. Y. Wang, P. Hui, and A. V. Vasilakos. 2013. "Understanding User Behavior in Online Social Networks: A Survey." *Ieee Communications Magazine* 51 (9):144-150.

Joa, David, Debashish Ghosh, Kurt Newman, Thayer S Allison, Elaine C Marino, Joy M Tarquin, Maryann Mangini, Yanghong Shao, and Mark V Krein. 2012. Unstructured data integration with a data warehouse. Google Patents.

Johne, Axel. 1994. "Listening to the Voice of the Market." *International Marketing Review* 11 (1):47-59. doi: doi:10.1108/02651339410057518.

Johnson, T. J., and B. K. Kaye. 2014. "Credibility of Social Network Sites for Political Information Among Politically Interested Internet Users." *Journal of Computer-Mediated Communication* 19 (4):957-974. doi: 10.1111/jcc4.12084.

Kaisler, S., F. Armour, J. A. Espinosa, and W. Money. 2013. "Big Data: Issues and Challenges Moving Forward." System Sciences (HICSS), 2013 46th Hawaii International Conference on, 7-10 Jan. 2013.

Kaplan, Andreas M., and Michael Haenlein. 2010. "Users of the world, unite! The challenges and opportunities of Social Media." *Business Horizons* 53 (1):59-68. doi: http://dx.doi.org/10.1016/j.bushor.2009.09.003.

Karami, Amir, Aryya Gangopadhyay, Bin Zhou, and Hadi Kharrazi. 2017. "Fuzzy Approach Topic Discovery in Health and Medical Corpora." *International Journal of Fuzzy Systems*:1-12.

Katal, Avita, Mohammad Wazid, and RH Goudar. 2013. "Big data: issues, challenges, tools and good practices." Contemporary Computing (IC3), 2013 Sixth International Conference on.

Kawabe, Takashi, Yoshimi Namihira, Kouta Suzuki, Munehiro Nara, Yoshitaka Sakurai, Setsuo Tsuruta, and Rainer Knauf. 2015. "Tweet credibility analysis evaluation by improving sentiment dictionary." Evolutionary Computation (CEC), 2015 IEEE Congress on.

Khobzi, H., and B. Teimourpour. 2015. "LCP segmentation: A framework for evaluation of user engagement in online social networks." *Computers in Human Behavior* 50:101-107. doi: 10.1016/j.chb.2015.03.080.

Khuc, Vinh Ngoc, Chaitanya Shivade, Rajiv Ramnath, and Jay Ramanathan. 2012. "Towards building large-scale distributed systems for twitter sentiment analysis." Proceedings of the 27th Annual ACM Symposium on Applied Computing, Trento, Italy.

Kiryakov, Atanas, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. 2004. "Semantic annotation, indexing, and retrieval." *Web Semantics: Science, Services and Agents on the World Wide Web* 2 (1):49-79.

Kitchin, Rob. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*: Sage.

Koh, Hian Chye, and Gerald Tan. 2011. "Data mining applications in healthcare." *Journal of healthcare information management* 19 (2):65.

Kohavi, Ron. 1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Proceedings of the 14th international joint conference on Artificial Intelligence.

Kontopoulos, E., C. Berberidis, T. Dergiades, and N. Bassiliades. 2013. "Ontology-based sentiment analysis of twitter posts." *Expert Systems with Applications* 40 (10):4065-4074. doi: 10.1016/j.eswa.2013.01.001.

Kopton, Isabella, Jens Sommer, Axel Winkelmann, René Riedl, and Peter Kenning. 2013. "Users' trust building processes during their initial connecting behavior in social networks: Behavioral and neural evidence." Proc. Int. Conf. Inform. Syst. 107, 1–12.

Krathwohl, David R. 1993. *Methods of educational and social science research: An integrated approach*: Longman/Addison Wesley Longman.

Kuang, L., X. Tang, M. Q. Yu, Y. J. Huang, and K. H. Guo. 2016. "A comprehensive ranking model for tweets big data in online social network." *Eurasip Journal on Wireless Communications and Networking* 2016 (1):46. doi: ARTN 4610.1186/s13638-016-0532-5.

Kumar, Akshi, and Teeja Mary Sebastian. 2012. "Sentiment analysis on twitter." *IJCSI International Journal of Computer Science Issues* 9 (3):372-378.

Kumar, K. P. Krishna, and G. Geethakumari. 2014. "Detecting misinformation in online social networks using cognitive psychology." *Human-centric Computing and Information Sciences* 4 (1):14. doi: 10.1186/s13673-014-0014-x.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. 2010. "What is Twitter, a social network or a news media?" Proceedings of the 19th international conference on World wide web.

Labrinidis, Alexandros, and H. V. Jagadish. 2012. "Challenges and opportunities with big data." *Proceedings of the VLDB Endowment* 5 (12):2032-2033. doi: 10.14778/2367502.2367572.

Lammerant, Hans, and Paul De Hert. 2016. "Visions of Technology." In *Data Protection on the Move*, 163-194. Springer.

Lavalle, S., E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz. 2011. "Big Data, Analytics and the Path From Insights to Value." *Mit Sloan Management Review* 52 (2):21-32.

Lavbič, Dejan, Slavko Žitnik, Lovro Šubelj, Aleš Kumer, Aljaž Zrnec, and Marko Bajec. 2013. "Traversal and relations discovery among business entities and people using semantic web technologies and trust management." Databases and Information Systems VII: Selected Papers from the Tenth International Baltic Conference, DB&IS 2012.

Lee, Kyumin, James Caverlee, and Steve Webb. 2010. "Uncovering social spammers: social honeypots + machine learning." Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, Geneva, Switzerland.

Li, Chenliang, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. "Topic Modeling for Short Texts with Auxiliary Word Embeddings." Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.

Liben-Nowell, D., and J. Kleiberg. 2008. "Tracing information flow on a global scale using internet chain-letter data." *National Academy of Sciences* 105 (12):4633-4638.

Lim, Ee-Peng, Hsinchun Chen, and Guoqing Chen. 2013. "Business Intelligence and Analytics." *ACM Transactions on Management Information Systems* 3 (4):1-10. doi: 10.1145/2407740.2407741.

Lior, Rokach. 2014. *Data mining with decision trees: theory and applications*. Vol. 81: World scientific.

Liu, Bing. 2012b. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5 (1):1-167.

Liu, Bing, and Lei Zhang. 2012. "A survey of opinion mining and sentiment analysis." In *Mining text data*, 415-463. Springer.

Liu, Dong, Li Wang, Jianhua Zheng, Ke Ning, and Liang-Jie Zhang. 2013. "Influence Analysis Based Expert Finding Model and Its Applications in Enterprise Social Network." Services Computing (SCC), 2013 IEEE International Conference on.

Louati, Amine, Joyce El Haddad, and Suzanne Pinson. 2014. "A Distributed Decision Making and Propagation Approach for Trust-Based Service Discovery in Social Networks." In *Group Decision and Negotiation. A Process-Oriented View*, edited by Pascale Zaraté, GregoryE Kersten and JorgeE Hernández, 262-269. Springer International Publishing.

Low, Yucheng, Joseph E Gonzalez, Aapo Kyrola, Danny Bickson, Carlos E Guestrin, and Joseph Hellerstein. 2014. "Graphlab: A new framework for parallel machine learning." *arXiv preprint arXiv:1408.2041*.

Lyu, Saixia, Jianxun Liu, Mingdong Tang, Yu Xu, and Jinjun Chen. 2015. "Efficiently predicting trustworthiness of mobile services based on trust propagation in social networks." *Mobile Networks and Applications* 20 (6):840-852. doi: 10.1007/s11036-015-0619-y.

Maalej, Maha, Achraf Mtibaa, and Faïez Gargouri. 2014. "Ontology-Based Context-Aware Social Networks." In *New Trends in Databases and*

Information Systems, edited by Barbara Catania, Tania Cerquitelli, Silvia Chiusano, Giovanna Guerrini, Mirko Kämpf, Alfons Kemper, Boris Novikov, Themis Palpanas, Jaroslav Pokorný and Athena Vakali, 287-295. Springer International Publishing.

Makice, Kevin. 2009. *Twitter API: Up and running: Learn how to build applications with the Twitter API*: " O'Reilly Media, Inc.".

Manuel Pérez-Martínez, Juan, Rafael Berlanga-Llavori, María José Aramburu-Cabo, and Torben Bach Pedersen. 2008. "Contextualizing data warehouses with documents." *Decision Support Systems* 45 (1):77-94. doi: http://dx.doi.org/10.1016/j.dss.2006.12.005.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H Byers. 2011. "Big data: The next frontier for innovation, competition, and productivity." *McKinsey Global Institute*.

March, S. T., and G. F. Smith. 1995. "Design and Natural-Science Research on Information Technology." *Decision Support Systems* 15 (4):251-266. doi: Doi 10.1016/0167-9236(94)00041-2.

Marz, Nathan, and James Warren. 2015. *Big Data: Principles and best practices of scalable realtime data systems*: Manning Publications Co.

Massa, Paolo, and Bobby Bhattacharjee. 2004. "Using Trust in Recommender Systems: An Experimental Analysis." In *Trust Management*, edited by Christian Jensen, Stefan Poslad and Theo Dimitrakos, 221-235. Springer Berlin Heidelberg.

Maynard, Diana, Valentin Tablan, Cristian Ursu, Hamish Cunningham, and Yorick Wilks. 2001. "Named Entity Recognition from Diverse Text Types." Recent Advances in Natural Language Processing 2001 Conference, Tzigov Chark.

McCord, M., and M. Chuah. 2011. "Spam Detection on Twitter Using Traditional Classifiers." In *Autonomic and Trusted Computing*, edited by JoseM Alcaraz Calero, LaurenceT Yang, FélixGómez Mármol, LuisJavier García Villalba, AndyXiaolin Li and Yan Wang, 175-186. Springer Berlin Heidelberg.

McLuhan, Marshall. 1994. *Understanding media: The extensions of man*: MIT press.

McPherson, M., L. Smith-Lovin, and J. M. Cook. 2001. "Birds of a feather: Homophily in social networks." *Annual Review of Sociology* 27 (1):415-444. doi: DOI 10.1146/annurev.soc.27.1.415.

Mendoza, Marcelo, Barbara Poblete, and Carlos Castillo. 2010. "Twitter Under Crisis: Can we trust what we RT?" Proceedings of the first workshop on social media analytics.

Michelson, Matthew, and Sofus A Macskassy. 2010. "Discovering users' topics of interest on twitter: a first look." Proceedings of the fourth workshop on Analytics for noisy unstructured text data.

Miles, Alistair, and Sean Bechhofer. 2009. SKOS simple knowledge organization system reference. In *Technical report*: W3C.

Miller, Z., B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang. 2014. "Twitter spammer detection using data stream clustering." *Information Sciences* 260:64-73. doi: 10.1016/j.ins.2013.11.016.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of machine learning*: MIT press.

Momeni, Elaheh, Claire Cardie, and Nicholas Diakopoulos. 2016. "A Survey on Assessment and Ranking Methodologies for User-Generated Content on the Web." *ACM Computing Surveys (CSUR)* 48 (3):41.

Morris, Meredith Ringel, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. "Tweeting is believing?: understanding microblog credibility perceptions." Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work.

Nadeau, David, and Satoshi Sekine. 2007. "A survey of named entity recognition and classification." *Lingvisticae Investigationes* 30 (1):3-26.

Narayan, Shashi, Srdjan Prodanovic, Mohammad Fazleh Elahi, and Zoë Bogart. 2010. "Population and Enrichment of Event Ontology using Twitter." *Information Management SPIM 2010*:31.

Naumann, Jennifer. 2013. "Do you trust my tweet? Return on Investment of Social Media: a study about@ Twitter investigating the effect of message intention in the followers’ level of trust." B.S., University of Twente.

Nepal, Surya, Cécile Paris, and Athman Bouguettaya. 2013. "Trusting the Social Web: issues and challenges." *World Wide Web* 18 (1):1-7. doi: 10.1007/s11280-013-0252-2.

News, Social Media. 2017. "Social Media Statistics Australia – January 2017." https://www.socialmedianews.com.au/social-media-statistics-australia-january-2017/

Nichols, Leah G. 2014. "A topic model approach to measuring interdisciplinarity at the National Science Foundation." *Scientometrics* 100 (3):741-754. doi: 10.1007/s11192-014-1319-2.

Nitzan, I., and B. Libai. 2011. "Social Effects on Customer Retention." *Journal of Marketing* 75 (6):24-38.

Obenshain, M. K. 2004. "Application of data mining techniques to healthcare data." *Infect Control Hosp Epidemiol* 25 (8):690-5. doi: 10.1086/502460.

Onan, Aytug, Serdar Korukoglu, and Hasan Bulut. 2016. "LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis." *Int. J. Comput. Linguistics Appl.* 7 (1):101-119.

Oren, Eyal, Knud Möller, Simon Scerri, Siegfried Handschuh, and Michael Sintek. 2006. "What are semantic annotations." *Relatório técnico. DERI Galway* 9:62.

Orenga-Rogla, S., and R. Chalmeta. 2016. "Social customer relationship management: taking advantage of Web 2.0 and Big Data technologies." *Springerplus* 5 (1):1462. doi: 10.1186/s40064-016-3128-y.

Paik, I., T. Tanaka, H. Ohashi, and Chen Wuhui. 2013. "Big Data Infrastructure for Active Situation Awareness on Social Network Services." Big Data (BigData Congress), 2013 IEEE International Congress on, June 27 2013-July 2 2013.

Pal, Aditya, Amaç Herdagdelen, Sourav Chatterji, Sumit Taank, and Deepayan Chakrabarti. 2016. "Discovery of topical authorities in instagram." Proceedings of the 25th International Conference on World Wide Web.

Papadopoulos, S., K. Bontcheva, E. Jaho, M. Lupu, and C. Castillo. 2016. "Overview of the Special Issue on Trust and Veracity of Information in Social Media." *Acm Transactions on Information Systems* 34 (3):14. doi: Artn 1410.1145/2870630.

Partners, NewVantage. 2017. Big Data Executive Survey 2016: Big Data Business Impact: Achieving Business Results through Innovation and Disruption.

Passant, Alexandre, Philipp Kärger, Michael Hausenblas, Daniel Olmedilla, Axel Polleres, and Stefan Decker. 2009. "Enabling trust and privacy on the social web." W3C workshop on the future of social networking.

Peffers, K., T. Tuunanen, M. A. Rothenberger, and S. Chatterjee. 2007. "A design science research methodology for Information Systems Research." *Journal of Management Information Systems* 24 (3):45-77. doi: 10.2753/Mis0742-1222240302.

Phillipps, T. 2013. "The Analytics Advantage We're just getting started. Deloitte. https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Deloitte-Analytics/dttl-analytics-analytics-advantage-report-061913.pdf

Podobnik, Vedran, Darko Striga, Ana Jandras, and Ignac Lovrek. 2012b. "How to calculate trust between social network users?" Software, Telecommunications and Computer Networks (SoftCOM), 2012 20th International Conference on.

Poria, S., E. Cambria, A. Gelbukh, F. Bisio, and A. Hussain. 2015. "Sentiment Data Flow Analysis by Means of Dynamic Linguistic Patterns." *Ieee Computational Intelligence Magazine* 10 (4):26-36. doi: 10.1109/Mci.2015.2471215.

Power, D. J. 2007. "A Brief History of Decision Support Systems." Accessed 24 July 2013. http://DSSResources.COM/history/dsshistory.html.

Prat, Nicolas, Isabelle Comyn-Wattiau, and Jacky Akoka. 2014. "Artifact Evaluation in Information Systems Design-Science Research-a Holistic View." Pacific Asia Conference on Information Systems.

Quercia, Daniele, Harry Askham, and Jon Crowcroft. 2012. "TweetLDA: supervised topic classification and link prediction in Twitter." the 4th Annual ACM Web Science Conference, Evanston, Illinois.

Quinlan, J Ross. 1993. "C4. 5: Programming for machine learning." *Morgan Kauffmann*:38.

Rainie, Lee, and Barry Wellman. 2012. *Networked: The new social operating system*: Mit Press.

Rajaraman, Anand, and Jeffrey David Ullman. 2011. *Mining of massive datasets*: Cambridge University Press.

Ramos, Juan. 2003. "Using tf-idf to determine word relevance in document queries." Proceedings of the First Instructional Conference on Machine Learning.

Ravikumar, Srijith, Kartik Talamadupula, Raju Balakrishnan, and Subbarao Kambhampati. 2013. "RAProp: ranking tweets by exploiting the tweet/user/web ecosystem and inter-tweet agreement." Proceedings of the 22nd ACM international conference on Conference on information &#38; knowledge management, San Francisco, California, USA.

Reddy, Kuldeep. 2013. "Novel Models and Architectures for Distributed Semantic Data Management." SEMAPRO 2013, The Seventh International Conference on Advances in Semantic Processing.

Rehurek, Radim, and Petr Sojka. 2010. "Software framework for topic modelling with large corpora." In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Reidenbach, R Eric. 2009. *Listening to the voice of the market: How to increase market share and satisfy current customers*: Productivity Press.

Resnick, Paul, Samuel Carton, Souneil Park, Yuncheng Shen, and Nicole Zeffer. 2014. "Rumorlens: A system for analyzing the impact of rumors and corrections in social media." Proc. Computational Journalism Conference.

Rizzo, G., and R. Troncy. 2011a. "Nerd: Evaluating named entity recognition tools in the web of data." Workshop on Web Scale Knowledge Extraction (WEKEX11).

Rizzo, Giuseppe, and Raphaël Troncy. 2011b. "Nerd: evaluating named entity recognition tools in the web of data." Workshop on Web Scale Knowledge Extraction (WEKEX11).

Robertson, S. E., and K. Sparck-Jones. 1976. "Relevance Weighting of Search Terms." *Journal of the American Society for Information Science* 27 (3):129-146. doi: DOI 10.1002/asi.4630270302.

Robertson, Stephen. 2004. "Understanding inverse document frequency: on theoretical arguments for IDF." *Journal of documentation* 60 (5):503-520.

Rogers, P, R Puryear, and J Root. 2013. "Infobesity: The enemy of good decisions." *Insights: Bain Brief* 11.

Ruan, Yefeng, and Arjan Durresi. 2016. "A survey of trust management systems for online social communities–Trust modeling, trust inference and attacks." *Knowledge-Based Systems* 106:150-163.

Saha, Barna, and Divesh Srivastava. 2014. "Data quality: The other face of Big Data." Data Engineering (ICDE), 2014 IEEE 30th International Conference on, March 31 2014-April 4 2014.

Saif, Hassan, Yulan He, and Harith Alani. 2011. "Semantic smoothing for Twitter sentiment analysis." 10th International Semantic Web Conference (ISWC 2011), Bonn, Germany, 23 - 27 Oct 2011.

Saif, Hassan, Yulan He, and Harith Alani. 2012. "Semantic Sentiment Analysis of Twitter." In *The Semantic Web – ISWC 2012*, edited by Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Manfred Hauswirth, JosianeXavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein and Eva Blomqvist, 508-524. Springer Berlin Heidelberg.

Salathé, Marcel, Duy Vu, Shashank Khandelwal, and David Hunter. 2013. "The dynamics of health behavior sentiments on a large online social network." *EPJ Data Science* 2 (1). doi: 10.1140/epjds16.

Sallam, RL, C Howson, Carlie J Idoine, Thomas W. Oestreich, James Laurence, and Joao Tapadinhas. 2017. Magic Quadrant for Business Intelligence and Analytics Platforms. Gartner.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. "A vector space model for automatic indexing." *Communications of the ACM* 18 (11):613-620.

Samuel, A. L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3):210-229. doi: 10.1147/rd.33.0210.

Saravanakumar, M., and T. SuganthaLakshmi. 2012. "Social Media Marketing." *Life Science Journal-Acta Zhengzhou University Overseas Edition* 9 (4):4444-4451.

Sashi, C. M. 2012. "Customer engagement, buyer-seller relationships, and social media." *Management Decision* 50 (1-2):253-272. doi: 10.1108/00251741211203551.

SAWYER, Rebecca, and CHEN Guo-Ming. 2012. "The Impact of Social Media on Intercultural Adaptation." *Intercultural Communication Studies* 21 (2).

Sayce, David. 2016. "10 Billions Tweets… number of tweets per day." http://www.dsayce.com/social-media/10-billions-tweets/.

Schmidt, Bernard, Diego Galar, and LIhui Wang. 2016. "Big Data in Maintenance Decision Support Systems: Aggregation of Disparate Data Types." Euromaintenance 2016 Conference Proceedings.

Schonhofen, Peter. 2006. "Identifying Document Topics Using the Wikipedia Category Network." Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.

Schubmehl, D, and D Vesset. 2014. "Unlocking the Hidden Value of Information." IDC Community. https://idc-community.com/groups/it_agenda/business-analytics-big-data/unlocking_the_hidden_value_of_informatio.

Shapiro, Matthew A, and Libby Hemphill. 2017. "Politicians and the Policy Agenda: Does Use of Twitter by the US Congress Direct New York Times Content?" *Policy & Internet* 9 (1):109-132.

Sharef, Nurfadhlina Mohd, Trevor Martin, Khairul Azhar Kasmiran, Aida Mustapha, Md. Nasir Sulaiman, and Masrah Azrifah Azmi-Murad. 2015. "A comparative study of evolving fuzzy grammar and machine learning techniques for text categorization." *Soft Computing* 19 (6):1701-1714. doi: 10.1007/s00500-014-1358-x.

Shen, W., J. Y. Wang, and J. W. Han. 2015. "Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions." *Ieee Transactions on Knowledge and Data Engineering* 27 (2):443-460. doi: 10.1109/Tkde.2014.2327028.

Shenoy, Aravind, and Anirudh Prabhu. 2016. "Social Media Marketing and SEO." In *Introducing SEO*, 119-127. Springer.

Sherchan, W., S. Nepal, and C. Paris. 2013. "A Survey of Trust in Social Networks." *Acm Computing Surveys* 45 (4). doi: Artn 4710.1145/2501654.2501661.

Shroff, Gautam, Lipika Dey, and Puneet Agarwal. 2013. "Socio-Business Intelligence Using Big Data " *Technical Trends*.

Shvachko, Konstantin, Hairong Kuang, Sanjay Radia, and Robert Chansler. 2010. "The hadoop distributed file system." Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on.

Sieber, Joan E. 2012. *The ethics of social research: Fieldwork, regulation, and publication*: Springer Science & Business Media.

Sikdar, S., Kang Byungkyu, J. O'Donovan, T. Hollerer, and S. Adah. 2013. "Understanding Information Credibility on Twitter." Social Computing (SocialCom), 2013 International Conference on, 8-14 Sept. 2013.

Silva, Arlei, Sara Guimarães, Wagner Meira Jr, and Mohammed Zaki. 2013. "ProfileRank: finding relevant content and influential users based on information diffusion." Proceedings of the 7th Workshop on Social Network Mining and Analysis.

Simon, Herbert A. 1996. *The sciences of the artificial*: MIT press.

Sirin, E., B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz. 2007. "Pellet: A practical OWL-DL reasoner." *Journal of Web Semantics* 5 (2):51-53. doi: 10.1016/j.websem.2007.03.004.

Song, Shuangyong, Qiudan Li, and Xiaolong Zheng. 2012. "Detecting popular topics in micro-blogging based on a user interest-based model." Neural Networks (IJCNN), The 2012 International Joint Conference on.

Sparck Jones, Karen. 1972. "A statistical interpretation of term specificity and its application in retrieval." *Journal of documentation* 28 (1):11-21.

Stevens, Robert. 2001. "What is an Ontology?" Accessed 3rd March. http://www.cs.man.ac.uk/~stevensr/onto/node3.html.

Stieglitz, Stefan, and Linh Dang-Xuan. 2013. "Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior." *Journal of Management Information Systems* 29 (4):217-248.

Sun, E., I. Rosenn, C. Marlow, and T. Lento. 2009. "Gesundheit! modeling contagion through facebook news feed." ICWSM 2009, San Jose, CA.

Sun, Guohao, Guanfeng Liu, Lei Zhao, Jiajie Xu, An Liu, and Xiaofang Zhou. 2014. "A Social Trust Path Recommendation System in Contextual Online Social Networks." In *Web Technologies and Applications*, edited by Lei Chen, Yan Jia, Timos Sellis and Guanfeng Liu, 652-656. Springer International Publishing.

Tess, Paul A. 2013. "The role of social media in higher education classes (real and virtual) – A literature review." *Computers in Human Behavior* 29 (5):A60-A68. doi: http://dx.doi.org/10.1016/j.chb.2012.12.032.

Theiler, James P, Neal R Harvey, Steven P Brumby, John J Szymanski, Steve Alferink, Simon J Perkins, Reid B Porter, and Jeffrey J Bloch. 1999. "Evolving retrieval algorithms with a genetic programming scheme." Imaging Spectrometry V.

Thomas, Edward, Jeff Z Pan, and Yuan Ren. 2010. "TrOWL: Tractable OWL 2 reasoning infrastructure." Extended Semantic Web Conference.

Thusoo, Ashish, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. 2009. "Hive: a warehousing solution over a map-reduce framework." *Proceedings of the VLDB Endowment* 2 (2):1626-1629.

Tien, J. M. 2013. "Big Data: Unleashing information." *Journal of Systems Science and Systems Engineering* 22 (2):127-151. doi: 10.1007/s11518-013-5219-4.

Toffler, Alvin. 1971. *Future shock*: Bantam.

Tole, Alexandru Adrian. 2013. "Big data challenges." *Database systems journal* 4 (3):31-40.

Tsarkov, Dmitry, and Ian Horrocks. 2006. "FaCT++ description logic reasoner: System description." International Joint Conference on Automated Reasoning.

Tsolmon, Bayar, and Kyung-Soon Lee. 2014. "A Graph-Based Reliable User Classification." In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, edited by Tutut Herawan, Mustafa Mat Deris and Jemal Abawajy, 61-68. Springer Singapore.

Turian, Joseph. 2013. Using AlchemyAPI for Enterprise-Grade Text Analysis. AlchemyAPI: Denver, CO, USA

Twitter. "Following rules and best practices." https://support.twitter.com/groups/56-policies-violations/topics/237-guidelines/articles/68916-following-rules-and-best-practices.

Twitter. 2009. "The twitter rules.". https://support.twitter.com/articles/18311-the-twitter-rules.

Van Kessel, Stijn, and Remco Castelein. 2016. "Shifting the blame. Populist politicians' use of Twitter as a tool of opposition." *Journal of Contemporary European Research* 12 (2).

Venable, J., J. Pries-Heje, and R. Baskerville. 2016. "FEDS: a Framework for Evaluation in Design Science Research." *European Journal of Information Systems* 25 (1):77-89. doi: 10.1057/ejis.2014.36.

Venable, John, Jan Pries-Heje, and Richard Baskerville. 2012. "A Comprehensive Framework for Evaluation in Design Science Research." In *Design Science Research in Information Systems. Advances in Theory and Practice*, edited by Ken Peffers, Marcus Rothenberger and Bill Kuechler, 423-438. Springer Berlin Heidelberg.

Venable, JohnR. 2013. "Rethinking Design Theory in Information Systems." In *Design Science at the Intersection of Physical and Virtual Design*, edited by Jan vom Brocke, Riitta Hekkala, Sudha Ram and Matti Rossi, 136-149. Springer Berlin Heidelberg.

Vicient, C., and A. Moreno. 2015. "Unsupervised topic discovery in micro-blogging networks." *Expert Systems with Applications* 42 (17-18):6472-6485. doi: 10.1016/j.eswa.2015.04.014.

Vitak, J. 2012. "The Impact of Context Collapse and Privacy on Social Network Site Disclosures." *Journal of Broadcasting & Electronic Media* 56 (4):451-470. doi: 10.1080/08838151.2012.732140.

von Alan, R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. 2004. "Design science in information systems research." *MIS quarterly* 28 (1):75-105.

Wang, Alex Hai. 2010. "Don't follow me: Spam detection in Twitter." Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on, 26-28 July 2010.

Wang, Chong, Bo Thiesson, Chris Meek, and David Blei. 2009. "Markov topic models." Artificial Intelligence and Statistics.

Weber, Sven. 2010. "Design Science Research: Paradigm or Approach?" Americas Conference on Information Systems.

Weerkamp, W., and M. de Rijke. 2012. "Credibility-inspired ranking for blog post retrieval." *Information Retrieval* 15 (3-4):243-277. doi: 10.1007/s10791-011-9182-8.

Weibel, Stuart, et al. 1998. "Dublin core metadata for resource discovery." *Internet Engineering Task Force RFC 2413* 222.

Weitzel, Leila, José Palazzo Moreira de Oliveira, and Paulo Quaresma. 2013. "Exploring trust to rank reputation in microblogging." International Conference on Database and Expert Systems Applications.

Weng, Jianshu, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. "Twitterrank: finding topic-sensitive influential twitterers." Proceedings of the third ACM international conference on Web search and data mining.

Wheeldon, Johannes, and Mauri K Ahlberg. 2011. *Visualizing social science research: Maps, methods, & meaning*: Sage.

Wongthongtham, Pornpit, and Bilal Abu-Salih. 2015. "Ontology and trust based data warehouse in new generation of business intelligence: State-of-the-art, challenges, and opportunities." Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on.

Wu, H., A. Arenas, and S. Gomez. 2017. "Influence of trust in the spreading of information." *Phys Rev E* 95 (1-1):012301. doi: 10.1103/PhysRevE.95.012301.

Wu, Jing, and Zheng Lin. 2005. "Research on customer segmentation model by clustering." Proceedings of the 7th international conference on Electronic commerce.

Xiao, Chunjing, Yuhong Zhang, Xue Zeng, and Yue Wu. 2013. "Predicting User Influence in Social Media." *Journal of Networks* 8 (11):2649-2655.

Yardi, Sarita, Daniel Romero, Grant Schoenebeck, and danah boyd. 2009. "Detecting spam in a Twitter network." *First Monday*. doi: 10.5210/fm.v15i1.2793.

Yen, Show-Jane, Yue-Shi Lee, Jia-Ching Ying, and Yu-Chieh Wu. 2011. "A logistic regression-based smoothing method for Chinese text categorization." *Expert Systems with Applications* 38 (9):11581-11590.

Yeniterzi, Reyyan, and Jamie Callan. 2014. "Constructing effective and efficient topic-specific authority networks for expert finding in social media." Proceedings of the first international workshop on Social media retrieval and analysis.

Yin, H. Z., B. Cui, L. Chen, Z. T. Hu, and X. F. Zhou. 2015. "Dynamic User Modeling in Social Media Systems." *Acm Transactions on Information Systems* 33 (3):10. doi: Artn 1010.1145/2699670.

Yin, J., A. Lampert, M. Cameron, B. Robinson, and R. Power. 2012. "Using Social Media to Enhance Emergency Situation Awareness." *Ieee Intelligent Systems* 27 (6):52-59. doi: Doi 10.1109/Mis.2012.6.

Yu, Xu, Liu Jianxun, Tang Mingdong, and Liu Xiaoqing. 2013. "An Efficient Trust Propagation Scheme for Predicting Trustworthiness of Service Providers in Service-Oriented Social Networks." Web Services (ICWS), 2013 IEEE 20th International Conference on, June 28 2013-July 3 2013.

Yuangang, Yao, Liu Hui, Yi Jin, Chen Haiqiang, Zhao Xianghui, and Ma Xiaoyu. 2014. "An automatic semantic extraction method for web data interchange." Computer Science and Information Technology (CSIT), 2014 6th International Conference on, 26-27 March 2014.

Zhai, Yanmei, Xin Li, Jialiang Chen, Xiumei Fan, and William K Cheung. 2014. "A novel topical authority-based microblog ranking." Asia-Pacific Web Conference.

Zhang, B., Q. Q. Song, J. H. Ding, and L. Wang. 2015. "A trust-based sentiment delivering calculation method in microblog." *International Journal of Services Technology and Management* 21 (4-6):185-198. doi: 10.1504/Ijstm.2015.073917.

Zhang, Dan, Dongsheng Wang, and Haiping Huang. 2013. "Application and Study on Sentiment-oriented Analysis on Social Semantic Network." 2013

International Conference on Advanced Computer Science and Electronics Information (ICACSEI 2013).

Zhang, W., Y. B. Cui, and T. Yoshida. 2017. "En-LDA: An Novel Approach to Automatic Bug Report Assignment with Entropy Optimized Latent Dirichlet Allocation." *Entropy* 19 (5):173. doi: ARTN 17310.3390/e19050173.

Zhao, L., T. Hua, C. T. Lu, and I. R. Chen. 2016. "A topic-focused trust model for Twitter." *Computer Communications* 76:1-11. doi: 10.1016/j.comcom.2015.08.001.

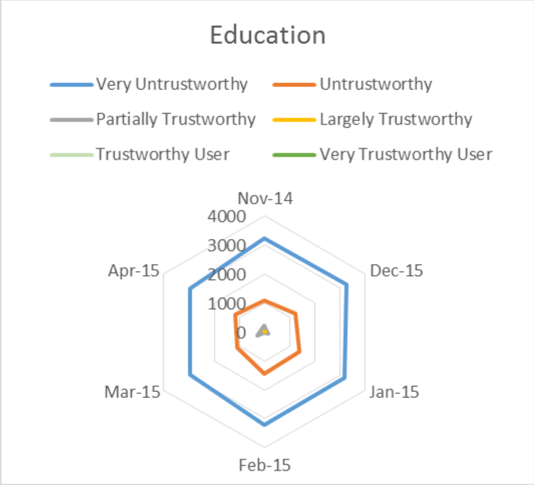
Zhou, Jingyu, Yunlong Zhang, and Jia Cheng. 2014. "Preference-based mining of top-K influential nodes in social networks." *Future Generation Computer Systems* 31:40-47.

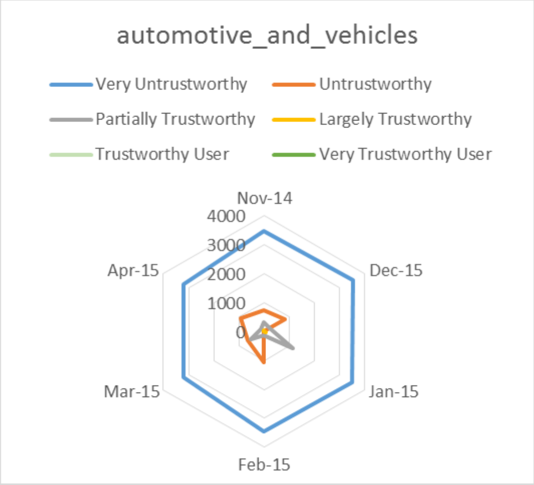
Zhu, Z. G., J. Q. Su, and L. P. Kong. 2015. "Measuring influence in online social network based on the user-content bipartite graph." *Computers in Human Behavior* 52:184-189. doi: 10.1016/j.chb.2015.04.072.

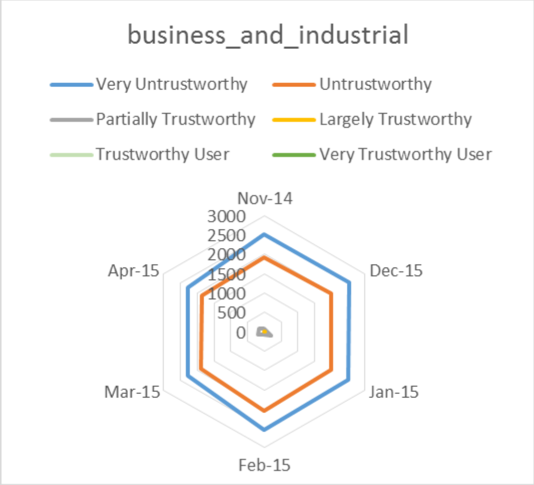
Zoghbi, S., I. Vulic, and M. F. Moens. 2016. "Latent Dirichlet allocation for linking user-generated content and e-commerce data." *Information Sciences* 367:573-599. doi: 10.1016/j.ins.2016.05.047.

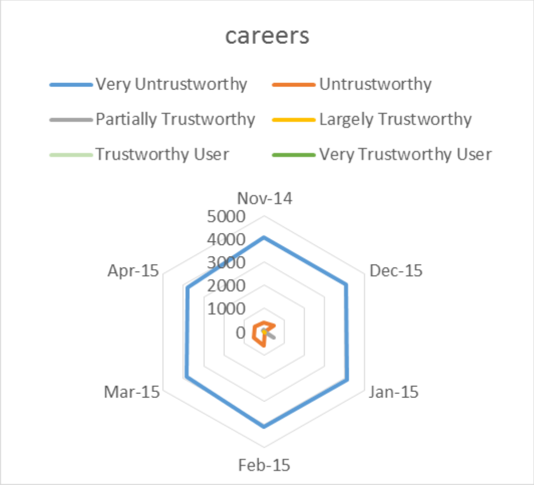
Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

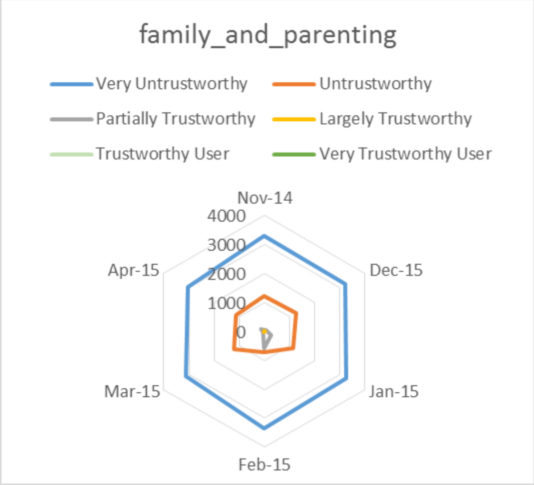
Appendix A. Miscellaneous Experimental Results

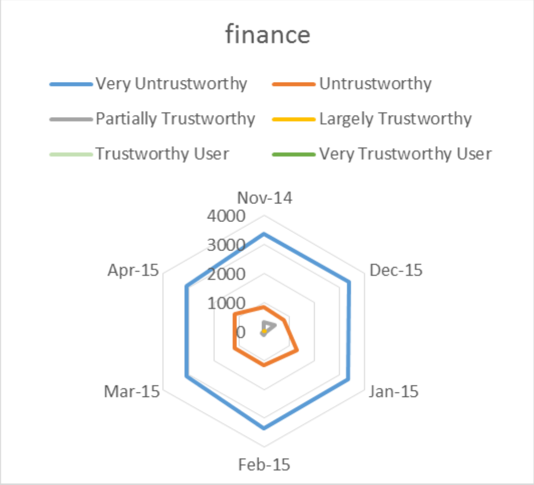












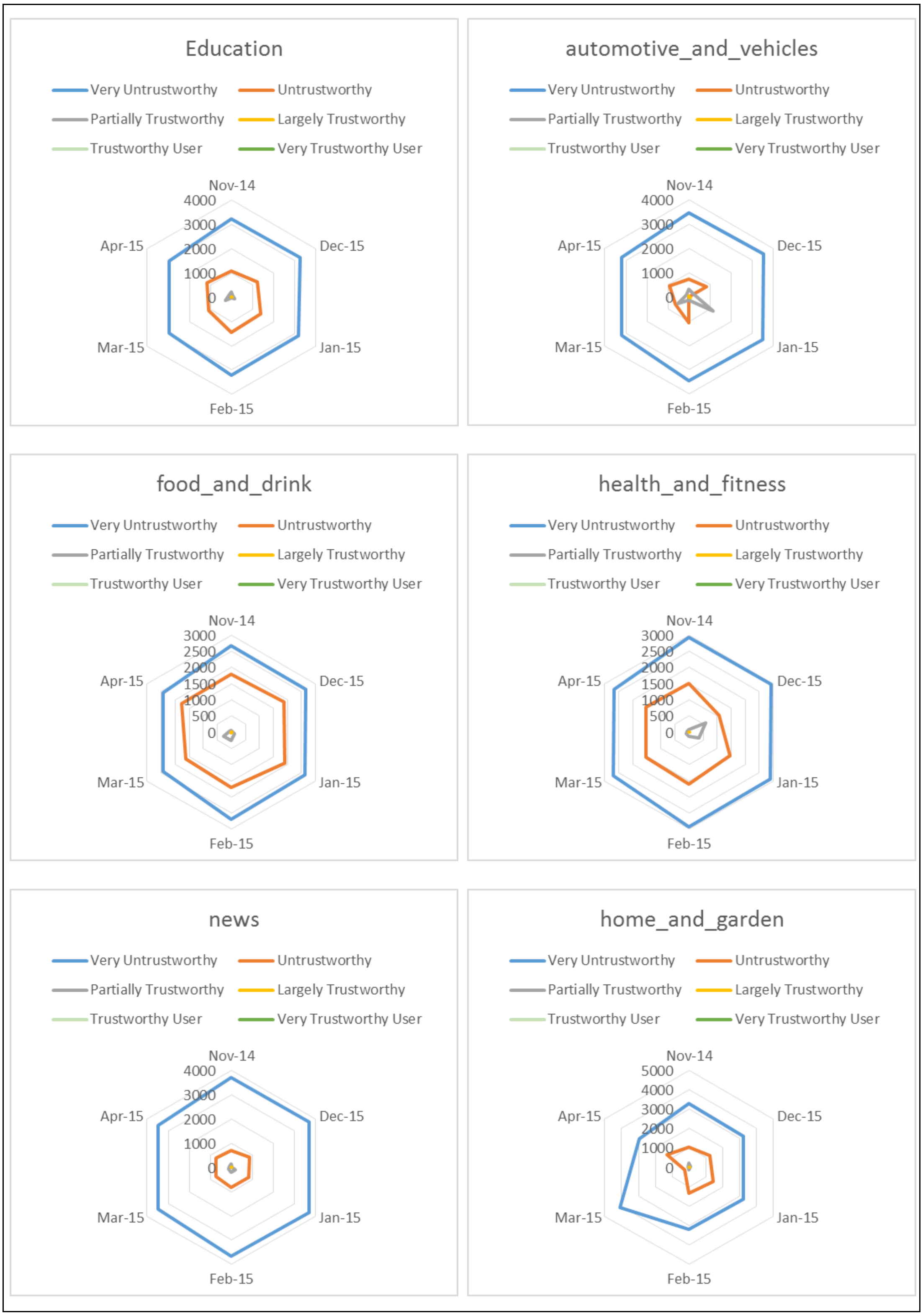


Figure A-1: Monthly users' trustworthiness levels in several other domains

Table A-1: Metadata of the top five trustworthy users in Technology and Computing

screen_name	edithyeung	wolf_gregor	johnjwall	commadelimited	JeremyKendall
domain_favorite_count	465	376	225	106	273
domain_replies_count	82	332	282	648	383
domain_retweet_count	328	344	117	37	162
followers_count	4861	4206	4180	3439	2542
friends_count	664	825	1436	241	1497
retweet_count	587	494	284	113	347
favorite_count	860	554	587	520	693
replies_count	210	610	690	2825	1261
count_domain_pos	661	1105	1415	671	613
count_domain_neg	649	1057	1368	639	560
sum_domain_pos	21.254	61.03	70.107	73.812	45.675
sum_domain_neg	-6.6321	-20.387	-21.318	-49.869	-21.583
TC	0.448	0.422	0.361	0.353	0.351
TC_normalized	5	5	5	4	4
Semantics/label	Very Trustworthy	Very Trustworthy	Very Trustworthy	Trustworthy	Trustworthy

Table A-2: Metadata of the top five trustworthy users in Art and Entertainment

screen_name	SpnMaisieDaisy	rinceya	commadelimited	MattDonnelly	DOEYROCK
domain_favorite_count	1821	1179	94	338	143
domain_replies_count	256	2232	570	103	109
domain_retweet_count	1017	148	14	370	184

screen_name	SpnMaisieDaisy	rinceya	commadelimited	MattDonnelly	DOEYROCK
followers_count	2818	2154	3439	3396	3480
friends_count	847	788	241	3272	348
retweet_count	4018	343	113	1178	416
favorite_count	6835	2520	520	916	295
replies_count	932	3973	2825	214	250
count_domain_pos	773	1763	601	435	361
count_domain_neg	750	1762	536	451	361
sum_domain_pos	46.274	210.508	72.642	16.2522	11.5197
sum_domain_neg	-27.465	-157.107	-24.91	-20.190	-5.812
TC	0.523	0.319	0.303	0.297	0.289
TC_normalized	5	4	3	3	3
Semantics/label	Very Trustworthy	Trustworthy	Largely Trustworthy	Largely Trustworthy	Largely Trustworthy

Table A-3: Metadata of the top five trustworthy users in Law, Govt and Politics

screen_name	EnglishVoice	IvorCrotty	Karen_Woodward	DrewCourt	wave3katie
domain_favorite_count	178	277	3	239	87
domain_replies_count	150	258	921	97	15
domain_retweet_count	326	472	0	220	177
followers_count	2303	2499	626	776	2658
friends_count	561	1288	435	436	743
retweet_count	767	3876	15	757	759
favorite_count	423	2556	45	1174	655
replies_count	349	1753	1683	608	166
count_domain_pos	652	495	420	570	195
count_domain_neg	666	539	494	570	195
sum_domain_pos	10.4089	19.781	47.806	10.933	1.386

screen_name	EnglishVoice	IvorCrotty	Karen_Woodward	DrewCourt	wave3katie
sum_domain_neg	-30.392	-35.814	-77.221	-10.442	-4.423
TC	0.403	0.3651	0.352	0.305	0.303
TC_normalized	5	5	5	4	4
Semantics/label	Very Trustworthy	Very Trustworthy	Very Trustworthy	Trustworthy	Trustworthy

Table A-4: Metadata of the top five trustworthy users in Sports

screen_name	nwipreps	citizencage	HowardWKYT	DionteSays	Bauzen
domain_favorite_count	2998	919	1876	325	881
domain_replies_count	299	1787	156	157	603
domain_retweet_count	2916	285	1510	457	525
followers_count	4920	1361	2179	2850	2378
friends_count	778	1725	474	806	108
retweet_count	3240	425	2598	1678	1097
favorite_count	3332	1437	3571	1218	1861
replies_count	327	3004	372	1335	2138
count_domain_pos	2523	1801	615	301	611
count_domain_neg	2497	1680	604	299	644
sum_domain_pos	80.049	228.782	29.052	17.699	44.054
sum_domain_neg	-12.746	-126.68	-16.086	-17.139	-69.256
TC	0.682	0.503	0.368	0.223	0.22
TC_normalized	5	4	3	2	2
Semantics/label	Very Trustworthy	Trustworthy	Largely Trustworthy	Partially Trustworthy	Partially Trustworthy

Table A-5: Training and Validation results of Cross Validation to Compare the Performance of Five Classifiers to Detect User Political Interest

	Fold	Training_accuracy	Validation_accuracy	
Logistic Classifier	1	0.993464052	0.960784314	
	2	0.989106754	0.980392157	
	3	0.991285403	0.960784314	
	4	0.986928105	1	
	5	0.986928105	1	
	6	0.986928105	1	
	7	0.993464052	0.960784314	
	8	0.989106754	1	
	9	0.991285403	0.980392157	
	10	0.989106754	0.980392157	
		0.989760349	0.982352941	Mean
		5.74328E-06	0.000265283	Variance
Support Victor Machine Classifier	1	0.986928105	1	
	2	0.984749455	1	
	3	0.976034858	1	
	4	0.982570806	1	
	5	0.976034858	1	
	6	0.976034858	1	
	7	0.984749455	1	
	8	0.997821351	0.803921569	
	9	0.978213508	1	
	10	0.984749455	0.980392157	
		0.982788671	0.978431373	Mean
		4.12472E-05	0.003417916	Variance
Decision Tree Classifier	1	0.991285384	1	
	2	0.989106774	1	
	3	0.991285384	1	
	4	0.991285384	0.960784314	
	5	0.984749436	0.921568627	

	Fold	Training_accuracy	Validation_accuracy	
	6	0.991285384	0.980392157	
	7	0.995642722	0.941176471	
	8	0.984749436	0.37254902	
	9	0.989106774	1	
	10	0.989106774	0.980392157	
		0.989760345	0.915686275	Mean
		9.54052E-06	0.033452518	Variance
Random Forest Classifier	1	0.984749436	1	
	2	0.989106774	0.960784314	
	3	0.978213489	1	
	4	0.980392158	1	
	5	0.978213489	0.921568627	
	6	0.980392158	0.980392157	
	7	0.982570827	0.960784314	
	8	0.991285384	0.666666667	
	9	0.986928105	1	
	10	0.980392158	1	
		0.983224398	0.949019608	Mean
		1.90335E-05	0.00947328	Variance
Boosted Trees Classifier	1	0.995642722	1	
	2	0.995642722	1	
	3	0.993464053	1	
	4	0.995642722	1	
	5	0.995642722	0.941176471	
	6	0.995642722	1	
	7	0.997821331	0.921568627	
	8	0.997821331	0.392156863	
	9	0.993464053	1	
	10	0.997821331	1	
		0.995860571	0.925490196	Mean
		2.32576E-06	0.032356786	Variance

Appendix B. Selected Publications

EMERALD PUBLISHING LIMITED LICENSE TERMS AND CONDITIONS

Jul 22, 2018

This Agreement between Mr. Bilal Abu Salih ("You") and Emerald Publishing Limited ("Emerald Publishing Limited") consists of your license details and the terms and conditions provided by Emerald Publishing Limited and Copyright Clearance Center.

License Number	4392921071427
License date	Jul 20, 2018
Licensed Content Publisher	Emerald Publishing Limited
Licensed Content Publication	Journal of Knowledge Management
Licensed Content Title	Twitter mining for ontology-based domain discovery incorporating machine learning
Licensed Content Author	Bilal Abu-Salih, Pornpit Wongthongtham, Kit Yan Chan
Licensed Content Date	Jun 11, 2018
Licensed Content Volume	22
Licensed Content Issue	5
Type of Use	Dissertation/Thesis
Requestor type	Academic
Licensed Content Author	Yes
Portion	Full article
Will you be translating?	No
Format	Print and electronic
Geographic Rights	World rights
Order Reference Number	
Requestor Location	Mr. Bilal Abu Salih Curtin University Perth, 6102 Australia Attn: Mr. Bilal Abu Salih
Publisher Tax ID	GB 665359306
Billing Type	Invoice
Billing Address	Mr. Bilal Abu Salih Curtin University Perth, Australia 6102 Attn: Mr. Bilal Abu Salih
Total	**0.00 AUD**
Terms and Conditions	

TERMS AND CONDITIONS

1. The publisher for this copyrighted material is Emerald Publishing Limited. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your RightsLink account and that are available at any time at http://myaccount.copyright.com).
2. Limited License. Publisher hereby grants to you a non-exclusive license to use this material. Licenses are for one-time use only with a maximum distribution equal to the number that you identified in the licensing process; any form of republication must be completed within 12 months from the date hereof (although copies prepared before then may be distributed thereafter).
3. Geographic Rights: Scope. Licenses may be exercised only in the geographic regions you identified in the licensing process.

4. Altering/Modifying Material: Not Permitted. You may not alter or modify the licensed material in any manner. For permission to translate the material into another language please contact permissions@emeraldinsight.com.

5. Reservation of Rights. Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

6. License Contingent on Payment. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

7. Emerald always informs its authors of requests to republish their article. In the unlikely event that any author objects to the granting of the license to republish, Emerald reserves the right to revoke said license. The licensee will be informed and the license fee reimbursed within 10 working days.

8. Copyright notice: Disclaimer: You must include the following copyright and permission notice in connection with any reproduction of the licensed material and shall ensure that every published article gives due prominence on the title page to the original author/s, the journal title, volume, issue, page numbers and the copyright designation "© Emerald Publishing Limited all rights reserved."

9. Warranties: None. Publisher makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.

10. Indemnity. You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of license. This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing. This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Jurisdiction: This license transaction shall be governed by and construed in accordance with the laws of the United Kingdom. You hereby agree to submit to the jurisdiction of the courts located in the United Kingdom for purposes of resolving any disputes that may arise in connection with this licensing transaction.

15. Publisher (STM Signatory): Emerald is a signatory of the STM Permission Guidelines. Where, as part of this licensing transaction, you have indicated that you/your publisher is part of the STM Permissions Guidelines (by selecting 'Publisher (STM Signatory)', you warrant that the information included in the request is accurate and complete. Emerald requires you to complete the RightsLink process to obtain your license as proof of permissions cleared. Reuses falling within the limits of the STM Permissions Guidelines will be subject to a zero-rated (no fee) license. As per the terms of the STM Permission Guidelines, any license granted will apply to all subsequent editions and editions in other languages. All licenses falling outside of the limits of the guidelines may be subject to a permissions fee.

16. Special Terms:

v 1.6

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Twitter Mining for Ontology-based Domain Discovery Incorporating Machine Learning

Abstract

Purpose
This paper aims to obtain the domain of the textual content generated by users of Online Social Networks (OSNs) platforms. Understanding a users' domain(s) of interest is a significant step toward addressing their domain-based trustworthiness through an accurate understanding of their content in their OSNs.

Design/Methodology/Approach
This study uses a Twitter mining approach for domain-based classification of users and their textual content. The proposed approach incorporates machine learning modules.

The approach comprises two analysis phases: (1) the time-aware semantic analysis of users' historical content incorporating five commonly used machine learning classifiers. This framework classifies users into two main categories: politics related and non-politics related categories. (2) In the second stage, the likelihood predictions obtained in the first phase will be used to predict the domain of future users' tweets.

Findings
Experiments are conducted to validate the mechanism proposed in the study framework, further supported by the excellent performance of the harnessed evaluation metrics. The experiments conducted verify the applicability of the framework to an effective domain-based classification for Twitter users and their content, as evident in the outstanding results of several performance evaluation metrics.

Research Limitations/Implications
For the purpose of the proof of concept, this study is limited to an on/off domain classification for content of OSNs. Hence, we have selected a politics domain due to Twitter's popularity as an opulent source of political deliberations. Such data abundance facilitates data aggregation and improves the results of the data analysis. For future work, we aim to develop a multi-domain based classification, leveraged by domain ontologies, semantic technologies and linked open data. Therefore, beside the politics domain, an analysis of other domains of interest will be further investigated in the future.

Furthermore, the currently implemented machine learning approaches assume that uncertainty and incompleteness do not affect the accuracy of the Twitter classification. In fact, data uncertainty and incompleteness may exist. In the future, we will formulate the data uncertainty and incompleteness into fuzzy numbers which can be used to address imprecise, uncertain and vague data. Based on the fuzzy numbers, a fuzzy based machine learning algorithm will be developed in order to estimate the effect of the uncertainty and incompleteness.

Practical Implications
This study proposes a practical framework comprising of significant implications for a variety of business-related applications such as the Voice of Customer (VoC)/Voice of Market (VoM), recommendation systems, the discovery of domain-based influencers, and opinion mining through tracking and simulation. In particular, the factual grasp of the domains of interest extracted at the user level or post level enhances the customer-to-business engagement. This contributes to an accurate analysis of customer reviews and opinions in order to improve brand loyalty, customer service, etc.

Originality/Value
The paper fills a gap in the existing literature by presenting a consolidated framework for Twitter mining that aims to uncover the deficiency of the current state-of-the-art approaches to topic distillation and domain

discovery. The overall approach is promising in the fortification of Twitter mining towards a better understanding of users' domains of interest.

Keywords: Domain Discovery; Twitter Mining; Ontology; Machine Learning; Domain-based Trustworthiness.

1. Introduction

The demand for real-time business intelligence and the popularity of social media has created a need for social business intelligence. Social business intelligence aims to reveal the fundamental factors derived from social perspectives, that determine an organisation's performance. People express their thoughts, feelings, activities, and plans, etc. via OSNs. Often, their posts link to product(s), service(s), event(s), society, or person(s), etc. and people in OSNs intuitively tend to seek and connect with like-minded people. This homophily results in building homogenous personal networks based on behaviours, interests, and feelings, etc.[1]. The rapid increase in unstructured social data has highlighted its importance as a means of acquiring deeper and more accurate insights into businesses and customers. In particular, OSNs are a medium for content makers to express and share their thoughts, beliefs, and domains of interest. This gives individuals access to a wider audience which positively affects their social rank and provides other benefits, such as gaining political support[2]. Therefore, the cornerstone of building users' online social profiles is a veritable understanding of their domains of interest.

Due to the open environment and limited restrictions of social media, rumours can spread quickly and false information can be broadcast rapidly. This may have adverse effects on businesses, political management, and public health etc. particularly if the false information is being published together with trustworthy information. However, if it is accurate information, this could be greatly beneficial to individuals and organisations as a means of acquiring value from social media data. Spam is a well-known category of low-quality content. Social spam content such as fake accounts, bulk messaging (sending the same post many times in a relatively short period of time), malicious links, and fake reviews lower the quality of experience of social community members [3]. Social media data is big, heterogeneous, and unstructured in its textual content, structured in its metadata, can be linked, and has different trust levels. Sherchan, et al. [4] defines "trust" as the measurement of confidence that a group of individuals or communities will behave in a predictable way. Trust in social media refers to the credibility of users and their shared content in a particular domain. Users are known to be trustworthy in a particular domain. However, this does not mean that their trustworthiness will have the same value in other domains. The trustworthiness of social media data is now crucial [5]. With such a vast volume of data interchanged within social media ecosystems, determining domain-based data credibility is considered a vital issue. The importance of domain-based trust in the social media context originates from affluent resources for market analysis e.g. the Voice of the Customer (VoC) and the Voice of the Market (VoM), recommendation systems, domain-based influencers' discovery, and the like. Hence, understanding users' domain(s) of interest is a significant step in addressing their domain-based trustworthiness through an accurate understanding of their content temporally in OSNs.

In this context, companies incorporate advanced social data analytics when designing effective marketing strategies and seek to leverage the interactive quality of OSNs. Thus, to create the required interaction with their customers, companies use many modern communication to attract customers and visitors to their online social platforms. Consequently, it is necessary for companies to analyse their customers' social content and classify the customers into appropriate categories based on their topics of interest, in order to deliver the right message to the right category.

Most of the existing approaches to this topic rely on bag-of-words techniques such as LDA [6]. However, despite the importance and popularity of these techniques for inferring the users' topics of interest, when it comes to the use of Twitter, there are three main shortcomings of such an approach; (1) the inability to consider the semantic relationships of the terms in the user's textual content; (2) the inadequacy of its application to a topic modelling technique using short text messages such as tweets; and (3) the high-level topic classifications that use these bag-of-words statistical techniques are inadequate and inferior[7].

On the other hand, incorporating semantic web consolidated tools such as AlchemyAPI™[1], offers a comprehensive list of taxonomies divided into hierarchies, where the high-level taxonomy represents the high-level domain and the deeper-level taxonomy provides a fine-grained domain analysis. For instance, "art and entertainment" is considered a high-level taxonomy in which "graphic design" is one of its deep-level taxonomies. LDA is unable to provide high-level topics such as "art and entertainment" from a corpus of tweets unless this term exists in the corpus. Semantic analysis, conversely, extracts semantic concepts and infers high-level domains through analysing the semantic hierarchy of each topic, leveraging an ontology; this is not possible when using an LDA technique.

The main challenge in obtaining the accurate domain of a tweet is the ability to accurately determine the classification of its textual content. This is due to the several features of linguistics such as: polysemy (where the same word has several meanings), homonymy (where words have the same spelling and pronunciation, but have different meanings), and contronymy (where the same word has contradictory meanings). This diversity in linguistics makes the process of determining the correct domain of interests from the short textual content of the tweet more difficult. Hence, it is essential to obtain an accurate understanding of the semantics of the tweet text in order to determine the user's domain of knowledge. This will assist in determining the topic/domain of the tweets that will be posted by the user in future. This paper aims to address this problem by proposing a comprehensive framework incorporating semantic analysis and machine learning.

Semantic analysis, through existing Ontologies and Linked Data, enables the eliciting of knowledge from social data, thereby enriching its textual content to deliver semantics and links each message with a particular domain. Machine Learning applications enable real-time predictions leveraging high quality and well-proven learning algorithms. Based on the current dominant position and high impact on business in several use cases, according to Gartner's recent report on emerging technologies[2], incorporating machine learning in particular enhances the decision-making process and provides valuable insights from large-scale data.

This study presents an approach to glean profound insights into users' domains of interest from their pervasive propagation of tweets. This is achieved through a systematic approach beginning by addressing the volume quality of social big data incorporating data generation and acquisition techniques, and then inferring the added value obtained from the data analysis. This aims to contribute to an advanced domain-based trustworthiness approach that is able to filter out unsolicited tweets and increase the value of content. To achieve this objective, this paper presents a consolidated framework leveraging former knowledge obtained from an analysis of the user's historical content. In this context, the politics domain is used to determine the user's interest in this domain. Hence, we propose an effective approach to classify Twitter users and their new updates according to two main categories; (i) **on-topic**: a user or tweet is classified under the politics domain; (ii) **off-topic**: a user or tweet is classified under the non-politics domain.

The proposed approach comprises two main analysis phases incorporating several semantic analysis tools and machine learning modules. In the first phase, the users' historical tweets are collected; their interest is examined over time thereby providing a prediction of the users' interest, taking the temporal factor into consideration. In the second phase, the outcome of the previous analysis is used as a primary input to forecast the domain of future tweet content. Users' classification is achieved through the use of well-known machine learning classifiers. A comparison is conducted to benchmark the performance of the incorporated machine learning modules.

[1] AlchemyAPI is accuired by IBM's Watson since 2015

[2] http://www.gartner.com/document/3383817?ref=solrAll&refval=175496307&qid=34ddf525422cc71383ee22c858f2238a, Visited in 25/10/2016.

The main contributions of this paper are summarised as follows:

- A time-aware framework incorporating comprehensive knowledge discovery tools and well-known machine learning algorithms is proposed for domain-based discovery, which is applicable to the Twittersphere platform and customisable to other OSNs.
- The proposed framework is able to perform classification tasks at the user level and tweet level.
- The conducted experiments verify the effectiveness and applicability of our model as evident in the outstanding results of several performance evaluation metrics.

The rest of this paper is organised as follows: Section 2 reviews the theoretical background and existing work related to tweet mining. The framework of the proposed approach is described in Section 3. Section 4 presents the various machine learning algorithms which are incorporated into the proposed framework. The detailed experiments conducted to classify Twitter users and their tweets are described in Section 5. In Section 6, the motivation for the research and the benchmark results are discussed with state of the art approaches. Finally, the paper is concluded by listing the contributions, the limitations, and the anticipated enhancements of the proposed framework.

2. Theoretical Background

Since the uprising of Web 2.0, the role of web browsers has changed to enable users to send and receive content that is leveraged by several online tools such as e-mail applications and chat forums to more recent and revolutionary electronic platforms such as OSNs. OSNs such as Facebook®, Twitter®, LiveBoon®, Orkut®, Pinterest®, Vine®, Tumblr®, Google Plus®, and Instagram® among others allow users to share videos, photos, and files, and have instant conversations. These platforms provide important means of growing and adhering between societies, bringing together concepts and visions, in addition to its active and distinctive role as an effective medium of social interaction. The dramatic increase in the impact of social data is a testimony to our growing digital lifestyles. Social data has emerged in industries and activities ranging from marketing and advertising to intelligence gathering and political influence. In fact, the extent of this revolution is continually spreading; it is about building data infrastructures that are needed to effectively digest the breeding of social data to achieve added value. This has motivated research communities to dig deep, to provide solutions and to develop platforms for potential use of these datasets in several applications (e.g. marketing [8] e-commerce [9], education [10], health [11], etc.). These endeavours include the recent efforts to understand the dynamic and unstructured nature of social content in an attempt to deliver the right content to its interested users [12], [13]. Further, social media has also been used to improve employees' productivity [14], knowledge sharing [8], [15], [16], and overall firm innovation performance [17].

The following section discusses the theoretical background for the current approaches to Twitter mining followed by an evaluation to these approaches and the proposed solutions.

2.1. Semantic Data Analysis

Berners Lee introduced the notion of the Semantic Web to facilitate the machine understanding of web language; this data can be used across several applications [18]. Ontology is defined as the formal explicit specification of a shared conceptualisation [19]. Incorporating Semantic Analysis in the area of social big data has generated a steady support from several research communities. Such endeavours attempt to untangle the ambiguity of the unstructured nature of social data content and discover the domain of knowledge through incorporating semantic analysis techniques to identify, annotate, and enrich entities embodied in social data content. In other words, incorporating semantic analysis ushers a better understanding of the contextual content of social media data through the extracting their semantic data. De Nart et al. [20] proposes a content-based approach to extract the main topics from the tweets. This approach is an attempt to understand the research communities' activities and their emerging trends. Chianese et al. [21] proposes a data-driven and ontology-based approach to identify cultural heritage key performance indicators as expressed by social network users. This approach can be used in different domains but is only relevant to user domains. [22] and [23] both apply ontology to create applications in crisis situations. The former ontology was designed to be used for earthquake evacuation to help people locate evacuation centres based on data posted on Twitter. The latter showed a geo-tagger that aims to process unstructured content and infer locations with the help of existing ontologies. In [24] the authors harness the ontology-driven approach to obtain Twitter users' interests however, their experiments have been conducted at the tweet level only, lacking a consideration of the user's domain of interest. Twitter mining through semantic analysis has been further extended to address social media trends [25], sentiment analysis [26], knowledge base and discovery [27], employment trends [28], event classification [29] and fundamentalism detection [30], among others.

2.2. Machine Learning for Data Classification and Topic Distillation

Topic distillation (a.k.a topic discovery, topic modelling, latent topic modelling or statistical topic modelling) is an automatic approach used to distil topics from a corpus of words embodied in a set of documents incorporating statistical techniques [6], [31], [32]. The primary reason for developing topic discovery techniques is to improve information retrieval particularly when searching large corpora of data and indexing.

These statistical-based techniques have also been used as other means of topic modelling and discovery in social data mining. Examples of such statistical-based techniques are: LDA (Latent Dirichlet Allocation) [33], Latent Semantic Analysis (LSA), and more recently, Fuzzy Latent Semantic Analysis (FLSA) [34]. LDA is based on an unsupervised learning model harnessed to identify topics from the distribution of words. LSA, an early topic modelling method, has also been extended to pLSA [35], and generates the semantic relationships based on a word-document co-occurrence matrix. FLSA supposes that the list of documents and their embodied words can be fuzzy clustered, where each cluster is represented by a certain topic. LDA and similar unsupervised techniques have been widely used in several modelling applications [36]-[41]. Vicient et al. [42] presented a methodology for unsupervised topic discovery through linking social media hashtags to terms of WordNet. Further, the authors of [43] harness in their approach statistical techniques that able to detect interpretable topics. Incorporating statistical techniques to benefit social data analysis approaches is also evident in the literature; Twitterrank [38] incorporates LDA technique to classify users interests through applying LDA modelling technique to the overall content of each user. Ito et al. [44] adopts LDA for topic discovery to validate the credibility of the content on Twitter. Xiao et al. [45] proposes an approach for predicting users influence in the social data context. They compute the topic distribution of users through the use of LDA technique.

2.3. Evaluation of Current Approaches

Inclusion of both user and tweet levels

The increasing use of Twitter has motivated researchers to develop several methods for discovering the main interest(s) of their users. Due to the ambiguity, shortness and nosiness of tweets [7], these endeavours are still in their infancy; Hence, extensive research in this area is vital [46]. Twitter tools [4], [26], [47] are focused on the exploration of user networks to obtain information for user interests and topics. These approaches only extract keywords to obtain a summary of Twitter data. However, the use of keywords only cannot fully cover user domains and may generate misleading user information. Therefore, the proposed approach in this study considers both the user level and tweet level, which involves semantics of words and accurate disambiguation for social networks study. The accurate classification of the users' interest assists in providing an accurate understanding of short textual content of future tweets. This benefits several applications, the aim of which is to obtain a correct domain-based trustworthiness of users and their content in OSNs.

Integration of different repositories

There have been two main research avenues in which domains of interest have been investigated and inferred from the textual content of users in OSNs. The first avenue focuses on the incorporation of ontologies, semantic web and linked data to enrich textual data and extract knowledge, thereby linking the textual data with a particular user domain. For instance, Michelson and Macskassy [7] use the DBpedia knowledge base to annotate entities in users' tweets, and extract the users' main interests by using the categories proposed on Wikipedia. Maio et al. [12] uses Wikipedia to infer users' topics of interest embodied onto their proposed ranking algorithm. Wikipedia has also been utilised as a knowledge base repository for topic discovery in [48], [49]. In addition to DBpedia, the current approach incorporates other knowledge base repositories such as Freebase, YAGO and OpenCyc. Furthermore, the study adopts and extends the BBC Politics ontology to capture politics domain knowledge.

Incorporation of domain ontology, semantic web, and machine learning

Statistical techniques have been used as another means of topic modelling and discovery in Twitter mining. The two dominant statistical techniques that have been used are LDA (Latent Dirichlet Allocation) [33], and Latent Semantic Analysis (LSA). LSA, an early topic modelling method that has been extended to pLSA [35], generates the semantic relationships based on a word-document co-occurrence matrix. LDA, is an extension of pLSI, and LDA is based on an unsupervised learning model in order to identify topics from the distribution of words. These approaches have been widely used in several modelling applications [36]-[39]. However, the high-level topic classifications that use these bag-of-words statistical techniques are inadequate and inferior [7]. Furthermore, the brevity and ambiguity of such short texts makes the process of topic modelling using these statistical models more difficult [50]. In addition, these methods do not consider the temporal factor. In other words, the users' knowledge evolves over time and their interest might be diverted elsewhere depending on their experience, work, study, or other factors. Hence, it is important to scrutinize users' interest over time to infer intrinsic topics of interest to users in OSNs. The approach of this study addresses these problems through the use of a systematic process which addresses temporally the domain of interest at the user level, and attempts to identify the domain not readily evident at the tweet level. The approach includes the use of domain ontology, semantic web technologies and machine learning, where domain ontology and semantic web attempt to extract the semantics of textual data in order to determine the domain of the textual data, and machine learning attempts to perform domain-based classification at the user and tweet levels.

Addressing key features of big data

There is a notable consensus among researchers that the traditional tools for collecting, analysing and storing data are no longer able to handle large amounts of big data. Therefore, more advanced, unconventional and adaptable technical solutions are required to address the challenges of managing a wide variety of big data islands, which expand exponentially through the huge data generated from tracking sensors, online social networks, transaction records, metadata and many other data fountains. Manyika et al. [51] lists some of the big data technologies such as Big Table, Cassandra (open Source DBMS), Cloud Computing, Hadoop (open Source framework for processing large sets of data), etc. Chen et al, [52] discusses the various open issues and challenges of big data, and lists the key technologies of big data. The incorporation of big data technology to facilitate domain discovery and to measure users' trustworthiness in OSNs is unavoidable, particularly regarding the nature of the content of social media which has a wide berth. This has interestingly attracted researchers of social trust to leverage the big data techniques to benefit their conducted experiments [53]-[55]. However, previous studies have failed to address the key features of big data such as volume (i.e. massive social data datasets), veracity (i.e. reputation of the data sources), and value (outcome product of the data analytics). Hence, this study starts with the characteristics of big data and sorts out issues related to these dimensions to better obtain the anticipated outcomes of social data analysis.

3. System Architecture

Figure 1 shows the architecture of the proposed framework which adopts a big data infrastructure. This framework comprises three main components, namely: (1) data collection and acquisition, (2) features extraction, and (3) the prediction module. The big data infrastructure at the School of Information Systems, Curtin University, is used as a distributed environment to facilitate data storage and analysis. This facility has a 6-node cluster, each with 2 TB Storage, 8 Core Processors, and 64 GB RAM.

The information flow through the proposed framework can be described in steps. As shown in Figure 1, steps 1 to 4 represent the processes required to achieve the predicted likelihood value of the user's interest in the politics domain. This is the first outcome value (Politics Likelihood / User Level) indicated by the red dotted line. Steps 5 to 9 follow and predict the politics domain-based likelihood value of a newly-posted user tweet. This is the second outcome value (Politics Likelihood / Tweet Level) indicated by the red dotted line. In the proposed framework, the user posts public content to the Twitter network, which facilitates data collection through the available application programming interfaces (APIs). The user's content is collected in two phases, namely, historical user's content and new user's content. The user's historical content represents the recent and former tweets which are collected in the first phase. The user's new content refers to their future tweets which will be collected during the second phase.

The collected historical tweets are pre-processed and passed to either the tweet features or user features extraction module. A list of user features is extracted and fed into a machine learning module to predict the politics domain likelihood value, where the domain likelihood indicates the user's interest in the politics domain. This domain likelihood is harnessed further and is added as another feature to the list of features extracted from the new user tweet after pre-processing during the second phase. The newly combined list of tweet features is fed into the machine learning module to predict the politics domain likelihood of the newly posted tweet. The following subsections explain the mechanism of each component of the proposed framework.

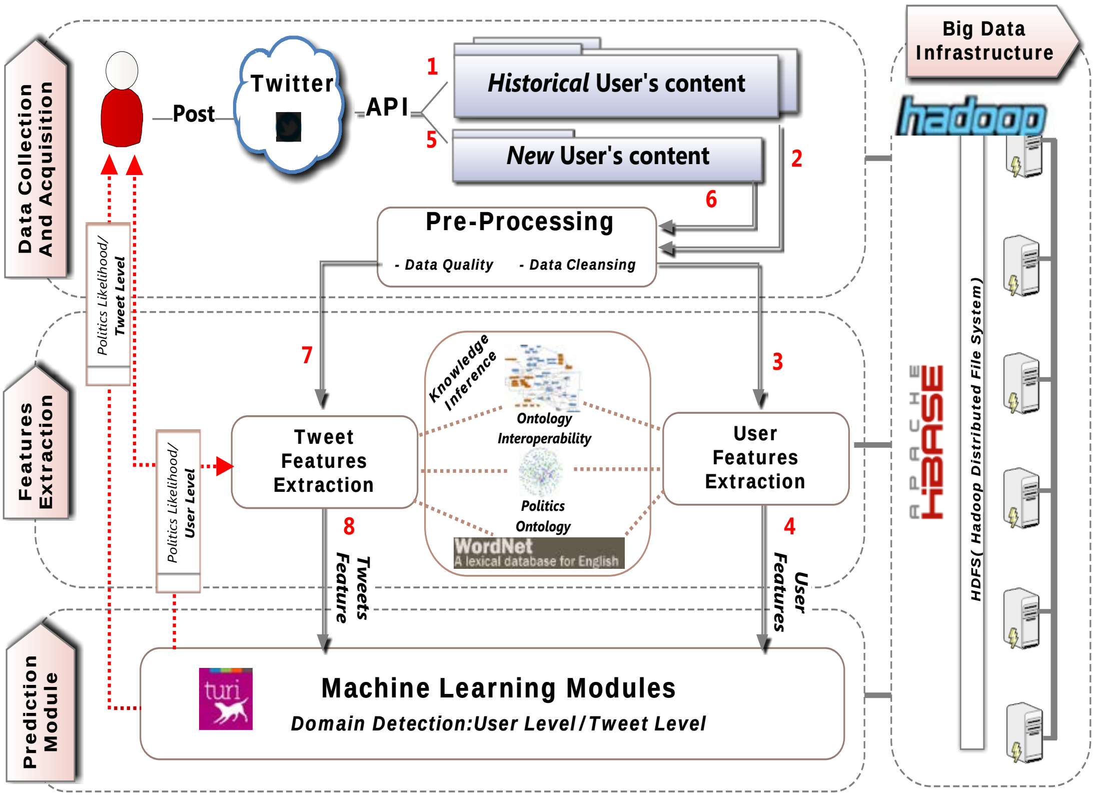


Figure 1: System Architecture

3.1. Data Collection and Acquisition

3.1.1. Data Generation and Selection

Since the establishment of Twitter™ in 2006, Twitter has provided a rich dataset of over 500 million tweets daily which is around 200 billion tweets a year [56]. Twitter mining is an emerging research field falling under the umbrella of data mining and machine learning. Twitter™ is the chosen subject of this paper because: (1) Twitter is a fertile medium for researchers in diverse disciplines, leveraging the vast volume of content; (2) Twitter facilitates data collection by providing easy access APIs to The Twitter sphere; (3) due to the economy and the ambiguity and brevity of a tweet's content, it is challenging to determine the accurate domain(s) to which the user's tweet is referring.

For the purpose of proof of concept, this study is limited to an on/off domain classification to content of OSNs. Hence, the politics domain has been selected for the following reasons: (1) Twitter has been intensively incorporated as an important arena by politicians to express and defend their policies, to practice electoral propaganda and to communicate with their supporters [57], (2) Twitter has raised considerable controversy regarding its usage as a platform to attack political opponents [58], (3) Twitter is characterised by its growing social base to include broad political social groups leveraged by ease of use, free access, and deregulated nature [59], (4) the amount of the political discourse in social content is overwhelming; over one-third of OSNS's users believe that they are worn-out by the quantity of the political content they encounter [60]. Such an abundance of data facilitates data aggregation and improves the outcome of data analysis. For future work, this study aims to develop a multi-domain-based classification, leveraged by domain ontologies, semantic technologies and linked open data. Hence, beside the politics domain, an analysis of other domains of interest may be further investigated in the future.

The dataset used for this study has been collected using Twitter's "User_timeline[3]" API method. This mechanism allows access to and retrieval of public users' content and metadata. The collection of the users' content was accomplished in two stages: (1) by collecting historical user content (up to "3,200" most recent tweets[4]). This dataset will be used to predict the user's interest in the politics domain in general; and (2) by collecting the new content of those users whose historical tweets were obtained in the first phase. This is used to predict the politics domain likelihood value of the new tweet. As will be described later, the dataset of the first stage is used to predict the user's interest in politics at the user level, i.e. to establish an understanding of the user's interest in the politics domain based on the user's past content. The politics domain likelihood value of the new user's tweet is predicted based on the analysis of its content, other than the politics interest likelihood value predicted at the user level.

3.1.2. Pre-processing Data

The veracity of data refers to the certainty, faultlessness and truthfulness of data [61]. Although reliability, availably and security of data's nascence and storage are significant, these factors do not guarantee data correctness and consistency. Appropriate data cleansing and integration techniques should be incorporated to ensure certainty of data. The data collected for the user's content, and historical and new tweets, are pre-processed by data quality enhancement and data cleansing techniques which are discussed below:

- **Data cleansing** of user content is conducted by using the following techniques: (1) all redundant content (i.e. same dataset crawled more than once) such as tweets or user data is eliminated with their metadata; (2) removing stop words; (3) removing URLs; (4) decoding all HTML entities to their applicable

[3] https://dev.twitter.com/rest/reference/get/statuses/user_timeline.

[4] This threshold is set by Twitter™ as the maximum number of recent tweets the twitter API is allowed to retrieve.

characters; (5) eliminating all HTML tags such as <p>, <a>, etc.; (6) removing punctuation marks, correcting encoding format, etc.

- **In data quality enhancement,** the list of Twitter handles (a.k.a. Twitter user/screen name such as @example), which are indicated in the user's tweets, is collected and replaced with the user's corresponding names. This is achieved through the Application Programming Interface (API) of Twitter's "lookup[5]". These handles are normally neglected or deleted when mining user's tweets. However, these handles are important because they are used by Twitter users to mention other Twitter users in their tweets, replies or re-tweets. Hence, it is essential to identify and ascertain the actual names of those users. This assists in the process of domain extraction. For example, a user shows an interest in the politics domain if she/he commonly indicates handles linked to politicians or political parties, in addition to publishing other politics-related content.

3.2. Features Extraction

The pre-processed dataset is passed to the features extraction modules. For the new users, the features of their content (historical tweets) are extracted in the "User Features Extraction" module. As for the new tweets of the already existing users, features are extracted in the "Tweet Features Extraction" module.

The aim of this study is to establish a fundamental ground for efficiently detecting the domain of interest of Twitter users, which will significantly contribute to a better understanding of the domain(s) of future users' tweets. As a proof of concept, the proposed system is validated by an application on the Politics domain, where the proposed system attempts to detect whether the domain of a tweet is or is not politics-related. This validation is based primarily on former knowledge about a user's political interests obtained by analysing the user's historical content. To do so, the following politics-domain knowledge inference approach is designed to extract the semantics of a user's tweets, thereby uncovering the user's domain of interest.

3.2.1 Politics Domain Knowledge Inference

In the feature extraction module, domain knowledge inference is the main process used to extract user and tweet features from pre-processed datasets. For the purpose of proof of concept, the study focuses on the politics domain, using politics ontology, WordNet, and ontology interoperability to infer politics knowledge.

Politics Ontology and WordNet®

The politics domain refers to the knowledge captured in politics ontology along with its knowledge base. BBC defines politics ontology as "an ontology which describes a model for politics, specifically in terms of local government and elections" [62]. The BBC Politics ontology conceptualises a politics model especially for the UK government and elections. It was originally designed to cope with UK local government and European Elections in May 2014. This study applies the BBC Politics ontology to Australian politics by further extending politics concepts. Figure 2 shows the BBC Politics ontology, while Figure 3 shows the Politics ontology used in this research. Furthermore, the study uses WordNet[6], which is a lexical dictionary used to construct relations between terms through synonymies. Synonyms (or synsets) are a set of interrelated terms or phrases which indicate the same semantic concept, such as the words "elections, public opinion poll, opinion poll, and ballot". All the synsets of the political concepts captured in politics ontology depicted in Figure 3 are examined, and only the synonyms applicable to the politics context are captured.

[5] https://dev.twitter.com/rest/reference/get/statuses/lookup.

[6] https://wordnet.princeton.edu/

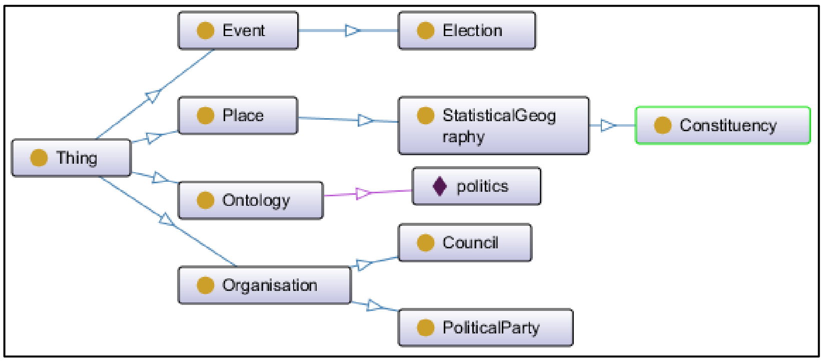
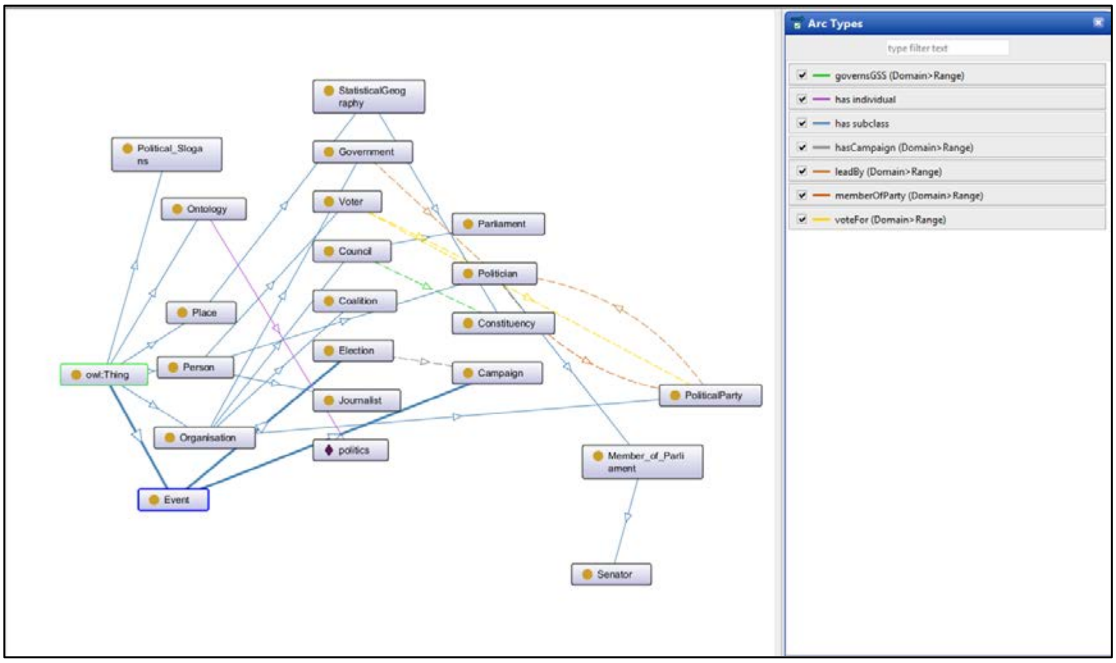


Figure 2: BBC Politics Ontology



☑ governsGSS (Domain>Range)
☑ has individual
☑ has subclass
☑ hasCampaign (Domain>Range)
☑ leadBy (Domain>Range)
☑ memberOfParty (Domain>Range)
☑ voteFor (Domain>Range)

Figure 3: BBC Politics Ontology Extension

Ontology Interoperability

The interlinking with other relevant entities defined in other datasets supports interoperability. The approach taken in this study addresses information interoperability by focusing on the equivalence links that direct the URI to refer to the same resource or entity. The politics ontology supports the equivalence links between the ontology components and the tweet data. The resources and entities are linked through the owl#sameAs relation. This implies that the subject URI and object URI resources are the same, and hence the data can be further explored.

In the interlinking process, we incorporate AlchemyAPI™ as a one-stop shop, leveraging access to a wide variety of linked data resources[7] through providing easy access APIs. These resources include but are not limited to: different vocabularies such as Upper Mapping and Binding Exchange Layer (UMBEL), Freebase (which is a community-curated database for well-known people, places, and things), YAGO high quality knowledge base, and DBPedia knowledge base, etc. These resources are used to help extend the knowledge base of the politics ontology by identifying (non-)Australian politicians and (non-)Australian political parties

[7] http://www.alchemyapi.com/products/alchemylanguage/linked-data.

from users' tweets. For example, at this stage, we capture "99,812" instances of "2009" politicians, and "48,704" instances of "59" political parties in the politics ontology.

3.2.2. User Level Features

The political interest of users is primarily measured by two main proposed factors: continuity and knowledgeability. Continuity refers to the frequent interest of a user in a certain domain. In other words, the user demonstrates an interest in the politics domain by tweeting or retweeting content in this domain over a relatively long period of time. Continuity is measured by counting the number of political entities identified from the user' tweets in each time period (such as every month, quarter, etc.). Knowledgeability (or Speciality) refers to the user's close acquaintance with the politics domain and also refers to the user's dedicated pursuit of the politics domain through a commitment such as work or study. Knowledgeability is measured by accumulating the distinct number of political entities annotated from the user's tweet, and the user's profile description. Table 1 shows the list of features used to classify whether the user's interest is **on-topic** or **off-topic**. **On-topic** refers to when the user demonstrates a continuous interest in the politics domain. **Off-topic** users are those whose Twitter content shows their non-interest in the politics domain.

Table 1: A List of User's Features

No	Features	Description
1	no_tweets, x_1	The total count of users' historical collected tweets up to 3,200 tweets.
2	unq_pol_entities, x_2	Total count of distinct/unique political entities extracted from all user's tweets
2	pol_entities_pre_QW_YYYY, x_3	Count of political entities annotated from the tweets posted before quarter 'W' of the year 'YYYY'
3	pol_entities_QW_YYYY, x_4	Count of political entities annotated from the tweets posted in quarter 'W' of the year 'YYYY'
4	pol_entities_QX_YYYY, x_5	Count of political entities annotated from the tweets posted in quarter 'X' of the year 'YYYY'
5	pol_entities_QY_YYYY, x_6	Count of political entities annotated from the tweets posted in quarter 'Y' of the year 'YYYY'
6	pol_entities_QZ_YYYY, x_7	Count of political entities annotated from the tweets posted in quarter 'Z' of the year 'YYYY'
7	profile_pol_entities, x_8	Count of political entities annotated from user's profile description
9	verified(Authentication Status), x_9	Authentication flag used for accounts of public interest (for example, politicians)

The features x_2 to x_8 as depicted in Table 1 are selected to primarily focus on users' ongoing interest in and knowledge about the politics domain by extracting the political entities from their tweets and by leveraging the knowledge-inference tools explained in the previous section. In particular, features x_2 to x_8 are proposed to address the political knowledgeability of users. Moreover, features x_3 to x_7 address the continuing interest of users in the politics domain. Features x_1 and x_9 are added to support the aforementioned features and will be discussed later in this paper.

Unq_pol_entities (x_2), listed in Table 1, refers to the number of distinct political entities extracted from the history of a user's tweets. Profile_pol_entities (x_8) represents the number of all political concepts that are identified in the users' profile description on their Twitter accounts. The former feature represents the diversity of the political concepts embodied in the users' tweets, and the latter feature, x_8, is used to examine the explicit indication of the user's interest in the politics domain, particularly if the user works in this domain. This is usually clearly indicated in their profile description.

The list of all political entities is counted periodically. The political entities extracted from the user content for each time period is used to scrutinise political interest temporally rather than scrutinising the tweets as a whole. Therefore, the collected historical tweets are divided into five groups, x_3 to x_7. Four groups, x_4 to x_7, indicate the four sequential and recent quarters (W, X, Y and Z), where 'Z' is the most recent quarter, and one group, x_3, indicates the rest of the tweets posted before the 'W' quarter. This mechanism is proposed because the user's interest(s) may change, and their knowledge may evolve over time. Hence, it is more efficient to examine the user's domain(s) of interest based on current and recent behaviours from the four time groups. Furthermore, some users only show a particular interest in the politics domain when popular political events are taking place. For example, a users' involvement in conversations during election campaigns does not necessarily indicate an interest in the politics domain generally, as the election is a trending topic only, on which users with dissimilar interests share their thoughts, and/or anticipations about the potential candidates.

The remaining two features listed in Table 1 are the no_tweets, and verified features. The no_tweets, x_1, relates to the number of collected historical tweets. This feature is important as a means of addressing the ratio between the number of political concepts accumulated for features x_2 to x_8 and the total number of tweets. For example, two users might archive the same number of distinct political concepts, although the number of tweets differs for each user. The verified feature, x_9, is the authenticated flag (i.e. blue verified badge ✓). Twitter may set this flag to '1' for users of public interest. Twitter currently offers this feature to help users find influential and high quality accounts in several domains.[8]

3.2.3. Tweet Level Features

In the previous section, the user's historical collected tweets were studied to obtain an accurate understanding of that user's interest in the politics domain. A list of features extracted from the content at the user level is formulated and will be used to predict the user's political interest (likelihood). On this backdrop, the likelihood of the user's interest in the politics domain would be a main driver facilitating an understanding of the domain of the users' future tweets. Table 2 summarises the list of features selected to predict the political likelihood at the tweet level.

Table 2: A List of Tweet Features

No	Features	Description
1	political_entities, x_{10}	Count of political entities extracted from the tweet
2	words_count, x_{11}	Count of tweet's words
2	political_perc, x_{12}	Computed as $\frac{x_{10}}{x_{11}}$
3	pol_entities_recent_quarter, x_{13}	Count of political entities annotated from the user's tweets posted in the most recent quarter
4	user_pol_likelihood, x_{14}	Political likelihood value

As shown in Table 2, political_entities (x_{10}) represents the number of political entities annotated from the tweet using the aforementioned knowledge discovery tools. Words_count (x_{11}) is the number of remaining words in the tweet after the cleansing process. Political_perc (x_{12}) represents the ratio between the number of political entities annotated in the tweet to the total words used. Despite its brevity, a tweet might discuss more than one topic; thus, x_{12} is proposed as an indicator of the weight of the politics domain in the tweet. Pol_entities_recent_quarter (x_{13}) represents the number of political entities from all tweets posted during the most recent quarter. This feature is included because it represents the user's most recent political (non-)interest. User_pol_likelihood (x_{14}) is the predicted value obtained from user analysis which signifies a user's general interest in the politics domain.

[8] https://blog.twitter.com/2016/announcing-an-application-process-for-verified-accounts-0.

Features x_{13}, and x_{14} are proposed to indicate the recent political interest of the user. These features assist in further understanding the actual context of the newly posted tweet, given their typically short length and ambiguity. Hence, users that have been predicted to be interested in the politics domain will likely post politics-related content in future posts. This will be discussed and demonstrated further in the experiments section (Section 5).

4. Machine Learning Module for Classification

This section provides an overview of well-known machine learning classification algorithms. Based on the user and tweet features, $\bar{x} = [x_1, x_2, \text{to } x_{14}]$, a machine learning module determines the likelihood of whether or not a user/tweet is in the politics domain, namely y, where the following commonly used implicit or explicit classifiers including logistic regression [63], decision tree [64] [65] [66], and support vector machine [67], are used for user based classification, and logistic regression is used for tweet-based classification. For demonstration purposes, this overview will consider the domain-based classification at the user level. Logistic regression [68], [69], decision tree [70], and support vector machine [71], [72] in particular have been used for text categorisations. Also these approaches are more narrow and computationally simpler than recently developed machine learning approaches such as the deep learning or deep networks approaches.

Development of a novel classifier is not the main research focus of this paper. Hence, the study attempt to implement a computationally simple but effective approach. Five commonly used approaches are used, namely: logistic regression, support vector machine, top-down inducing based decision tree, random forest-based decision tree, and gradient-boosting-based decision tree.

4.1. Logistic Classifier

Logistic regression is commonly used for conducting binary classification tasks [63]. In logistic regression, the likelihood of whether the user is in the politics domain is determined by a logistic function consisting of a linear summation of $x_1$to x_9. The logistic function is given as:

$$f^{LR}(\bar{x}) = P(y=1|\bar{x}) = \frac{1}{1+\exp\left(-\left(b_0 + \sum_{i=1}^{14} b_i \cdot x_i\right)\right)} \qquad (1)$$

In the study, b_0, b_1, to , b_{14} are the logistic coefficients, which are determined by maximizing the likelihood when $y=1$, which indicates that the user is definitely in the politics domain. Unlike linear regression which has normally distributed residuals, ordinary least square regression cannot be applied to determine the logistic coefficients. Hence, to determine b_0, b_1, to, b_{14}, Newton's method is used. Newton's method begins with tentative logistic coefficients and it adjusts the coefficients slightly to see whether they can be improved. It repeats this iterative process until the process converges. A user is classified in the politics domain, when the value of $f^{LR}(\bar{x})$ in (1) is large than 0.5. Otherwise, the user is classified as being in the non-politics domain.

4.2. Support Victor Machine

Support vector machine (SVM) is commonly used for conducting binary classification tasks [67] particularly involving with the confusion matrix analysis (true-positive and false negative). SVM is relatively new and was designed for applications involving text categorization and recognition (see for example [71], [72]).

In SVM, the user is classified as either being in the politics or the non-politics domains, based on the following formulation:

$$f^{SVM}(\bar{x}) = \operatorname{sgn}(D(\bar{x})) \qquad (2)$$

where $D(\bar{x}) = \sum_{i=1}^{14} w_i \varphi(x_i) + b$; (3)

and $\operatorname{sgn}(D(\bar{x})) = \begin{cases} 0 \text{ if } D(\bar{x}) < 0 \\ 1 \text{ if } D(\bar{x}) \geq 0 \end{cases}$ (4)

φ is the transform function which is correlated to the kernel function and w_i with $i = 1,2,$ to14 and b represents the SVM parameters. The five common kernel functions are; linear function, homogeneous polynomial, inhomogeneous polynomial, gaussian radial basis function and hyperbolic tangent. The kernel function is generally determined by a trial and error method. After the kernel function has been determined, w_i and b are reformulated as a quadratic programming problem, which is solved by the gradient descent algorithm. When the value of $f^{SVM}(\bar{x})$ in (2) is equal to 1, the user is classified as being in the politics domain. Otherwise, the user is classified as being in the non-politics domain.

4.3. Decision Tree Classifier

A Decision tree is a classifier which can express a recursive partition of the instance space. A decision tree is a flow-chart-like structure, where each internal (or non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The highest node in the tree is the root node. Figure 4 illustrates how a decision tree is used to determine whether a user is in the politics domain. The study considers a simple decision tree with four features, x_2, x_3, x_7, and x_8. The red branches of the decision tree indicate that the user is in the politics domain; this occurs if any of following three conditions are met: (1) if x_2 is larger than 100 and x_3 is larger than 50, (2) if x_2 is less than 100, x_7 is larger than 80 and x_8 is larger than 50, or (3) if x_2 is less than 100, x_7 is less than 80 and x_8 is larger than 100.

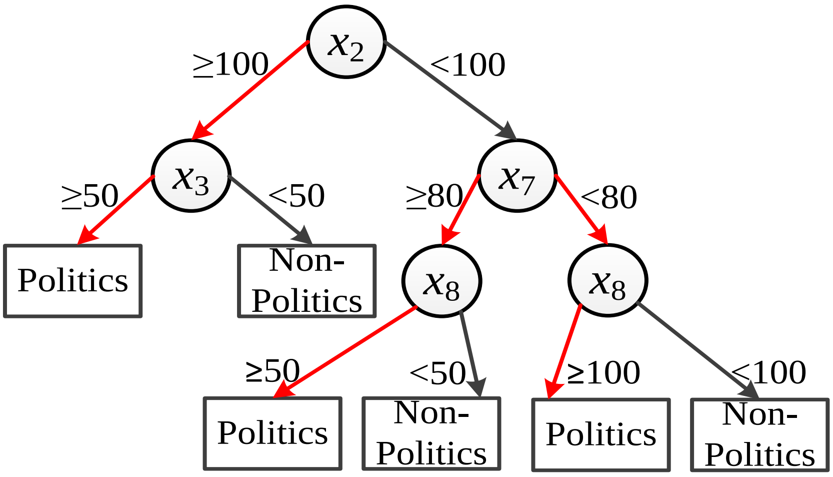


Figure 4: Example of a Decision Tree for the Politics Domain

Compared with logistic regression and SVM, decision trees are very intuitive and easy to interpret. In addition, empirical results have demonstrated that decision trees outperform SVM and logistic regression on 11 benchmark problems, in terms of ten classification metrics [73]. Three commonly-used approaches, namely top-down inducing C4.5 [64], random forest [65], and gradient boosting [66] are used to develop decision trees for determining whether a user is in the politics domain. In top-down inducing C4.5, the decision tree is constructed from the top to the bottom, based on a divide-and-conquer mechanism. The top-down inducing C4.5 trains the samples based on the splitting measures. After the selection of an appropriate split, which results in a minimum classification error, each node further subdivides the training samples into smaller subsets of samples, until the split gains satisfy the splitting measure. In a random forest, multiple trees are generated based on randomly selected subspaces of features. The trees generalise their classification in complementary ways and their combined classification attempts to improve each single tree. In gradient boosting, a base decision classifier is expanded by adding additional branches to the base of the tree. The expansion continues until no further improvement can be obtained by adding an additional branch.

5. System Evaluation

In Sections 3 and 4, a system framework is proposed to detect the domain-based interest of users/tweets by incorporating machine learning. This section evaluates the effectiveness of the proposed system framework.

5.1. Datasets Collection and Ground Truth

To evaluate the proposed system framework, a list of Australian Twitter users and their public content was collected and pre-processed as discussed in section 3.1. The tentative list of users who are potentially interested in the politics domain was selected from the following sources: (1) a list of Members of Parliament and Senators indicated on the official website of the Parliament of Australia.[9] (2) members and subscribers of three politics-based Australian Twitter lists,[10] and (3) miscellaneous sources.[11] Due to the lack of online sources indicating those users interested in politics in OSNs, the aforementioned lists are selected because it is assumed that these people are interested in the politics domain as is evident later in the paper.

Users who are assumed to have little or no interest in the politics domain were tentatively selected from two collected datasets: (1) members of various Australian Twitter lists established to discuss sports, information technology, and other non-politics domains; (2) a list of Australian users who achieved the highest trustworthiness values in all domains except "news, government and politics", extracted from an on-going project, the preliminary approach of which has been described in previous work [5]. The tentative selection criterion was established based on the user's profile description, choosing users who indicate a non-politics interest.

The collected and cleansed tweets of each user is then carefully examined to obtain an accurate understanding of the user's domain of interest, thereby establishing a truth dataset for developing and validating the proposed system framework at the user level. In this dataset, users are labelled and assigned to two categories: (1) **on-topic** users who show a particular interest in the politics domain and (2) **off-topic** users who demonstrate no or minimal interest in the politics domain. Table 3 shows a tentative list of collected users, and the actual number of users selected for the ground truth, based on an examination of all tweets.

Table 3: Ground Truth - User level

	#Collected users (tentative list)	**Ground Truth**
on-Topic	310	**227**
off-Topic	350	**283**

The collected users of the ground truth dataset indicated in Table 3 are analysed with their historical tweets to develop the prediction model. This is used to predict the likelihood of users in the politics domain.

The next phase involves conducting experiments at the user level to predict the politics classification of the new users' tweets. Therefore, another dataset is collected which contains new tweets posted by already-

[9] http://www.aph.gov.au/.

[10] https://twitter.com/latikambourke/lists/australian-journalists/subscribers; https://twitter.com/lizziepops/lists/politics/members ; https://twitter.com/smh/lists/federal-politicians.

[11] http://earleyedition.com/2009/04/22/australias-top-100-journalists-and-news-media-people-on-twitter; Wikipedia: Australian political journalists : https://en.wikipedia.org/wiki/Category:Australian_political_journalists.

examined users. The new tweets are examined and a subset of the tweets is selected to construct the ground truth for conducting experiments at the tweet level. The selection was based on four criteria; (1) tweets indicating a **politics** domain, and posted by **politics** users; (2) tweets indicating a **politics** domain, and posted by **non-politics** users; (3) tweets indicating a **non-politics** domain, and posted by **politics** users; (4) tweets indicating a **non-politics** domain, and posted by **non-politics** users. These four criteria are chosen to support the prediction model which will be constructed at the tweet level. Table 4 shows the total number of tweets collected based on the four selection criteria.

Table 4: Ground Truth - Tweets Level

	Politics users	**Non-politics users**
Politics tweets (on-topic)	150	125
Non-politics tweets (off-topic)	105	100

The proposed system framework is implemented in the Turi Graphlab Create™ which is used for these experiments using the Python programming environment. Turi Graphlab Create is used as it is scalable and can therefore accommodate relatively huge datasets. The proposed system framework is used to conduct the experiment at the user level with the nine features (x_1 to x_{14}) illustrated in Table 1 and the five classifiers discussed in section 3.3, logistic regression (LR), support vector machine (SVM), top-down inducing based decision tree (TD-DT), random forest-based decision tree (RF-DT) and gradient-boosting-based decision tree (GB-DT). Turi Graphlab Create is also used to conduct experiments at the tweet level with the features listed in Table 2. 10-fold cross validation is used on the datasets to evaluate the generalisation capability of the proposed system framework which is embedded with the five classifiers.

At the user level analysis, and as depicted in Figure 5, the proposed system framework can be used to determine (classify) whether or not a user is interested in politics. The circled ones are classified as the politics-interested users and the non-circled ones are the users who are not interested in politics. Four scenarios are illustrated by the classification as:

1. True-positives (TP), which indicate the number of actual politics users that are classified correctly as politics users;
2. False-positives (FR) which indicate the number of non-politics users that are classified incorrectly as politics users;
3. False-negative (FN) which indicate the actual politics users that are classified incorrectly as non-politics users; and
4. True-negative (TN) which indicate the non-politics users that are classified correctly as non-politics users.

These four scenarios can also be shown in the confusion matrix (Table 5) which depicts the performance of the prediction. The model illustrated in Figure 5 is also applicable to the tweets classification which is the second analysis phase in the proposed approach.

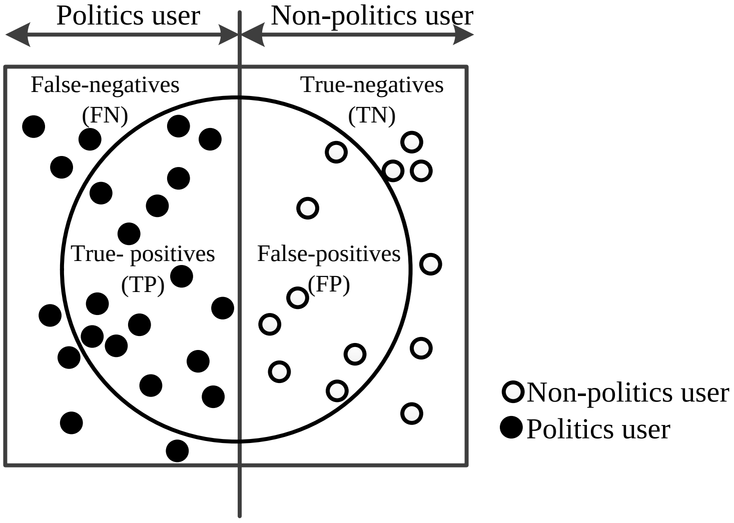


Figure 5: Classification of Politics/Non-politics Users

Table 5: Confusion Matrix

		Prediction	
		on-topic	**off-topic**
True	**on-topic**	TP	FN
	off-topic	FP	TN

In Graphlab Create ™, the confusion matrix is often a table used to provide further details on the true and false predictions. This table comprises three columns:

(1) Target_label: the classification label of the ground truth. It represents the **on-topic** and **off-topic** label in this study;
(2) Predicted_label: the classifier prediction label; and
(3) Count: the number of times the predicted_label matches the target_label.

The evaluation has been performed by using the following metrics to evaluate the classification performance in predicting whether or not the user/tweet is in the politics domain.

Accuracy indicates the correctness of the incorporated classifier in making the correct prediction. This is essentially the ratio between the correct predictions (i.e. $\mathrm{TP}+\mathrm{TN}$) and the total predictions ($\mathrm{FN}+\mathrm{TP}+\mathrm{FP}+\mathrm{TN}$). This is computed as:

$$Accuracy = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{FN}+\mathrm{TP}+\mathrm{FP}+\mathrm{TN}} \tag{5}$$

Log-loss (logarithmic loss) is a fine-grained classification evaluation metric. This value is computed by the negative of the accumulation of the log probability of each sample, normalised by the number of samples:

$$Log\text{-}Loss = -\frac{1}{n}\sum_{i\epsilon 1,..N}(y_i \log(P_i) + (1-y_i)\log(1-P_i)) \tag{6}$$

Where $\mathrm{y_i}$ is the i-th target value, and $\mathrm{P_i}$ is the i-th predicted probability. This metric is used because the likelihood probability is addressed to predict the **on-topic** or **off-topic** likelihood of the user or tweet.

Precision, Recall and F-score are metrics commonly used to evaluate classification performance. Precision, Recall and F-score are shown in (7), (8) and (9) respectively.

$$\mathrm{Pr}\,ecision = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{7}$$

$$\mathrm{Re}\,call = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{8}$$

$$F\text{-}score = 2 \cdot \frac{\mathrm{Precision}\ \cdot\ \mathrm{Recall}}{\mathrm{Precision}\ +\ \mathrm{Recall}} \tag{9}$$

Precision indicates the ratio between the number of actual politics users/tweets that are classified correctly, and the total number of correct and incorrect classifications of politics users/tweets. Recall indicates the ratio between the number of actual politics users/tweets that are classified correctly, and the total number of actual politics users/tweets. Hence, high precision indicates that the classifier is capable of generating substantially more relevant predictions for actual politics users/tweets than the irrelevant ones. High recall indicates that the classifier is capable of generating most of the relevant predictions for actual politics users/tweets. Precision with a value of '1' indicates that every prediction is the actual politics user/tweet but it does not mean that all the actual politics users/tweets are retrieved; while a recall score with a value of '1' indicates that all predictions are actual politics users/tweets but it does not indicate the number of non-politics predictions that are retrieved. Hence, the F-score is used to provide the trade-off between precision and recall.

5.2. Domain Detection – User Level

The aforementioned features in Table 1 are analysed for each user where tweets are divided temporally into five groups to address the temporal dimension. The second and third columns in Table 6 show the feature values with respect to the on-topic samples and off-topic samples respectively, where **on-topic samples** represent the list of users interested in the politics domain and **off-topic samples** show the users who did not have an interest in the politics domain. For the on-topic samples, the i-th feature is denoted as x_i on-topic. For the off-topic samples, the i-th feature is denoted as x_i off-topic. ARD (Absolution Relative Difference) in (10) is used to indicate the relative difference between the **on-topic samples** and the **off-topic samples.**

$$\mathrm{ARD} = 100 \times \mathrm{Abs}\left(\frac{x_i^{\mathrm{on\_topic}} - x_i^{\mathrm{off\_topic}}}{x_i^{\mathrm{on\_topic}} + x_i^{\mathrm{off\_topic}}}\right) \tag{10}$$

The higher the ARD value, the higher the impact of the corresponding feature used to discriminate **on-topic** and **off-topic** users. For example, an ARD of x_8 equal to 100 indicates that x_8 is highly significant in identifying the (non-)interested users in the politics domain by examining their profile description. This evidence will be discussed later.

Table 6 Dataset Statistics – User Level

	on-topic samples	**off-topic samples**	**ARD**
Total #Users	227	283	
Total #Tweets, x_1	499,475	611,014	10.044
Total #unq_pol_entities, x_2	14,818	2,833	67.9
Total #pol_entities_pre_Q3_2015, x_3	110,128	8,770	85.248
Total #pol_entities_Q3_2015, x_4	18,492	869	91.023
Total #pol_entities_Q4_2015, x_5	14,842	522	93.205
Total #pol_entities_Q1_2016, x_6	21,562	601	94.577

Total #pol_entities_Q2_2016, x_7	39,712	1,218	94.048
Total #profile_pol_entities, x_8	**237**	**0**	**100**
Total #Verified, x_9	167	94	27.969

As depicted in Table 6, the political entities detected in features x_2 to x_8 for **on-topic** users are much greater than the entities detected for the **off-topic** users. This is because **on-topic** users have shown an extensive interest in the politics domain through their content on Twitter.

To evaluate the effectiveness of the proposed system framework embedded with the five classifiers (LR, SVM, TD-DT, RF-DT and GB-DT) 10-fold cross-validations were used. In the cross-validations, the total observations (i.e. 510 users) are randomly split into two datasets, namely the training dataset (which is 80% of the total sample) and the validation dataset (which is 20% of the total sample). Table 7 illustrates the main settings and parameters used to train each of the five classifiers in the proposed system framework.

Table 7: Classifiers Settings

Classifier	Main settings	Parameters
LR	Hyperparameters- L1 penalty	0
	Hyperparameters-L2 penalty	0.01
	Solver	Newton-Raphson
	Solver iterations	9
SVM	Solver	L-BFGS[12]
	Predefined number of iterations	10
	Hyperparameters Mis-classification penalty	1
TD-DT	Number of trees	1
	Max tree depth	6
RF-DT	Number of trees	10
	Max tree depth	6
GB-DT	Number of trees	10
	Max tree depth	6

Table 8 depicts the confusion table used to quantify the performance of each classifier. It can be seen that the LR performs better in the classification task of this study; of the 107 samples used to validate each algorithm, only 2 samples were incorrectly classified by LR. However, all other classifiers, TD-DT and RF-DT algorithms for example, wrongly classified more samples in the prediction validations. Nevertheless, the classification performance of the incorporated algorithms is acceptable. These methods can generally perform effectively in terms of this domain classification problem.

Table 8: Confusion Table

Target_ label	Predicted_ label	LR	SVM	TD-DT	RF-DT	GB-DT
on-topic	on-topic	**59**	58	41	57	48
off-topic	off-topic	**46**	46	52	45	45
on-topic	off-topic	**2**	3	2	4	3
off-topic	on-topic	**0**	0	3	1	1

Table 9 shows the evaluation performance metrics of each classifier, where the means and variances for the 10 fold cross-validations are given. The metric means that the non-bracketed values and the metric variances

[12] L-BFGS: is a Limited memory of Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization algorithm. This is a robust solver for datasets with many coefficients

are the bracketed values. It can be seen from Table 9 that LR achieves the better metric means for the five classification metrics among all of the methods where Accuracy, Precision, Recall and F1_score are "the larger-the-better" and Log_loss is "the smaller-the-better". The metric variances generated by LR are generally the smallest. Therefore, LR can yield the best and most robust classification when compared to the other four methods.

Despite the classifier's convergence on the four metrics (i.e. Accuracy, Precision, Recall, and F1-score), LR is generally better than the other four methods, particularly regarding log_loss. This indicates that the predicted likelihoods of the validation dataset using FR closely match with the assigned labels. TD-DT on the other hand is generally the poorest method when compared with the others.

Table 9: Performance Comparison of Five Classifiers to Detect User Political Interest

	Accuracy	Log_loss	Precision	Recall	F1_score
LR	0.9824 (0.0002653)	0.0406	1.0000	0.9672	0.9833
SVM	0.9784 (0.003417916)	0.5781	1.0000	0.9508	0.9748
TD-DT	0.9157 (0.033453)	0.4816	0.9318	0.9535	0.9425
GB-DT	0.9255 (0.032357)	0.1321	0.9831	0.9508	0.9667
RF-DT	0.9490 (0.009473)	0.2321	0.9828	0.9344	0.9580

Note: Accuracy, Precision, Recall and F1_score are "the larger-the-better". Log_loss is "the smaller-the-better".

Table 10 shows the highest estimated coefficient values calculated for each feature using LR. It shows that "profile_pol_entities, x_8" is the highest estimated coefficient. This is consistent the results illustrated in Table 8 where x_8 has the highest impact when compared with the other features. This is due to the importance of this feature in distinguishing the user's interest in the politics domain. In particular, users whose profile descriptions include politics-related entities such as a parliament member, political journalist, etc., are likely to suggest the politics domain in their tweets.

Table 10: Highest Positive Coefficients- User Level

Feature	Value
profile_pol_entities, x_8	8.601
verified, x_9	2.162
unq_pol_entities, x_2	0.144
pol_entities_Q4_2015, x_5	0.02

In addition, the t-test [74] was used to evaluate the significance of the hypothesis that the accuracy means obtained by the best method LR are higher than those obtained by the other methods (SVM, TD-DT, RF-DT and GB-DT). The t-values between LR and the other methods are shown in Table 11. Based on the t-distribution table, if the t-value is higher than 1.699, the significance is 95% confidence, which means that the accuracies obtained by the LR are higher than those obtained by the other methods with a 95% confidence level. The t-value can be determined by:

$$t\text{-value} = \frac{\mu_2 - \mu_1}{\sqrt{\left(\sigma_1^2 + \sigma_2^2\right)/N}}, \qquad (11)$$

where μ_1 is the mean accuracy obtained by the LR and μ_2 is for the other methods, σ_1^2 is the accuracy variance obtained by the LR and σ_2^2 is for the other compared methods. N_1 is equal to 10 as this is a 10-fold cross validation. In general, the results indicate that there is no significant difference between LR and the other tested methods, although better accuracies can generally be obtained by the LR.

Table 11: T-values Between LR and the Other Tested Methods

	LR and SVM	**LR and TD-DR**	**LR and RF-DR**	**LR and GB-DR**
T-values	0.20842	1.1487	1.0703	0.99622

Therefore, the decision trees obtained by TD-DT can be interpreted and explained to executives of the user domain, as the accuracies obtained by TD-DT are similar to those obtained by LR. Figure 6 illustrates the resultant decision tree of the TD-DT classifier generated by Graphlab Create. It is evident that a feature is selected as a root node in TD-DT if this feature achieves the lowest classification error among the other features by applying the same dataset. The values associated with each leaf node in Figure 6 represent the “margins” which are a form of prediction showing the distance of samples from the decision boundary. The greater the distance, the higher the confidence in the classifier’s prediction that the user is interested in the politics domain. These margins can be converted to likelihood values (predictions) by applying the sigmoid function to the margins.

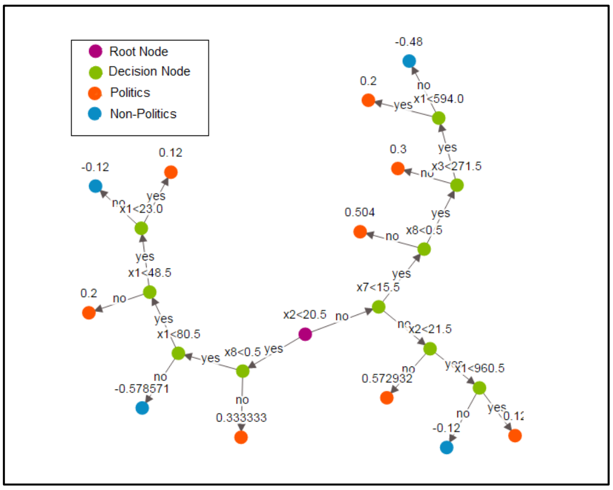


Figure 6: Decision Tree created by TD-DT

As depicted in Figure 6, feature x_2 (fuchsia node) has been selected by the classifier as the root node at which to split the tree. To evaluate this tree, we start with the root node and follow the correct path through the decision nodes (green nodes) until we approach the leaf node (red/blue node) which indicates whether the user is interested in politics (red node) or not (blue node). For example, consider the two observations provided in Table 12; one indicates a user interested in politics (**@SenatorWacka**) and one does not show an interest in this domain (**@LabGallerie**). This Table shows the margins and the associated predictions for each sample. To apply the tree represented generated for **@SenatorWacka,** we start with the root node **“**$x_2 < 20.5$**”** is **no** because $\mathbf{x_2}$ SenatorWacka $= 42$, "$x_7 < 15.5$" is **no** because $\mathbf{x_7}$ SenatorWacka $= 20$, **“**$x_2 < 20.5$” is **no** because $\mathbf{x_2}$ SenatorWacka $= 42$. This leads to a red leaf with the value of “0.572932” which represents a user who is interested in the politics domain. The application of the same tree on @**LabGallerie** leads to a blue leaf with a value of -0.578571, which indicates a non-politics user. This is evident in both users, whose classification labels match with the resulting predictions.

Table 12: Margins and Predictions of Two Samples

Twitter_ID	**x1**	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**	**Label (1:Politics, 0:non-Politics**	**Margins**	**Predictions**
@SenatorWacka	880	42	468	12	3	1	20	1	1	1	0.572932	0.639440
@LabGallerie	1498	4	19	1	2	7	3	0	0	0	-0.578571	0.359261

5.2.1. A Comparison with LDA and SLA

As discussed, LDA and SLA are statistically well-known models used for several topic modelling applications. In this section, we describe an experiment used to benchmark the applicability of our model at the user level against these two models, to identify a user's main topic of interest. Gensim's python implementation [75] of LDA and SLA is used. The collected historical tweets of two politicians' accounts (i.e. @sarahinthesen8 and @stephenjonesALP) have been fed to the three models: LDA, SLA and our model incorporating a Politics Knowledge Inference. The experimental settings for LDA and SLA are set to one topic modelling, and the extracted terms indicate the 25 most contributed terms to this topic. In our approach, we extract the top 25 frequently annotated entities from the users tweets. Table 13 and Table 14 show the top 25 terms/entities extracted using the three approaches for @sarahinthesen8 and @stephenjonesALP respectively.

Table 13: Top entities/terms Extracted using LDA, SLA and our Approach For @sarahinthesen8

LDA	**LSA**	**Politics Knowledge Inference**	
		Entity	SubType
refuge	refuge	Government of Australia	Organization
young	young	Australian Greens	Political Party
sarah	sarah	Member of Parliament	Politician
hanson	hanson	Elections	Event
nauru	nauru	Australian Labor Party	Political Party
children	children	Parliament	Organization
detent	detent	Liberal Party of Australia	Political Party
govt	govt	Malcolm Turnbull	Politician
australia	australia	Peter Dutton	Politician
green	green	Tony Abbott	Politician
abbott	abbott	Politics	Ontology
today	today	Sarah Hanson-Young	Politician
asylum	asylum	Electorate	Voter
manu	manu	Council	Organization
aust	aust	Politician	Person
people	people	inequality	Political_Slogan
senate	senate	Coalition	Political_Slogan
seeker	seeker	Joe Hockey	Politician
abuse	abuse	George Brandis	Politician
news	news	Liberal National Party of	Political Party
minister	time	Queensland	Political_Slogan
time	minister	welfare	Politician
dutton	dutton	Barnaby Joyce	Politician
turnbull	turnbull	Nick McKim	Politician
australian	australian	Kristina Keneally	Politician
		Simon Birmingham	

Table 14: Top Entities/terms Extracted Using LDA, SLA and our Approach For @ stephenjonesALP

LDA	**LSA**	**Politics Knowledge Inference**	
		Entity	SubType
illawarra	illawarra	Member of parliament	Politician
qt	qt	Elections	Event
today	today	Parliament	Organisation
great	great	Australian Labor Party	Political Party
mp	mp	Government of Australia	Organisation
stephen	stephen	Liberal Party of Australia	Political Party
good	good	Coalition	Political Slogan
post	post	Tony Abbott	Politician
school	school	Council	Organisation
abbott	abbott	Anthony Albanese	Politician
jone	jone	Politics	Ontology
photo	photo	Julia Gillard	Politician
auspol	auspol	Electorate	Voter
parliament	day	Greg Combet	Politician
day	jame	Sharon Bird	Politician
jame	parliament	Joe Hockey	Politician
big	big	Mark Butler	Politician
support	support	Malcolm Turnbull	Politician
nbn	nbn	Kate Ellis	Politician
house	house	Barack Obama	Politician
facebook	facebook	Joel Fitzgibbon	Politician
time	time	Jamie Briggs	Politician
fb	fb	Australian Greens	Political Party
australia	australia	Steven Ciobo	Politician
purser	purser	Greg Hunt	Politician

The list of the top contributed terms identified using 1-topic modelling for each user incorporating LDA and SLA illustrates the inadequacy of these approaches in identifying a high-level topic. On the other hand, with the top 25 entities annotated for both users using our approach, the high-level topic (i.e. politics) is highly noticeable. In our proposed system framework, each entity is linked with a specific class in the ontology. The knowledge obtained for each entity can be enriched to facilitate the overall semantic interlinking which leads to a better understanding of the domain of knowledge. Interlinking and enrichment are not applicable to LDA and SLA. Furthermore, all the top entities annotated using our proposed system framework indicate politics entities, although some of the most frequently occurring terms extracted using LDA and SLA are politics entities. In a nutshell, the outcome of this experiment shows the applicability and effectiveness of our proposed framework.

5.3. Domain Detection – Tweet Level

Table 15 shows the statistics of the dataset used for this experiment at the tweet level. The new tweets are collected from the list of users indicated in the previous section. These tweets represent the new tweets posted after quarter 2, 2016. Hence, the tweet-level experiments are conducted on the set of tweets which have not been included in the users historical tweets as discussed in the previous section.

The features shown in Table 2 are formulated for each tweet. **On-topic samples** in Table 1 represent the list of tweets labelled as politics tweets. **Off-topic samples** show the list of tweets labelled as non-political tweets. The ARD value is calculated for each feature. Table 15 shows the statistics calculated for the ground truth which is used to classify tweets according to a particular domain. It is evident that the calculated ARD for the two mean values of x_5 is the smallest value due to the noticeable convergence of $\boldsymbol{x_5}$ in both categories. This is because a user who has been classified as belonging to the politics domain does not necessarily post all of his/her future tweets in this domain. Likewise, a user who has been classified as a non-politically interested user may show an interest in this domain in future tweets. Nevertheless, x_5 is most likely to distinguish the ambiguous political entities annotated from the textual content of a tweet, thereby helping to accurately ascertain the tweet's domain. This will be discussed further in this section.

Table 15 Dataset Statistics – Tweet Level

	on-topic samples	**off-topic samples**	**ARD**
Total #Tweets	255	225	
Total #political_entities(x_1)	880	71	85.068
Total #words_count(x_2)	3,762	2,391	22.282
Average political_perc, (x_3)	0.249	0.033	76.596
Total #pol_entities_recent_quarter(x_4)	65,049	37,248	27.177
Average user_pol_likelihood, x_5	**0.638**	**0.563**	**6.245**

Due to the ability of the LR to detect the domain of interest at the user level, LR is further used to classify tweets in this phase with the same set of parameters listed in Table 7. To validate the efficiency of the proposed approach, 10-fold cross validation is performed where the 480 samples are randomly split into a training dataset (80%) and a validation dataset (20%). To further validate the effectiveness of the proposed approach, another experiment was conducted which excluded user_pol_likelihood, x_5 from the feature sets. This is to measure the significance of this feature to predict a tweets domain. Table 16 shows the confusion table used to quantify the performance of the LR classifier in each experiment, where Exp.1 refers to the first experiment conducted incorporating all features listed in Table 2. Exp.2 refers to the second experiment conducted on the same dataset excluding x_5.

Table 16: Confusion Table –Tweet Level

Target label	Predicted label	Exp.1	Exp.2
on-topic	on-topic	**58**	**56**
off-topic	off-topic	**42**	**39**
on-topic	off-topic	**0**	**3**
off-topic	on-topic	**0**	**2**

As depicted in the confusion matrix in Table 16, Exp.1 achieved better results than Exp. 2; incorporating all features including the past user's political prediction (x_5) leads to zero incorrect classifications. However, eliminating x_5 from the list of features results in five out of 100 incorrect classifications. This is confirmed by the comparison of the performance results of the two experiments illustrated in Table 17.

Table 17: Performance Comparison of Two Experiments – Tweet Level

	Accuracy	Log_loss	Precision	Recall	F1_score
Exp.1	1	0.01	1	1	1
Exp.2	0.95	0.072	0.949	1	0.957

Despite the convergence in each metric listed in Table 17, the predicted likelihoods of the validation dataset incorporating all features closely match the assigned labels.

Table 18: Highest Positive Coefficients- Tweet Level

Exp.1		**Exp.2**	
Feature	**Value**	**Feature**	**Value**
political_perc, x_3	24.86	political_perc, x_3	25.126
user_pol_likelihood, x_5	12.095	political_entities, , x_1	1.823
political_entities, , x_1	5.409	words_count, x_2	0.825
words_count, x_2	0.623	pol_entities_recent_quarter, x_4	0.009

Table 18 shows the highest estimated coefficient values calculated for each tweets feature in each of the conducted experiments. It is evident that political_perc (x_3**)** obtained the highest coefficient value in Exp.1 and Exp.2. This is due to the impact of the tweets political weight, indicating the tweets domain. This feature is supported by considering the number of political entities (x_1) and the total number of words in the tweet, x_2. User_pol_likelihood (x_5**)** obtained the second highest estimated coefficient after conducting Exp.2. This is due to the significance of incorporating former knowledge about the user's political interest in the process of predicting the domain of their future tweets.

Table 19 elucidates further the significance of incorporating x_5. Table 19 shows two real tweets of the ground truth dataset; one is labelled "politics" and the other is labelled "non-politics", posted by two Twitter users (i.e. @tamaleaver, non-politics user, and @peterjblack, politics user). The list of features included in Table 2 is calculated for each tweet. As depicted in Table 19, features x_1, x_2, and x_3 obtained the same values for each tweet. This exacerbates the process of obtaining the correct domain by considering only the number of political entities and counting the words in each tweet. It is evident that features x_4 and x_5 are important for identifying the tweets domain due to their significance for the classification task.

Assumption: It is argued that the annotated political entity of a tweet posted by a user who has already been predicted to be interested in the politics domain, and who has included a relatively large number of political entities annotated in their tweets, is likely to indicate an actual political concept. Likewise, a user who has not shown an interest in politics in the past, is not likely to indicate politics-related content in future tweets. This helps to eliminate the ambiguity for those entities which might have dissimilar meanings in several contexts. Moreover, this is applicable to all domains of knowledge.

Therefore, despite obtaining one political entity (Labour/Labor) for each tweet in Table 19, these tweets convey two different messages which are unrelated in terms of context.

Table 19: Features Extracted for Two Tweets Posted by Two Users (Politics and Non-politics)

Twitterer	Tweet	x_1	x_2	x_3	x_4	x_5	Label
@tamaleaver	"Researching microcelebrity: Methods, access and **labour**, Jonathan Mavroudis".	1	7	1/7	4	**0.019**	**non-politics**
@peterjblack	"**Labor** could support 'self-executing' same-sex marriage plebiscite".	1	7	1/7	927	**0.98**	**politics**

6. Discussion

The rapid growth of enterprise needs in conjunction with an increase in the volume of modern data repositories, and the nature of the data that can be stored, have made traditional statistical methods insufficient to meet all data analysis requirements. This has necessitated the development of advanced data analytics to extract useful knowledge from such vast data volumes.

In light of the general perception of advanced data analytics, companies incorporate advanced social data analytics to build effective marketing strategies, leveraging the interactive quality of OSNs. Thus, to create the required interaction with their customers, companies use many modern forms of communication to attract customers and visitors to their online social platforms. Consequently, it is necessary for companies to analyse the customers' social content and classify the customers into appropriate categories, in order to deliver the right message to the right category. Segmentation is the first step towards effective marketing to classify customers according to their interests, needs, geographical locations, purchasing habits, life style, financial status and level of brand interaction. If companies succeed in building effective clusters of customers, and thus determine the basic criteria for each cluster in making their buying decisions, companies will be able to take clear actions to implement them. For example, companies can identify the most optimal products/services captured for each segment of customers. This fine-grained analysis leads to maximisation of a customer's satisfaction with a company through designing and manufacturing several, segments-oriented products.

Unstructured data is produced exponentially. This necessitates further efforts to absorb such datasets in order to understand its context. Textual content (a.k.a natural language text) is considered the largest amongst all sources of information [76]. The wealth of free-form textual social data has attracted the attention of researchers in an attempt to disclose hidden knowledge regarding textual content. This problem has been untangled through the emergence of text mining technology, an extension of data mining, which aims to detect rules, patterns and trends from textual data such as tweets, HTML webpages, instant messages and emails [77], [78]. Natural language text is very ambiguous, and this is evident particularly when it comes to the continuous occurrences of the named entities. Hence, indicating and inferring key entities such as a person's name and profession, location of cities and countries, products, companies, specialised terms, etc. from the text can significantly enhance several business processes and techniques such as knowledge base population, topics distillation, keyword searches, and information integration [46]. Therefore, there is a need for an approach to derive knowledge from social big data. This approach enhances the overall comprehension of the processed textual datasets, and delivers knowledge in the form of unambiguous results through providing metadata which aids in accurately interpreting and understanding related data.

Twitter is designed to track public figures and news, and provide a platform for users to follow their friends and associates. The "maximum 140 characters" quality has made Twitter particularly important and widespread; however, this feature constricts the size of published content for each user which is needed to conduct an adequate analysis. This paper presents an effective approach to address two main related problems: (1) the sporadic quality of tweets which entangles bag-of-words statistical techniques and (2) the problematic nature of obtaining a factual understanding of the contextual meaning of users social content. The most well-known approaches for inferring a users topics of interest are LDA-related techniques. Despite their popularity, they fail to address several key issues, namely; (1) the number of topics to be discovered is set as a parameter in the experiment, thus it is hard to identify the optimal number which represents the adequate number of topics extracted from the document [79], (2) the topics extracted by these models do not contemplate the temporal aspects. A document's corpus evolves through time and subsequently so does its themes [80], (3) these models are considered as monolingual topic models, hence they do not differentiate idioms of the same language [81]; and (4) these models are unable to infer topics from short text such as tweets [50].

Incorporating ontologies, semantic web and linked data enriches textual data and the extraction of knowledge, thereby linking the textual data with a particular user domain. This approach is better able to address the temporal factor, and at harnessing advanced machine learning techniques to perform domain-based classification. For example, by recurrence to the benchmark comparison conducted in section 5.2.1, if a user is interested in finding Twitter users who discuss "Australian Political Parties", through implementing an LDA technique, this user could find "Sarah Hanson-Young@sarahinthesen8" and "Stephen Jones@stephenjonesALP" amongst the search results. This is possible only if "Sarah" and "Stephen" indicated "Australian Political Parties" explicitly in their content alongside tweets pointing out their declared political party (i.e. "Australian Greens" and "Australian Labor Party" respectively).

Moreover, LDA retrieves search results that neglect the temporal dimension; users knowledge evolves over time and their interest might be diverted elsewhere depending on their experience, work, study, or other factors. Leveraging domain ontology and semantic web tools facilitates the building of conceptual hierarchies and the process of populating the domain ontology with instances extracted from user tweets. Therefore, "Australian Greens" and "Australian Labor Party" are annotated in the knowledge base as a subset of "Australian Political Parties". This hierarchy extends the knowledge obtained from social data by adding semantics to its textual content. Unlike LDA and other unsupervised statistical approaches, we incorporate supervised machine learning techniques to perform the classification task for the already semantically-enriched temporally-segmented textual content. This, as indicated in the conducted experiments, validates the applicability of veritably classifying users based on their domains of interest which has an intrinsic impact on several applications. For example, adding a user-domain dimension when calculating trust in social media helps to provide a fine-grained trust analysis. In this context, the notion of domain-based trust for the data extracted from the unstructured content (such as social media data) is significant. This is achieved by calculating trustworthiness values which correspond to a particular user in a particular domain. This issue will be addressed in our future work as indicated in the following section.

7. Conclusion and Future Work

This paper presents the preliminary stages of a research project intended to provide a methodology for social business intelligence incorporating the notion of trust, semantic web analysis and machine learning applications [82]. The importance of trust in the context of OSNs is indicated by the numerous resources available for market analysis, listening to the Voice of Customer (VoC), and by sentiment analysis – all of which are major resources that feed business intelligence applications.

The semantic extraction of the textual content of OSNs represents a further step towards understanding the factual context of a user's content. One of the major challenges of OSN analysis is to better understand the domain of knowledge in which the user is interested. This problem is exacerbated by: (1) inconsistent user behaviour (a user's interest can evolve and change over time), and (2) the brevity and economy of tweets' content. Therefore, this paper presents a consolidated approach to addressing this problem by means of semantic analysis and the application of machine learning.

The proposed framework comprises two analysis phases: (1) the time-aware semantic analysis of users historical content incorporating five well-known machine learning classifiers. This classifies users into two main categories; politics-interested and non-politics-interested. (2) The prediction likelihood values obtained in the first phase have been harnessed to predict the domain of the users future tweets. The experiments conducted to evaluate this framework validate the applicability and effectiveness of better understanding the domain of Twitter content at the user and tweet levels. This is evident through the notable performance of the machine learning experiments conducted at both the user and tweet levels.

Through experiments conducted using the Twitter platform as one of the dominant OSNs, this work provides the essential groundwork for a better understanding of user interest in several domains of knowledge. This is achieved by incorporating domain-based ontologies and semantic web analysis to gain a better familiarity with user interests. This facilitates the process of measuring user credibility in each domain of knowledge. The following are the possible enhancements and research directions to be addressed in our anticipated future work:

- Beside politics, a domain-based analysis of several domains of knowledge will be conducted to gain a more comprehensive insight into each domain. This is to facilitate the development of several domain-based ontologies leveraged by semantic web technologies and Linked Open Data.
- A domain-based trustworthiness approach will be developed based on the factual understanding of the users main interests.
- Machine learning will be harnessed to achieve the abovementioned research objectives through multi-classification applications, to predict the likelihood of user interest in several domains of knowledge.
- Semantic analysis and trust will be integrated for social business intelligence applications, which will enhance the quality and accuracy of data stored in data warehouses. This will dramatically affect the decision-making process as well as the quality of extracted reports.

REFERENCES

[1] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology,* vol. 27, pp. 415-444, 2001.

[2] L. Rainie and B. Wellman, *Networked: The new social operating system*: Mit Press, 2012.

[3] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, Geneva, Switzerland, 2010.

[4] W. Sherchan, S. Nepal, and C. Paris, "A Survey of Trust in Social Networks," *ACM Comput. Surv.,* vol. 45, 2013.

[5] B. Abu-Salih, P. Wongthongtham, and D. Zhu, "A Preliminary Approach to Domain-Based Evaluation of Users' Trustworthiness in Online Social Networks," in *2015 IEEE International Congress on Big Data*, 2015, pp. 460-466.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research,* vol. 3, pp. 993-1022, 2003.

[7] M. Michelson and S. A. Macskassy, "Discovering users' topics of interest on twitter: a first look," in *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, 2010, pp. 73-80.

[8] V. Bolotaeva and T. Cata, "Marketing opportunities with social networks," *Journal of Internet Social Networking and Virtual Communities,* vol. 2010, pp. 1-8, 2010.

[9] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons,* vol. 53, pp. 59-68, 1// 2010.

[10] P. A. Tess, "The role of social media in higher education classes (real and virtual) – A literature review," *Computers in Human Behavior,* vol. 29, pp. A60-A68, 9// 2013.

[11] M. Salathé, D. Vu, S. Khandelwal, and D. Hunter, "The dynamics of health behavior sentiments on a large online social network," *EPJ Data Science,* vol. 2, 2013.

[12] C. De Maio, G. Fenza, M. Gallo, V. Loia, and M. Parente, "Time-aware adaptive tweets ranking through deep learning," *Future Generation Computer Systems,* 2017/07/25/ 2017.

[13] C. De Maio, G. Fenza, V. Loia, and F. Orciuoli, "Unfolding social content evolution along time and semantics," *Future Generation Computer Systems,* vol. 66, pp. 146-159, 2017/01/01/ 2017.

[14] V. Scuotto, M. Del Giudice, and K. Obi Omeihe, "SMEs and Mass Collaborative Knowledge Management: Toward Understanding the Role of Social Media Networks," *Information Systems Management,* vol. 34, pp. 280-290, 2017/07/03 2017.

[15] V. Scuotto, M. Del Giudice, M. R. d. Peruta, and S. Tarba, "The performance implications of leveraging internal innovation through social media networks: An empirical verification of the smart fashion industry," *Technological Forecasting and Social Change,* vol. 120, pp. 184-194, 2017/07/01/ 2017.

[16] A. Ardichvili, M. Maurer, W. Li, T. Wentling, and R. Stuedemann, "Cultural influences on knowledge sharing through online communities of practice," *Journal of knowledge management,* vol. 10, pp. 94-107, 2006.

[17] V. Scuotto, M. Del Giudice, and E. G. Carayannis, "The effect of social networking sites and absorptive capacity on SMES'innovation performance," *The Journal of Technology Transfer,* vol. 42, pp. 409-424, 2017.

[18] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american,* vol. 284, pp. 28-37, 2001.

[19] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?," *International journal of human-computer studies,* vol. 43, pp. 907-928, 1995.

[20] D. De Nart, D. Degl'Innocenti, M. Basaldella, M. Agosti, and C. Tasso, "A Content-Based Approach to Social Network Analysis: A Case Study on Research Communities," in *Digital Libraries on the Move: 11th Italian Research Conference on Digital Libraries, IRCDL 2015, Bolzano, Italy, January 29-30, 2015, Revised Selected Papers*, D. Calvanese, D. De Nart, and C. Tasso, Eds., ed Cham: Springer International Publishing, 2016, pp. 142-154.

[21] A. Chianese, F. Marulli, and F. Piccialli, "Cultural Heritage and Social Pulse: A Semantic Approach for CH Sensitivity Discovery in Social Media Data," in *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, 2016.

[22] I. S. M. Iwanaga, N. The-Minh, T. Kawamura, H. Nakagawa, Y. Tahara, and A. Ohsuga, "Building an earthquake evacuation ontology from twitter," in *2011 IEEE International Conference on Granular Computing (GrC)*, 2011, pp. 306-311.

[23] L. Ghahremanlou, W. Sherchan, and J. A. Thom, "Geotagging Twitter Messages in Crisis Management," *The Computer Journal,* p. bxu034, 2014.

[24] A. Kumar and A. Joshi, "Ontology Driven Sentiment Analysis on Social Web for Government Intelligence," presented at the Proceedings of the Special Collection on eGovernment Innovations in India, New Delhi AA, India, 2017.

[25] R. Kaushik, S. Apoorva Chandra, D. Mallya, J. N. V. K. Chaitanya, and S. S. Kamath, "Sociopedia: An Interactive System for Event Detection and Trend Analysis for Twitter Data," in *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics: ICACNI 2015, Volume 2,* A. Nagar, D. P. Mohapatra, and N. Chaki, Eds., ed New Delhi: Springer India, 2016, pp. 63-70.

[26] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of Twitter," *Information Processing & Management,* vol. 52, pp. 5-19, 2016.

[27] M. Sendi, M. N. Omri, and M. Abed, "Possibilistic interest discovery from uncertain information in social networks," *Intelligent Data Analysis,* vol. 21, 2017.

[28] Y. Mehta and S. Buch, "Semantic proximity with linked open data: A concept for social media analytics," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 2016, pp. 337-341.

[29] S. Romero and K. Becker, "Improving the classification of events in tweets using semantic enrichment," presented at the Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, 2017.

[30] H. Saif, T. Dickinson, L. Kastler, M. Fernandez, and H. Alani, "A Semantic Graph-Based Approach for Radicalisation Detection on Social Media," in *The Semantic Web: 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 – June 1, 2017, Proceedings, Part I*, E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler, and O. Hartig, Eds., ed Cham: Springer International Publishing, 2017, pp. 571-587.

[31] G. Anthes, "Topic models vs. unstructured data," *Communications of the ACM,* vol. 53, pp. 16-18, 2010.

[32] C. Wang, B. Thiesson, C. Meek, and D. Blei, "Markov topic models," in *Artificial Intelligence and Statistics*, 2009, pp. 583-590.

[33] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research,* vol. 3, pp. 993-1022, 2003.

[34] A. Karami, A. Gangopadhyay, B. Zhou, and H. Kharrazi, "Fuzzy Approach Topic Discovery in Health and Medical Corpora," *International Journal of Fuzzy Systems,* pp. 1-12, 2017.

[35] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50-57.

[36] Y. Chen, B. Yu, X. Zhang, and Y. Yu, "Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 2016, pp. 1-5.

[37] L. G. Nichols, "A topic model approach to measuring interdisciplinarity at the National Science Foundation," *Scientometrics,* vol. 100, pp. 741-754, 2014.

[38] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 261-270.

[39] S. Asharaf and Z. Alessandro, "Generating and visualizing topic hierarchies from microblogs: An iterative latent dirichlet allocation approach," in *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, 2015, pp. 824-828.

[40] D. Quercia, H. Askham, and J. Crowcroft, "TweetLDA: supervised topic classification and link prediction in Twitter," presented at the the 4th Annual ACM Web Science Conference, Evanston, Illinois, 2012.

[41] A. Onan, S. Korukoglu, and H. Bulut, "LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis," *Int. J. Comput. Linguistics Appl.*, vol. 7, pp. 101-119, 2016.

[42] C. Vicient and A. Moreno, "Unsupervised topic discovery in micro-blogging networks," *Expert Systems with Applications,* vol. 42, pp. 6472-6485, 2015.

[43] M. H. Alam, W.-J. Ryu, and S. Lee, "Hashtag-based topic evolution in social media," *World Wide Web,* pp. 1-23, 2017.

[44] J. Ito, J. Song, H. Toda, Y. Koike, and S. Oyama, "Assessment of tweet credibility with LDA features," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 953-958.

[45] C. Xiao, Y. Zhang, X. Zeng, and Y. Wu, "Predicting User Influence in Social Media," *JNW,* vol. 8, pp. 2649-2655, 2013.

[46] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," *IEEE Transactions on Knowledge and Data Engineering,* vol. 27, pp. 443-460, 2015.

[47] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychological Review,* vol. 110, pp. 145-172, 2003.

[48] P. Schonhofen, "Identifying Document Topics Using the Wikipedia Category Network," presented at the Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 2006.

[49] M. M. Hassan, F. Karray, and M. S. Kamel, "Automatic Document Topic Identification using Wikipedia Hierarchical Ontology," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, 2012, pp. 237-242.

[50] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic Modeling for Short Texts with Auxiliary Word Embeddings," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 165-174.

[51] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, *et al.*, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Global Institute,* 2011.

[52] M. Chen, S. Mao, Y. Zhang, and V. M. Leung, "Open Issues and Outlook," in *Big Data*, ed: Springer International Publishing, 2014, pp. 81-89.

[53] D. Lavbič, S. Žitnik, L. Šubelj, A. Kumer, A. Zrnec, and M. Bajec, "Traversal and relations discovery among business entities and people using semantic web technologies and trust management," in *Databases and Information Systems VII: Selected Papers from the Tenth International Baltic Conference, DB&IS 2012,* 2013, p. 164.

[54] J. Herzig, Y. Mass, and H. Roitman, "An author-reader influence model for detecting topic-based influencers in social media," in *Proceedings of the 25th ACM conference on Hypertext and social media,* 2014, pp. 46-55.

[55] M. Smith, C. Szongott, B. Henne, and G. v. Voigt, "Big data privacy issues in public social media," in *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, 2012, pp. 1-6.

[56] D. Sayce. (2016). *10 Billions Tweets… number of tweets per day*. Available: http://www.dsayce.com/social-media/10-billions-tweets/

[57] M. A. Shapiro and L. Hemphill, "Politicians and the Policy Agenda: Does Use of Twitter by the US Congress Direct New York Times Content?," *Policy & Internet,* vol. 9, pp. 109-132, 2017.

[58] S. Van Kessel and R. Castelein, "Shifting the blame. Populist politicians' use of Twitter as a tool of opposition," 2016.

[59] Y. Halberstam and B. Knight, "Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter," *Journal of Public Economics,* vol. 143, pp. 73-88, 2016.

[60] M. Duggan. (2016, 15/09/2017). *The Political Environment on Social Media*. Available: http://www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/

[61] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing big data issues in scientific data infrastructure," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, 2013, pp. 48-55.

[62] BBC. (20014, 21/09/2016). *BBC Politics Ontology*. Available: http://www.bbc.co.uk/ontologies/politics

[63] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression* vol. 398: John Wiley & Sons, 2013.

[64] J. R. Quinlan, "C4. 5: Programming for machine learning," *Morgan Kauffmann,* p. 38, 1993.

[65] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, 1995, pp. 278-282.

[66] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics,* pp. 1189-1232, 2001.

[67] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144-152.

[68] M. Al-Tahrawi, "Arabic text categorization using logistic regression," *International Journal of Intelligent Systems and Applications,* vol. 7, pp. 71-78, 2015.

[69] S.-J. Yen, Y.-S. Lee, J.-C. Ying, and Y.-C. Wu, "A logistic regression-based smoothing method for Chinese text categorization," *Expert Systems with Applications,* vol. 38, pp. 11581-11590, 2011.

[70] N. M. Sharef, T. Martin, K. A. Kasmiran, A. Mustapha, M. N. Sulaiman, and M. A. Azmi-Murad, "A comparative study of evolving fuzzy grammar and machine learning techniques for text categorization," *Soft Computing,* vol. 19, pp. 1701-1714, June 01 2015.

[71] B. Altınel, M. Can Ganiz, and B. Diri, "A corpus-based semantic kernel for text classification by using meaning values of terms," *Engineering Applications of Artificial Intelligence,* vol. 43, pp. 54-66, 2015/08/01/ 2015.

[72] L. Dong, N. Feng, P. Quan, G. Kong, X. Chen, and Q. Zhang, "Optimal kernel choice for domain adaption learning," *Engineering Applications of Artificial Intelligence,* vol. 51, pp. 163-170, 2016.

[73] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 161-168.

[74] G. E. Box, J. S. Hunter, and W. G. Hunter, *Statistics for experimenters: design, innovation, and discovery* vol. 2: Wiley-Interscience New York, 2005.

[75] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

[76] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *Journal of emerging technologies in web intelligence,* vol. 1, pp. 60-76, 2009.

[77] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*: Cambridge university press, 2007.

[78] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management,* vol. 33, pp. 464-472, 2013.

[79] W. Zhang, Y. Cui, and T. Yoshida, "En-LDA: An Novel Approach to Automatic Bug Report Assignment with Entropy Optimized Latent Dirichlet Allocation," *Entropy,* vol. 19, p. 173, 2017.

[80] R. Alghamdi and K. Alfalqi, "A Survey of Topic Modeling in Text Mining," 2015.

[81] S. Zoghbi, I. Vulić, and M.-F. Moens, "Latent Dirichlet allocation for linking user-generated content and e-commerce data," *Information Sciences,* vol. 367, pp. 573-599, 2016.

[82] B. Abu-Salih, P. Wongthongtham, S. Beheshti, and B. Zajabbari, "Towards a methodology for social business intelligence in the era of big social data incorporating trust and semantic analysis," in *Second International Conference on Advanced Data and Information Engineering (DaEng-2015), ed. Bali, Indonesia: Springer*, 2015.

Copyright Clearance Center RightsLink®

Home | Create Account | Help

IEEE Requesting permission to reuse content from an IEEE publication

Title:	A Preliminary Approach to Domain-Based Evaluation of Users' Trustworthiness in Online Social Networks
Conference Proceedings:	2015 IEEE International Congress on Big Data
Author:	Bilal Abu-Salih
Publisher:	IEEE
Date:	June 2015

LOGIN

If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials.

Already **a RightsLink user** or want to learn more?

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK | CLOSE WINDOW

A Preliminary Approach to Domain-based Evaluation of Users' Trustworthiness in Online Social Networks

Bilal Abu-Salih[1], Pornpit Wongthongtham[1], Seyed-Mehdi-Reza Beheshti[1,2], and Dengya Zhu[1]
School of Information Systems, Curtin University, Perth, Australia[1]
School of Computer Science and Engineering, University of New South Wales, Sydney, Australia[2]
bilal.abusalih@curtin.edu.au, p.wongthongtham@curtin.edu.au, sbeheshti@cse.unsw.edu.au, d.zhu@curtin.edu.au

***Abstract*— Online Social Networks (OSNs) are a fertile medium through which users can unleash their opinions and share their thoughts, activities and knowledge of various topics and domains. This medium allows legitimate users as well as spammers to publish their content, leveraging the open environment and fewer restrictions associated with OSNs. Hence, it is essential to evaluate users' credibility in various domains and accordingly make judgements about potentially influential users in a particular domain(s). Most of the existing trustworthiness evaluation approaches of users and their posts in OSNs are generic-based approaches. There is a lack of domain-based trustworthiness evaluation mechanisms. In OSNs, discovering users' influence in a specific domain has been motivated by its significance in a broad range of applications such as personalized recommendation systems and expertise retrieval. The aim of this paper is to present a preliminary approach to evaluating domain-based user's trustworthiness in OSNs. We provide a novel distinguishing measurement for users in a set of knowledge domains. Domains are extracted from the user's content using semantic analysis. In order to obtain the level of trustworthiness, a metric incorporating a number of attributes extracted from content analysis and user analysis is consolidated and formulated considering temporal factor. The approach presented in this paper is promising since it provides a fine-grained trustworthiness analysis of users and their domains of interest in the OSNs.**

Keywords-component; Domain-Based Trust, Online Social Networks, Information Retrieval, Semantic Analysis

I. Introduction

Online Social Networks (OSNs) have been defined by Nepal [1] as the system compounds of certain tools, applications and platforms that sustain the online social interactions of people and communities. Examples of such web-based social media include, but are not limited to, Facebook®, LinkedIn® and Twitter®. These Web Social Networks have thrown open the doors of platforms for people to unleash their opinions and build new kinds of social interactions based on these virtual communities. OSNs provide fertile grounds for legitimate users as well as spammers to publish their content leveraging of the open environment and less restrictions which OSNs facilitate. The spamming activities in social platforms increased dramatically [2]. The spammer's activities comprise abuse in utilizing OSNs' features and tools; spammers send annoying messages to legitimate users; their contents include malicious links, and hijack popular topics [3]. Spammers post contents for various topics, and they duplicate posts [2]. Further, to propagate their vicious activities, spammers abuse other OSNs' features such as hashtags, and mention other users and Link-shortening services [4]. Various approaches for factorization of such entities have been discussed in [5]. These features can be considered as Cross-Cutting Aspects [6] in analysing OSNs.

For example, since there are over 200 million active users of Twitter [7], a significant question arises regarding the quality and trustworthiness of the massive data that is being published every minute by users of such virtual environments. Sherchan et al. [8] defined Trust as the measurement of confidence where a group of individuals or communities behave in a predictable way. The significance of Trust is evident in multiple disciplines such as computer science, sociology, and psychology. Most of the current trustworthiness evaluation approaches of users and their posts in OSNs are generic-based approaches [9, 10] [11-14]. There is a lack of evaluation mechanisms that incorporate domain-based trustworthiness. In OSNs, discovering users' Influence in a specific domain has been motivated by its significance in a broad range of applications such as personalized recommendation systems [13] and expertise retrieval [15].

Domains are these areas of people's expertise, knowledge or specialization [16]. The Semantic Web (SW) was introduced by Berners Lee who provided a new vision for the next web where data is given semantic meanings via data annotation and manipulation in a machine-readable format [17]. By incorporating semantic web technology, this resolves the issue of ambiguity of data and provides metadata which helps related data to be accurately interpreted and understood. In this paper, we incorporate AlchemyAPI[1] as a domain knowledge inference API to analyse the dataset and enrich its textual content in order to provide semantics of textual data and link each message with particular taxonomies; thus, useful knowledge will be inferred for further analysis. AlchemyAPI

[1] http://www.alchemyapi.com/

resolves text disambiguation by incorporating Linked Data[2] such as (DBpedia, Freebase, etc). These open RDF datasets are used by AlchemyAPI to annotate textual content using URIs and infer its semantics accordingly.

Distinguishing users in a set of domains is another significant aspect. For convenience, *distinguishing* and *discriminating* are interchangeably used in this paper. The idea of discrimination was proposed in Information Retrieval (IR) through applying $tf.idf$ formula [18]. "The intuition was that a query term which occurs in many documents is not a good discriminator"[19]. This implies that a term which occurs in many documents decreases its weight in general as this term does not show the particular document of interest to the user [20]. We incorporate this heuristic aspect into our model to evaluate trustworthiness of users in the OSNs platforms. Consequently, we argue that a user who posts in all domains has a low trustworthiness value in general. This argument is justified based on the following facts: (i) There is no one person who is an expert in all domains [21]; (ii) A user who posts in all domains does not declare to other users which domain(s) (s)he is interested in. In OSNs, a user shows to other users which domain (s)he is interested in by posting wide range of contents in that particular domain; (iii) There is a potential that this user is a spammer due to the behaviour of spammers posting tweets about multiple topics [2]. This could end up by tweets being posted in all domains which is not a legitimate users' behaviour.

Moreover, we investigate a metric incorporating a number of attributes to measure users' behaviours in social networks. The key attributes are obtained from content and user analysis. We focus on the twitter platform as it provides a vast amount of diversity in users' contents in various domains; however, the proposed technique can be certainly applied to other social networks. This paper provides a fine-grained trustworthiness analysis of users and their domains of interest in the OSNs. To the best of our knowledge, this work is the first to evaluate a knowledge-based distinguishing mechanism for users in OSNs. The major contributions of this paper are summarized as follows:

- We provide a novel discriminating measurement for users in a set of knowledge domains. Domains are extracted from the user's content using semantic analysis.
- We consolidate and formulate a metric incorporating a number of attributes extracted from content/user analysis to obtain the level of trustworthiness. We provide a holistic trustworthiness approach based on three main dimensions: (i) distinguishing OSNs' users in the set of their domains of knowledge; (ii) feature analysis of users' relation and their contents; (iii) time-aware trustworthiness evaluation.

The rest of this paper is organized as follows: Section II reviews the related work of Trust and Credibility in OSNs. The framework of the proposed approach is described in Section III. Section IV presents the method used for domain knowledge inference. The idea of user's discrimination in a set of their knowledge domains is illustrated in Section V. Section VI demonstrates the mathematical definitions for ranking users based on content and user analysis. Section VII presents the incorporation of the temporal factor in the trustworthiness evaluation mechanism.

II. Literature Review

Trust evaluation in the social media ecosystem is still immature; hence, extensive research is required in this area [8]. There are some approaches to measuring trustworthiness in social media [9, 10, 13, 14, 22-25] [26] [27] . Podobnik et al. [9] proposed a model that calculates trust between friends in a network graph based on weights of the edges between user's connected friends in Facebook. Agarwal and Bin [10] suggested a methodology to measure the trustworthiness of a social media user by using a heterogeneous graph in which each actor in the twitter domain was presented as a vertex type in the graph. The level of trustworthiness was calculated using a backward propagation process. The paper, on the other hand, omits to consider a weighting scheme and temporal factor. Each edge type should be evaluated at different trustworthiness levels; hence, a weighting scheme should be applied. Trustworthiness values vary over time; therefore, the temporal factor should be assimilated. Authors of [11] incorporated a number of attributes; indegree(#followers), retweets, and mentions to measure users' trustworthiness. Brown et al. [12] adopted K-shell algorithm to measure users influence. The algorithm takes a graph of followers/following relationship as an input and evaluates the k-shell level which forms users' ranking. Arlei et al. [13] investigated the influence of social media users and the relevance of their contents in information diffusion data. Tsolmon and Lee's [14] work measured the credibility of Twitter users. Parameters of the Following-Ratio (#follower/#following) and Retweet-Ratio (total retweet of user/total tweets) are used to extract well-known users using the HITS Algorithm; However, they do not take the topic or subject factor into consideration; the classification has been computed in general. Users will have a certain reputation in one domain but that does not always apply to any other domain. The user's reliability should be domain-driven.

Adding a user-domain dimension when calculating trust in social media is an important factor. In this context, in our previous works [28, 29] we highlighted the notion of trust for the data extracted from the unstructured content (such as social media data) in order to calculate trustworthiness values which correspond to a particular user in a particular domain. The literature of trust in social media shows a lack methodologies for measuring domain-based Trust. Ontology represents the core of the domain where the knowledge is shared amongst different entities within the system that may include people or software agents[30]. Recent research has been undertaken to evaluate users' influence in specific topics. Authors of [23] presented a method to discover experts in topic-specific authority networks. They applied a modified version of the HITS Algorithm for more topic-specific network analysis. However, attributes such as (followers/following/friends counts, likes/favourites counts,

[2] http://www.w3.org/DesignIssues/LinkedData.html

etc) were not addressed to infer user reliability. Herzig et al. [31] Proposed an Author-Reader Influence (ARI) model that estimates a user content’s attraction (i.e content’s uniqueness and relevance). In [32] the paper addresses the problem of selecting top-k expert users in social group based on their knowledge about a given topic. In [33], the authors built a model to discover popular topics by analysing users’ relationships and their interests. Jiyeon and Sung-Hyon [34] analysed the flow of information amongst users of social networks to discover “dedicators” who influence others by their ideas and specific topics. Further work has been undertaken to discover experts and influential users in social networks such as [35]. One of the top cited works in topic-based user ranking is Twitterrank [22]. Authors of Twitterrank incorporated topic-sensitive PageRank to infer topic-specific influential users of twitter. However, they did not consider the temporal factor.

Moreover, Twitterrank as well as the mentioned topic-based trustworthiness approaches incorporates a bag-of-words technique called Latent Dirichelet Allocation (LDA) [36] for topic modelling. LDA is an unsupervised machine learning probabilistic model which extracts latent topics by presenting each topic as a words distribution. This statistical mechanism does not consider the semantic relationships of terms in a document [37]. For example, AlchemyAPI offers a comprehensive list of taxonomies divided into five hierarchies where the high-level taxonomy represents the high-level domain and the deeper-level taxonomy provides a fine-grain domain analysis. For instance, “art and entertainment” is considered a high-level taxonomy in which “graphic design” is one of its deep-level taxonomy. LDA is unable to provide high-level topics such as “art and entertainment” from a corpus of posts or tweets unless this term exists in the corpus. Semantic analysis, on the other hand, extracts semantic concepts and infers high-level domains through analysing the semantic hierarchy of each topic leveraging an ontology, which is not possible using LDA technique. Other techniques such as on-line analytical processing of graphs [38] can be used to analyse OSNs.

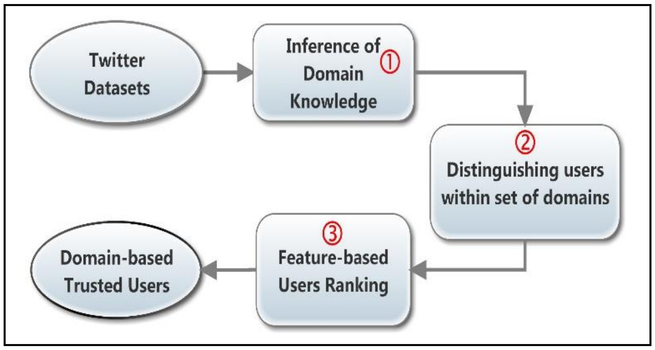

Twitter Datasets
Inference of Domain Knowledge
①
②
Distinguishing users within set of domains
③
Feature-based Users Ranking
Domain-based Trusted Users

Figure. 1. A Framework of the proposed approach

III. Domain-Based Users’ Trustworthiness Evaluation Approach

Figure 1 depicts the framework of the proposed approach. Twitter datasets will be collected using the TwitterAPI. Each tweet will pass via the domain knowledge inference module (Step①). AlchemyAPI will be incorporated in this module to infer tweets taxonomies. Users will be identified based on their tweets distribution within the corresponding set of domains in Step②. A metric incorporating a number of attributes based on user analysis and content analysis is investigated in Step③. The output of this approach is domain-based trustworthiness users in OSNs. The following sections provide further details of the modules used in our approach.

IV. Inference Of Domain Knowledge

In this context, AlchemyAPI is used as a domain knowledge inference tool to analyse the twitter dataset and enrich its textual content in order to provide semantics of textual data and link each message with a particular taxonomy so useful knowledge will be inferred for further analysis. AlchemyAPI is a powerful tool and outperforms other entities’ recognition and semantic mapping tools such as DBPedia Spotlight[3], Extractiv[4], OpenCalais[5] and Zemanta[6] [39].

Table 1 shows an example of taxonomies being extracted using AlchemyAPI from real twitter pages of people specialized in their domains. Scores in Table 1, are calculated using AlchemyAPI, convey the correctness degree of an assigned Taxonomy/Domain to each Twitterer.

TABLE I. SET OF ENTITIES AND CONCEPTS EXTRACTED USING ALCHEMYAPI

Twitterer	Taxonomy	Score
@jimmyfallon	/art and entertainment/humour	0.630
@stocktwits	/finance/investing/stocks	0.763
@alisonjardine	/art and entertainment/visual art and design/painting	0.811
@SHAQ	/sports/basketball	0.703
@DrOz	/health and fitness/disease/cold and flu	0.577

As shown in Table 1, Most of the Jimmy Fallon (**@jimmyfallon**) tweets for example are humour-related topics. Thus, a high score is assigned to the “/art and entertainment/ humour” taxonomy which reflects his experience in this particular domain.

In the implementation of our approach, we intend to analyse and infer taxonomies of each user’s tweet instead of analysing the user’s timeline tweets as one block in order to provide a fine-grained tweet data analysis.

V. DISTINGUISHING DOMAIN-BASED OSNS USERS

As we mentioned, domain-influential users of the OSNs are those users who post widely in a particular domain(s). If

[3] http://dbpedia.org/spotlight/

[4] http://wiki.extractiv.com/w/page/29179775/Entity-Extraction

[5] www.opencalais.com

[6] www.zemanta.com

the user usually tweets in a broad range of domains, this implies that she is not a domain-based influential user due to the fact that there is no knowledgeable person in all domains of our life. From this perspective, we incorporate the traditional Term Frequency-Inverse Document Frequency (TF-IDF) technique which is used in Information Retrieval as a statistical measure to evaluate the importance of a term to a document in a corpus of texts [40]. IDF is a core component of TF_IDF and it is used as a discriminating measure to infer the term's importance in a certain document(s) [18]. In our context, we incorporate this model to distinguish domain-based influential users of OSNs among others. Hence, we argue that in OSNs, a user u whose posts in general are discussing a particular domain(s), u gets a higher distinguishing value in this domain(s) and overcomes other users who post usually in a broad range of domains.

Table 2 shows a synthetic dataset of twitter and the domains in which each user is interested as evident from the previous analysis phase. Data in Table 2 shows the Tweet Frequency (TF) of the content posted by each user in each domain. For example, 23% of U1's tweets were about the politics domain, and almost two thirds of her tweets are IT focused. U2, on the other hand, seems interested in health topics; almost 90% of her tweets discussed health issues, and this tentatively emphasizes her importance in this particular domain. U3 shows an interest in all domains which reduces her importance in all domains accordingly. Fractions in Table 2 provide fine-grain tweets analysis; a user's tweet can discuss multiple topics.

TABLE II. TWEET FREQUENCY OF USERS IN EACH PARTICULAR DOMAIN (TF)

Users	Domains					Total Tweets
	Politics	*Arts*	*IT*	*Sports*	*Health*	
U1	138	0	387	15	60	600
U2	0	0	26	31.2	462.8	520
U3	290.25	56.25	67.5	31.5	4.5	450
U4	2.5	50	0	47.5	0	100
U5	90	337.5	139.5	333	0	900

Data in Table 2 provide insights into the users' interests and domain knowledge; however, the tiny numbers of tweets for a user in a set of domains may end up dropping the overall discriminating value of this user in all domains. These small fractions should be considered due to the following: (i) incorrect domain assignment may occur to a tweet in the domain analysis phase which assigns a user's tweet to an unrelated domain(s); (ii) users may deviate from their domain of expert to discuss general, unrelated or trending topics. Hence, to provide more precise and reasonable results, we propose a fine-tuning parameter which is used as a thresholding value when counting the total number of tweets for each user in each domain. Moreover, data in Table 2 should be normalized to some practical values for further analysis. Thus, we incorporate and customize the sub-linear equation of TF as follows:

$$wf_{t,d} = \begin{cases} 1 + log(tf_{t,d}), \; if \; tf_{t,d} > x \\ 0, \; Otherwise \end{cases} \quad (1)$$

Where $wf_{t,d}$ is the normalized version of TF, x is a thresholding parameter. Table 3 shows the normalized values of tf ($wf_{t,d}$).

TABLE III. NORMALIZED TERM FREQUENCY $\boldsymbol{wf_{t,d}}$

Users	Domains				
	Politics	*Arts*	*IT*	*Sports*	*Health*
U1	3.14	0.00	3.59	2.18	2.78
U2	0.00	0.00	2.41	2.49	3.67
U3	3.46	2.75	2.83	2.50	1.65
U4	1.40	2.70	0.00	2.68	0.00
U5	2.95	3.53	3.14	3.52	0.00

Domain Frequency (DF) is the total numbers of domains that a user u is interested in. Inverse Domain Frequency (IDF) is used to distinguish users amongst domains as follows:

$$idf_t = log(N/(df + 0.5)) \, ; \, where \quad (2)$$

N = the total number of domains in the collection.
df = the domain frequency for each user.

idf_t assigns to a user u a weight in a domain d that is: (i) Highest when a user u has all tweets within a tiny number of domains. (ii) Lower when a user u has fewer tweets in a particular domain(s) or has tweets in a wide range of domains (iii) Lowest when a user u tweets in all domains since this user does not declare which domain (s)he is interested in. Table 4 shows the Domain Frequency (DF) and Inverse Domain Frequency (idf_t) for users of Table 2.

TABLE IV. DOMAIN FREQUENCY (DF) AND INVERSE DF (IDF)

Users	DF	IDF
U1	4	0.097
U2	3	0.222
U3	5	0
U4	3	0.222
U5	4	0.097

The last step at this phase is to combine the results of normalized term frequency $wf_{t,d}$ (users' interest in each domain) with the inverse domain frequency idf_t (total number of domains that users interested in) as follows:

$$W_{t,d} = wf_{t,d} * idf_t \quad (3)$$

Where $W_{t,d}$ is the discrimination value for each user in each domain. Table 5 is the outcome of applying $W_{t,d}$ on data of Tables 2 and 4. It is interesting to note that U2 achieves a higher distinguishing value in the sports domain than U5, although U5 posted more tweets in the sports domain. This emphasizes the importance of U2 in this particular domain. This importance is evident since U2 focuses on fewer domains which distinguish against this user in these domains including sports.

TABLE V. $W_{t,d} = wf_{t,d} * idf_t$

Users	Domains				
	Politics	*Arts*	*IT*	*Sports*	*Health*
U1	0.30	0.00	0.35	0.21	0.27
U2	0.00	0.00	0.54	0.55	0.81
U3	0.00	0.00	0.00	0.00	0.00
U4	0.31	0.60	0.00	0.59	0.00
U5	0.29	0.34	0.31	0.34	0.00

VI. FEATURE-BASED USER RANKING

Although applying $wf_{t,d} * idf_t$ distinguishes users in a set of their domains of knowledge, this formula is insufficient to extract socially domain-based reliable users of social networks. Thus, in the context of twitter, we investigate a metric incorporating a number of attributes to measure users' behaviours in social networks. The key attributes are obtained from content and user analysis and are defined as follows:

Definition 1. Domain-based Retweet Ratio (DR) refers to the total count of retweets of user's contents in each domain to the total count of retweets of user's contents in all domains. It can be calculated as follows:

$$DR_{u,d} = \frac{Total\ Retweets\ of\ user's\ contents\ in\ domain\ d}{Total\ Retweets\ of\ user's\ contents} \quad (4)$$

Definition 2. Domain-based Likes Ratio (DL) refers to the total number of likes/Favourites count to the users' content in each domain to the total number of likes/Favourites of user's contents in all domains. It is represented as:

$$DL_{u,d} = \frac{Total\ Likes\ of\ user's\ content\ in\ domain\ d}{Total\ Likes\ of\ \ user's\ contents} \quad (5)$$

Definition 3. Domain-based Replies Ratio (DP) refers to the total number of replies to the user's tweets in each domain to the total number of replies to all user's contents in all domains. It can be calculated as follows:

$$DP_{u,d} = \frac{Total\ Replies\ to\ user's\ tweets\ in\ domain\ d}{Total\ Replies\ to\ user's\ contents} \quad (6)$$

Definition 4. The Twitter Follower-Friends Ratio (TFF) refers to the total number of user's followers to the total number of users' friends or whom a user follows. Twitter applies certain rules to band the aggressive following behaviour; twitter defines the aggressive following as "indiscriminately following hundreds of accounts just to garner attention" [41]. Twitter limits the total number of users that a user can follow to 2,000 users. Any addition to this number requires an addendum to the list of followers first; hence, the follower-following relationship remains balanced. The dramatic increase of friends that a user u follows compared to the steadiness in the number of followers is considered to be suspicious behaviour, and such a user is most likely to be a spammer [2, 42]. We incorporate the reputation feature proposed in [2] to measure the relative reputation of a user by analysing the follower-following relationship as follows:

$$TFF_u = \frac{Total\ follower\ ofuser\ u}{Total\ follower\ ofuser\ u + Total\ friends\ of\ user\ u} \quad (7)$$

The above equations represent domain-based social trustworthiness indicators of a user in a social network. We incorporate these attributes with the discriminating measure from the previous section to formulate the initial holistic domain-based trustworthiness formula as follows:

$$DT_{u,d} = TFF_u + w^u_{t,d} \times \left(\alpha * DR_{u,d} + \beta * DL_{u,d} + \gamma * DP_{u,d}\right) \quad (8)$$

Where $DT_{u,d}$ represents the user u's trustworthiness in domain d, $w^u_{t,d}$ is the distinguishing value of user u in domain d, while α, β, γ are introduced to adjust the significance of each ratio ($where\ \alpha + \beta + \gamma = 1$); It is apparent that "Retweet" has much higher influence than "Favorite" in twitter context. In general, when a user u retweets a user v's tweet, this implies that u trusts v in this tweet more than user w who is satisfied by "like/favorite" of that tweet.

TABLE VI. DOMAIN-BASED RETWEET RATIO (DR)

Users	Domains				
	Politics	*Arts*	*IT*	*Sports*	*Health*
U1	0.25	0.00	0.07	0.29	0.40
U2	0.00	0.00	0.68	0.01	0.31
U3	0.22	0.03	0.16	0.06	0.52
U4	0.03	0.50	0.00	0.48	0.00
U5	0.37	0.12	0.30	0.21	0.00

TABLE VII. DOMAIN-BASED LIKES RATIO (DL)

Users	Domains				
	Politics	*Arts*	*IT*	*Sports*	*Health*
U1	0.24	0.00	0.17	0.25	0.34
U2	0.00	0.00	0.25	0.36	0.39
U3	0.14	0.02	0.14	0.29	0.31
U4	0.27	0.05	0.00	0.67	0.00
U5	0.06	0.26	0.64	0.05	0.00

TABLE VIII. DOMAIN-BASED REPLIES RATIO (DP)

Users	Domains				
	Politics	*Arts*	*IT*	*Sports*	*Health*
U1	0.35	0.0	0.13	0.36	0.17

U2	0.0	0.0	0.20	0.47	0.32
U3	0.14	0.02	0.14	0.29	0.31
U4	0.36	0.17	0.00	0.47	0.00
U5	0.22	0.13	0.32	0.33	0.00

TABLE IX. THE TWITTER FOLLOWER-FRIENDS RATIO(TFF)

Users	**Follower-Friends Ratio**
U1	0.84
U2	0.56
U3	0.25
U4	0.32
U5	0.66

Tables 6 – 9 show examples for the definitions provided in this section based on the synthesis dataset provided in section IV. Tables 6-8 represents the domain based retweet ratio, domain based likes ratio, and domain-based replies ratio correspondingly. Table 9 shows the users' follower to friends ratio. Values of Table 10 represent the domain-based users' trustworthiness indicators by applying Eq. (8) on the ratio tables and table 5. The significance of each ratio (i.e. $\alpha + \beta + \gamma$) are initiated as (0.4, 0.2, 0.4) respectively.

TABLE X. DOMAIN-BASED USERS' TRUSTWORTHINESS

Users	Domains				
	Politics	***Arts***	***IT***	***Sports***	***Health***
U1	0.93	0.84	0.88	0.91	0.92
U2	0.56	0.56	0.78	0.77	0.73
U3	0.25	0.25	0.25	0.25	0.25
U4	0.39	0.49	0.32	0.62	0.32
U5	0.73	0.71	0.78	0.74	0.66

The results of table 10 emphasizes the significance of incorporating the distinguishing factor into the trustworthiness evaluation mechanism. For example, trustworthiness of U2 is still higher than U5 in *Sports* domain although U5 has a higher reputation indicator (i.e TFF value). On the other hand, U1 has the highest trustworthiness value in *Sports* domain although her twitter frequency in this particular domain is the least. This is due to her great reputation indicator. U3 has an equal trustworthiness value in all domains which is basically equivalent to her TFF ratio. The intuition is that if a user posts in all domains her trustworthiness value will be evaluated based on the reputation indicator which is a focal factor in the trustworthiness evaluation approach.

VII. TIME-AWARE TRUSTWORTHINESS EVALUATION

Although Eq. (8) evaluates users' trustworthiness in domains of knowledge, the users' behaviours may change over time. It follows that trustworthiness values vary over time; hence, the temporal factor should be assimilated. The temporal factor is significant due to the following observations: (i) At time t a highly active user u is likely to be more trustworthy than a user v whose vivacity is low, considering both users hold the same trustworthiness values at time $t-1$. (ii) Similarly, if a user u has shown a dramatic decrease over time in one or more of (DR, DL, DP and TFF) ratios, this implies a reduction in the u's trustworthiness value and vice versa.(iii) Spammers' behaviours are unsteady as they are not legitimate users although they pretend to be. Hence, their "temporal patterns of tweeting may vary with frequency, volume, and distribution over time" [43] . We address the temporal dimension as follows:

$$TDT_{u,d} = \frac{\sum_{t=1}^{I} F(t) \times DT_{u,d_t}}{I} \quad (9)$$

Where $TDT_{u,d}$ is the new time-aware domain-based user u trustworthiness in domain d, I is the number of twitter datasets which corresponds to different time periods. $F(t)$ is a function which takes the timestamp of corresponding $DT_{u,d}$ and returns the weight of $DT_{u,d}$ where more recent the dataset, the highly weight is assigned. $F(t)$ could return 1 for all datasets if they are considered to hold the same trustworthiness values.

VIII. CONCLUSION

In this paper, we propose a preliminary approach to evaluate a domain-based users' trustworthiness in OSNs. In the context of twitter, we investigate a number of factors to infer domain-based users' trustworthiness: (i) applying semantic analysis to discover domain knowledge; (ii) a customized version of TF-IDF weighting mechanism is incorporated to reflect the importance of a user in a particular domain(s); (iii) a metric incorporating a number of attributes extracted from content analysis and user analysis is consolidated and formulated. (iv) time-aware trustworthiness evaluation is considered to analysis user's behaviour over time.

In future, we will be extending this work by proposing a graph-based model. Users' credibility values should be propagated amongst the entire network; thus, we will study the link structure between users of the social network as a whole. Therefore, an enhanced version of Twitterrank [22] will be proposed that takes into consideration the temporal factor to infer domain-based, socially well-known users in OSNs. Moreover, we are in the process of implementing the proposed algorithm using twitter real datasets. Further, we are working on enhancing our framework to take into consideration Big Data infrastructure in retrieving, storing and data analysis.

REFERENCES

[1] S. Nepal, C. Paris, and A. Bouguettaya, "Trusting the Social Web: issues and challenges," World Wide Web, pp. 1-7, 2013/08/10 2013.

[2] A. H. Wang, "Don't follow me: Spam detection in Twitter," in Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on, 2010, pp. 1-10.

[3] M. McCord and M. Chuah, "Spam Detection on Twitter Using Traditional Classifiers," in Autonomic and Trusted Computing. vol. 6906, J. A. Calero, L. Yang, F. Mármol, L. García Villalba, A. Li, and Y. Wang, Eds., ed: Springer Berlin Heidelberg, 2011, pp. 175-186.

[4] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," Information Sciences, vol. 260, pp. 64-73, 2014.

[5] S.-M.-R. Beheshti, S. Venugopal, S. H. Ryu, B. Benatallah, and W. Wang, "Big data and cross-document coreference resolution: Current state and future opportunities," CoRR abs/1311.3987, 2013.

[6] S.-M.-R. Beheshti, B. Benatallah, and H. Motahari-Nezhad, "Enabling the Analysis of Cross-Cutting Aspects in Ad-Hoc Processes," in Advanced Information Systems Engineering. vol. 7908, C. Salinesi, M. Norrie, and Ó. Pastor, Eds., ed: Springer Berlin Heidelberg, 2013, pp. 51-67.

[7] T. Guardian. (2012, 09 July 14). Twitter active users pass 200 million. Available: http://www.theguardian.com/technology/2012/dec/18/twitter-users-pass-200-million

[8] W. Sherchan, S. Nepal, and C. Paris, "A Survey of Trust in Social Networks," ACM Comput. Surv., vol. 45, 2013.

[9] V. Podobnik, D. Striga, A. Jandras, and I. Lovrek, "How to calculate trust between social network users?," in Software, Telecommunications and Computer Networks (SoftCOM), 2012 20th International Conference on, 2012, pp. 1-6.

[10] M. Agarwal and Z. Bin, "Detecting Malicious Activities Using Backward Propagation of Trustworthiness over Heterogeneous Social Graph," in Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, 2013, pp. 290-291.

[11] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," ICWSM, vol. 10, pp. 10-17, 2010.

[12] P. E. Brown and J. Feng, "Measuring user influence on twitter using modified k-shell decomposition," in Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[13] A. Silva, S. Guimarães, W. Meira Jr, and M. Zaki, "ProfileRank: finding relevant content and influential users based on information diffusion," in Proceedings of the 7th Workshop on Social Network Mining and Analysis, 2013, p. 2.

[14] B. Tsolmon and K.-S. Lee, "A Graph-Based Reliable User Classification," in Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). vol. 285, T. Herawan, M. M. Deris, and J. Abawajy, Eds., ed: Springer Singapore, 2014, pp. 61-68.

[15] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si, "Expertise Retrieval," Foundations and Trends in Information Retrieval, vol. 6, pp. 127-256, 2012.

[16] B. Hjørland and H. Albrechtsen, "Toward a new horizon in information science: domain-analysis," J. Am. Soc. Inf. Sci., vol. 46, pp. 400-425, 1995.

[17] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," Scientific american, vol. 284, pp. 28-37, 2001.

[18] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," Journal of the American Society for Information science, vol. 27, pp. 129-146, 1976.

[19] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," Journal of documentation, vol. 60, pp. 503-520, 2004.

[20] J. Ramos, "Using tf-idf to determine word relevance in document queries," in Proceedings of the First Instructional Conference on Machine Learning, 2003.

[21] D. Gentner and A. L. Stevens, Mental models. Hillsdale, N.J: L. Erlbaum Associates, 1983.

[22] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in Proceedings of the third ACM international conference on Web search and data mining, 2010, pp. 261-270.

[23] R. Yeniterzi and J. Callan, "Constructing effective and efficient topic-specific authority networks for expert finding in social media," in Proceedings of the first international workshop on Social media retrieval and analysis, 2014, pp. 45-50.

[24] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," in Proceedings of the 19th international conference on World wide web, 2010, pp. 591-600.

[25] S. Sikdar, K. Byungkyu, J. O'Donovan, T. Hollerer, and S. Adah, "Understanding Information Credibility on Twitter," in Social Computing (SocialCom), 2013 International Conference on, 2013, pp. 19-24.

[26] V. Podobnik, D. Striga, A. Jandras, and I. Lovrek, "How to calculate trust between social network users?," ed: IEEE, 2012, pp. 1-6.

[27] K.-Y. Jeong, J.-W. Seol, and K. Lee, "Follower Classification Based on User Behavior for Issue Clusters," in Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). vol. 285, T. Herawan, M. M. Deris, and J. Abawajy, Eds., ed: Springer Singapore, 2014, pp. 143-150.

[28] P. Wongthongtham, "Ontology and Trust based Data Warehouse in New Generation of Business Intelligence: State-of-the-art, challenges, and opportunities," Journal of Universal Computer Science, 2014.

[29] B. A. Salih, P. Wongthongtham, S.-M.-R. Beheshti, and B. Zajabbari, "Towards A Methodology for Social Business Intelligence in the era of Big Social Data incorporating Trust and Semantic Analysis," in Second International Conference on Advanced Data and Information Engineering (DaEng-2015), ed. Bali, Indonesia: Springer, 2015.

[30] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, "What are ontologies, and why do we need them?," IEEE Intelligent systems, vol. 14, pp. 20-26, 1999.

[31] J. Herzig, Y. Mass, and H. Roitman, "An author-reader influence model for detecting topic-based influencers in social media," in Proceedings of the 25th ACM conference on Hypertext and social media, 2014, pp. 46-55.

[32] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, and G. Vesci, "Choosing the right crowd: expert finding in social networks," in Proceedings of the 16th International Conference on Extending Database Technology, 2013, pp. 637-648.

[33] S. Song, Q. Li, and X. Zheng, "Detecting popular topics in micro-blogging based on a user interest-based model," in Neural Networks (IJCNN), The 2012 International Joint Conference on, 2012, pp. 1-8.

[34] J. Jang and S.-H. Myaeng, "Discovering Dedicators with Topic-Based Semantic Social Networks," in ICWSM, 2013.

[35] D. Liu, L. Wang, J. Zheng, K. Ning, and L.-J. Zhang, "Influence Analysis Based Expert Finding Model and Its Applications in Enterprise Social Network," in Services Computing (SCC), 2013 IEEE International Conference on, 2013, pp. 368-375.

[36] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.

[37] M. Michelson and S. A. Macskassy, "Discovering users' topics of interest on twitter: a first look," in Proceedings of the fourth workshop on Analytics for noisy unstructured text data, 2010, pp. 73-80.

[38] S.-M.-R. Beheshti, B. Benatallah, H. R. Motahari-Nezhad, and M. Allahbakhsh, "A framework and a language for on-line analytical processing on graphs," in Web Information Systems Engineering-WISE 2012, ed: Springer, 2012, pp. 213-227.

[39] G. Rizzo and R. Troncy, "Nerd: evaluating named entity recognition tools in the web of data," in Workshop on Web Scale Knowledge Extraction (WEKEX11). 2011.

[40] A. Rajaraman and J. D. Ullman, Mining of massive datasets: Cambridge University Press, 2011.

[41] Twitter. Following rules and best practices. Available: https://support.twitter.com/groups/56-policies-violations/topics/237-guidelines/articles/68916-following-rules-and-best-practices

[42] Twitter. (2009). The twitter rules. Available: https://support.twitter.com/articles/18311-the-twitter-rules

[43] S. Yardi, D. Romero, G. Schoenebeck, and d. boyd, Detecting spam in a Twitter network, 2009.