

# AN APPROACH FOR TIME-AWARE DOMAIN-BASED ANALYSIS OF USERS' TRUSTWORTHINESS IN BIG SOCIAL DATA

Bilal. Abu-Salih<sup>1</sup>, Pornpit. Wongthongtham<sup>1</sup>, Dengya. Zhu<sup>1</sup>, Shihadeh. Alqrainy<sup>2</sup>

<sup>1</sup> Curtin University, <sup>2</sup> Albalqa Applied University

{bilal.abusalih, Pornpit. Wongthongtham, d.zhu}@curtin.edu.au, alqrainy62@gmail.com

## Abstract

In Online Social Networks (OSNs) there is a need for better understanding of social trust in order to improve the analysis process and mining credibility of social media data. Given the open environment and fewer restrictions associated with OSNs, the medium allows legitimate users as well as spammers to publish their content. Hence, it is essential to measure users' credibility in various domains and accordingly define influential users in a particular domain(s). Most of the existing approaches of trustworthiness evaluation of users 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 an approach to analysing 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 by considering temporal factor. We show the accuracy of the proposed algorithm by providing a fine-grained trustworthiness analysis of users and their domains of interest in the OSNs using big data Infrastructure.

**Keywords:** Domain-Based Trust, Online Social Networks, Information Retrieval, Semantic Analysis

## 1. INTRODUCTION

The ability to harness the ever increasing amounts of business-related data will enable us to understand what is happening in the world. In this context, 'Big Data' is one of the biggest buzzwords these days ([Dumbill, 2012](#)). It has an impact on the Business Intelligence applications. In particular, generating huge metadata (e.g. trust, security, and privacy) for imbuing the business data with additional semantics, the adoption of social media, the digitalization of business artifacts (e.g. files, documents, reports, and receipts), and using sensors (e.g. smart sensors in credit cards) will generate part of the big data. Thus, understanding and analyzing the semantics of the big data is a goal of enterprises today. Big data can be described by some

characteristics. These include and are not limited to: (i) *Volume* refers to the vast increase in the data growth where proper tools and techniques are required to manage such huge blocks of data; (ii) *Veracity* ([Marz, 2013](#)) refers to the accuracy, correctness and trustworthiness of data; (iii) *Variability* ([Fan & Bifet, 2013](#)) refers to variance in meaning; (iv) *Value* measures the quality and significance of data (new insights) ([Kaisler, Armour, Espinosa, & Money, 2013](#)).

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. There are massive amount of data generated from OSNs in term of texts, images, videos, etc. OSNs have been defined by Nepal ([Nepal, Paris, & Bouguettaya, 2013](#)) 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 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.

For example, since there are over 320 million monthly active users of Twitter<sup>1</sup>, 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. (Sherchan, Nepal, & Paris, 2013) 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 (Agarwal & Bin, 2013; Podobnik, Striga, Jandras, & Lovrek, 2012a) (Brown & Feng, 2011; Cha, Haddadi, Benevenuto, & Gummadi, 2010; Silva, Guimarães, Meira Jr, & Zaki, 2013; Tsolmon & Lee, 2014). 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) and expertise retrieval (Balog, Fang, de Rijke, Serdyukov, & Si, 2012).

Domains are these areas of people's expertise, knowledge or specialization (Hjørland & Albrechtsen, 1995). The Semantic Web (SW) was introduced by Berners-Lee who provided a new vision for the next web where data is given semantics via data annotation and manipulation in a machine-readable format (Berners-Lee, Hendler, & Lassila, 2001). 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*<sup>2</sup> 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* resolves text disambiguation by

incorporating *Linked Data*<sup>3</sup> 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 (S. E. Robertson & Jones, 1976). "The intuition was that a query term which occurs in many documents is not a good discriminator" (S. 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). We incorporate this heuristic aspect into our model to measure 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 (Gentner & Stevens, 1983); (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 (Wang, 2010). This could end up by tweets being posted in all domains which is not a legitimate users' behaviour.

Moreover, the users' behaviours may change over time. It follows that trustworthiness values vary over time; hence, the temporal factor should be assimilated. 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 measure a knowledge-based distinguishing mechanism for users in OSNs.

The major contributions of this paper are summarized as follows:

<sup>1</sup> <https://about.twitter.com/company>, Accessed 27 Nov. 2015

<sup>2</sup> <http://www.alchemyapi.com/>

<sup>3</sup> <http://www.w3.org/DesignIssues/LinkedData.html>

- 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.
- We develop a distributed data processing solution to facilitate data storing and 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 approach used for data collection and storage. The data analysis phase is described in Section V. The experimental results are illustrated in Section VI.

## 2. LITERATURE REVIEW

Trust evaluation in the social media ecosystem is still immature; hence, extensive research is required in this area ([Sherchan et al., 2013](#)). There are some approaches to measuring trustworthiness in social media ([Agarwal & Bin, 2013](#); [Kwak, Lee, Park, & Moon, 2010](#); [Podobnik et al., 2012a](#); [Sikdar, Byungkyu, O'Donovan, Hollerer, & Adah, 2013](#); [Silva et al., 2013](#); [Tsolmon & Lee, 2014](#); [Weng, Lim, Jiang, & He, 2010](#); [Yeniterzi & Callan, 2014](#)) ([Podobnik, Striga, Jandras, & Lovrek, 2012b](#)) ([Jeong, Seol, & Lee, 2014](#)) Agarwal and Bin ([Agarwal & Bin, 2013](#)) 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. Arlei et al. ([Silva et al., 2013](#)) investigated the influence of social media users and the relevance of their contents in

information diffusion data. Tsolmon and Lee's ([Tsolmon & Lee, 2014](#)) 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 ([Abu Salih, Wongthongtham, Beheshti, & Zajabbari, 2015](#); [Abu Salih, Wongthongtham, Beheshti, & Zhu, 2015](#); [Wongthongtham & Abu Salih, 2015](#)) 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 ([Chandrasekaran, Josephson, & Benjamins, 1999](#)). Recent research has been undertaken to evaluate users' influence in specific topics. Authors of ([Yeniterzi & Callan, 2014](#)) 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, etc) were not addressed to infer user reliability. Herzig et al. ([Herzig, Mass, & Roitman, 2014](#)) proposed an Author-Reader Influence (ARI) model that estimates a user content's attraction (i.e. content's uniqueness and relevance). In ([Bozzon, Brambilla, Ceri, Silvestri, & Vesci, 2013](#)) the paper addresses the problem of selecting top-k expert users in social group based on their knowledge about a given topic. Jiyeon and Sung-Hyon ([Jang & Myaeng, 2013](#)) 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 ([Liu, Wang, Zheng, Ning, & Zhang, 2013](#)). One of the top cited works in topic-based user ranking is Twiterrank ([Weng et al., 2010](#)). Authors of Twiterrank incorporated topic-sensitive PageRank to infer topic-specific influential users of twitter. However, they did not consider the temporal factor.

Moreover, Twiterrank as well as the mentioned topic-based trustworthiness approaches incorporates a

bag-of-words technique called Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) 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 (Michelson & Macskassy, 2010). 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.

### 3. FRAMEWORK OF THE PROPOSED APPROACH

Figure 1 depicts the framework of the proposed approach. Twitter datasets are collected using the TwitterAPI. Each tweet will pass via the domain knowledge inference module. AlchemyAPI is incorporated in this module to infer tweets taxonomies. Big data infrastructure is used for data storage. A metric incorporating a number of attributes based on user analysis and content analysis is investigated in the trust evaluation approach. The output of this approach is domain-based trustworthiness values for users of

OSNs. The following sections provide further details of the modules used in our approach.

### 4. DATA ACQUISITION AND STORAGE

This paper focuses on the data generated from Twitter micro-blogging. We have chosen Twitter due to the following reasons: (i) Twitter platform has been studied broadly in the research communities (Chen, Madhavan, & 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. The developed prototype can then be applied to other social media platforms.

Twitter API (Makice, 2009) was utilized to retrieve batches of tweets. This was attained by developing and deploying a PHP script incorporating *User timeline* API method to access Twitter platform 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 et al., 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 in 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 paper is to measure the users’ trustworthiness hence user-driven tweets collection is the suitable approach.

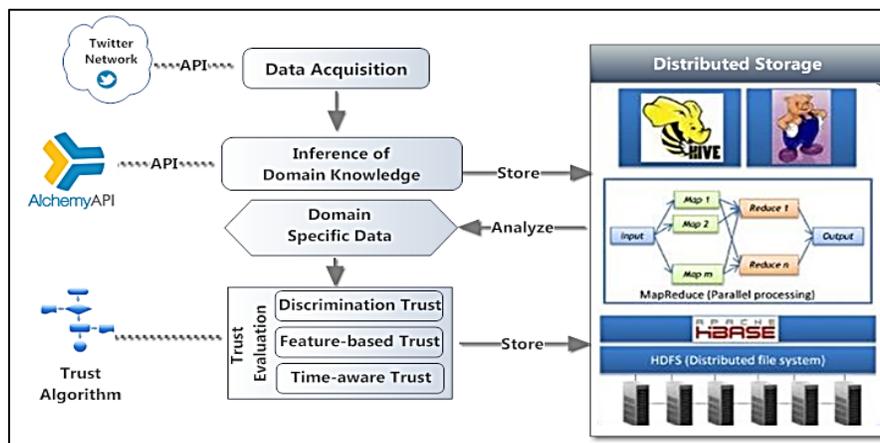


Figure 1. The Architecture of the proposed approach

*Volume* is one of the Big Data features. 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 storage in this chain provides distributed and parallel data processing infrastructure based on the *Hadoop*<sup>4</sup> platform for Big Data. Hadoop: is a distributed computing platform for data processing. It is an open source project developed by Apache<sup>TM</sup> to provide scalable, reliable and fault tolerant framework for Big Data. We utilize the Big Data infrastructure at the Information Systems School - Curtin University for data storage. This is a 10 nodes Big Data cluster; 6×(64 GB RAM, 2 TB Storage, 8 Core Processor) and 4×(16 GB RAM, 1.2 TB Storage, 8 Core Processor).

We designed and implemented *Hive*<sup>5</sup> tables in this distributed environment to facilitate storing information of users and tweets as well as the resultant data from the analysis phase.

## 5. DATA ANALYSIS

This is the focal area of our approach. In this stage, datasets are collected from the distributed environment using NoSQL Language. The collected datasets are processed via two main modules of this research; (i) Knowledge inference; and (ii) trustworthiness evaluation.

### 5.1 Inference of Domain Knowledge

*Variability* (Fan & Bifet, 2013) is an important Big Data dimension. Variability refers to variance in meaning. Incorporating semantic analysis will reduce the ambiguity of the data thus decrease the variability of big data (Hitzler & Janowicz, 2013). In this context, *AlchemyAPI* is used as a domain knowledge inference tool to analyse and enrich the textual message of tweets in order to provide semantics of textual data and obtain each message's taxonomy. This then will be inferred for further analysis. *AlchemyAPI* is a powerful tool and outperforms other entities' recognition and semantic mapping tools such as *DBpedia Spotlight*<sup>6</sup>, *Extractiv*<sup>7</sup>, *OpenCalais*<sup>8</sup> and *Zemanta*<sup>9</sup> (Rizzo & Troncy, 2011). In addition, in March 2015 IBM has acquired *Alchemy* for IBM's development of next generation cognitive computing applications (IBM, 2015).

*AlchemyAPI* offers a technique for analysing the text and providing as a maximum three related taxonomies with the corresponding scores and confident values. *Scores* are calculated using *AlchemyAPI*, range from 0 to 1, and convey the correctness degree of an assigned Taxonomy/Domain to the processed text. *Confident* is a flag calculated by *AlchemyAPI* as well, associated with each response, indicates whether *AlchemyAPI* is confident with the output or not. Hence, if confident parameter comes with "no" value, then *AlchemyAPI* is not certain with the resultant taxonomy.

### 5.2 Trustworthiness Evaluation

*Value* of Big Data (Kaisler et al., 2013) measures the quality and significance of data with new insights. This phase will address this significant big data aspect. Acquiring substantial and valuable information from data in big data scale is a vital task. In OSNs, there is a need for better understanding of social trust in order to improve the analysis process and mining credibility from social media data. Trust in social media refers to the credibility of users and their posted and shared content in a particular domain. For example, since there are over 320 million monthly active users of Twitter, a significant question arises regarding the quality and trustworthiness of the massive data is being published every second by users of such virtual environments.

Evaluating users' trustworthiness is not a trivial task. To achieve this goal, multiple diverse aspects should be considered in order to provide a comprehensive solution to a certain limit. Although spammers could not be stopped, the proper understanding of their behaviour is noteworthy. Further, discovering users' influence in a particular domain has been motivated by its significance in a broad range of applications such as personalized recommendation systems and expertise retrieval. In our approach, we have discussed some dimensions to measure the trustworthiness of users in the social media; hence, discovering domain-influential users of the OSNs. In this section we will go over these dimensions to gain a better understanding of users' profiles and behaviours; thus constructing a comprehensive formula to measure their trustworthiness accordingly.

<sup>4</sup> <https://hadoop.apache.org/>

<sup>5</sup> <https://hive.apache.org/>

<sup>6</sup> <http://dbpedia.org/spotlight/>

<sup>7</sup> <http://wiki.extractiv.com/w/page/29179775/Entity-Extraction>

<sup>8</sup> [www.opencalais.com](http://www.opencalais.com)

<sup>9</sup> [www.zemanta.com](http://www.zemanta.com)

### A. 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 the user usually tweets in a broad range of domains, this implies that this user 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 ([Rajaraman & Ullman, 2011](#)). 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) ([S. E. Robertson & Jones, 1976](#)). 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 usually post in a broad range of domains.

Table 1 shows a list of real twitterers with the corresponding count of tweets. This table also includes the domains in which each user shows the most interest, the total number of tweets about that particular domain and the percentage of #Tweets for the top domain to Total number of tweets

**Definition 1.** Tweet Frequency ( $tf_{u,d}$ ): refers to the total number of tweets  $t$  posted by a user  $u$  where a domain  $d$  was inferred.

Data in Table 2 shows the Tweet Frequency ( $tf_{u,d}$ ) of the content posted by each user in each domain. For example, [@fitnfun](#) seems interested in “health and fitness” topics; almost 47% of her tweets discussed health and fitness issues. On the other hand, 2.7% of ([@Morgancomputers](#))’s tweets were about “health and fitness” domain, and 82% of his tweets are “Technology and computing” focused. And this tentatively emphasizes his importance in this particular domain. ([@CulturalSavage](#)) shows an interest in all domains which reduces her importance in all domains accordingly.

Twitterer	#Total Tweets	Top domain of Interest	#Tweets for top domain	Percentage (%)
@Morgancomputers	339	Technology and computing	279	0.82
@fitnfun	328	Health and fitness	153	0.47
@CulturalSavage	2354	Art and entertainment	555	0.26
@GreenStGoods	302	Food and drink	100	0.33
@spokanechicago	522	Sports	378	0.72

Table 1: list of real twitterers for the demonstration purposes

Twitterer	Law govt and politics	Art and entertainment	Technology and computing	Sports	Health and fitness
@CulturalSavage	96	555	171	135	262
@fitnfun	4	9	8	15	153
@GreenStGoods	10	57	9	12	41
@Morgancomputers	2	12	279	4	9
@spokanechicago	19	97	20	378	9

Table 2: Tweet Frequency of users in each particular domain ( $tf_{u,d}$ )

Twitterer	Law govt and politics	Art and entertainment	Technology and computing	Sports	Health and fitness
@CulturalSavage	2.982	3.744	3.233	3.130	3.418
@fitnfun	0.000	0.000	0.000	2.176	3.185
@GreenStGoods	0.000	2.756	0.000	2.079	2.613
@Morgancomputers	0.000	2.079	3.446	0.000	0.000
@spokanechicago	2.279	2.987	2.301	3.577	0.000

Table 3: Normalized Tweet Frequency ( $wf_{u,d}$ )

Twitterer	Domain Frequency ( $df_u$ )	Inverse $df_u$ ( $idf_u$ )
@CulturalSavage	5	0
@fitnfun	2	0.398
@GreenStGoods	3	0.222
@Morgancomputers	2	0.398
@spokanechicago	4	0.097

Table 4: Domain Frequency ( $df_u$ ) and Inverse  $df$  ( $idf_u$ )

Twitterer	Law gov't and politics	Art and entertainment	Technology and computing	Sports	Health and fitness
@CulturalSavage	0	0	0	0	0
@fitnfun	0	0	0	0.866	1.268
@GreenStGoods	0	0.612	0	0.462	0.580
@Morgancomputers	0	0.828	1.371	0	0
@spokanechicago	0.221	0.290	0.223	0.347	0.000

Table 5:  $W_{u,d} = wf_{u,d} * idf_{u,d}$ 

Data in Table 3 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 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, 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 3 should be normalized to some practical values for further analysis. Thus, we incorporate and customize the sub-linear equation of Term Frequency as follows:

$$wf_{u,d} = \begin{cases} 1 + \log(tf_{t,d}), & \text{if } tf_{t,d} > x \\ 0, & \text{Otherwise} \end{cases} \quad (1)$$

Where  $wf_{u,d}$  is the normalized version of Tweet Frequency,  $x$  is a thresholding parameter (equals to 10 by experiments). Table 4 shows the normalized values of  $tf$  ( $wf_{u,d}$ ).

It is intuitive that Tweet Frequency is insufficient to show which domain the user is interested in; thus, we have to incorporate statistically the domain-level by addressing the number of domains the user has tweeted about.

**Definition 2.** Domain Frequency ( $df_u$ ): is the total numbers of domains that a user  $u$  is interested in. Inverse Domain Frequency ( $idf_u$ ): is used to distinguish users amongst domains as follows:

$$idf_u = \begin{cases} \log(N/(df)), & \text{if } df > 0 \\ 0, & \text{Otherwise} \end{cases} \quad (2)$$

Where  $N$  = the total numbers of domains in the collection,  $df_u$  = the domain frequency for each user  $u$ . Table 4 shows the Domain Frequency ( $df_u$ ) and Inverse Domain Frequency ( $idf_u$ ) for users of Table 3.

As shown in Table 7, **@CulturalSavage** achieves the lowest  $idf_u$  value because he posted tweets about all domains. On the other hand, **@fitnfun** and **@Morgancomputers** achieves the highest  $idf_u$  values as they focused on less number of domains.

The last step of this phase is to combine the results of normalized term frequency  $wf_{u,d}$  (users' interest in each domain) with the inverse domain frequency  $idf_u$  (distinguish users amongst domains of interest) as follows:

$$W_{u,d} = wf_{u,d} * idf_u \quad (3)$$

Where  $W_{u,d}$  is the discrimination value for each user in each domain.  $W_{u,d}$  assigns to a user  $u$  a weight in domain  $d$  that is: (i) highest when a user  $u$  has 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 8 is the outcome of applying  $W_{u,d}$  on data of Tables 6 and 7. It is interesting to note that *@GreenStGoods* achieves a higher distinguishing value in the “Art and entertainment” domain than *@spokanechicago*, although *@spokanechicago* posted more tweets in the “Art and entertainment” domain. This emphasizes the importance of *@GreenStGoods* in this particular domain. This importance is evident since *@GreenStGoods* focuses on fewer domains which distinguish against this user in these domains including “Art and entertainment”. *@CulturalSavage*, on the other hand, achieves the lowest weighting values as (s)he has posted tweets about all domains and did not declare which domain (s)he is interested in. Thus, the overall weight of *@CulturalSavage* has dropped in all domains accordingly.

### B. FEATURE-BASED USER RANKING

Although applying  $wf_{u,d} * idf_u$  distinguishes users in a set of their domains of knowledge, this formula is insufficient to extract socially domain-based reliable users of OSNs. Thus, in the context of twitter, we investigate a metric incorporating a number of attributes to measure users’ behaviours in OSNs. The key attributes are obtained from content and user analysis and are defined as follows:

**Definition 3.** Domain-based Retweet Ratio ( $DR_{u,d}$ ) refers to the total count of retweets for user  $u$ ’s contents in each domain  $d$  to the total count of retweets for user  $u$ ’s contents in all domains. It can be calculated as follows:

$$DR_{u,d} = \frac{\text{Total Retweets for user's contents in domain } d}{\text{Total Retweets for user's contents}} \quad (4)$$

**Definition 4.** Domain-based Likes Ratio (DL) refers to the total number of likes/Favourites count for the users’ content in each domain to the total number of likes/Favourites for user’s contents in all domains. It is represented as:

$$DL_{u,d} = \frac{\text{Total Likes for user's content in domain } d}{\text{Total Likes for user's contents}} \quad (5)$$

**Definition 5.** Domain-based Replies Ratio ( $DP_{u,d}$ ) refers to the total number replies to the user’s tweets in each domain to the total number replies to all user’s contents in all domains. It can be calculated as follows:

$$DP_{u,d} = \frac{\text{Total Replies to user's tweets in domain } d}{\text{Total Replies to user's contents}} \quad (6)$$

**Definition 6.** The Twitter Follower-Friends Ratio ( $TFF_u$ ) 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” (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). We incorporate the reputation feature proposed in (Wang, 2010) to measure the relative reputation of a user by analysing the follower-following relationship as follows:

$$TFF_u = \frac{\text{Total follower of user } u}{\text{Total follower of user } u + \text{Total friends of user } u} \quad (7)$$

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:

$$DT_{u,d} = TFF_u + W_{u,d} \times (\alpha * DR_{u,d} + \beta * DL_{u,d} + \gamma * DP_{u,d}) \quad (8)$$

Where  $DT_{u,d}$  represents the user  $u$ ’s trustworthiness in domain  $d$ ,  $W_{u,d}$  is the distinguishing value of user  $u$  in domain  $d$ , while  $\alpha$ ,  $\beta$ ,  $\gamma$  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.

Tables 6 – 9 show examples for the definitions provided in this section based on the crawled data for the real twitterers. 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 8. The significance of each ratio (i.e.  $\alpha + \beta + \gamma$ ) is initiated as (0.4, 0.2, 0.4) respectively.

The results of table 10 emphasize the significance of incorporating the distinguishing factor into the trustworthiness evaluation mechanism. For example, trustworthiness of *@GreenStGoods* is still higher than

**@spokanechicago** in “Art and entertainment” domain although **@GreenStGoods** did not get any likes, replies or retweet for her tweets. Further, **@GreenStGoods** has the second highest trustworthiness value in “Technology and computing” domain although her tweet frequency in this particular domain is low. This is due to her great reputation indicator.

**@CulturalSavage** 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 measured based on the reputation indicator which is a focal factor in the trustworthiness evaluation approach.

Twitterer	Total Retweet	Law govt and politics		Art and entertainment		Technology and computing		Sports		Health and fitness	
		#Retweet	%	#Retweet	%	#Retweet	%	#Retweet	%	#Retweet	%
@CulturalSavage	8,040	29	<b>0.004</b>	145	<b>0.018</b>	75	<b>0.009</b>	53	<b>0.007</b>	83	<b>0.010</b>
@fitnfun	445	4	<b>0.009</b>	0	<b>0</b>	0	<b>0</b>	1	<b>0.002</b>	205	<b>0.461</b>
@GreenStGoods	0	0	<b>0</b>	0	<b>0</b>	0	<b>0</b>	0	<b>0</b>	0	<b>0</b>
@Morgancomputers	23	0	<b>0</b>	0	<b>0</b>	21	<b>0.913</b>	0	<b>0</b>	0	<b>0</b>
@spokanechicago	10	0	<b>0</b>	1	<b>0.100</b>	0	<b>0</b>	4	<b>0.400</b>	0	<b>0</b>

Table 6. Domain-based Retweet Ratio ( $DR_{u,d}$ )

Twitterer	Total Likes	Law govt and politics		Art and entertainment		Technology and computing		Sports		Health and fitness	
		#Likes	%	#Likes	%	#Likes	%	#Likes	%	#Likes	%
@CulturalSavage	2135	84	<b>0</b>	476	<b>0</b>	116	<b>0</b>	109	<b>0</b>	250	<b>0</b>
@fitnfun	15	0	<b>0</b>	1	<b>0</b>	0	<b>0</b>	1	<b>0.866</b>	7	<b>1.268</b>
@GreenStGoods	0	0	<b>0</b>	0	<b>0.612</b>	0	<b>0</b>	0	<b>0.462</b>	0	<b>0.580</b>
@Morgancomputers	3	0	<b>0</b>	0	<b>0.828</b>	3	<b>1.371</b>	0	<b>0</b>	0	<b>0</b>
@spokanechicago	38	3	<b>0.221</b>	5	<b>0.290</b>	1	<b>0.223</b>	25	<b>0.347</b>	2	<b>0.000</b>

Table 7. Domain-based Likes Ratio ( $DL_{u,d}$ )

Twitterer	Total Replies	Law govt and politics		Art and entertainment		Technology and computing		Sports		Health and fitness	
		#Replies	%	#Replies	%	#Replies	%	#Replies	%	#Replies	%
@CulturalSavage	4221	205	<b>0</b>	985	<b>0</b>	320	<b>0</b>	171	<b>0</b>	540	<b>0</b>
@fitnfun	12	1	<b>0</b>	0	<b>0</b>	0	<b>0</b>	0	<b>0.866</b>	7	<b>1.268</b>
@GreenStGoods	0	0	<b>0</b>	0	<b>0.612</b>	0	<b>0</b>	0	<b>0.462</b>	0	<b>0.580</b>
@Morgancomputers	20	0	<b>0</b>	2	<b>0.828</b>	7	<b>1.371</b>	0	<b>0</b>	0	<b>0</b>
@spokanechicago	25	0	<b>0.221</b>	6	<b>0.290</b>	0	<b>0.223</b>	11	<b>0.347</b>	0	<b>0.000</b>

Table 8. Domain-based Replies Ratio ( $DP_{u,d}$ )

Twitterer	#Followers	#Friends	Follower-Friends Ratio
@CulturalSavage	2046	1026	<b>0.666</b>
@fitnfun	202	362	<b>0.358</b>
@GreenStGoods	149	33	<b>0.819</b>
@Morgancomputers	221	42	<b>0.84</b>
@spokanechicago	59	401	<b>0.128</b>

Table 9. The twitter Follower-Friends Ratio ( $TFF_u$ )

Twitterer	Law gov and politics	Art and entertainment	Technology and computing	Sports	Health and fitness
@CulturalSavage	0.666	0.666	0.666	0.666	0.666
@fitnfun	0.358	0.358	0.358	0.378	1.437
@GreenStGoods	0.819	0.819	0.819	0.819	0.819
@Morgancomputers	0.840	0.914	2.991	0.840	0.840
@spokanechicago	0.141	0.304	0.132	0.734	0.128

Table 10. Domain-Based Users' Trustworthiness ( $DT_{u,d}$ )

### C. TIME-AWARE TRUSTWORTHINESS EVALUATION

Although Eq. (8) measures 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 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 ( $DR_{u,d}$ ,  $DL_{u,d}$ ,  $DP_{u,d}$  and  $TFF_u$ ) 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, Romero, Schoenebeck, & boyd, 2009). We address the temporal dimension as follows; (i) divide 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 Trustworthiness Evaluation section. The only feature that is common in these chunks is the Twitter Follower-Friends  $TFF_u$  Ratio. This is because we have one snapshot for  $TFF_u$  which represents the follower to friends relationship at the time we crawled twitter.

$$TDT_{u,d} = \frac{\sum_{t=1}^I I \times DT_{u,d,t}}{\sum I} \quad (9)$$

Where  $TDT_{u,d}$  is the new time-aware domain-based user  $u$  trustworthiness in domain  $d$ ,  $I$  is a number assigned to each collection of tweets that corresponds to different time periods. We divided the crawled tweets into six chunks where each chunk compounds of the tweets and related metadata of a particular month.  $DT_{u,d,t}$  is the domain-based trustworthiness value for a user  $u$  in a domain  $d$  for the user behavior at period  $t$  as calculated based on Eq. (8). Where  $DT_{u,d,1}$  refers to the evaluation of the domain based trustworthiness of user  $u$  at the first month, and this value should be assigned the lowest weight.  $DT_{u,d,I}$

is the domain based trustworthiness value for user  $u$  in the latest month which reflect the recent behavior of user  $u$  thus the highest weight is assigned accordingly.

## 6. EVALUATION AND DISCUSSION

In the previous section we have proposed a new mechanism in analysing the trustworthiness of users in the online social networks. We have selected five real twitter users to present our approach. This section shows the experiments conducted to evaluate the proposed approach.

### 6.1 Crawled dataset

Three evaluation criteria were identified to select twitterers for conducting experiments: (i) A web-based tool Topsy (<http://www.topsy.com>) was used to select the top influential users in each domain listed in Table 1. (ii) Randomly selection of six users whose profile descriptions did not show a particular domain of interest; or the descriptions exhibited a wide range of interests. (iii) Randomly five users who have high  $TFF_u$  values were added to the list of users.

We crawled users' tweets using twitterAPI as illustrated in Section 4. The crawled dataset was cleansed as follows: Firstly we removed a tweet and its metadata from the dataset if AlchemyAPI was not able to infer any domain for that particular tweet. This could happen when the tweet is very short, or the content is unclear or nonsense, or the tweet was written in a non-English language. Currently English language contents are the only contents supported by AlchemyAPI in their taxonomy inference technique. Secondly we only select tweets which their domain(s) have acquired the score above 0.4 and confident value of 'yes' as thresholds. In other words, we omit tweets which their domains have the score below 0.4 or confident value of 'no'. We select these thresholds after noting that the retrieved domains are closely related to the tweets' context when the score is above 0.4 and confident value equals 'yes'.

These rules are proposed to increase the quality and correctness of the retrieved domains thus satisfy the

*veracity* aspect of Big Data and improve the trustworthiness values accordingly.

Criterion	Twitterer	#Tweets	Domain Freq.(DF)	#Followers	#Friends	#Replies	#Retweets	#Fav
Tech and Computing	<b>cloud_geek</b>	2,347	10	3,167	3,097	5	61	212
Tech and Computing	<b>computingclouds</b>	2,786	15	13,710	672	40	3,448	590
Tech and Computing	<b>joetalik</b>	2,320	12	1,125	108	16	167	376
Tech and Computing	<b>thecloudnetwork</b>	2,916	6	38,991	38,476	11	1,009	635
Tech and Computing	<b>yogeshmalik</b>	2,103	21	22,490	1,828	111	155	230
Art and Entertainment	<b>buzzinghot</b>	2,793	15	2,906	2,887	2	46	44
Art and Entertainment	<b>freeelsa</b>	3,074	13	1,021	924	1	15	41
Art and Entertainment	<b>maxthreshold</b>	1,452	16	10,856	11,887	15	4,356	879
Art and Entertainment	<b>MikeRussEntsUK</b>	2,695	22	3,550	2,215	154	301,072	493
Health and Fitness	<b>dc_trainer</b>	2,975	17	4,741	2,065	211	523	718
Health and Fitness	<b>feelhealthynow</b>	2,966	14	23,740	22,507	5	299	208
Health and Fitness	<b>lifetimesuccess</b>	2,907	21	51,546	449	229	259	526
Health and Fitness	<b>lifetosuccess</b>	2,299	13	42,933	43,602	4	146	113
Health and Fitness	<b>my_health_tips</b>	1,267	14	1,834	1,511	1	22	51
Law, govt and politics	<b>breakingnewz</b>	2,650	17	3,547	0	30	283	213
Law, govt and politics	<b>fouyehaiti</b>	1,434	6	4,911	1,797	60	340	132
Law, govt and politics	<b>infocussa</b>	1,921	15	1,638	212	2	42	34
Law, govt and politics	<b>ninews</b>	2,597	21	4,020	2	16	234	102
Law, govt and politics	<b>theragingqueen</b>	2,348	14	1,321	462	10	34	49
Sports	<b>dtnsports</b>	2,219	1	1,161	82	2	35	64
Sports	<b>mPulseFootball</b>	3,110	13	706	2	1	26	29
Sports	<b>palacetickets</b>	1,529	9	1,512	1,918	2	13	28
Sports	<b>pfzap</b>	1,950	17	4,106	4,001	3	13	29
Sports	<b>WishFeeder</b>	3,171	9	493	10	1	52	33
High_TFF	<b>commadelimited</b>	2,449	23	3,510	250	4,199	62,329	901
High_TFF	<b>jacksonwest</b>	1,428	23	2,437	245	2,086	32,689	1,267
High_TFF	<b>megtripp</b>	1,485	21	6,466	213	583	245,016	954
High_TFF	<b>ronxo</b>	2,185	21	13,235	536	3,314	54,245	3,302
High_TFF	<b>tsand</b>	1,429	21	3,887	199	1,254	51,893	978
Multi_domains	<b>andrewyb</b>	1,628	21	2,259	1,900	390	461,964	443
Multi_domains	<b>fn</b>	1,520	23	2,564	2,361	70	42,193	120
Multi_domains	<b>jrotem</b>	1,266	22	1,902	912	877	366,597	705
Multi_domains	<b>mayagirl</b>	953	21	1,496	2,001	433	48,260	655
Multi_domains	<b>rich1</b>	1,128	22	2,040	767	314	38,548	1,034
Multi_domains	<b>theRab</b>	2,289	22	6,840	6,475	588	962,432	819

Table 11: Evaluation dataset with the list of users and their metadata

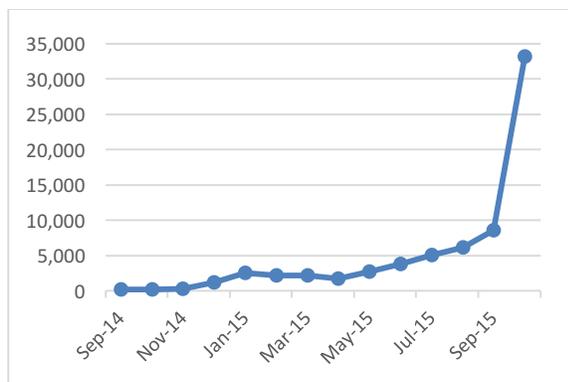


Figure 2. Tweets Distribution

Figure 2 displays the distribution of the crawled tweets posted between Sep 2014 and the Oct 2015. It is noticeable that almost two thirds of the collected tweets were posted during the last three or four months of the collection period. This is because Twitter currently does not allow crawling more than 3,200 of the recent users' tweets of their profile timeline. Hence, some active users (e.g. dc\_trainer, dtnsports, etc.) posted most of their recent 3,200 tweets during the last four months.

## 6.2 Experimental Results

Table 11 shows the evaluation criteria with the corresponding list of users and their metadata, and the

number of tweets after cleansing process. Our trustworthiness approach addresses the temporal dimension in the trustworthiness formula, thus tweets with all related metadata were extracted from the dataset for the time period between (01-07-2015) and (31-10-2015). We divided the selected subset into four chunks, where each chunk includes user’s tweets and their metadata of a particular month. We chose four months only to facilitate the prototype implementation, and this particular time period has been selected because it reflects the most recent users’ behavior. We calculated the time-aware domain-based trust for the list of users in Table 11 as explained in the steps discussed in Section 5.2. For those users who tweeted 3,200 tweets within the last two three month

Figures 3[a-e] represent the five selected domains. Each figure shows the time-ware domain-based trustworthiness values ( $TDT_{u,d}$ ) for the top three users who achieved the highest  $TDT_{u,d}$  and three other users whose  $TDT_{u,d}$  values were the lowest for that particular domain.

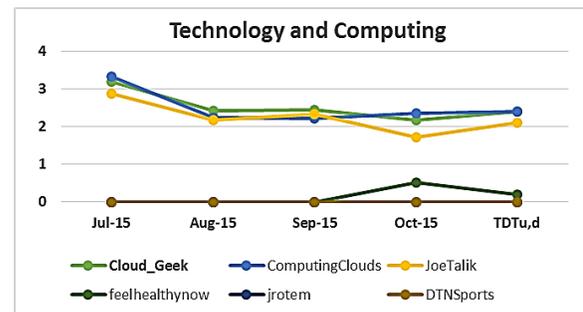
Figure (3.a) shows the list of highest and lowest  $TDT_{u,d}$  values in “Technology and Computing” domain. @cloud\_geek and @computingclouds have achieved the highest  $TDT_{u,d}$  values which reflect their apparent interest in this particular domain. It is obvious that @feelhealthynow, @jrotem, and @DTNSports obtained the lowest values. This is because @feelhealthynow and @DTNSports were selected as influencers in “Health” and “Sports” domains and their tweets reflected such interest. @jrotem is the common denominator in the outcomes of this experiment. This is because @jrotem has acquired the lowest  $TDT_{u,d}$  value in almost all domains. This is evident since this user has been selected as her profile’s description did not declare a particular domain of interest. Further, her domain frequency (DF) which was exposed using our approach was value of “22” domains which has affected her trustworthiness values in all domains accordingly. This example endorses the importance of applying the distinguishing mechanism in evaluating the trustworthiness of users in OSNs.

Figure (3.b) displays highest and lowest  $TDT_{u,d}$  values in “Art and entertainment” domain. The top three users were selected based on  $High\_TFF$  criterion and they are not from the list chosen as *entertainment* influencers users. Although the list of  $High\_TFF$  posted in wide range of domains and this reduces their trustworthiness in one hand; however their metadata ( $TFF_u$ ,  $DR_{u,d}$ ,  $DL_{u,d}$  and  $DP_{u,d}$ ) have increased their trustworthiness values on the other hand.

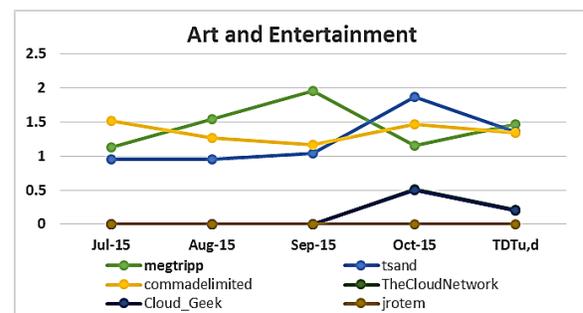
In “Health and Fitness” (Figure 3.c) @my\_health\_tips gained the highest  $TDT_{u,d}$  value despite the unsteadiness in her  $TDT_{u,d}$  values during the selected four month. This emphasizes the significance of addressing the temporal dimension in the trustworthiness formula.

Figure (3.d) shows the list of highest and lowest  $TDT_{u,d}$  values in “Sports” domain. @DTNSports achieved the highest value in this domain. There were two factors assisted this user to achieve this dominant position; Firstly, this user posted all of her tweets about only “Sports” domain thus her domain frequency was value of “1”. Secondly, her  $TFF_u$  value was the highest amongst all users who have been selected as influencers in *Sports* domain.

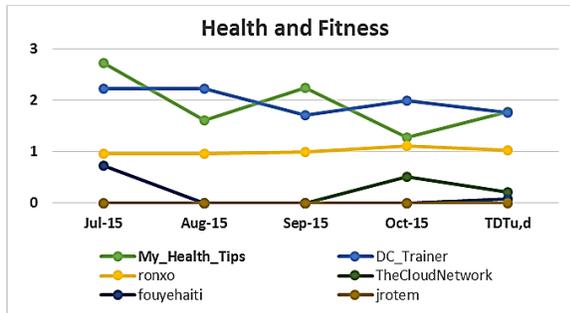
Figure (3.e) displays highest and lowest  $TDT_{u,d}$  values in “Law, Govt and Politics” domain. Although @fouyehaiti gained the highest  $TDT_{u,d}$  value in this domain, this user mainly posted tweets with the hashtag #Haiti and most of her tweets are politics related to Haiti issue. In the future work we will extend the number of selected users in this domain and other domains to discover more domain-based influential users in OSNs.



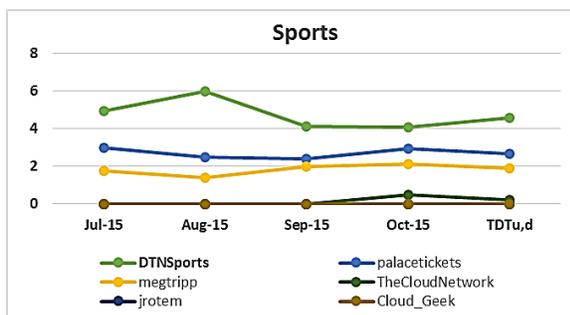
Figures 3.a: Highest and lowest  $TDT_{u,d}$  in Technology and Computing



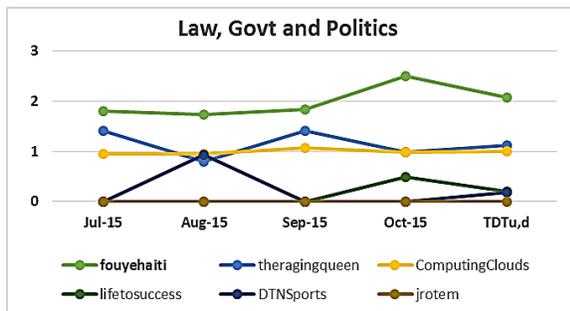
Figures 3.b: Highest and lowest  $TDT_{u,d}$  in Art and entertainment



Figures 3.c: Highest and lowest  $TDT_{u,d}$  in Health and Fitness



Figures 3.d: Highest and lowest  $TDT_{u,d}$  in Sports



Figures 3.e: Highest and lowest  $TDT_{u,d}$  in Law, govt and politics

## 7. CONCLUSION

This paper presents a novel approach to measure time-aware 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. A Big Data infrastructure is utilized to store and evaluate the crawled dataset. The experimental results shows that the proposed mechanism is promising to analyse the users' trustworthiness and infer domain-based influencers in OSNs.

In future, we will be extending this work by crawling a larger dataset and 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 Twiterrank ([Weng et al., 2010](#)) will be proposed that takes into consideration the temporal factor to infer domain-based, socially well-known users in OSNs.

## REFERENCES

Abu-Salih, B., Wongthongtham, P., Beheshti, S.-M.-R., & Zajabbari, B. (2015). Towards A Methodology for Social Business Intelligence in the era of Big Social Data incorporating Trust and Semantic Analysis [Press release]

Abu-Salih, B., Wongthongtham, P., Beheshti, S.-M.-R., & Zhu, D. (2015). A Preliminary Approach to Domain-Based Evaluation of Users' Trustworthiness in Online Social Networks. Paper presented at the Big Data (BigData Congress), 2015 IEEE International Congress on.

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

Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., & Si, L. (2012). Expertise Retrieval. Foundations and Trends in Information Retrieval, 6(2-3), 127-256.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 28-37.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

- Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M., & Vesci, G. (2013). Choosing the right crowd: expert finding in social networks. Paper presented at the Proceedings of the 16th International Conference on Extending Database Technology.
- Brown, P. E., & Feng, J. (2011). Measuring user influence on twitter using modified k-shell decomposition. Paper presented at the Fifth International AAAI Conference on Weblogs and Social Media.
- Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM*, 10, 10-17.
- Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent systems*, 14(1), 20-26.
- Chen, X., Madhavan, K., & Vorvoreanu, M. (2013). A Web-Based Tool for Collaborative Social Media Data Analysis. Paper presented at the Cloud and Green Computing (CGC), 2013 Third International Conference on.
- Dumbill, E. (2012). Planning for big data: " O'Reilly Media, Inc."
- Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. *SIGKDD Explor. Newsl.*, 14(2), 1-5. doi:10.1145/2481244.2481246
- Gentner, D., & Stevens, A. L. (1983). *Mental models*. Hillsdale, N.J: L. Erlbaum Associates.
- Herzig, J., Mass, Y., & Roitman, H. (2014). An author-reader influence model for detecting topic-based influencers in social media. Paper presented at the Proceedings of the 25th ACM conference on Hypertext and social media.
- Hitzler, P., & Janowicz, K. (2013). Linked Data, Big Data, and the 4th Paradigm. *Semantic Web*, 4(3), 233-235.
- Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: domain-analysis. *J. Am. Soc. Inf. Sci.*, 46(6), 400-425. doi:10.1002/(sici)1097-4571(199507)46:6<400::aid-asi2>3.0.co;2-y
- IBM. (2015). IBM Acquires AlchemyAPI, Enhancing Watson's Deep Learning Capabilities. Retrieved from <http://www-03.ibm.com/press/us/en/pressrelease/46205.wss>
- Jang, J., & Myaeng, S.-H. (2013). Discovering Dedicators with Topic-Based Semantic Social Networks. Paper presented at the ICWSM.
- Jeong, K.-Y., Seol, J.-W., & Lee, K. (2014). Follower Classification Based on User Behavior for Issue Clusters. In T. Herawan, M. M. Deris, & J. Abawajy (Eds.), *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)* (Vol. 285, pp. 143-150): Springer Singapore.
- Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, 7-10 Jan. 2013). Big Data: Issues and Challenges Moving Forward. Paper presented at the System Sciences (HICSS), 2013 46th Hawaii International Conference on.
- Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? Paper presented at the Proceedings of the 19th international conference on World wide web.
- Liu, D., Wang, L., Zheng, J., Ning, K., & Zhang, L.-J. (2013). Influence Analysis Based Expert Finding Model and Its Applications in Enterprise Social Network. Paper presented at the Services Computing (SCC), 2013 IEEE International Conference on.
- Makice, K. (2009). Twitter API: Up and running: Learn how to build applications with the Twitter API: " O'Reilly Media, Inc."
- Marz, N. (2013). *Big Data: Principles and best practices of scalable realtime data systems*: O'Reilly Media.
- Michelson, M., & Macskassy, S. A. (2010). Discovering users' topics of interest on twitter: a first look. Paper presented at the Proceedings of the fourth workshop on Analytics for noisy unstructured text data.
- Nepal, S., Paris, C., & Bouguettaya, A. (2013). Trusting the Social Web: issues and challenges. *World Wide Web*, 1-7. doi:10.1007/s11280-013-0252-2
- Podobnik, V., Striga, D., Jandras, A., & Lovrek, I. (2012a). How to calculate trust between social network users? Paper presented at the Software,

Telecommunications and Computer Networks (SoftCOM), 2012 20th International Conference on.

Podobnik, V., Striga, D., Jandras, A., & Lovrek, I. (2012b). How to calculate trust between social network users? (pp. 1-6): IEEE.

Rajaraman, A., & Ullman, J. D. (2011). Mining of massive datasets: Cambridge University Press.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. Paper presented at the Proceedings of the First Instructional Conference on Machine Learning.

Rizzo, G., & Troncy, R. (2011). Nerd: evaluating named entity recognition tools in the web of data. Paper presented at the Workshop on Web Scale Knowledge Extraction (WEKEX11).

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation*, 60(5), 503-520.

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3), 129-146.

Sherchan, W., Nepal, S., & Paris, C. (2013). A Survey of Trust in Social Networks. *ACM Comput. Surv.*, 45(4). doi:10.1145/2501654.2501661

Sikdar, S., Byungkyu, K., O'Donovan, J., Hollerer, T., & Adah, S. (2013, 8-14 Sept. 2013). Understanding Information Credibility on Twitter. Paper presented at the Social Computing (SocialCom), 2013 International Conference on.

Silva, A., Guimarães, S., Meira Jr, W., & Zaki, M. (2013). ProfileRank: finding relevant content and influential users based on information diffusion. Paper presented at the Proceedings of the 7th Workshop on Social Network Mining and Analysis.

Tsolmon, B., & Lee, K.-S. (2014). A Graph-Based Reliable User Classification. In T. Herawan, M. M. Deris, & J. Abawajy (Eds.), *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)* (Vol. 285, pp. 61-68): Springer Singapore.

Twitter. Following rules and best practices. Retrieved from <https://support.twitter.com/groups/56-policies-violations/topics/237->

[guidelines/articles/68916-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. Retrieved from <https://support.twitter.com/articles/18311-the-twitter-rules>

Wang, A. H. (2010, 26-28 July 2010). Don't follow me: Spam detection in Twitter. Paper presented at the Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). Twiterrank: finding topic-sensitive influential twitterers. Paper presented at the Proceedings of the third ACM international conference on Web search and data mining.

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

Yardi, S., Romero, D., Schoenebeck, G., & boyd, d. (2009). Detecting spam in a Twitter network.

Yeniterzi, R., & Callan, J. (2014). Constructing effective and efficient topic-specific authority networks for expert finding in social media. Paper presented at the Proceedings of the first international workshop on Social media retrieval and analysis.

## Authors



**Bilal Abu Salih** received his B.Sc. degree in Computer Science from Qatar University, Qatar, and the M.Sc degree in Computer Science from Al-Balqa Applied University, Jordan. He is now pursuing the Ph.D. degree in the School of Information Systems, Curtin University, Australia. His current research interests include Big Social Data analytics, Trust, Semantic Analytics and Distributed Computing.

**Pornpit Wongthongtham**

received the BSc degree in Mathematics, the MSc degree in Computer Science, and the PhD degree in Information Systems. She is currently an academic staff at the Curtin Institute for Computation (CIC) and also at the school of Information Systems, Curtin University, Perth, Australia. Her research interests include Semantic Analytics, Big Social Data, Ontology Engineering, Semantic Web Technology, and the like.



**Dengya Zhu** is an adjunct research fellow at Curtin University of Technology and a data analyst at the Australian Taxation Office. He has a broad range

of experience in government, industry, and academic research. Dr Zhu's research interests include information retrieval, data mining, machine learning, natural language processing, big data, sentiment analysis, open source software and software development. His research projects are usually practically oriented to address real world issues. He has published a number of papers in his research areas.



**Shihadeh Alqrainy** received his B.Sc. and M.Sc. degrees in Computer Science from Yarmouk University, Jordan, in 1985 and 2002, respectively. During 1987-1999, he was a

computer instructor and training manager of one of the biggest computer institutes in K.S.A. On completing his M.Sc. Degree in 2002, he started working as a full time lecturer at Albalqa Applied University in Jordan. He got his Ph.D. degree in Artificial Intelligence and Software Engineering from DeMontfort University, Leicester, UK, in 2008. His prime research area of interest within Artificial Intelligence is Natural Language

Processing and Understanding. Currently he is as an associate professor of Artificial Intelligence and Software Engineering and the chairman of the Software Engineering Department at Albalqa Applied University since 2012.