

Copyright © 2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Twitter sentiment mining: A multi domain analysis

Saeideh Shahheidari
Dept. of Information System
University of Malaya
Kuala Lumpur, Malaysia
saeideh@um.edu.my

Hai Dong
School of Information Systems
Curtin University of Technology
Perth, Australia
Hai.Dong@cbs.curtin.edu.au

Md Nor Ridzuan Bin Daud
Dept. of Artificial Intelligence
University of Malaya
Kuala Lumpur, Malaysia
ridzuan@um.edu.my

Abstract— Microblogging such as Twitter provides a rich source of information about products, personalities, and trends, etc. We proposed a simple methodology for analyzing sentiment of users in Twitter. First, we automatically collected Twitter corpus in positive and negative tweets. Second, we built a simple sentiment classifier by utilizing the Naive Bayes model to determine the positive and negative sentiment of a tweet. Third, we tested the classifier against a collection of users' opinion from five interesting domains of Twitter, i.e., news, finance, job, movies, and sport. The experimental results show that it is feasible to use Twitter corpus alone to classify new tweet for a certain domain applications.

Keywords- *Opinion mining, sentiment analysis, text mining, classifier, social media*

I. INTRODUCTION

Nowadays, microblogging becomes increasingly popular among people as people are able to share their opinions about every aspect of their life through this platform [1]. Opinion mining is a procedure to extract knowledge from the opinions that people share in web forums, blogs, discussion groups, and comment boxes. In addition, opinion mining uses text mining and natural language processing techniques to make computer understand the expression of emotions. However, its main concern is to extract sentimental and emotional expressions from unstructured text [2]. Identifying the best method for classification is a critical task for sentiment analysis. Many of the approaches rely on database for sentiment analysis [3, 4]. Therefore, the main objective of this paper is to design a simple but faster sentiment classifier by using the Twitter corpus as the database for storing sentiment patterns. Our second objective is to study in which domain sentiment can be best classified with this new classifier. By implementing this approach, we can reduce the development time of building an accurate sentiment classifier for a selected domain. To realize these two objectives, we made use of the Naive Bayes algorithm to classify the polarity of people's opinion about different subjects on Twitter. It is expected that the time and complexity used for classifying the sentimentality of tweets would be dramatically reduced by means of this approach.

This paper consists of the following parts. In Section 2, we provide a literature review of the related research. We also discuss prior works on sentiment analysis and their application for blogging and microblogging. In Section 3, we describe the process of collecting the Twitter corpora. In Section 4, we describe the linguistic analysis of the obtained corpus and show how to train a sentiment classifier and our experimental evaluations. Finally, we conclude our work in Section 5.

II. PREVIOUS LITERATURE

In this section, we only focus on the opinion mining research that conveys people's positive and negative sentiments and the sentiment classification research.

A. Review of Opinion Mining

Opinion mining or sentiment mining is the area of research that attempts to design automatic systems to determine human opinions from text written in natural languages. In general context, textual information can be divided into two types: facts and opinions. Facts are used to define the objective phrases of events, entities, and properties. Opinions are subjective expressions that people usually use to express positive or negative sentiments [5].

Opinion mining tasks can be divided into three main steps:

- Using both data mining and natural language processing techniques [6] to extract product features suggested by customers.
- Identifying if an opinion sentence which is either positive or negative by performing three subtasks: identifying the adjectives which are used to express opinions by using natural language processing techniques; determining semantic orientation (positive or negative) for each adjective by WordNet [7]; and deciding the opinion orientation for each sentence.
- Summarizing the result from previous tasks [6].

Determining whether a piece of text is subjective or objective is the fundamental problem of opinion mining [2]. Subjective expression is used to express the factual status of opinions, beliefs, and emotions in text [8], while objective

expression conveys information according to the intention of the writer [2]. Opinion mining is able to identify if a subjective sentiment is positive, neutral, or negative. Opinion mining includes natural language processing and text analytics. A natural linguistic processor splits text into sentences and assigns tags, such as noun, verb, and adjective, to each word. With the purpose of analyzing sentiments in text, researchers used different techniques, such as polarity tags, semantic orientation, link based patterns, document citations, fuzzy pattern matching, stemming, phrase patterns, punctuation, and stylistic measures [9]. Generally, opinion mining contains two tasks. The first task is sentiment analysis that identifies the polarity of text (positive, negative, and neutral opinion). The second task is to rank the objectivity and subjectivity of the text.

B. Review of Emotion Mining

Emotion is the mental statement with psychological attribute. Ekman [10] divided emotions into 6 groups: happiness, sadness, anger, fear, disgust, and surprise. Emotion mining is divided into 3 scales. The first scale states that the text is either positive or negative. The second scale discusses the richness of the text. The third scale identifies its power or excitement [11]. There are few techniques for extracting emotion from the text. These techniques can be classified into three categories:

- Keyword Spotting
- Lexical Affinity measures
- Statistical Natural Language Processing techniques

The first category relies on lexicon or a dictionary of words. It recognizes emotional words from text according to a lexicon. This technique is effective as it is simple and has time-saving. However, if the sentence structure is too complex, it cannot recognize emotions from text. A typical example is WordNet-Affect [4], which clusters words into several groups of synonyms (called “synsets”) by using WordNet. The other example is SentiWordNet [12], which divides WordNet synsets into two scales – a positive/negative scale and a subjective/objective scale. SentiWordNet is a powerful tool for opinion mining because it is able to recognize positive, negative or objective in text by means of numerical scores [12]. The approach for measuring emotional weights for natural language text is based on keyword spotting. In particular, for emotional weight estimation, this approach divides text into words and relationships among subject, verb, and object [13]. SentiWordNet needs a large-scale database for testing the accuracy of the synset measures. Therefore, some approaches combine SentiWordNet and WordNet-Affect [3]. Some researchers used WordNetSynset to obtain different meanings of a word because a word can be used in different situations and convey different opinions. WordNet is also used to identify adjective and semantic orientation of text [14].

The second technique uses Lexical Affinity measures. This technique is better than the first category. It assigns to each word a probabilistic affinity for definite feelings. For

instance, the word “intelligent” has a 90% probability of reflecting a positive event. Alike to keyword spotting, it depends on sentence corpus and performing poorly when analyzing complex text, for example, “This was not intelligent at all!” For measuring the probability of each emotion, it relies on text corpus used in training. The calculation measures each word as being the ratio of emotional senses over the total senses the word may have. WordNet-Affect and WordNet measure the total number of senses and the number of emotional senses. This method is easy to enumerate how many synsets the word has. Cai et al., [15] presented such a model which combines information measures and domain knowledge by using WordNet to extract concepts from text.

The third technique employs Statistical Natural Language Processing. It uses machine learning algorithms to recognize the relationship between lexicon affinity and words’ co-occurrence frequencies [16,17]. However, it is difficult to be applied in social media sites because of informal language and lack of structured sentences.

C. Review of Sentiment Classification

There are two approaches for classifying sentiment: the knowledge based approach and the machine learning approach [18].

- In the knowledge-based approach, predefined affect dictionaries of opinion words are employed to seek the input words and find its effects [19].
- The machine learning approach uses statistical model for predicting classification of sentiment in input documents, which is also based on vocabulary [20,21].

III. METHODOLOGY

Our methodology comprises three phases – data collection, sentiment analysis studio, and tree tagger analysis. In the rest of this section, the technical details of the three phases will be introduced.

A. Data Collection

Twitter provides an excellent site for sentiment classification and analysis trail. A successful Twitter text classification needs to handle a diverse set of short-text messages with abbreviations, and slangs, in which not much sound grammar is used. By using Twitter API, we collected a corpus of tweets and formed a dataset of two classes: positive and negative sentiments. We followed the same procedure as in [22], by querying Twitter for two types of emotions: happy emotion (e.g. ☺) and sad emotion (e.g. ☹). The two types of retrieved corpora are used to train a classifier to recognize positive and negative sentiments. For data testing, we collected a collection of users’ opinion from Twitter covering about five domains, i.e., news, finance, job, movies, and sport. For each domain, we carefully collected 100 tweets.

B. Sentiment Analysis Studio

We designed the framework of Sentiment Analysis Studio, which is used to evaluate people’s sentiments or feelings toward entities, events, and their properties, using statistical text analysis (Fig. 1).

- **Training Corpora**

Training Corpora comprises 10,000 positive and 10,000 negative tweets, which include emotion lexicon for each positive and negative. We used an emotion corpus for detecting opinions. The positive and negative tweets are all obtained from Twitter web site.

- **Statistical Model**

This step discovers the polarities of the emotions from training corpora and uses various statistical algorithms. Here we used the Naive Bayes classification algorithm where the algorithm is trained on a labeled data set. The Naive Bayes classification algorithm employs a supervised learning method as well as a statistical method for classification. This model analyzes the sentiment of tweets for determining probability of positive and negative. The Naive Bayes model is expressed in (1).

$$p(W|F) = \frac{p(W)p(F|W)}{p(F)} \quad (1)$$

The equation can be written as:

$$\text{Posterior} = \frac{\text{prior} \cdot \text{likelihood}}{\text{evidence}} \quad (2)$$

- **Polarity Keywords**

The polarity keywords help to determine whether the expressed sentiment is positive or negative. Polarity is used to define the strength of the expression of sentiment, which is determined by associated with predefined positive and negative tweets in Training Corpora. For example, if a keyword is specified as outstanding, it indicates that the keyword has a higher strength than the ones specified as very good.

- **Test**

Sentiments are measured according to the statistical model that we developed. This step returns the probability of positive and negative.

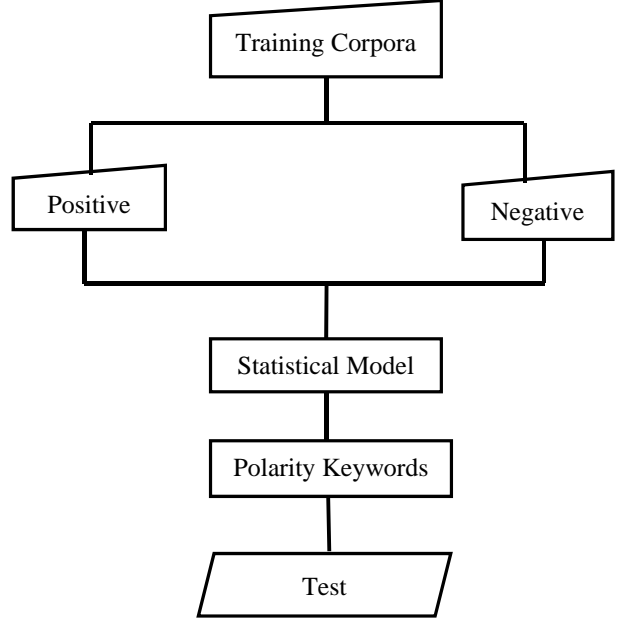


Figure 1. SAS Architecture

C. Tree Tagger Analysis

We used Tree Tagger for English tweets to assign tag for all of the tweets in the corpus. To calculate a difference of tags distributions between positive and negative sets of text we have compared positive and negative of tags distributions according to this formula:

$$S_{P,N} = \frac{T_p - T_n}{T_p + T_n} \quad (3)$$

where T_p and T_n are number of tag occurrences in positive and negative respectively.

We can observe that subjective tweets include personal pronouns (PP, PP\$), while objective tweets contain proper nouns (NPS, NP, NNS). Moreover, in subjective tweets, users are more likely to depict themselves as first person or address the audience as second person (VBP), and use simple past (VBD) or simple verbs (VB) for verb, which is explained by the frequent use of modal verb (MD), while verbs in objective tweets are common in the form of third person (VBZ). JJS stands for superlative adjective for expressing emotions and opinions and comparative adjectives (JJR) states facts and information. In subjective tweets, users use adverbs (RB) for giving an emotional color to verb.

IV. RESULT & ANALYSIS

Accordingly, we collected the comments shared by users. We analyzed them in order to measure the subjectivity expressed by all the Twitter users. Finally, the average of the subjectivity measures of all messages showed that users are more interested in movies, sport, and job.

A. SAS and Human for Polarity of tweets

We manually calculated the probabilistic percentage of positive/negative each tweet based on polarity lexicon of tweets and compared it with the accuracy percentage of SAS. The results are shown in Table 1.

B. Polarity of Lexicons

SAS classifies and calculates the probabilistic percentage of positive/negative for each tweets based on polar lexical. Out of all experiments more than 90% lexicon of tweets are available in Training Corpora for job, movies, and sport. The result shows that the news domain has the obviously negative token (about 52%) in Table 2.

C. Values for Positive vs. Negative

The graphs present the tag distributions between positive and negative tweets on each domain.

The abbreviations of the Penn Treebank tags are as follows:

RP = Particle, PP = Prepositional Phrase, PDT= Predetermine, POS = Possessive, CC = Coordinating conjunction, WRB = Wh-adverb, RBS = Adverb-superlative, RBR = Adverb-comparative, DT=Determiner, WDT= Wh-determiner, VBD = Verb-past tense, VB = Verb-base form, MD = Modal, VBP = Verb- non-3rd person singular present, VBZ = Verb-3rd person singular present, JJ = Adjective, JJS = Adjective-superlative, JJR = Adjective-comparative, NN = Noun, NNS = Noun- plural, NP = Noun Phrase.

Another indicator of negative text is verbs in the past tense (VBD), since most of users gave negative tweets with regard to defeat or frustrated in the domains of finance and job.

- **Finance Tag Values**

We can see from Fig. 3 that POS tags are not distributed evenly in positive and negative. Users are more likely to use JJS to express positive opinions towards Finance. The status of comparative adjectives (JJR) shows that users are more likely to express facts and provide information.

- **Job Tag Values**

The experimental results show that users' negative sentiments towards job are much obvious than their positive sentiments. In addition, it is found that users are more likely to express negative opinions about job, by observing the status of superlative adjective (JJS).

- **Movies Tag Values**

We can see from Fig. 4 that POS tags are distributed evenly in positive, where superlative adjective (JJS)

was more often used to expressed positive emotions and opinions.

TABLE 1. POLARITY OF TWEETS

| Domain | SAS (positive) | SAS (Negative) | Human (TP) | Human (TN) |
|---------|----------------|----------------|------------|------------|
| Finance | 48% | 52% | 39% | 38% |
| Job | 90% | 22% | 75% | 17% |
| Movies | 88% | 12% | 77% | 11% |
| News | 45% | 55% | 35% | 48% |
| Sport | 90% | 10% | 82% | 8% |

TABLE 2. POLARITY OF LEXICONS

| Domain | Lexicon (Exist) | Lexicon (Positive) | Lexicon (Negative) |
|---------|-----------------|--------------------|--------------------|
| Finance | 82% | 40% | 42% |
| Job | 99% | 81% | 18% |
| Movies | 99% | 87% | 12% |
| News | 88% | 36% | 52% |
| Sport | 93% | 83% | 10% |

- **News Tag Values**

We can see from Fig. 5 that POS tags are not distributed evenly in positive and negative. Users more often used superlative adjective (JJS) to express negative emotions and opinions.

- **Sport Tag Values**

Fig. 6 shows that majority of users on Twitter provided positive sentiments regarding sports. Moreover, people are more likely to provide subjective texts by using simple past tense (VBD) and base form of verb (VB).

In summary, we observed that the average of the subjectivity measures of all tweets showed that users are more likely to give positive tweets for movies and sport to express their interest and give negative tweets for job and finance to express their frustration. All of graphs show objective tweets because users more often use proper nouns (NP, NPS).

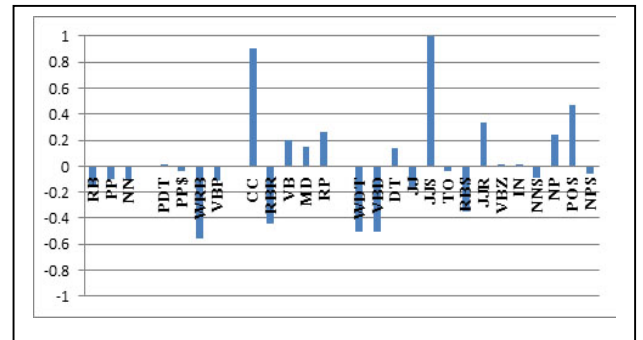


Figure 2. Finance Values for Positive vs. Negative

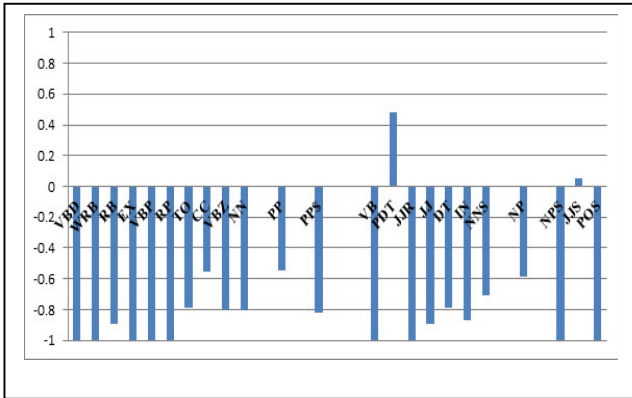


Figure 3. Job Values for Positive vs. Negative

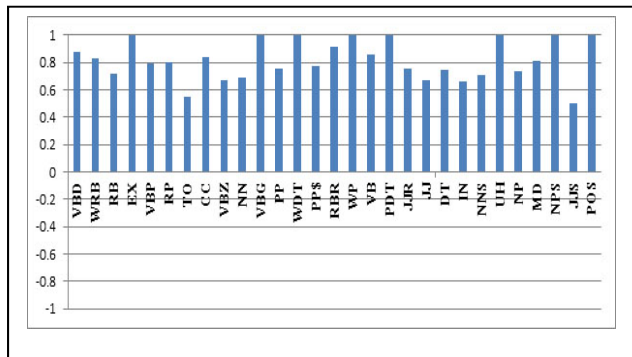


Figure 4. Movies Values for Positive vs. Negative

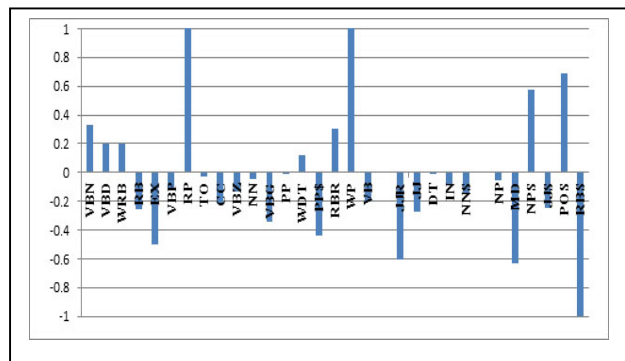


Figure 5. News Values for Positive vs. Negative

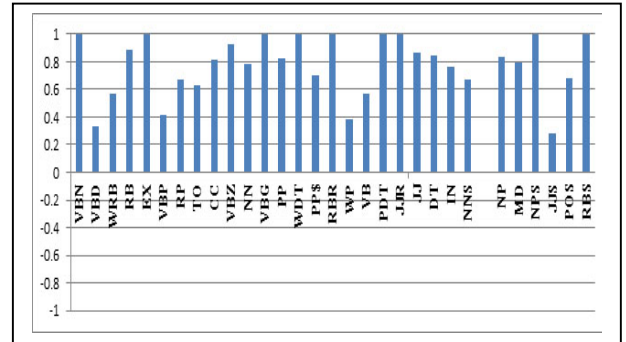


Figure 6. Sport Values for Positive vs. Negative

V. CONCLUSION

Social networking sites are increasingly encouraged and attracting attention of people for creating social relation and collaboration in society. Sentiment mining on Twitter web site will become more and more popular as statistics show that more and more people are using this social network site for discussing their opinions. The large amount of information contained in Twitter web site makes it an attractive source of data for opinion mining and subjectivity classification.

In our research, we propose a sentiment analysis for users in Twitter to determine positive and negative aspects of tweets. We used SAS for classifying sentiments. The classifier is based on the Naive Bayes classifier. This work uses Tree Tagger for English tweets to assign tag to all of the tweets. In addition, the difference of tags distributions between positive and negative tweets in each domain is calculated.

We conclude that the sentiment analysis application on twitter is very challenging. The main reason is that most sentiment includes encrypted words. Users write in a different way than the intended word, but pronounce similarly. The main challenge for this approach is to use existing tweets to classify new tweets. It is a good approach to solve informal and unstructured language of online social network. Moreover, time and complexity on classifying the sentimentality of tweets are reduced instead of creating sentiment mining database. Sentiment mining database is complex, difficult and time consuming to spend majority of CPU time waiting for I/O activity.

In doing this research we have some limitations. One of them is that most of the sentences in twitter website are meaningless. We have limited collecting data only to particular domains (such as finance, movies, sport, job, and news). We focus on English sentence, although Twitter has many international users.

In the future work, we will analyze sentiment structure using a parser, which would yield better result to assess positive and negative sentiments, because the same word may have different meaning in each sentence depending on usage. The other challenge is to use multi-language processors to handle complexity in classification.

REFERENCES

- [1] Pak, A. and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. in Proceedings of LREC. 2010: p. 1320–1326.
- [2] Khan, K., B. Baharudin, and A. Khan. Mining opinion from text documents: A survey. in Digital Ecosystems and Technologies, 2009. DEST'09. 3rd IEEE International Conference on. 2009: IEEE: p. 217–222.
- [3] Chaumartin, F., *A knowledge-based system for headline sentiment tagging*. In Proceedings of SemEval-2007, June 2007: p. 422-425.
- [4] Valitutti, C.S.a.A., *WordNet-Affect: an affective extension of WordNet*. in Proceedings of 4th International Conference on Language Resources and Evaluation, 2004: p. 1083–1086.
- [5] Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of Natural Language Processing, 627-666: p. 1–32.
- [6] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining: p. 168–177.
- [7] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database*. International journal of lexicography, 3(4), 235-244.
- [8] Jindal, N., & Liu, B. (2006). Mining comparative sentences and relations. Paper presented at the Proceedings of the National Conference on Artificial Intelligence: p. 1331–1336
- [9] Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B., et al. (2003). Recognizing and organizing opinions expressed in the world press. Paper presented at the Working Notes-New Directions in Question Answering (AAAI Spring Symposium Series) : p. 1–8.
- [10] Ekman, P. (1992). An argument for basic emotions. Cognition & Emotion, 6(3-4), p. 169-200.
- [11] Yassine, M. and H. Hajj. A framework for emotion mining from text in online social networks. in Data Mining Workshops (ICDMW), 2010 IEEE International Conference on. 2010: IEEE: p. 1136–1142.
- [12] Sebastiani, A.E.a.F., SentiWordNet: a publicly available lexical resource for opinion mining. in Proceedings of the 5th International Conference on Language Resources and Evaluation, 2006: p. 417–422.
- [13] Ma, C., H. Prendinger, and M. Ishizuka, Emotion estimation and reasoning based on affective textual interaction. Affective computing and intelligent interaction, 2005: p. 622-628.
- [14] Bergler, A.A.a.S., Mining Word-Net for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06), 2006: p. 209–216.
- [15] Cai, L., & Hofmann, T. (2003). Text categorization by boosting automatically extracted concepts. Paper presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval: p. 182–189.
- [16] Alm, C.O., D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005: Association for Computational Linguistics: p. 579–586.
- [17] T. Wilson, J.W., and R. Hwa, Just how mad are you? Finding strong and weak opinion clauses. In Proceedings of AAAI-04, 21st Conference of the American Association for Artificial Intelligence, 2004: p. 761–769.
- [18] Yu, B., S. Kaufmann, and D. Diermeier. Exploring the characteristics of opinion expressions for political opinion classification. in Proceedings of the 2008 international conference on Digital government research. 2008: Digital Government Society of North America: p. 82–91.
- [19] Melville, P., W. Gryc, and R.D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009: ACM: p. 1275–1284.
- [20] Dave, K., S. Lawrence, and D.M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. in Proceedings of the 12th international conference on World Wide Web. 2003: ACM: p. 519-528.
- [21] Hatzivassiloglou, V. and K.R. McKeown. Predicting the semantic orientation of adjectives. in Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. 1997: Association for Computational Linguistics: p. 174-181.
- [22] Go, A., R. Bhayani, and L. Huang, Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009: p. 1-12.
- [23] Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Paper presented at the Proceedings of the ACL Student Research Workshop: p. 43–48.