SindhUniv. Res. Jour. (Sci. Ser.) Vol. 48(3): 695-700 (2016)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Discovering Twitter Sentiments using Correlations among Multiple Terms

A. SHAIKH[++], N. A. MAHOTO, N.U.-HUDA*, M. A. UNAR**

Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Sindh,

Received 12thMarch 2016 and Revised 24thAugust 2016

Abstract: Twitter micro-blogging website has taken attention of re- searchers, since its birth in 2006. A number of research efforts has been devoted to study different aspects of data published on Twitter. These studies apply several mechanisms and tools to capture, store and analyze Twitter data. The experts and analysts may look into the sentiments of the online users within their comments published on Twitter and under- stand their views towards certain subjects or area of interest. This paper presents a novel lexicon-based sentiment analysis approach to detect polarities of Twitter users through correlations among multiple terms. The proposed approach is the fusion of polarity and topicality for a certain topic. Twitter data based on a certain topical subject are collected in a document, called *Twitter Dataset Document (TDD)*. Association analysis technique has been exploited to build a directed graph of the multiple associated terms present in the TDD. Then, the top ranked terms are used in sentiment analysis with the help of SentiWordNet dictionary. The users' sentiments are computed based on correlations among multiple terms to facilitate the analysts in understanding the topic oriented expressions of the Twitter users. The results reveal the effectiveness and usefulness of the proposed approach on real text data (i.e., tweets collected in a document).

Keywords: Knowledge discovery, Data mining, Social Network Analysis, Sentiment Analysis

1. INTRODUCTION

The emergence of social networks and modern communication has resulted in massive collection of unstructured data (Wu *et al.,* 2014). The growth in data volumes, in recent years, has taken attention of researchers towards analyzing such data (also known as *Big Data*) in different aspects and in different application domains (Wu *et al.,* 2014), (Librinidis and Jagadish, 2012), (Mahoto, *et al.,* 2014). Significant research has been devoted in textual data analysis using data mining techniques such as association rule analysis and clustering (Cagliero and Mahoto, 2013), (Mahoto, *et al.,* 2014). Data mining extracts useful information from large databases. To get insight information from unstructured data published on social network websites, data mining techniques have profitably been applied.

In recent years, a number of research works has been carried out on the datasets coming from social networks using well-established and well- known data mining techniques. For example, topical trends of online community have been detected from twitter data (Mahoto *et al.,* 2014); Twitter Generalized Rule Visualizer (TGRV) is developed to effectively visualize the high-level associations contained in tweets (small text postings on Twitter[1]) (Cagliero and Mahoto, 2013). Another example, during presidential debate among two major USA political parties (i.e., the Democratic Party and the Republican Party) in October 2012 more than 10 million tweets were published in only 2 hours (Sharp, 2012). These tweets reflected the public interests, for instance, discussions about policies and medical vouchers. Such type of discussions provide new doors to understand the public interests in real time and lead towards the emergence of *Big Data*, which is beyond the capabilities of traditional applications for exploratory analysis.

Particularly, social network data is an excellent data source for the research on big data (Goonetilleke et al., 2014). Since its birth in 2006, Twitter micro-blogging website produces over 500 million tweets a day and over 200 million active users publish tweets (*Tweet is a short text message published on Twitter website*) on monthly basis[2].A wide research has been carried out in order to understand the inter- actions between online communities on Twitter and to explore human behavior from different aspects (Goonetilleke *et al.,* 2014). Sentiment analysis focuses on detecting user emotions contained in unstructured social network data with the help of techniques like natural language processing and data mining. The availability of

sentiments in comments, feedback and user critiques can effectively be utilized for several purposes. For instance, a production company may be interested in getting extent of product acceptance and determining policy design to improve quality of the products.

[++]Corresponding author. Email: anoud.shaikh@faculty.muet.edu.pkTel.: +92-22-2771351 ext. 6907

*Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Sindh, Pakistan

**Department of Computer Systems Engineering, Mehran University of Engineering and Technology, Jamshoro, Sindh, Pakistan

Business organizations and service providers seek the knowledge about their products and services to improve product and service quality by means of mining social network data (Kouloumpis, Wilson and Moore, 2011). However, the challenging issue in micro-blogging is the tremendous range of topics covered. People tweet, share and exchange their thoughts and opinions about anything and everything, more often on social network websites. Therefore, it is difficult to mine Twitter sentiments about a certain topic. Besides, sentiments can either be categorized in polarities: positive and negative, or set of scale values (i.e, very good, good, satisfactory, bad, very bad) (Prabowo and Thelwall, 2009). Therefore, it is an interesting open research area and challenging to understand the user sentiments towards certain topics hidden in massive amount of social network data.

This paper presents a novel and robust lexicon-based sentiment analysis approach to get Twitter sentiments about a topical interest. The proposed approach allows fusion of polarity and topicality and crawls big data from twitter stream about interested topics into a document. Well-established data mining techniques are applied to mine the correlations among multiple terms available in the crawled tweet document. A directed graph is generated from the top correlations among terms available in the tweet document. The ranking of the terms is measured by popular PageRank algorithm (Brin and Page, 1998), and top ranked terms participate in the sentiment scoring of the tweet document. Lexicon-based approach (especially, dictionary-based approach) of sentiment analysis is applied to measure the sentiment scores about the top ranked terms. In particular, textual data published on Twitter is considered for the experiments to demonstrate the significance of the proposed approach in the real-life context. Experiments show the effectiveness of the proposed approach in discovering twitter sentiments for a certain topic at document-level. To the best of our knowledge, detecting sentiments based on correlations among multiple terms in a given document of tweets has not been addressed earlier. However, graph-based summarizer proposed in (Baralis et al., 2013) uses the similar approach and discovers the correlations among multiple terms in multi-document summarization. This paper, on the other hand, detects positive, negative and neutral sentiments from tweets.

The rest of the paper is organized as follows. Section 2 presents the architecture of the proposed approach and describes its main phases. Section 3 assesses the effectiveness of the proposed approach in detecting sentiments of real-life Twitter datasets and reports the results. Finally, conclusions are drawn in Section 5.

2. DISCOVERING TWITTER SENTIMENTS

This section addresses the architecture of the proposed approach for understanding the feelings and attitude of Twitter users towards certain subjective topics. The entire architectural picture of the system is shown in **(Fig. 1),** which depicts that tweets are collect into a Twitter Dataset Document using Topical Crawler by means of Twitter API (Application Programming Interface). These tweets are then preprocessed to transform into the format for further processing. The details of each step presented in Fig. 1 are described in the subsequent sections.

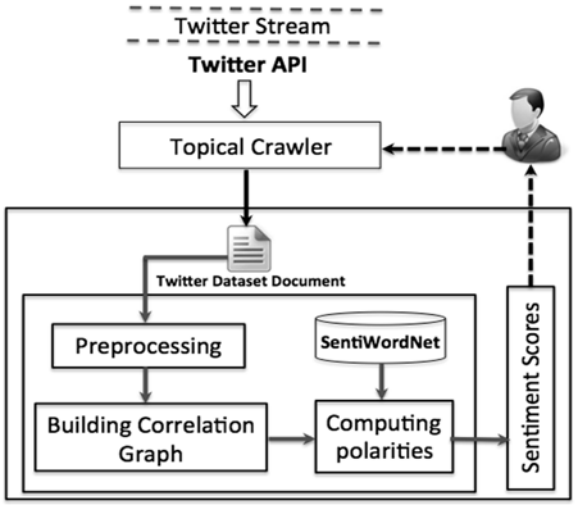


Fig. 1 Main phases of the proposed approach

1.1 Twitterdata retrieval and preprocessing

This phase aims at collecting and preprocessing the text data published on Twitter micro-blogging website. The collection of Twitter messages (i.e., tweets) over a certain topic is crawled using the Twitter Application Programming Interfaces (APIs). Twitter APIs are general-purpose utilities that allow retrieving the tweets from the Twitter. However, these tweets are unsuitable for the analysis due to their formats. Moreover, retrieving tweets over a certain topic in close timestamp (i.e., last 24 hours) is another challenging task. The proposed approach applies *Focused Crawler*, which automatically collects the tweets over a certain topic by setting filtering parameters. The filtering parameters define the tweet contents (i.e, keywords contained in tweets) and geo-location (i.e., user location). Focused Crawler regularly monitors Twitter Stream and retrieves the tweets in accordance with filtered parameters.

The crawled tweets are stored in a Twitter Dataset Document (c.f. Definition 1). Since, textual data in Twitter Dataset Document (TDD) is unsuitable for the

further analysis, a preprocessing step is required. Especially, preprocessing step maps the raw tweet data into a suitable format using preliminary data cleaning procedures. In particular, numbers, stop-words, non-acsii characters and special symbols are eliminated. Later, basic stemming algorithm based on WordNet (Miller, 1995) is applied to manage the tweets into their root stems (i.e., words). Likewise, each tweet transaction is mapped into its distinct stems.

Definition 1: *Twitter Dataset Document. Let TS be a set of tweets. The Twitter Dataset Document TDD associated with TS is set of twitter transactions* TDD_i*, such that for each* $ts_i \in TDD$*. Each transaction in Twitter Dataset Document* $TDD_i = \{t_1, t_2, t_3, \dots, t_n\}$ *is a set of terms (i.e., words) related to* ts_i*.*

For instance, consider a tweet "*This is a tweet comment!*" is modeled into a tweet transaction comprises of 3 distinct terms {This, tweet, comment} and TDD is composed of set of tweet transactions.

1.2 Building correlations among multiple terms

The discovery of the correlations among multiple terms of the TDD is obtained using well-known data mining technique, namely, association rule mining (Agrawal, Imielinski and Swami, 1993). The correlation or association rule (c.f. Definition 2) implicates the relationship between the terms (i.e., words) of the considered dataset. To extract strong correlations, association rule support (c.f. Definition 3) and association rule confidence (c.f. Definition 4) are considered as the correlation measures.

Definition 2:*Association rule. Let A and B be two terms in* TDD *such that* $A \cap B = \emptyset$*. An association rule denoted as* $A \Rightarrow B$ *is an implication between A and* B*. The Ais called antecedent and B is the consequence of the rule.*

Definition 3:*Association rule support. Let* TDD *be a Twitter Dataset Document and* $A \Rightarrow B$ *be an association rule (i.e., correlation). The support of* $A \Rightarrow B$ *in* TDD *is defined as the frequency of* $A \cup B$ *in* TDD*.*

Definition 4:*Association rule confidence. Let* TDD *be a Twitter Dataset Document and* $A \Rightarrow B$ *be an association rule. Let* $s(A \cup B)$*be support of the rule* $A \Rightarrow B$ *in* TDD*. The confidence of* $A \Rightarrow B$*is given by,*

$$conf(A \Rightarrow B) = \frac{s(A \cup B)}{s(A)} \quad (1)$$

Where, $s(A \cup B)$ *is the frequency of A and B together in* TDD*, and* $s(A)$ *is the frequency of termA in* TDD*.*

For example, $\{tweet\} \Rightarrow \{comment\}$ represents the correlation among terms (i.e., words) of the TDD.

The confidence of a rule is the conditional probability of the correlation $A \Rightarrow B$ that measures the estimation of the implication strength of term B given A. For example, consider the correlation $\{tweet\} \Rightarrow \{comment\}$ (*support=10% and confidence=50%*) indicates the implication between the terms *tweet* and *comment* having 10% frequency in the considered data. In other words, the terms *tweet* and *comment* co-occurred 10% in the data, where as 50% confidence indicates *comment* occurred half of the times in which *tweet* occurred.

Given a TDD, a minimum support threshold *minsup* and a minimum confidence threshold *minconf*, the proposed approach entails discovering correlations among multiple terms present in the TDD, which satisfy both *minsup* and *minconf*.

An association rule can be depicted as a graph (Graph is a representation of set of objects in which some objects are paired with links), in which antecedent and consequence of the association rule (i.e., correlation) represent nodes of the graph and correlation measures (i.e., association rule support and confidence) represent links of the graph. The proposed approach constructs a correlation graph (c.f. Definition 5) from the discovered correlations among multiple terms of the considered dataset. Further, correlation graph nodes (i.e., words) are ranked according to their pagerank score, which is computed by applying popular PageRank algorithm (Brin and Page, 1998) used in google search engine. The higher ranked nodes are used as input to the next phase that is described in the following section.

Definition 5:*Correlation graph. Let* $C = \{c_1, c_2, c_3, \dots, c_n\}$ *be the set of correlations among multiple terms in* TDD*. A correlation graph G is a Graph of ordered pairs* $G = (N, E)$*, where N is the set of vertices (i.e. stems or words) such that* $N \in C \in TDD$ *and E is the set of association rule support and confidence values.*

1.3 Computing polarities of multiple terms

The higher ranked nodes of the correlation graph reveal that the multiple terms are converged to them, thus, these nodes are essential in understanding and exploration of the TDD.The sentiment score of these essential terms (i.e., nodes or words) is calculated using SentiWordNet dictionary (Baccianella, Esuli and Sebastiani, 2010). The dictionary returns more than one value for a given word, therefore, normalized word score is computed as given below.

$$word_score_{type} = \sum_{i=1}^{n} \frac{score_i}{|n|} \quad (2)$$

where $type = \{positive, negative, neutral\}$, and $|n|$ is count of scores in SentiWordNet for a given word.

The summation of the word scores represents the sentiments (i.e., Positive, Negative, Neutral) of a given TDD. If summation of positive scores is greater than the total score of other two polarities (i.e., Negative and Neutral), then the document would be regarded as having positive sentiments of the users towards the considered dataset. Similarly, sum of all negative word scores if higher, then the dataset would be considered as having negative sentiments.

1. EXPERIMENTAL VALIDATION

To evaluate the applicability of the proposed approach, experiments has been performed on five real datasets crawled from Twitter. All the experiments are performed on a 2.26 GHz Intel Core 2 Due system with 4 GB RAM.

Table 1Characteristics of datasets

Dataset	No. of transactions	Min. words	Max. words	Avg. words	No. of distinct words
Dataset$_1$	1159	1	107	16.06	577
Dataset$_2$	1273	1	68	13.59	730
Dataset$_3$	1079	2	71	14.18	702
Dataset$_4$	1387	1	96	9.89	442

1.4 Evaluated datasets

Focused Crawler retrieves twitter messages over a period [20 March 2015, 5 April 2015]. The tweets are results of selected keywords covering different topics (i.e., health, education and social events). Focused Crawler retrieved tweets and stored them into TDD. The considered datasets (TDDs) for the experiments are summarized in **(Table 1).**

1.5 Detected Sentiments

This section presents the computational results of the considered real-life datasets for analyzing the applicability of the proposed approach.

Fig. 2 describes the detected sentiment scores of the considered datasets. The experiments are performed at varying support and confidence values of the correlations among multiple terms. As expected, results show that by increasing the support value, fewer score are detected due to the fact that less number of words has higher frequency in the considered dataset. Furthermore, misspelled words and abbreviations (i.e., ‘r‘ instead of ‘are‘, ‘u‘ instead of ‘you‘ etc) are not taken into consideration in the experimental results.

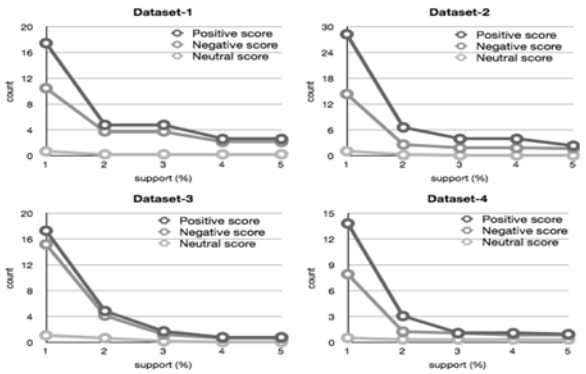


Fig. 2 Experimental results

The performance evaluation of the proposed approach is depicted in Fig. 3. The total number of multiple terms is higher, when lowering the support value and vice-versa. Likewise, processing time is inversely proportional to the support value. For instance, lower support values take more time and higher support values take lesser processing time.

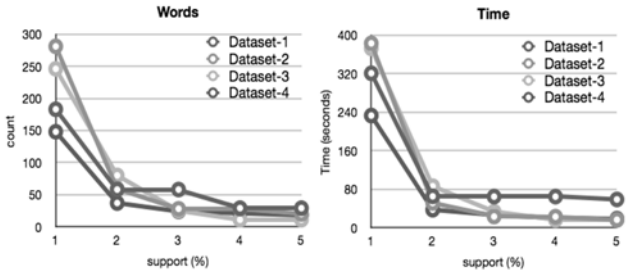


Fig. 3. Performance evaluations

Business organizations and service providers may benefit from the sentimental information of Twitter data for different purposes. For instance, mobile service providers may seek the people’s attitude towards their network coverage and service provisions and then tailor the issues to gain user trust. Similarly, public as well as private organizations may get first hand feedback for their services, products, and development projects from online user community. This feedback may effectively be used in policy designing and decision making strategies.

2. RELATED WORKS

Exploration of sentiments has greatly attracted attention of researchers. The sentiment analysis ranges from document level to phrase and word level (Pang and Lee, 2008), (Kouloumpis, Wilson and Moore, 2011). Analyzing tweet is similar to sentence level sentiment analysis, but tweets make this task more challenging due to presence of special characters in the tweet messages. A number of efforts have been devoted towards sentiment analysis of Twitter data in last decade.

Analysis of opinions fall into two strategies, namely, *linguistic analysis* presented in (Turney, 2002) and *machine learning* described in (Pang and Lee and Vaithyanathan, 2002). These studies are often referred to as unsupervised learning and supervised learning approaches (Martinez-Camara *et al.,* 2014). A wide range of study is discussed on sentiment analysis in (Pang and Lee, 2008), (Liu, 2010), (Tsytsarau and Palpanas, 2012). Sentiment analysis of Twitter data using two models: feature based and tree kernel based compared to state-of-the-art unigram baseline is reported in (Agarwal *et al.,* 2011), and concluded two models outperformed. The experiments are carried out on two data sources 1) manually annotated Twitter data, and 2) acronym dictionary that is collected from online repositories.

The significance of public mood having greater impact over social and economical circumstances is addressed in (Bollen, Pepe and Mao, 2009). They reported that the importance of social as well as political events directly affect public mood, who express it in tweets. A semi-automatic and combined approach of sentiment analysis that unites three approach 1) rule-based classification, 2) supervised learning and 3) machine learning is described in (Prabowo and Thelwall, 2009), which considered movie reviews, product reviews and MySpace comments as data repositories. The proposed approach (Prabowo and Thelwall, 2009) allows provision to each classifier to contribute to other classifiers for get- ting effective level of results.

The combination of polarities and topicality is addressed in online document collection (i.e., UseNet, online message broads, etc) (Hurst and Nigam, 2003) with the help of statistical machine learning classifier. The proposed system (Hurst and Nigam, 2003) harvested data in a certain domain, and then selected the sentences about a given topic. These sentences, later, have been classified according to their polarities: positive or negative. The basic idea of fusion the polarity and the topicality is similar to our proposed approach, but the fundamental difference between the system in (Hurst and Nigam, 2003) is that it uses classifier to assign polarities by training the classifier. Whilst, our approach exploits graph structure and finds correlations among multiple terms for detecting sentiments of Twitter data. However, graph-based structure and finding correlations among multiple terms is addressed in (Baralis et al., 2013) for multi-document summarization. Furthermore, our approach applies lexicon-based sentiment analysis and uses Senti Word Net dictionary to compute polarities; where as system in (Hurst and Nigam, 2003) builds corpus for the considered domain.

3. CONCLUSIONS

This paper introduced a novel approach of detecting user expressions over a certain topic posted on Twitter website. The approach combines the topicality and polarity about a certain topic contained in tweets. The tweets are retrieved and stored in Twitter Dataset Document for sentiment analysis. The proposed approach uses graph-based correlations among the top ranked multiple terms (i.e., words) in the TDD to compute their sentiments. A popular PageRank algorithm calculates the ranking of the words. Furthermore, SentiWordNet dictionary is utilized in computing the sentimental score for the correlated multiple terms. The performance and applicability of the approach is evaluated on a real- life Twitter datasets. The experimental results show usefulness of the approach in real-life context for both public as well as private organizations.

REFERENCES:

Agarwal, A., B Xie,. I. Vovsha, O. Rambow, R. Passonneau, (2011). Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, 30–38.

Agrawal, R., T Imielin´ski,. A. Swami, (1993). Mining association rules between sets of items in large databases. SIGMOD Rec. 22(2), pp. 207–216. DOI 10.1145/ 170036.170072. URL http://doi.acm.org/10.1145/170036.170072

Baccianella, S., A. Esuli, F.Sebastiani, (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: LREC, vol. 10, 2200- 2204

Baralis, E., L. Cagliero, N. Mahoto, A. Fiori, (2013). Graphsum: discovering correlations among multiple terms for graph-based summarization. Information Sciences, pp. 96–109

Bollen, J., Pepe, A., Mao, H. (2009). Modeling public mood and emotion: Twitter senti- ment and socio-economic phenomena. arXiv preprint arXiv:0911.1583

Brin, S., L. Page, (1998). The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems 30(1), 107–117

Cagliero, L., N.A. Mahoto, (2013). Visualization of high-level associations from twitter data. In: Packaging Digital Information for Enhanced Learning and Analysis: Data Visualization, Spatialization, and Multidimensionality: Data Visualization, Spatialization, and Multidimensionality, Information Science Reference (IGI Global). 164–181.

Goonetilleke, O., T. Sellis, X., Zhang, S Sathe,. (2014). Twitter analytics: a big data management perspective. ACM SIGKDD Explorations Newsletter 16(1), 11–20

Hurst, M. F., K. Nigam, (2003). Retrieving topical sentiments from online document col- lections. In: Electronic Imaging 2004International Society for Optics and Photonics, 27–34.

Kouloumpis, E.,. (2011). Twitter sentiment analysis: The good the bad and the omg! ICWSM 11, 538–541

Labrinidis, A., H.Jagadish, (2012). Challenges and opportunities with big data. Proceedings of the VLDB Endowment 5(12), 2032–2033

Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing 2, 627–666

Mahoto, N.A.,(2014). Association rule mining in social net- work data. In: Communication Technologies, Information Security and Sustainable Development, Communications in Computer and Information Science: IMTIC 13, vol. 414, pp. 149–160. Springer International Publishing (2014). DOI 10.1007/978-3-319-10987-9 14

Mart´ınez-C´ E., Amara, L.A., Montejo A.R -Raez,. (2014). Sentiment analysis in twitter. Natural Language Engineering 20(01), 1–28

Miller, G.A. (1995). Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41

Pang, B., L.Lee, (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2(1-2), 1–135

Pang, B., L. Lee, S. Vaithyanathan, (2002). Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Association Computational Linguistics.Vol.10, 79–86.

Prabowo, R., M.Thelwall, (2009). Sentiment analysis: A combined approach. Journal of Informetrics 3(2), 143–157

Sharp, A. (2012). Twitter blog, dispatch from the denver debate (2012). https://blog.twitter.com/2012/dispatch-from-the-denver-debate (Accessed:27-04-2015)

Tsytsarau, M., T. Palpanas, (2012). Survey on mining subjective data on the web. Data Mining and Knowledge Discovery 24(3), 478–514

Turney, P.D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsu- pervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics . 417–424

Wu, X., X Zhu,.G.Q., Wu, W.Ding, (2014). Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1), 97–107