Sentiment Analysis using Optimised Feature Sets in Different Facebook/Twitter Dataset Domains with Big Data

Mohammed Ibrahim Al-mashhadani^{1,*}, Kilan M. Hussein², Enas Tariq Khudir³,
Muhammad ilyas⁴

¹Department of Electrical and Electronics engineering, Altinbas university Istanbul Turkey

²Diyala university Electronic computer center, Iraq

³University of Technology

⁴Electrical and Electronics Engineering, Altinbas University, Turkey

*Corresponding Author: Mohammed Ibrahim Al-mashhadani

DOI: <https://doi.org/10.52866/ijcsm.2022.01.01.007>

Received October 2021; Accepted November 2021; Available online January 2022

ABSTRACT: Nowadays, in many real-life applications, sentiment analysis plays a vital role in the automatic prediction of human activities, especially on online social networks (OSNs). Since the last decade, the research on opinion mining and sentiment analysis has been growing with the increase in volume of online reviews available over the social media networks, such as Facebook OSNs. Sentiment analysis falls under the data mining domain research problem. Sentiment analysis is a type of text mining process to determine subjective attitude, such as sentiments from written texts, and hence has become a main research interest in domains of natural language processing and data mining. The main task of sentiment analysis is to classify human sentiments with the objective of classifying the sentiment or emotion of end users based on their specific text on the OSNs. Several research approaches have been designed for sentiment analysis, in which the factors of accuracy, efficiency and speed have been used to evaluate the effectiveness of sentiment analysis methods. The Map-Reduce framework under the domain of big data is used to minimise the speed of execution and efficiency recently with many data mining methods. The sentiment analysis of messages on Facebook OSNs is more challenging compared with those of other sites because of the misspellings and slang words in their Twitter dataset. In this study, the different solutions that have been recently presented are discussed in detail. Then, a new approach for sentiment analysis based on hybrid feature extraction and multi-class support vector machine methods is proposed. The algorithms have been designed using big data techniques in view of optimising the performance of sentiment analysis.

Keywords: Data Mining; Opinion Mining; Online Social Networks; Features; Classification; Sentiment Analysis; Big data

1. INTRODUCTION

Online social network (OSN) websites, such as Facebook and Twitter, are common nowadays; they have allowed individuals to express their opinions, share their thoughts or report real-time events at ever-increasing rates. Facebook has solely become the most popular website destination since the last decade; as the data of Facebook has been increasing continuously, it is a good example of big data [1, 2]. At present, Facebook has more than 300 million active users, and the messages sent every day is approximately 600 million according to its official website. The proposition of studying the opinions of people toward political events [3], popular movies or products [3] hinges on many companies and organisations being curious about these data. However, as users utilize informal language in Twitter or Facebook and the number of

characters is restricted (e.g. 140 characters for each tweet in Twitter), knowing user understanding based on their opinion and adopting an appropriate method of performance analyses are difficult. Besides, the existence of sarcasm in OSN messages renders the task even more challenging. Here, we describe sarcasm as using remarks that clearly mean the opposite of what is being said as a way to hurt someone's feelings or humorously criticise something [4].

Sarcasm can be regarded as a vast concept, as it is deeply associated with language and common knowledge. The enhancement and refinement of available systems of automatic sentiment analysis are the basic focus of sarcasm; thus, these two aspects are also utilised. Whilst some people express more sarcasm compared with others, sarcasm in general is common, although it is difficult to recognise. Individuals typically convey sarcasm in their daily lives to make jokes and be humorous besides criticising or making remarks about certain ideas, events or individuals [5]. Consequently, sarcasm is vastly used in OSNs, specifically microblogging websites such as Twitter. Incidentally, even state-of-the-art methods of opinion mining and sentiment analysis tend to have a low performance when analysing OSN data [6].

The numerous methods of sentiment analysis use specialised lexica (i.e. a list of negative and positive words) and are usually combined with other approaches, such as machine learning (ML) [1], to assign sentiment scores to texts. In the majority of those lexicas, the words are related to their preceding polarity, suggesting that the out-of-context words conjure something negative or otherwise. For instance, the word 'wonderful' has positive associations (i.e. prior polarity), whereas 'horrible' has negative associations [7]. The methods that use a lexica of prior polarity are famous because they do not need to disambiguate the word sense to give a perceptual score to the word, and they are usually mostly independent of a domain. The prior polarity lexica may be categorised into two groups: the manually built lexica (hiring of expert annotators, such as linguists, or crowd-sourcing annotations on web platforms, such as Mechanical Turk,) and the lexica automatically obtained from pre-existing knowledge. The first type of lexica provides a high degree of precision, but its coverage is low; the opposite is true for the second type [7].

Sentiment visualisation has a significant impact on helping people understand sentiment classification results. The majority of the research on sentiment classification has been concluded with passively produced tables or charts that produce the final portions of each class (i.e. neutral–positive–negative ratio) [8]. When people have difficulty gaining insights into the voluminous classified list of unstructured tweets, it means the sentiment visualisation has been poorly designed. Thus, for the classification results of 'President Debate', we use a 'river' graph to anticipate sentiment intensity based on colour gradations (i.e. the sentiment polarity is illustrated by the ups and downs of the river flow along the timeline) [9]. In other words, with graphical representations, people can clearly recognise not only the tendency of a topic but also the sentiment peak and its bottom. The sentiment analysis process discussed in this study originates from the data mining domain whose common steps include pre-processing (e.g. noise removal), feature extraction (e.g. positive and negative features) and classification (e.g. SVM and Naïve Bayes). The volume of tweets or online messages in OSNs are continuously increasing. Hence, analytic methods, such as performance-efficient sentiment classification, are required to be able to cope with the large volume of OSN messages [10].

This research aims to present the recent existing techniques of sentiment analysis along with their comparative analytic studies. Recent problems are also discussed, and finally, a roadmap to ensure an efficient sentiment analysis using the big data framework is presented. In Section II, recently proposed techniques are discussed. In Section III, the comparative analysis is presented. In Section IV, the current research problems and the proposed new method to solve such problems are elaborated. In Section V, the conclusion and future work are presented.

2. LITERATURE REVIEW

In this section, the different methods of sentiment analysis that have been recently adopted for feature extraction and classification, amongst others, are discussed with respect to the efficiency of emotion analysis of OSN users.

Turney et al. (2002)

In [1], the bag-of-words method was proposed for sentiment analysis. However, the correlations between respective words were not considered. A document of a collection of words was merely produced.

Kamps et al. (2004)

In [2], the emotional contents of certain words over different dimensions as suggested by WordNet (a lexical database) were determined. A distance metric for WordNet was produced, and the semantic orientations of adjectives were specified. The words that have been connected by synonym relations are currently part of the WordNet database.

Baroni et al. (2009)

In [3], a system was introduced to overcome the difficulties of implementing the lexical substitution task based on word space model formalism. The local contexts of the word besides its general distribution were presented.

Balahur et al. (2012)

In [4], the conceptual representation of texts for storing the semantics and structure of real events for a certain domain was demonstrated using EmotiNet. The finite state automata concept utilised in EmotiNet could identify emotional responses triggered by actions.

Barbosa et al. (2010)

In [5], a two-step method for automatic sentiment analysis was proposed for tweet classification. A noisy training dataset was utilised to reduce the labelling effort in the development of classifiers. The tweets were firstly classified according to objective and subjective types. Then, the subjective tweets were classified as either negative and positive.

Celikyilmaz et al. (2010)

In [6], the pronunciation-based approach of word clustering was proposed for the normalisation of noisy tweets. The words with similar pronunciation were clustered and assigned a common token. Text processing methods, such as assigning similar tokens for HTML links, numbers, target organisation names and user identifiers, were utilised in the normalisation. Following the normalisation task, the lexicons of polarity were identified on the basis of probabilistic models. The classification was implemented using the BoosTexter classifier, and the lexicons of polarity were used as the features. Their research attained a decrease in error rate.

Wu et al. (2011)

In [7], the sentiment analysis of Twitter messages was proposed on the basis of influence probability models. Findings indicate that if the @username is found in the tweet's body, then it influences both the action and the probability. Any tweet that starts with @username is a re-tweet representing the influenced action, and it plays a role in influencing the probability. The strong association between those probabilities was demonstrated.

Pak et al. (2010)

In [8], a Twitter corpus was designed to automatically collect tweets and annotate emoticons by using the Twitter API. Apart from the corpus, a sentiment classifier was built on the basis of the multi-nominal Naïve Bayesian classifier. N-gram and POS-tags were utilised as the features. However, errors may arise in this approach. The tweet emotion data in the training set were solely labelled on the basis of the emoticons' polarity. Furthermore, the training dataset only included tweets with emoticons, indicating low efficiency.

Xia et al. (2011)

In [9], an ensemble framework was utilised for sentiment classification. An ensemble model was built by combining a variety of feature datasets and classification approaches. Two types of feature datasets and three base classifiers were utilised to form the ensemble model. The two types of feature datasets were derived using word-relation and part-of-speech information. Naïve Bayesian, support vector machines (SVMs) and maximum entropy were chosen as the base classifiers. A variety of ensemble approaches, such as the weighted combination, fixed combination and meta-classifier combination, were employed for sentiment classification, hence the relatively high accuracy.

Neethu M S et al. (2013)

In [10], Twitter posts concerning electronic products, such as laptops, mobile devices, and so on, were analysed using the ML method. By performing sentiment analysis on a certain domain, the effects of the domain information on the sentiment classification could be identified. A novel feature vector was designed to classify tweets as positive or negative and extract peoples' opinion about a product. Then, existing classifiers were compared with the proposed classifier to determine the latter's precision, recall and accuracy performance.

Shenghua Liu et al. (2014)

In [11], the first-ever semi-supervised approach of topic adaptive sentiment classification (TASC) was proposed. The TASC method was initialised using a classifier that was built with common characteristics and mixed labelled data about a variety of the topics. Minimal hinge loss was observed in terms of the adaptation of unlabelled data and features, which included topic-associated sentiment words, authors' sentiments and sentiment connections derived from the '@' mentions of tweets (also called 'topic adaptive characteristics' in their work). Text and non-text features were obtained and naturally split into two sets for the co-training. The TASC learning approach could update the topic-adaptive features according to the collaborative selection of unlabelled data. This scheme allowed the selection task to obtain more reliable tweets, suggesting the improved performance of TASC. Consequently, an adaptation model along a timeline (TASC-t) was designed for dynamic tweets.

Rui Xia et al. (2015)

In [12], a novel method called dual sentiment analysis (DSA) was introduced to solve sentiment classification problems. Firstly, an innovative data expansion approach was developed by creating a sentiment-reversed review for each training and test review. Secondly, the DSA was designed to benefit from the reverse and the original training reviews (in pairs). The learning of sentiment classifier and the dual prediction approach to classify the test reviews comprised the two parts of a single review. Thirdly, the DSA model was extended from polarity (positive-negative) classification to three-class (positive-negative-neutral) classification by considering the neutral reviews. Finally, a corpus-based approach was designed for the construction of a pseudo- antonym dictionary, thus overcoming the dependency of the DSA on an external

antonym dictionary for the reversion of the review. The novel approach, which can be described as different from the other methods, is illustrated in Figure1.

R. Suresh Ramanujam et al. (2015)

In [13], firstly, a sentiment analysis using the big data of online product reviews of customers was conducted. Then, the Map-Reduce framework was integrated into the sentiment analysis method to reduce the computational resources. Secondly, a novel approach for handling neutral reviews based on the lexicon method was introduced to overcome the restriction of the classification process.

FIGURE 1. Dual sentiment analysis architecture [12]

Figure 2 shows the working scheme of the big data-based sentiment analysis process.

FIGURE 2. Hadoop-based data processing [13]

Mondher Bouazizi et al. (2016)

In [14], the importance of using sarcasm on OSNs was initially investigated. Then, a novel method to detect sarcastic tweets was introduced. Finally, the detected information was utilised to enhance sentiment analysis performance and accuracy.

3. ANALYSIS

Table 1 shows a comparison of the recent methods and their respective purpose, methods and performance metrics.

Table 2 shows a comparative analysis of recent studies about minimising computational resources for sentiment analysis.

Three performance metrics are commonly used to determine the efficiency and effectiveness of sentiment analysis methods.

Table 1. COMPARATIVE ANALYSIS OF RECENT SENTIMENT ANALYSIS METHODS

Author and Year	Methodology	Dataset	Performance Metrics
Alexandra Balahur et .al. (2012)	Emotion DetectionDetection using model of Appraisal Theory	Emotion Antecedents and Reactions (ISEAR) corpus	Precision, Recall, Accuracy.
Rui Xia et.al.et al. (2011)	Ensemble Feature Extraction. Ensemble Classification.	Cornell movie-review, book, DVD datasets	Accuracy
Neethu M S et.al.et al. (2013)	Efficient feature vector and classifiers.	Twitter	Precision, Recall, Accuracy.
Shenghua Liu et .al. (2014)	Semi-supervised topic-adaptive sentiment classification	Twitter	Accuracy, precision, recall, F-score.
Rui Xia et .al. (2015)	Dual Sentiment Analysis. Dual prediction algorithm	Multi-domain dataset	Accuracy
R. Suresh Ramanujam et .al. (2015)	Big-D data, Map-Reduce, Lexicon based method.	Twitter	Hourly sentiment analysis
Mondher Bouazizi et .al. (2016)	Pattern based Sarcasm Detection	Twitter	Precision, Recall, Accuracy.

Table 2. COMPARATIVE ANALYSIS RECENT SENTIMENT ANALYSIS METHODS

Author and Year	Big -Data Framework	Related Metrics
Alexandra Balahur et .al. (2012)	No	No
Rui Xia et.al.et al. (2011)	No	No
Neethu M S et.al.et al. (2013)	No	No
Shenghua Liu et .al. (2014)	No	No
Rui Xia et .al. (2015)	No	No
R. Suresh Ramanujam et .al. (2015)	Yes	Only Hourly sentiment analysis
Mondher Bouazizi et .al. (2016)	No	No

Accuracy Rate: This metric is used to compute the classification correctness of an input dataset. From a technical standpoint, accuracy is the measure of fraction computation of the entire correctly classified samples with respect to the total number of samples.

Precision Rate: This metric is used to measure the ratio of extracted tweets considered to be relevant. It computes the total number of test samples successfully classified as sarcastic over the entire sample classified as sarcastic.

Recall Rate: This metric is used to measure the ratio of relevant sarcastic tweets that have been retrieved. Figure 3 shows the comparative analysis of accuracy (%) of the different methods.

4. RESEARCH GAPS AND METHODOLOGY

4.1 RESEARCH GAPS

- Tables 1 and 2 show the limitations of the techniques. Nowadays, with the ever-increasing volume of online datasets, a scalable and robust approach for analysing more tweets with less computation resources is needed. The existing sentiment analysis methods simply focus on improving efficiency in terms of the accuracy of classification or solving domain-specific problems, including topic aware classification.
- Figure 3 shows a comparison of the accuracy performance of the different recently reported techniques. The accuracy performance of ‘correct classification of messages or tweets’ is poor.
- The previous methods are not generalised. Most of the past studies evaluated their respective methods by considering Twitter datasets only or by focusing on specific domains only.

4.2 EFFICIENCY OF THE METHOD

Here, we present a roadmap for efficient sentiment analysis to solve the aforementioned problems. All of the steps can be accomplished using the optimised Map-Reduce framework and the novel scheduling technique design. The objective of the proposed approach is to improve the scalability, robustness, effectiveness and efficiency of sentiment analysis that use big data as the framework. The contributions of this study via the proposed method are as follows.

FIGURE 3. Accuracy graph

- An approach for optimised resource scheduling was designed for the Map-Reduce model. In this manner, the quality of service of big data for sentiment analysis can be improved.
- An innovative and adaptive pre-processing approach was designed to address the dataset diversity problem.
- A novel hybrid feature extraction algorithm was designed by considering the different sets of features to be extracted from the input text messages. In this manner, accuracy performance can be optimised.
- Designing the optimisation technique based modified SVM classifier for classification purpose.

Finally evaluating the efficiency through the consideration of performance metrics such as recall, precision, F1-measure, accuracy, processing time, memory utilisation, CPU utilisation, and so on.

5. CONCLUSION AND FUTURE WORK

Nowadays, online sentiment analysis is widely used by firms as a technique to gather product recommendations and movie analysis, amongst others. In this study, we presented the importance of sentiment analysis and discussed the different techniques that have been proposed since the last decade. A detailed comparative analysis of those techniques is presented in this study. The problems associated with those methods were also analysed. Finally, the different contributions of the proposed methodology were explored. For future work, we will focus on contributing a design to solve the aforementioned research problems.

ACKNOWLEDGEMENTS

The first author thanks the reviewers for their useful suggestions that improved the presentation of this paper.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] P. D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417424, Association for Computational Linguistics, 2002.
- [2] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, Using wordnet to measure semantic orientations of adjectives, 2004.
- [3] D. Pucci, M. Baroni, F. Cutugno, and A. Lenci, Unsupervised lexical substitution with a word space model, in Proceedings of EVALITA workshop, 11th Congress of Italian Association for Artificial Intelligence, Citeseer, 2009.
- [4] A. Balahur, J. M. Hermida, and A. Montoyo, Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model, Affective Computing, *IEEE Transactions on*, 2012; vol. 3, no. 1, pp. 88101.
- [5] L. Barbosa and J. Feng, Robust sentiment detection on twitter from biased and noisy data, in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 3644, Association for Computational Linguistics, 2010.

- [6] A. Celikyilmaz, D. Hakkani-Tur, and J. Feng, Probabilistic model-based sentiment analysis of twitter messages, in Spoken Language Technology Workshop (SLT), 2010 IEEE, pp. 7984, IEEE, 2010.
- [7] Y. Wu and F. Ren, Learning sentimental influence in twitter, in Future Computer Sciences and Application (ICFCSA), 2011 International Conference on, pp. 119122, IEEE, 2011.
- [8] A. Pak and P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in Proceedings of LREC, 2010; vol. 2010.
- [9] R. Xia, C. Zong, and S. Li, Ensemble of feature sets and classification algorithms for sentiment classification, *Information Sciences: and International Journal*, 2011; vol. 181, no. 6: pp. 11381152.
- [10] M.S. Neethu, R. Rajasree, "Sentiment Analysis in Twitter using Machine Learning Techniques", 4th ICCCNT 2013 July 4 - 6, 2013, Tiruchengode, India, *IEEE*.
- [11] Shenghua Liu, Xueqi Cheng, Fuxin Li, and Fangtao Li, "TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets", *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, MANUSCRIPT*, 2014.
- [12] Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li, "Dual Sentiment Analysis: Considering Two Sides of One Review", *IEEE Transactions on Knowledge and Data Engineering*, 2015.
- [13] R. Suresh Ramanujam, R. Nancyamala, Nivedha, "SENTIMENT ANALYSIS USING BIG DATA", 2015 INTERNATIONAL CONFERENCE ON COMPUTATION OF POWER, ENERGY, INFORMATION AND COMMUNICATION
- [14] Mondher Bouazizi, Tomoaki Ohtsuki, "A Pattern-Based Approach for Sarcasm Detection on Twitter", *IEEE Transactions on Knowledge and Data Engineering*, 2016.