Hybrid Approach for Twitter Sentiment Analysis Using Majority Voting Based Ensemble Technique

Poonam Vashisht[1], Vishal Gupta[2]
UIET, Panjab University, Chandigarh (India)
vashisthpoonam1@gmail.com[1], vishal@pu.ac.in[2]

***Abstract*—** Twitter micro-blogging site is one of the most used platforms by people for expressing the views and opinions, making, twitter as immense dataset for capturing sentiments. In this research paper, we are using ensemble learning method majority voting with classifiers-SVM, Naïve Bayes, Decision Tree, KNN and Random Forest to automatically classify the tweets into the positive and negative sentiment text. The approach can help various users who need sentiments classification techniques for various applications like finding reviews about any movie, place, restaurant, products, etc. for any businessman to monitor the performance of their brands, etc. Our research has used majority voting ensemble learning method with classifiers on different twitter datasets and showed that it has improved the performance of individual classifiers for the sentiment analysis.

***Keywords*—** *Opinion Mining, Sentiment Analysis, Twitter, Machine Learning, Ensemble Method*

I. INTRODUCTION

Twitter is one of the popular micro-blogging sites [1] used by huge population to comment or write about any issue, topic or trend like social issues, political happenings, daily routines of a person, movie, fashion, etc. It provides the best data collection for analyzing views and opinions of people using sentiment classification techniques. In this paper, we are trying to identify tweets with positive and negative polarity by using ensembles of multiple base classifiers like SVM, KNN, NB, Decision Tree, and Random Forest.

The organization of the research paper follows as: -the following segment of the literature survey addresses the related literature works; segment 3, the proposed approach has described and explained the whole approach. Segment 4 provides the result, and the last segment concludes the whole paper.

II. LITERATURE SURVEY

Sentiment analysis or opinion mining is used for finding of the polarity or attitude of the opinion of users towards any subject, object or event by analyzing their comments or reviews on various social media platforms, blogging sites, news sites, etc. Extensive research has been carried out in this area. Mantyla et al. [2] have reviewed on evolution, research topics, venues, etc. regarding the sentiment classification research field. A detailed survey of the basics of sentiment classification – tasks, approaches, and applications has been published by Ravi and Ravi [3]. Another detailed survey is provided by Medhat et al. [4] on SA algorithms and applications. A survey of comparative analysis of existing approaches on the Twitter dataset is covered by Kharade and Sonawane [5].

Most of the research papers on sentiment classification have applied various approaches like machine learning, lexicon-based, hybrid approach, etc. for performing SA [6, 7, 8, 9, and 10]. In machine learning approach many scholars have used single classifiers, and some have combined a number of classifiers to achieve best results for polarity determination. Some of the researchers tend to combine various classifiers by ensemble methods to make the best use of available classifier for a given set of the sentiment classification problem. The combined classifier produces a generalized decision boundary for classification input [11]. Lin & Koltz [12], Rodriguez-Penago et al. [13] have made use of the ensemble method - majority voting in the research work. Clarke et al. [14] have made use of weighted voting ensemble to train NB classifier. Hassan et al. [15] have presented an approach to combining different features and classification parameters. Da Silva et al. [16] has calculated the average of classification output of SVM, NB, Random Forest & LR for finding the final result of SA. Catal & Nangir [17] used meta-classifier, NB and SVM CVparameter selection on majority voting algorithm for SA of Turkish dataset. Foud et al. [18] used IG feature selection method and ensemble model with SVM, NB & LR for SA of twitter dataset. However, it does not assure that the ensemble of classifiers would always provide better results than state-of-the-art classifier but reduce the selection risk of classifiers with poor efficiency.

III. THE PROPOSED APPROACH

A. *Preprocessing Step*

The main objective of this step is to preprocess the data to perform sentiment analysis. The input tweet text is processed by using natural language processing techniques which includes steps naming tokenization, stop word removal, expansion of acronym/abbreviations and stemming. The first step involves the splitting of the input text into tokens. The second step involves stop words removal, further removing or replacing of abbreviations or acronyms using acronym dictionary and stemming is the last step which reduce or derive the words to their word stem, root or base form.

B. *Feature extraction process*

In this process, the processed text is used to extract various features for analyzing the sentiments reflected in the tweets. The different types of features for the proposed system are: BoW-Uni, it contains all the distinct unigrams and the

number of positive & negative words in the tweets are counted as identified by opinion lexicon Liu et al. [19]. The lexicon-based features (Lex_features) include the positive & negative scores of all nouns, adjectives, adverbs, and verbs present in the tweets using SWN [24]. The third set of features is the scores of emoticons (Emo) in the tweets [20]. The statistical features includes two types of ratios: the first one is the ratio of total count of words with positive score i.e. positive word or negative score i.e. negative word present in a tweet to the total count of words in the tweet and the second one is the total score of the positive words or the negative words in the tweet to the total weight of the words of the tweet.

Table 1: Feature set of the proposed method

Features set for the proposed sentiment model	
BoW-Uni	Total count of unigrams in the tweet. Count of the positive and negative words in the tweets
Lexicon features set (Lex_features)	Positive SWN score [24] of the nouns in the tweet. Negative SWN score of the nouns in the tweet. Positive SWN score of the verbs in the tweet. Negative SWN score of the verbs in the tweet. Positive SWN score of the adjectives in the tweet. Negative SWN score of the adjectives in the tweet. Positive SWN score of the adverbs in the tweet. Negative SWN score of the adverbs in the tweet.
Emoticon (smiley) features set (Emo)	Sentiment scores of the emoticons in the tweet [23].
Proposed Statistical features set	The ratio of number of positive or negative words to the number of words in the tweet. The ratio of the score of positive or negative words to the total weight of the words in the tweet.

C. Sentiment Analysis Approach

The proposed sentiment analysis approach is designed on the ensemble learning of five classifiers Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN), the Random Forest and the Decision Tree classifiers using majority voting ensemble methods.

Support Vector Machine

The SVM attempts to find a hyper-plane that separates the classes by providing maximum distance between them.

Naïve Bayes

The working of the classifier is according to the Bayes theorem of probability for the prediction of classes of an unknown dataset with an assumption that one feature present in a class is independent of the other features in the class.

KNN

The KNN classifier predicts the class of an object which is most common among its K-nearest neighbors.

Decision Tree

The decision tree represents a structure where internal nodes hold attributes that most efficiently splits the sample set into subsets of same kinds and each leaf node holds a class label.

Random Forest

This classifier first starts building multiple decision trees and then perform majority voting to find the most suitable and accurate result.

Ensemble method

The ensemble learning approach uses multiple learning algorithms called as base learners. It is done to build a more robust system which includes predictions from all the base learners. Voting ensemble method: It is a meta-algorithm used in decision-making process by application of different rules like majority voting, average voting, maximum voting, etc. Majority voting predicts by considering maximum votes from multiple models predictions while predicting the outcomes of the classification problem.

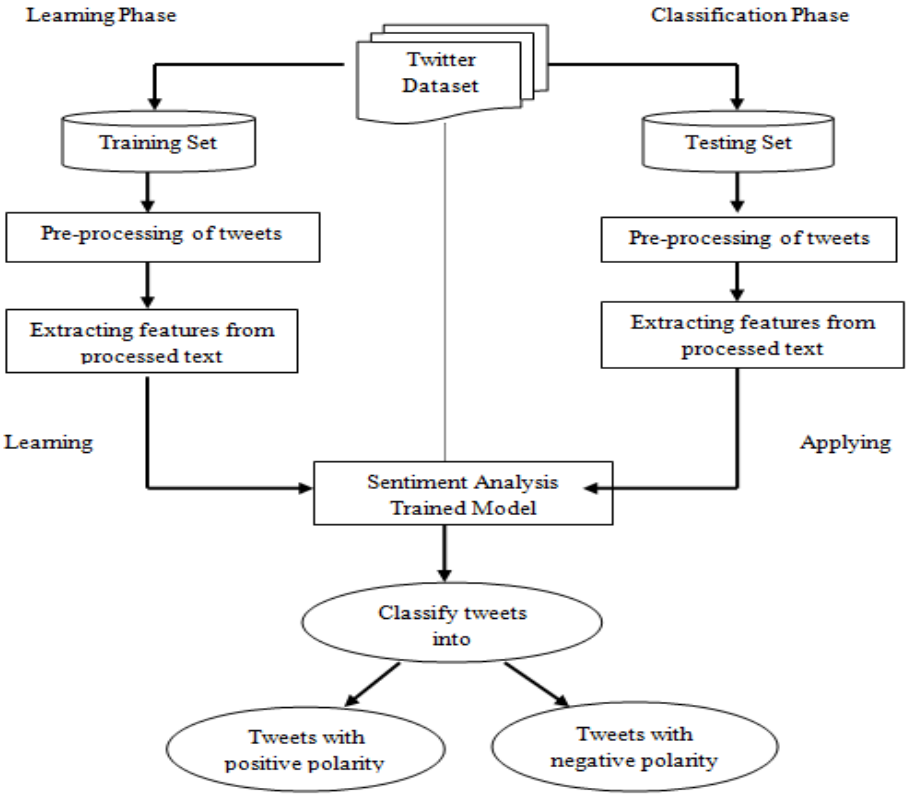

Figure 1: Flow diagram of the proposed sentiment analysis approach

D. Dataset explanation

Stanford Twitter Sentiment Corpus [21] contains positive and negative tweets collected using Twitter API. Due to computational limitation two samples of the dataset are considered with 1000 and 3000 tweets named as Stanford-1K and Stanford- 3K. The Sanders dataset [22] consists of almost 5512 tweets which are manually annotated as positive, negative or neutral and irrelevant concerning the topics. The different topics are Apple, Microsoft, Google, and Twitter. The Health Care Reforms dataset [23] was constructed by crawling of the tweets with the hash-tag \# "hcr" in March 2010 with manual annotations by the authors as positive, negative neutral irrelevant and unsure. The dataset was split into three sets (training, development, and test).

Algorithm of proposed system for sentiment analysis:

Step 1: The data from standard datasets: Stanford twitter dataset, Sanders twitter dataset and HCR dataset are divided into training set and testing set.

Step 2: Perform pre-processing on text from training and testing sets which includes: tokenization, stop word removal, expansion of acronym/abbreviations and stemming.

Step 3: The preprocessed text is used to extract features (BoW-Uni, Lex_features, Emo, and proposed statistical features) for analyzing the sentiments reflected in the tweets.

Step 4: The extracted features are used to train the proposed sentiment analysis model. The proposed model performs in the following manner:

Step 4.1: The features are fed into classifiers: Support Vector Machine (SVM), Naïve Bayes (NB), K- Nearest Neighbor (KNN), Random Forest, and Decision Tree.

Step 4.2: The majority voting ensemble method is then applied on the predictions made by the classifiers.

Step 4.3: The final prediction is made by considering maximum votes from different classifiers.

Step 5: The proposed trained model is applied on the testing set to classify the tweets into positive and negative class.

IV. RESULT

This section presents the result of proposed sentiment analysis system. Table 2 presents the comparison of results for different feature sets. Following feature sets are implemented to achieve the improved performance for the sentiment analysis:

- BoW-Uni: It includes the total count of unigrams in the tweet and also the count of the positive & negative words in the tweets.
- Lexicon features set: It includes the positive and negative SWN score [24] of the nouns, verbs, adjectives and adverbs in the tweet.

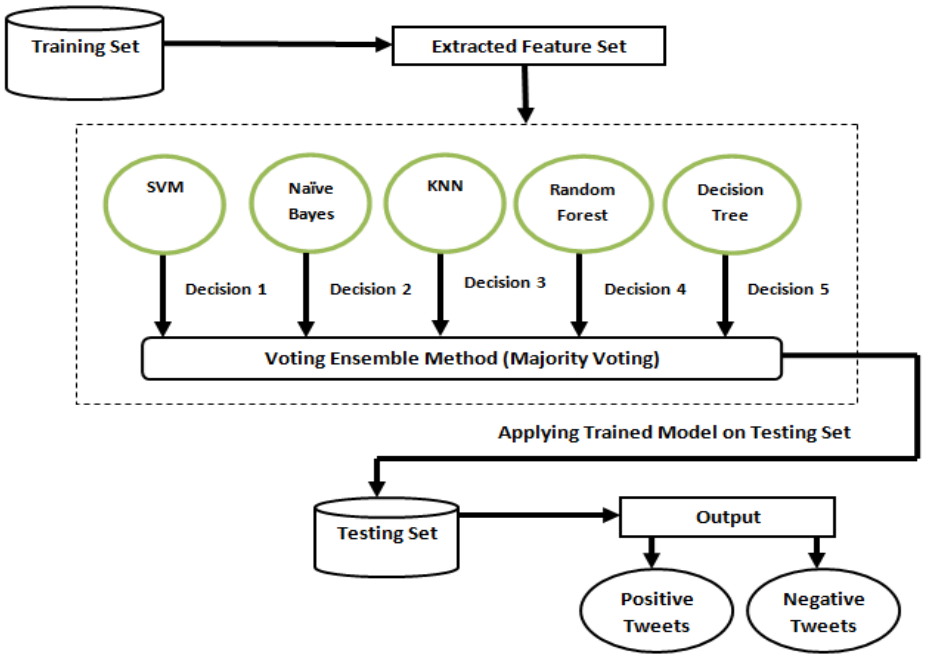

Figure 2: The proposed ensemble learning using majority voting approach for sentiment analysis

- Emoticon (smiley) features set: It includes the scores of the emoticons in the tweet.
- Proposed Statistical feature sets: The two different ratio-based statistical features are proposed here: the first one is the ratio of total count of words with positive score i.e. positive word or negative score i.e. negative word present in a tweet to the total count of words in the tweet and the second one is the total score of the positive words or the negative words in the tweet to the total weight of the words of the tweet. It shows that the proposed statistical feature set + BoW-Uni + Lex_features + Emo have improved the accuracy by 6%. It has performed well for Stanford and HCR datasets.

However, no accuracy improvement is observed for Sanders dataset. Table 3 presents the accuracy comparison results of proposed approach with the classifiers: Support Vector Machine (SVM) , Naïve Bayes (NB), K-Nearest Neighbor (KNN), Random Forest , and Decision Trees on three different datasets namely: Stanford twitter (1K and 3K tweets) dataset, Sanders twitter dataset and HCR dataset. The result in the table shows that theproposed ensemble learning approach has improved accuracy by 6% as compared to state-of-the-art classifiers. The proposed approach has performed significantly better for Stanford and Sanders datasets but a marginal improvement is obtained for HCR dataset.

V. CONCLUSION

This work presents the voting based ensemble learning approach to classify the sentiments in tweets. The proposed work has processed the three different datasets naming: Stanford, sanders and HCR datasets to calculate unigrams and statistical measures. The obtained feature sets are further processed through voting based ensemble learning approach. The majority voting based ensemble learning approach is implemented with Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN), Random Forest, and Decision Tree classifiers.

The proposed approach has classified tweets into positive and negative tweets. However, it does not identify the neutral tweets. The results show that the proposed ensemble learning approach has improved accuracy by 6% as compared to state-of-the-art classifiers. The proposed approach has performed significantly better for Stanford and Sanders datasets but a marginal improvement is obtained for HCR dataset. The proposed statistical feature set + BoW-Uni + Lex_features + Emo have improved the accuracy by 6%. It has performed well for Stanford dataset and HCR dataset in comparison to Sanders Dataset.

The majority voting based ensemble learning approach increases the classification accuracy. However, the proposed approach is tested on small datasets. In future it can be extended to large datasets. The proposed approach does not classify neutral tweets, a mechanism for neutral tweets identification can also be added to the proposed scheme. It is applied only on the English language tweets; one can try ensemble learning models for sentiment analysis on other language tweets also.

Table 2: Accuracy obtained using various features set [18]

	Accuracy (%)			
Feature Set	**Stanford-1K Dataset**	**Stanford-3K Dataset**	**Sanders Dataset**	**HCR Dataset**
BoW-Uni[18]	73.9	76	93.53	84.58
BoW-Uni + Lex_features[18]	77.9	76.53	93.73	83.91
BoW-Uni + Emo[18]	74.5	75.27	93.33	84.75
BoW-Uni + PoS[18]	74	75.6	93.53	84.41
BoW-Uni + Lex_features + Emo[18]	78.7	76.57	93.73	84.75
BoW + Lex + PoS[18]	77.3	77.27	**93.94**	84.41
BoW-Uni + Emo + PoS[18]	74.4	75.63	93.34	84.58
BoW-Uni + Lex_features + Emo + PoS[18]	78.7	77	93.53	84.75
Proposed Statistical features +BoW-Uni + Lex_features +Emo	**86.59**	**93.2**	**92.11**	**85.55**

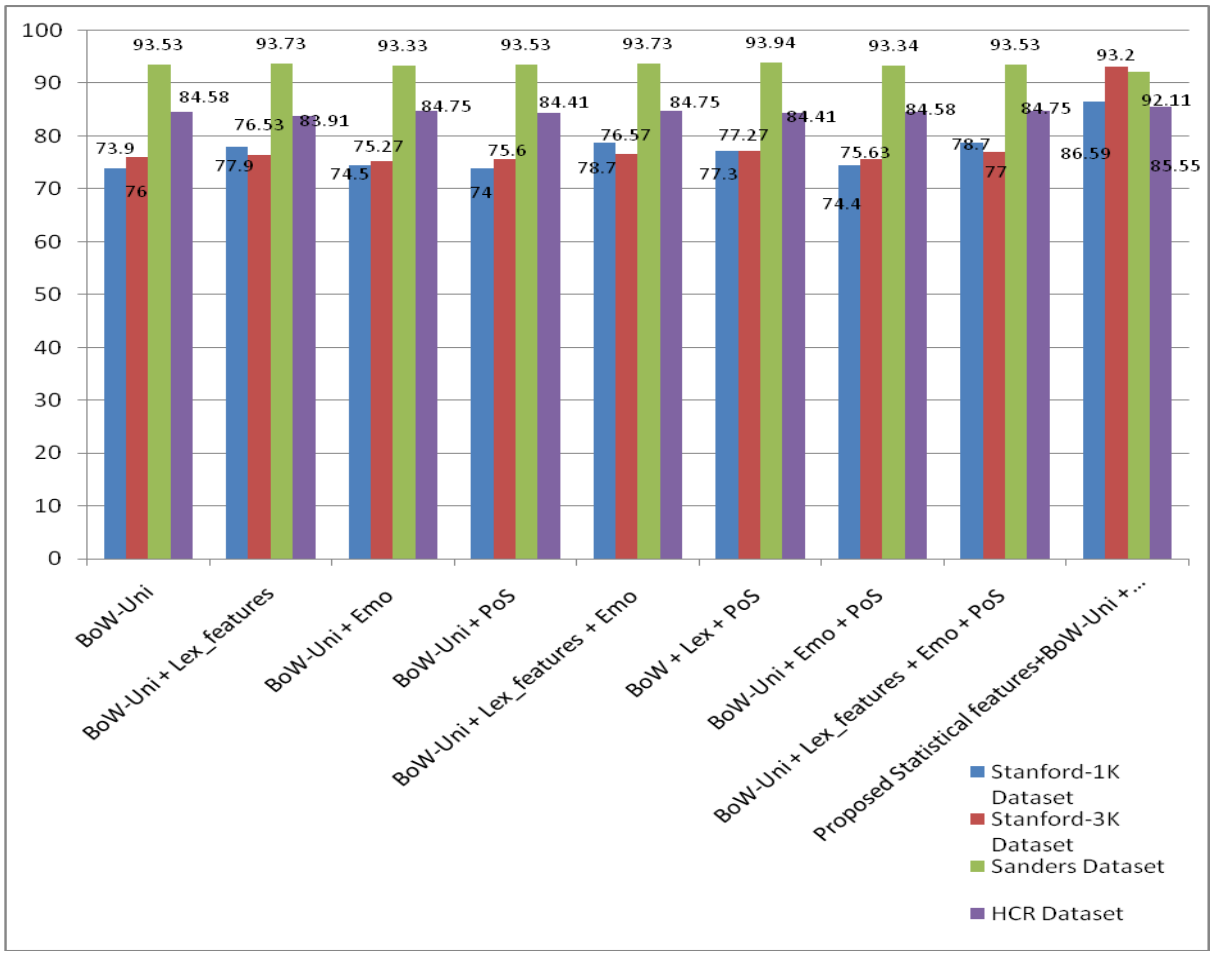

Figure 3: Accuracy (%) result comparison of various features set

Table 3: Classification Accuracy comparison of proposed approach with state-of-the-art classifiers

	Accuracy (%)			
Classifiers	**Stanford-1K Dataset**	**Stanford-3K Dataset**	**Sanders Dataset**	**HCR Dataset**
Decision Tree	61.6	73.56	71.25	69.81
KNN	58.4	70.63	73.63	77.96
Naïve Bayes	62.8	81.36	79.28	74.26
SVM	77.6	86.69	85.36	84.26
Random Forest	68.2	78.63	71.86	77.96
Proposed Ensemble Learning Approach	**86.59**	**93.2**	**92.11**	**85.55**

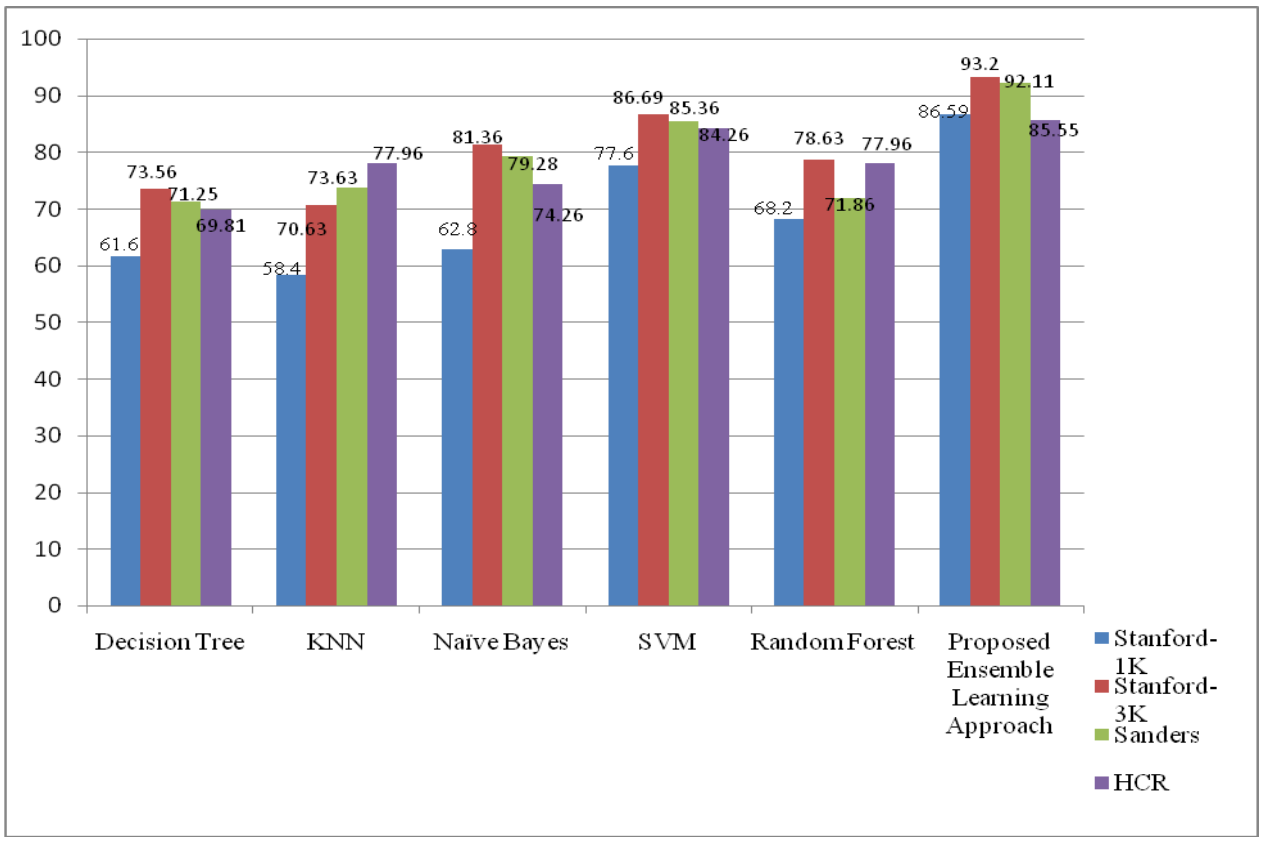

Figure 4: Classification Accuracy (%) comparison of proposed approach with state-of-the-art classifiers

REFERENCES

[1] Ritter, S. Clark, Mausam, O. Etzioni, Named entity recognition in tweets: an experimental study, Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP'11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 1524–1534.

[2] Mäntylä, Mika V., Daniel Graziotin, and Miikka Kuutila. "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers." Computer Science Review 27 (2018): 16-32.

[3] Ravi, Kumar, and Vadlamani Ravi. "A survey on opinion mining and sentiment analysis: tasks, approaches, and applications." Knowledge-Based Systems 89 (2015): 14-46.

[4] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal 5.4 (2014): 1093-1113.

[5] Kharde, Vishal, and Prof Sonawane. "Sentiment analysis of Twitter data: a survey of techniques." arXiv preprint arXiv:1601.06971 (2016).

[6] Suresh, Hima, and S. Gladstone Raj. "A Fuzzy Based Hybrid Hierarchical Clustering Model for Twitter Sentiment Analysis." International Conference on Computational Intelligence, Communications, and Business Analytics. Springer, Singapore, 2017.

[7] Pandey, Avinash Chandra, Dharmveer Singh Rajpoot, and Mukesh Saraswat. "Twitter sentiment analysis using hybrid cuckoo search method." Information Processing & Management 53.4 (2017): 764-779.

[8] Fernández-Gavilanes, Milagros, et al. "Unsupervised method for sentiment analysis in online texts." Expert Systems with Applications 58 (2016): 57-75.

[9] Pollacci, Laura, et al. "Sentiment Spreading: An Epidemic Model for Lexicon-Based Sentiment Analysis on Twitter." Conference of the Italian Association for Artificial Intelligence. Springer, Cham, 2017.

[10] Huang, Arthur, David Ebert, and Parker Rider. "You Are What You Tweet: A New Hybrid Model for Sentiment Analysis." International Conference on Machine Learning and Data Mining in Pattern Recognition. Springer, Cham, 2017.

[11] Kuncheva, Ludmila I. Combining pattern classifiers: methods and algorithms. John Wiley & Sons, 2004.

[12] Lin, Jimmy, and Alek Kolcz. "Large-scale machine learning at twitter." Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012.

[13] Rodríguez-Penagos, Carlos, et al. "FBM: Combining lexicon-based ML and heuristics for Social Media Polarities." Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Vol. 2. 2013.

[14] Clark, Sam, and Rich Wicentwoski. "Swatcs: Combining simple classifiers with estimated accuracy." Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Vol. 2. 2013.

[15] Hassan, Ammar, Ahmed Abbasi, and Daniel Zeng. "Twitter sentiment analysis: A bootstrap ensemble framework." Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013.

[16] Da Silva, Nadia FF, Eduardo R. Hruschka, and Estevam R. Hruschka Jr. "Tweet sentiment analysis with classifier ensembles." Decision Support Systems 66 (2014): 170-179.

[17] Catal, Cagatay, and Mehmet Nangir. "A sentiment classification model based on multiple classifiers." Applied Soft Computing 50 (2017): 135-141.

[18] Fouad, Mohammed M., Tarek F. Gharib, and Abdulfattah S. Mashat. "Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble." International Conference on Advanced Machine Learning Technologies and Applications. Springer, Cham, 2018.

[19] Liu, Bing, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web." Proceedings of the 14th international conference on World Wide Web. ACM, 2005.

[20] Stanford Twitter Sentiment Corpus. http://help.sentiment140.com/for-students. Accessed May 2018

[21] Sanders Dataset. http://www.sananalytics.com/lab/. Accessed May 2018

[22] Health Care reforms dataset https://bitbucket.org/speriosu/updown. Accessed May 2018

[23] Hogenboom, Alexander, et al. "Exploiting emoticons in sentiment analysis." Proceedings of the 28th annual ACM symposium on applied computing. ACM, 2013.

[24] Senti Word Net Lexicon http://sentiwordnet.isti.cnr.it/. Accessed May 2018