Sentiment Classification of Twitter Data: A Review

Paridhi Pravin Nigam[1], Prof. Dinesh D. Patil [2], Prof. Yogesh S. Patil[3]

[1]P.G. Student, Department of Computer Engineering, Shri Sant Gadge Baba College of Engineering & Technology, Near Z.T.C., Bhusawal - 425203, Maharashtra, India.
[2]Associate professor & Head, Department of Computer Engineering, Shri Sant Gadge Baba College of Engineering & Technology, near Z.T.C., Bhusawal - 425203, Maharashtra, India.
[3]Assistant professor, Department of Computer Engineering, Shri Sant Gadge Baba College of Engineering & Technology, near Z.T.C., Bhusawal - 425203, Maharashtra, India.

---***---

Abstract - *Twitter is a popular microblogging service where users create tweets which sometimes express opinions about different topics. Sentiment analysis of twitter data is useful for companies that want to monitor the public sentiment of their brands also for consumers who want to research the sentiment of products before purchase. Different approaches for finding sentiments have their own advantages and disadvantages. There are number of different methods for finding sentiment of twitter data. Thus this paper is an overview of works done by the researchers in the area of twitter sentiment analysis.*

KeyWords: Twitter, Sentiment classification, Training dataset, supervised learning, Classifiers.

1. INTRODUCTION

A lot of work has been done in the field of Twitter sentiment analysis till date. Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity [4]. Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral" [10]. Most of these techniques use Machine Learning algorithms with features such as unigrams, n-grams, Part-Of-Speech (POS) tags [9]. However, the training datasets are often very large, and hence with such a large number of features, this process requires a lot of computation power and time. Supervised learning is based on labeled dataset and thus the labels are provided to the model during the process. These labeled dataset are trained to get meaningful outputs when encountered during decision-making. The success of both of these learning methods mainly depends on the selection and extraction of the specific set of features used to detect sentiment [2]. The machine learning approach applicable to sentiment analysis mainly belongs to supervised classification. In a machine learning techniques, two sets of data are needed: Training Set and Test Set [5]. A number of machine learning techniques have been formulated to classify the tweets into classes. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) have achieved great success in sentiment analysis. Machine learning starts with collecting training dataset. Next we train a classifier on the training data. Once a supervised classification technique is selected, an important decision to make is to select feature [11]. They can tell us how documents are represented. Thus the method of sentiment analysis is almost same for different strategies. An example of sentiment analysis is an advice of items proposed by a recommendation system can be predicted by taking into account considerations such as positive or negative opinions about those items.

This paper is organized as follows. We discuss related work in Section 2. After analyzing the researches done we will make a review of these theories and implementations in Section 3. In Section 4 we conclude the paper.

2. RELATED WORK

The micro-blogging tool Twitter is well-known and increasingly popular. Twitter allows its users to post messages, or 'Tweets' of up to 140 characters each time, which are available for immediate download over the Internet. Tweets are extremely interesting to marketing since their rapid public interaction can either indicate customer success or breakage. Consequently, the content of tweets and identifying their sentiment polarity as positive or negative is a most trending topic.

Jiang et al. [2011] was the first to propose targeted Sentiment analysis on Twitter, who emphasize the importance of targets by showing that 40% of sentiment analysis errors are caused by not considering them in classification [1]. They incorporate 7 rule-based target-dependent features into a model with traditional target independent Sentiment analysis features, which give a significant improvement. [7] Further, Mitchell et al. [2013] apply a sequence labeling model to simultaneously detect entities and predict opinions towards them.

Pak and Paroubek [13] proposed a model to classify the tweets as objective, positive and negative. They created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Using that corpus, they developed a sentiment classifier based on the multinomial Naive Bayes method that uses features like N-gram and POS-tags. The training set they used was less efficient since it contains only tweets having emoticons.

Alec Go et al [12] introduced a completely unique approach for mechanically classifying the sentiment of Twitter messages. These messages area unit classified as either positive or negative with relation to a question term. This is often helpful for customers United Nations agency need to analysis the sentiment of product before purchase, or firms that need to watch the general public sentiment of their brands. There's no previous analysis on classifying sentiment of messages on micro blogging services like Twitter.

Go et al. (2009) used distant supervision to classify sentiment of Twitter [3]. Emoticons have been used as noisy labels in training data to perform distant supervised learning (positive and negative). Three classifiers were used: Naïve Bayes, Maximum Entropy and Support Vector Machine, and they were able to obtain more than 80% accuracy on their testing data.

Aisopos et al. (2011) divided tweets in to three groups using emoticons for classification. If tweets contain positive emoticons, they will be classified as positive and vice versa. Tweets without positive/ negative emoticons will be classified as neutral. However, tweets that contain both positive and negative emoticons are ignored in their study. Their task focused on analyzing the contents of social media by using n-gram graphs, and the results showed that n-gram yielded high accuracy when tested with C4.5, but low accuracy with Naïve Bayes Multinomial (NBM) [6].

One of the earliest works which used supervised method to solve sentiment classification problem is [14]. In this paper, authors used three machine learning techniques to classify sentiment of movie review documents. To implement these machine learning techniques on movie review documents, they used the standard bag of features frame work. They test several features to find optimal feature set [9]. Unigrams, bigrams, adjective and position of words were used as features in these techniques. The results show that the best performance is achieved when the unigrams are used in SVM classifier. Overall, text classification using machine learning is a well studied field (Manning and Schuetze 1999). (Pang and Lee 2002) researched the effects of various machine learning techniques Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) in the specific domain of movie reviews. They were able to achieve an accuracy of 82.9% using SVM and a unigram model [14].

J. Read in (Read, 2005) used emoticons such as ":-)" and ":-(" to form training set for the sentiment classification. For this purpose, the author collected texts containing emoticons from Usenet newsgroups [8]. The dataset was divided into "positive" (texts with happy emoticons) and "negative" (texts with sad or angry emoticons) samples. Emoticon constrained classifiers: SVM and Naive Bayes were able to obtain up to 70% of accuracy on the test set [8].

Luoet. al. [15] highlighted the challenges and efficient techniques to mine opinions from Twitter tweets. Spam and wildly varying language makes opinion retrieval within Twitter challenging task.

Thus after studying various strategies implemented for finding sentiment of twitter data, we can thus find the accuracy of these methods in order to find the better one at a particular situation.

3. REVIEW RESULT

Table 1: Selected previous work done in sentiment classification of data

PAPER	METHOD	FEATURES	DATASET	TYPE
[1]	Naive Bayes, SVM, Maximum Entropy	Unigram, Bigram, POS, EFWS, Subjectivity.	Twitter review	Distant Supervised
[2]	K-NN classifier	BOW, Corpus, Porter Stemmer rule.	Twitter data	Supervised
[3]	Naïve Bayes, Maximum Entropy and SVM	Emoticons	Twitter review	Distant supervised
[8]	SVM and Naive Bayes	Emoticons	Twitter review	Supervised
[13]	Naïve Bayes	Emoticons, Corpus, N-gram, POS tags	Twitter review	Supervised
[14]	SVM ,Naive Bayes, Maximum Entropy	Unigrams, bigrams, bag of features	Movie review	Supervised

Through table 1 we can review the previous work done on sentiment classification of data and its overall performance thus predicting the best performance done by which classifier and the features used, thus reducing the work for classifier. Hence further dividing into types as- supervised learning is the data mining task of inferring a function from labeled training data while unsupervised learning is that of trying to find hidden structure in unlabeled data.

4. CONCLUSION

Sentiment analysis has many applications in information systems, including review classification, summarization, benefit to customers from feedbacks, fall down and rise of company structure, opinions tracking in online discussions and etc. This paper tries to make a bunch of all researches till now done on this topic. Also the innovations grouped in different concepts in order to categorize research done using which strategy to understand it better. Further, more research is needed to improve methods and techniques introduced in this paper.

ACKNOWLEDGEMENT

I feel great pleasure in submitting this paper on **"Sentiment Classification of Twitter Data: A Review"**. I wish to express true sense of gratitude towards my Principal **Dr. R. P. Singh** and special thanks to my guide **Prof. Yogesh S. Patil** and **H.O.D.**, **Prof. Dinesh D. Patil** who at every discrete step in preparation of this paper contributed his valuable guidance and help to solve every problem that arose. Also I thank all the researchers and authors of reference papers referred by us. This paper is a review of various researches done in this field. Also, most likely I would like to express my sincere gratitude towards my parents for always being there when I needed them the most. With all respect and gratitude, I would like to thank all the people, who have helped me directly or indirectly. I owe my all success to them.

REFERENCES

[1] Tapan Sahni, Chinmay Chandak, Naveen Reddy Chedeti, Manish Singh "Efficient Twitter Sentiment Classification using Subjective Distant Supervision", 2017 IEEE 9th International Conference on Communication Systems and Networks (COMSNETS), 548-553.

[2] Paridhi Pravin Nigam , Dinesh D. Patil " Twitter sentiment classification using supervised lazy learning method", International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE) Vol. 6, Issue 6, June 2018. Pp-6230-6235.

[3] Alec Go, Richa Bhayani, and Lei Huang. "Twitter Sentiment Classification using Distant Supervision" CS224N Project Report, Stanford, pages 1-12. 2009.

[4] Bing Liu. Sentiment Analysis and Subjectivity. In Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boc. 2010.

[5] https://en.wikipedia.org/wiki/Training,_test,_and_validation_sets

[6] N.Saranya, Dr. R.Gunavathi "A Study on Various Classification Techniques for Sentiment Analysis on Social Networks" International Research Journal of Engineering and Technology (IRJET), pp- 1332-1337.

[7] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification", in Proceedings of the Association for Computational Linguistics: Human Language Technologies – Vol.1, Portland, Oregon, 2011, pp. 151–160.

[8] J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2005.*

[9] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 3644. 2010*

[10] Ayushi Dalmia, Manish Gupta*, Vasudeva Varma. Twitter Sentiment Analysis The good, the bad and the neutral! *IIIT-H at SemEval, 2015.*

[11] Bhawna Nigam "Document Classification Using Expectation Maximization with Semi Supervised Learning" International Journal on Soft Computing (IJSC) Vol.2, No.4,pp- 37-44, November 2011.

[12] Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(2009), p.12.

[13] A.Pak and P. Paroubek., "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326.

[14] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, 2002.

[15] Vishal A. Kharde , S.S. Sonawane," Sentiment Analysis of Twitter Data: A Survey of Techniques" International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016, pp-5-15.