LITERATURE REVIEW ON SENTIMENT ANALYSIS OF TWITTER DATA

Sumneet Kaur[1], Aman Puri[2], Zorawar Singh Jaiswal[3], Yashi Jain[4]
[1,2,3] Department of Electronics and Communication Engineering
[4]Department of Computer Science and Engineering
Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi

Abstract : Sentiment analysis is an evolving field of study which involves the process of evaluating and distinguishing the opinions or emotions expressed in a given text. Twitter promotes unregulated communication by providing an easily accessible medium where millions of people tweet everyday to contribute their thoughts and viewpoints to the world. This paper presents a review on the techniques of Sentiment Analysis on the Twitter Data. With this paper, we present a brief review of all the work done on twitter sentiment analysis so far and elaborate the models and their methodologies used. We have surveyed all the papers published in this field and focused on the recent approach so as to facilitate the development of promising avenues of future projects and research.

Index Terms **- sentiment analysis, twitter data**

I. INTRODUCTION

Any written sentence can be broadly classified into positive, negative or neutral. Context, tone, emotion, etc plays an important role in determing the view point of the writer. It is a basic human psychology to get influenced by the love or hatred we receive on social media. Sentiment analysis is a self-explanatory term which means to analyse the sentiments of the texts. Sentiment analysis on a large data set helps in concluding a general public opinion which can be sometimes be used to analyse customer feedback. Sentiment Analysis is carried at 3 levels: document, sentence and aspect. People these days are in a habit of making short and frequent posts on microblogging sites. Twitter is an open platform where people are allowed to express their particular beliefs and emotions. We have targeted Twitter due to a number of reasons, the major one being its vast audience. More than 300 million people use Twitter every month. Marketers view Twitter and other social media sites as a great opportunity to reach out to their customers. Other reasons to choose Twitter are its unbiased and unambiguous nature.

In order to help them classify all the feedback and opinions that the people have expressed, sentiment analysis is essential. The procedure for sentiment analysis involves the following major steps:

A. Data Collection and Cleaning

To obtain the tweets, we need to first get Twitter API access and necessary keys. After the necessary data has been obtained, we procced to the first step of data classification which is data cleaning and converting it into the useful format. This is important because the tweets have a lot of noise factor. Since the tweets have a restriction of maximum number of characters, people tend to use slang words or mix languages which lead to unfavorable dataset. The preprocessing techniques include: Tokenisation and Removal of non-English Tweets, URL, targets, stop words and hashtags [1]. Once the data has been cleaned up, we perform its conversion into a data frame.

B. Common learning algorithms

After we have the required dataset, we need to follow a sentiment analysis technique, [2] states that there are two major sentiment analysis methodologies, one is machine learning and the other is Lexicon-based Approach. The former approach is based on supervised classification algorithm. The later one involves keeping a threshold value which varies with respect to the polarity of the tokens. There are a lot of ways to obtain the polarity of sentence: Natural Language Processing (NLP), Support Vector Machine (SVM), Case-Based Reasoning (CBR), Artificial Neural Network (ANN).

This paper covers the comparison and analysis of all the research and methodologies that have been used to implement sentiment analysis on Twitter data in the pastdecade.

REVIEW OF SENTIMENTANALYSIS ON TWITTER DATA

It is of no surprise that Twitter can also be responsible for commendation or defamation of a brand or company since it is very convenient for users to post their personal liking and preference in the form of online reviews. Bernard *et al.*[3] showed in their research how the sentiments of people fluctuate from week to week and the struggle of brands in maintaing a positive image in front of its potential customers. A simple supervised machine learning approach was developed by Ted [4] to break down meaningless words and provide cleaner datasets.

Varsha *et al.*[5] proposed the use of Parts of Speech (POS)- specific prior polarity feature. They also introduced the tree kernel model based methodology in their paper in order to remove the repetititive features. Tony *et al.* [6] suggested to use support vector machines (SVMs) in order to obtain an efficient system of sentiment analysis. The results show that indeed a hybrid SVM yields a score with better accuracy. On a side note, we may also conclude that through their research it was evident that adding Osgood values didn't do anything good to the performance score but introducing Turney value did actually help the result accuracy.

In the paper "Sentiment analysis on Twitter data" [7], Apoorv *et al.* introduced new features and experimented combinations of various models which were: Unigram model, Tree kernel model, 100 Senti- features model, Kernel plus Senti-features and

Unigram plus Senti-features and compared the accuracy score in each case over a provided data set. Their results proved that using tree kernel and feature based models perform better than the unigram baseline.

G. Vinodhini and RM. Chandrasekaran [8] focused on the challenges and problems prevelant in this field along with a comparison of analysis done on movie reviews and product reviews. They had also concluded that most of the researchers prefer to use the movie reviews dataset but it's not right to judge which dataset will give better performance result.

A recent work on Twitter movie review sentiment analysis has been done by Kiruthika*et al.* [9]. They extracted the twitter data using the traditional method of twitter API after building the required application on the developer site. Thereafter, they performed a sentiment analysis of Twitter data about movies using supervised learning approach.

They used feature based opinion mining approach to analyse various aspects of movie reviews on twitter. They extracted twitter data of six movies from the Twitter API, preprocessed it and then applied various models on the same. Hence a system of supervised learning and POS tagger was proposed which weighed the sentiment orientation of tweets which reviewed those movies.

Changhua*et al.* [10] used support vector machine (SVM) and conditional random field (CRF) to explore the emotion classification. After training the classifiers with the common emotion words, they presented their results in the form of precision, recall and F-Score. Their research was carried at the document level as well as the sentence level. In the later one, they compared the performance of CRF classifier and that of the SVM classifier. The results showed that CRF outperformed the SVM. Another interesting finding of this research was that the emotion conveyed in the last sentence generally described the entire emotion of document which meant that people usually like to conclude their opinionated text in such a form that it addressed their real motive of writing the entirepiece.

A problem of mixed reviews was expressed in the paper represented by Kushal *et al.* [11]. The author expressed his concern over the classification of reviews which contain both positive and negative sentiments. These reviews often end up reducing the performance score due to incorrect categorization. They also concluded that Amazon reviews are likely to give better results than twitter reviews when applied under the same machine learning algorithm for sentiment prediction because their length is comparatively longer.

Akshi and Teeja [12] identified the aggregation of the positive/negative values of the sentences by combining two methods, namely corpus based (for adjectives) and dictionary based (for verbs and adverbs). Ankit *et al.*[13] suggested that SVM outperforms ANN in text categorization. They have expressed the need to use supervised algorithm to analyse the sentiments instead of the Vector Quantization which comes under the category of unsupervised algorithm. Neelima and Ela [14] proposed the use of Bayes Classifier and Maximum Entropy classifier in Twitter Sentiment analysis and made a comparison between the two results.

Bhumika *et al.* [15] successfully compared the accuracies of the following models: DAN2, SVM, Bayesian Logistic Regression, Naïve Bayes, Random Forest Classifier, Neural Network, Maximum Entropy and Ensemble classifier. The last two classifiers gave the highest performance rate. They also concluded that the efficiency of classifier is inversely proportional to the number of classes made.

CONCLUSION

Broadly, there are three techniques for sentiment classification, namely machine learning approach, lexicon based approach and hybrid approach. The hybrid approach is a combination of both the machine learning and lexicon methodology. From the above discussion, it is conspicuous that the best way to perform the sentiment analysis is the hybrid approach. For small datasets, Naive Bayesian work perfectly fine as well. NLP gives enhanced results when compared to Naïve Bayes. HMM (Hidden Markov Model) which is also good for text analysis. The best results are however obtained only when we use ensemble methods which involve clubbing of multiple classifiers together. The dependencies of the analysis remains on the context of the topic being explored. For different objectives, different types of approaches are required. Twitter sentiment analysis holds a wide scope of improvement in terms of its performance and accuracy rate.

S. no.	Year of publication	Author's Name	Main finding
1.	2014	Ayushi Dalmia *et al.*	Twitter sentiment analysis using end to end system at phrase level and message level.
2.	2014	Aliza Sarlan and Shuib Basri	Customer's positive and negative comments were represented in a pie chart and html page
3.	2009	Bernard J. Jansen *et al.*	They showed how the sentiments of customers fluctuate from week to week
4.	2002	Ted Pederson	An approach to break down meaningless words was developed
5.	2015	Varsha *et al.*	The use of Parts of Speech (POS)-polarity feature was addressed
6.	2004	Tony *et al.*	The paper suggested the use of support vector machines (SVMs) in order to obtain an efficient system of sentiment analysis.
7.	2011	Apoorv *et al.*	Their results proved that using tree kernel and feature based models perform better than the unigram baseline.
8.	2012	G. Vinodhini and RM. Chandrasekaran	A comparison of analysis was done on movie reviews and product reviews.
9.	2016	Kiruthika*et al.*	A system of supervised learning and POS tagger was proposed which weighed the sentiment orientation of tweets which reviewed those movies.
10.	2007	Changhua*et al.*	A comparison was made in the performance of CRF classifier and that of the SVM classifier and the results showed that CRF outperformed the SVM.
11.	2003	Kushal *et al.*	Retrieval techniques for feature extraction were developed.
12.	2012	Akshi and Teeja	Semantic orientation determination using corpus and dictionary based methodologies was made.
13.	2017	Ankit *et al.*	Preprocessing and classification of tweets for sentiments.
14.	2015	Neelima and Ela	Multilingual text analysis was performed.
15.	2017	Bhumika *et al.*	Use of Machine Learning on Python for twitter sentiment analysis.

REFERENCES

[1] Ayushi Dalmia, Mayank Gupta, Arpit Kumar Jaiswal, Chinthala Tharun Reddy, "Sentiment Analysis in Twitter", 2014

[2] Aliza Sarlan, Shuib Basri, "Twitter sentiment analysis", International Conference on Information Technology and Multimedia Nov 18 – 20, 2014

[3] Bernard J. Jansen, Mimi Zhang, Kate Sobel, Abdur Chowdury, "Micro-blogging as Online Word of Mouth Branding", CHI 2009, April 4–9, 2009

[4] Ted Pederson, "A Baseline Methodology for Word Sense Disambiguation", Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, pp 126- 135, 2002

[5] Varsha Sahayak, Vijaya Shete, Apashabi Pathan "Sentiment Analysis on Twitter Data", International Journal of Innovative Research in Advanced Engineering, Issue 1, Volume 2, Jan 2015

[6] Tony Mullen, Nigel Collier, "Sentiment analysis using support vector machines with diverse information sources" In Proc. EMNLP. 412-418, 2004

[7] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau "Sentiment Analysis of Twitter Data", Proceedings of the Workshop on Languages in Social Media, Page 30-38, June 23 - 23, 2011

[8] G. Vinodhini, RM. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012

[9] Kiruthika M., Sanjana Woona, Priyanka Giri, "Sentiment Analysis of Twitter Data", International Journal of Innovations in Engineering and Technology, 2016

[10] Changhua Yang, Kevin Hsin-Yih Lin, Hsin- Hsi Chen, "Emotion Classification Using Web Blog Corpora", International Conference on Web Intelligence, 2007

[11] Kushal Dave, Steve Lawrence, David M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", Proceedings of the 12th international conference on World Wide Web Pages 519-528, May 20 - 24, 2003

[12] Akshi Kumar, Teeja Mary Sebastian, "Sentiment Analysis on Twitter", International Journal of Computer Science Issues, Vol. 9, Issue 4, July 2012

[13] Ankit Pradeep Patel, Ankit Vithalbhai Patel, Sanjaykumar Ghanshyambhai Butani, Prashant B. Sawant, "Literature Survey on Sentiment Analysis of Twitter Data using Machine Learning Approaches", International Journal for Innovative Research in Science & Technology, Volume 3, Issue 10, March 2017

[14] Neelima, Dr. Ela Kumar, "Indisent Analysis in Twitter using Machine Learning Methods", International Journal of Innovative Research in Computer and Communication Engineering,Vol. 3, Issue 7, July 2015

[15] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, "Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python", International Journal of Computer Applications Volume 165, May 2017