JES Journal of Engineering Sciences

Vol 13, Issue 04, APRIL /2022

ISSN NO:0377-9254

Machine Learning Based Algorithms On Tweets For Analysing Women Safety In Indian Cities

Bantu Sreemani Subhash[1], M.V. Krishna Vamsi[2], Y. Rewanth[3], Kondapureddy Sairam Reddy[4], G. Karthika[5]

#1##2,#3,#4, Student, Department of CSE, GITAM (deemed to be University), Gandhi nagar Rushikonda Visakhapatnam530045 Andhra Pradesh, INDIA

#5 Assistant Professor, Department of CSE, GITAM (deemed to be University), Gandhi nagar Rushikonda Visakhapatnam530045 Andhra Pradesh, INDIA

ABSTRACT_ Nowadays ladies are experiencing a lot of violence such as harassment in locations in numerous cities. This begins from stalking which then leads to abusive harassment or additionally known as abuse assault. In this paper we in general center of attention on the position of social media which can be used to promote the protection of ladies in India, given the one of a kind reference to the participation of many social media web sites or functions such as Twitter, Facebook and Instagram platforms. This paper additionally focuses on growing the obligations amongst the frequent humans on the a number of components of Indian cities so that the protection of ladies round them is ensured. Tweet on the Twitter software consists of the textual content messages, audio data, video data, images, smiley expressions and hash-tags. This tweet content material can be used to examine amongst the humans and hence can teach them in order to take strict movements if tweets are abusive to girls and subsequently can punish such human beings if the harassment is made. Applications which consist of hash-tags, such as Twitter and Instagram, can be used to unfold the messages throughout the whole globe and make the ladies sense free to specific their views and feelings. By this we can be aware of the nation of their idea when they go out for work or journey in a public transportation or surrounded by means of nameless guys and whether or not it feels they are impervious or not..

1.INTRODUCTION

Twitter in this cutting-edge generation has emerged as a last microblogging social community consisting over hundred million customers and generate over 5 hundred million messages acknowledged as 'Tweets' each day. Twitter with such a large target market has magnetized customers to emit their standpoint and judgemental about each current difficulty and subject matter of internet, consequently twitter is an informative source for all the zones like institutions, groups and organizations. On the twitter, customers will share their opinions and standpoint in the tweets section. This tweet can solely include one hundred forty characters, for this reason making the customers to compact their messages with the assist of abbreviations, slang, shot forms, emoticons, etc. In addition to this, many human beings categorical their opinions by using the use of polysemy and sarcasm also. Hence twitter language can be termed as the unstructured. From the tweet, the sentiment in the back of the message is extracted. This extraction is finished by using the usage of the sentimental evaluation procedure. Results of the sentimental evaluation can be used in many areas like sentiments related to a unique manufacturer or launch of a product, examining public opinions on the authorities policies, humans ideas on women, etc. In order to operate classification of tweets and analyze the outcome, a lot of find out about has been accomplished on the records acquired with the aid of the twitter. We additionally evaluate some research on desktop mastering in this paper and lookup on how to operate sentimental evaluation the usage of that area on twitter data. The paper scope is confined to laptop mastering algorithm and models. Staring at female and passing feedback can be positive sorts of violence and harassments and these practices, which are unacceptable, are normally regular particularly on the phase of city life. Many researches that have been carried out in India suggests that girls have pronounced sexual harassment and different practices as

mentioned above. Such research have additionally proven that in famous metropolitan cities like Delhi, Pune, Chennai and Mumbai, most female sense they are risky when surrounded by way of unknown people. On social media, human beings can freely specific what they experience about the Indian politics, society and many different thoughts. Similarly, female can additionally share their experiences if they have confronted any violence or sexual harassment and this brings harmless humans collectively in order to stand up in opposition to such incidents. From the evaluation of tweets textual content series bought through the twitter, it consists of names of humans who has burdened the ladies and additionally names of girls or harmless humans who have stood towards such violent acts or unethical behaviour of men and consequently making them uncomfortable to stroll freely in public. The facts set of the tweet will be used to system the computing device gaining knowledge of algorithms and models. This algorithm will function smoothening the tweet facts through doing away with zero values. Using Laplace and porter's theory, a technique is developed in order to analyze the tweet records and cast off redundant facts from the facts set. Huge numbers of human beings have been attracted to social media platform such as Twitter, Facebook, Instagram. People categorical their sentiments about society, politics, women, and so on by using the textual content messages, emoticons and hash-tags thru such platforms. There are some strategies of sentiment that can be categorized like desktop leaning based totally and lexicon based totally learning.

2.LITERATURE SURVEY

2.1 Barbosa, Luciano, and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data." Proceedings of the 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, 2010.

In this research, we offer a method for automatically detecting feelings in Twitter messages (tweets) that takes into account specific features of how tweets are written as well as meta-information about the words that make up these messages. In addition, we use sources of noisy labels as training data. A few sentiment detection websites provided these noisy labels based on twitter data. Our investigations show that because our features can capture a more abstract representation of tweets, our method is more effective than prior ones and also more robust when dealing with skewed and noisy data, which is what these sources deliver.

2.2 Agarwal, Apoorv, Fadi Biadsy, and Kathleen R. Mckeown. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009

We present a classifier to predict contextual polarity of subjective phrases in a sentence. Our approach features lexical scoring derived from the Dictionary of Affect in Language (DAL) and extended through WordNet, allowing us to automatically score the vast majority of words in our input avoiding the need for manual labeling. We augment lexical scoring with n-gram analysis to capture the effect of context. We combine DAL scores with syntactic constituents and then extract ngrams of constituents from all sentences. We also use the polarity of all syntactic constituents within the sentence as features. Our results show significant improvement over a majority class baseline as well as a more difficult baseline consisting of lexical n-grams.

2.3 Bermingham, Adam, and Alan F. Smeaton. "Classifying sentiment in microblogs: is brevity an advantage?." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

Microblogging has become a popular method for Internet users to publish thoughts and information in real-time. Automated sentiment analysis of microblog posts is of interest to many, allowing monitoring of public sentiment towards people, products and events, as they happen. The short length of microblog documents means they can be easily published and read on a variety of

platforms and modalities. This brevity constraint has led to the use of nonstandard textual artefacts such as emoticons and informal language. The resulting text is often considered "noisy". It is reasonable to assume that the short document length introduces a succinctness to the content. The focused nature of the text and higher density of sentiment-bearing terms may benefit automated sentiment analysis techniques. On the other hand, it may also be that the shorter length and language conventions used mean there is not enough context

3.PROPOSED SYSTEM

In this paper author is describing concept to analyse women safety using social networking messages and by applying machine learning algorithms on it. Now-a-days almost all peoples are using social networking sites to express their feelings and if any women feel unsafe in any area then she will express negative words in her post/tweets/messages and by analysing those messages we can detect which area is more unsafe for women's.
In propose work author using TWEEPY package from python to download tweets from twitter but every time INTERNET will not available to download tweets online so we downloaded MEETOO tweets on women safety and safe inside dataset folder. Application will read this tweets to detect women's sentiments.
Author using NLTK (natural language tool kit) to remove special symbols and stop words from tweets and to make them clean.
Author using TEXTBLOB corpora package and dictionary to count positive, negative and neutral polarity and the tweets which has polarity value less than 0 will consider as negative as and greater than 0 and less than 0.5 will consider as neutral and polarity greater than 0.5 will consider as positive.

1) Data extraction: First step involved in analysis of sentiment is the collection of information from the social network website like twitter. This helps in extracting the tweet message but this message also includes extra data like tweets likes, dislikes and comments.

2) Text Cleaning: Once the data is extracted from the twitter source as the datasets, this information has to be passed to the classifier. The classifier cleans the dataset by removing redundant data like stop words, emoticons in order to make sure that non textual content is identified and removed before the analysis.

3) Sentiment Analysis: After the classifier cleans the dataset, the data is ready for the sentimental analysis process. Machine learning and Lexicon based learning and Hybrid learning are some of the approaches of sentimental analysis. There are also some other approaches such as Nero Linguistic Programming and Natural Language Processing. Training the dataset and then testing that trained dataset involves in machine learning approach. Training data and Testing data are useful for the classifier to perform the algorithm. Maximum Entropy, Naives Bayes classification, Bayesian Networks and Network Support Vector Machine are some of the algorithm which can be used to train the classifier. Testing data is used to identify the efficiency of the sentiment classifier. In case of Lexicon based leaning, training dataset is not used. This approach uses a built-in dictionary in which words associated with sentiments of human are present. The third approach, which is the Hybrid learning, combines both machine leaning approach and lexicon learning approach in order to improve the performance of classifier.

4) Sentiment Classification: At this step, the dataset is ready for the classification. Each and every sentence of the tweet will be examined and opinion will be formed accordingly for subjectivity. Subjective expression sentences are retained and those of objective expression sentences are rejected. Techniques like Unigrams, Negation, Lemmas and so on are used at different levels of sentimental analysis. Sentiments can be distinguished broadly into two groups – Positive and Negative. At this point of sentimental analysis, each of the subjective sentences which will be retained are classified into good, bad or like, dislike or positive and negative.

5) Output Presentation: To generate useful and meaningful information out of the raw data, sentimental analysis plays vital role. Once the algorithm is completed, the

outcome of the analysis can be visualized by creating different types of graphs. Bar graphs, Time series and Pie charts are some of the examples which can be used to display the output. To measure the sentiment of the tweets in terms of Positive and Negative, Bar graphs can be used. Similarly, to measure in terms of likes, dislikes, average length of tweet for a certain period, Time series can be used. To obtain the initial source of the tweet, pie charts can be used.

5.RESULTS AND DISCUSSION

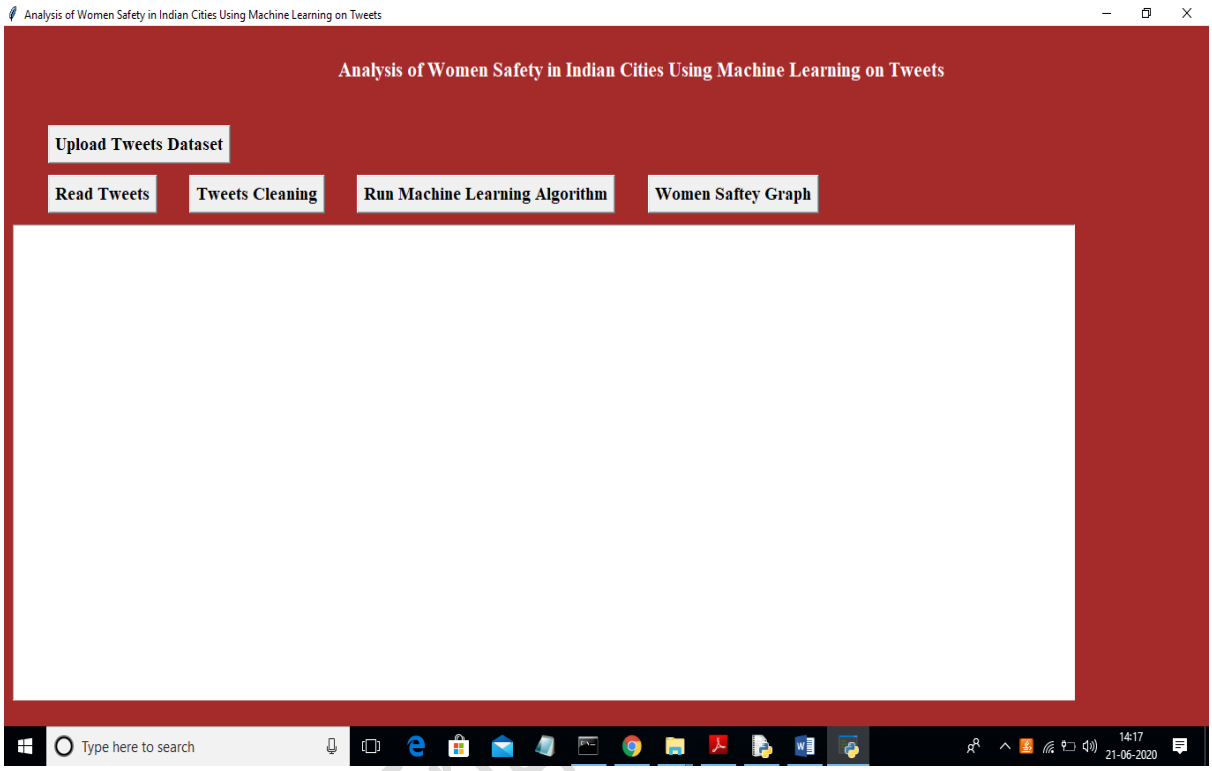

Fig 5.1 In above screen click on 'Upload Tweets Dataset' button and upload tweets

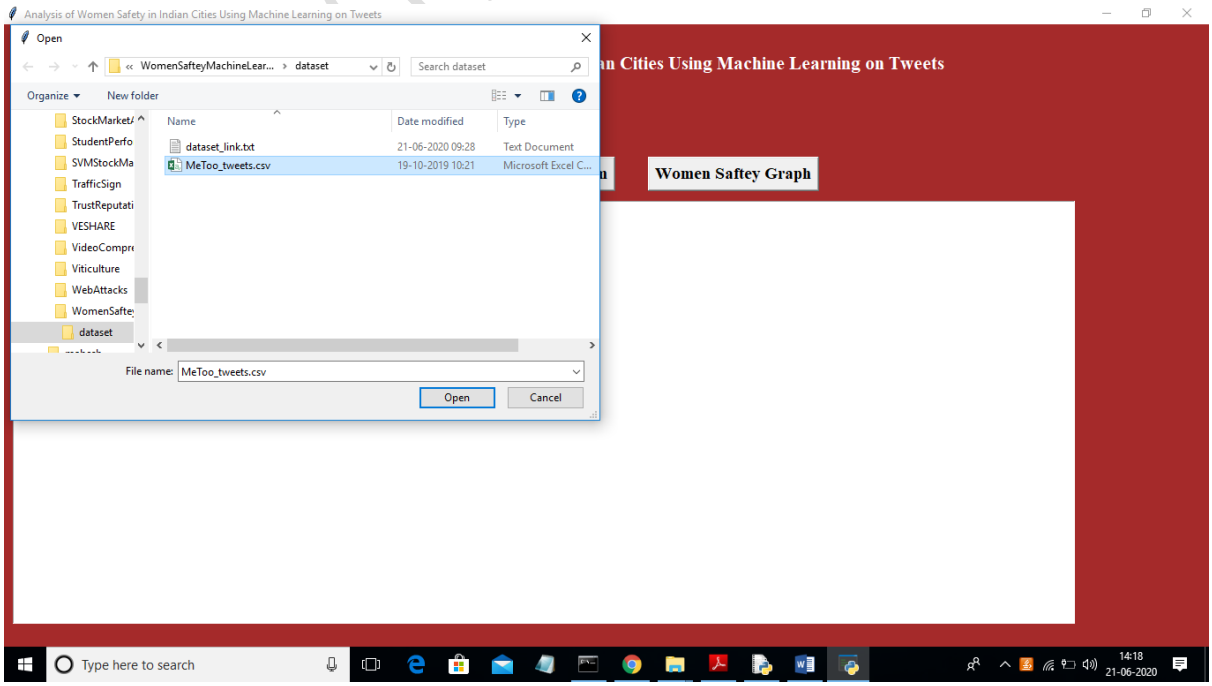

Fig 5.2 In above screen uploading MeeToo_tweets.csv file and then click on 'Open' button to load dataset and to get below screen

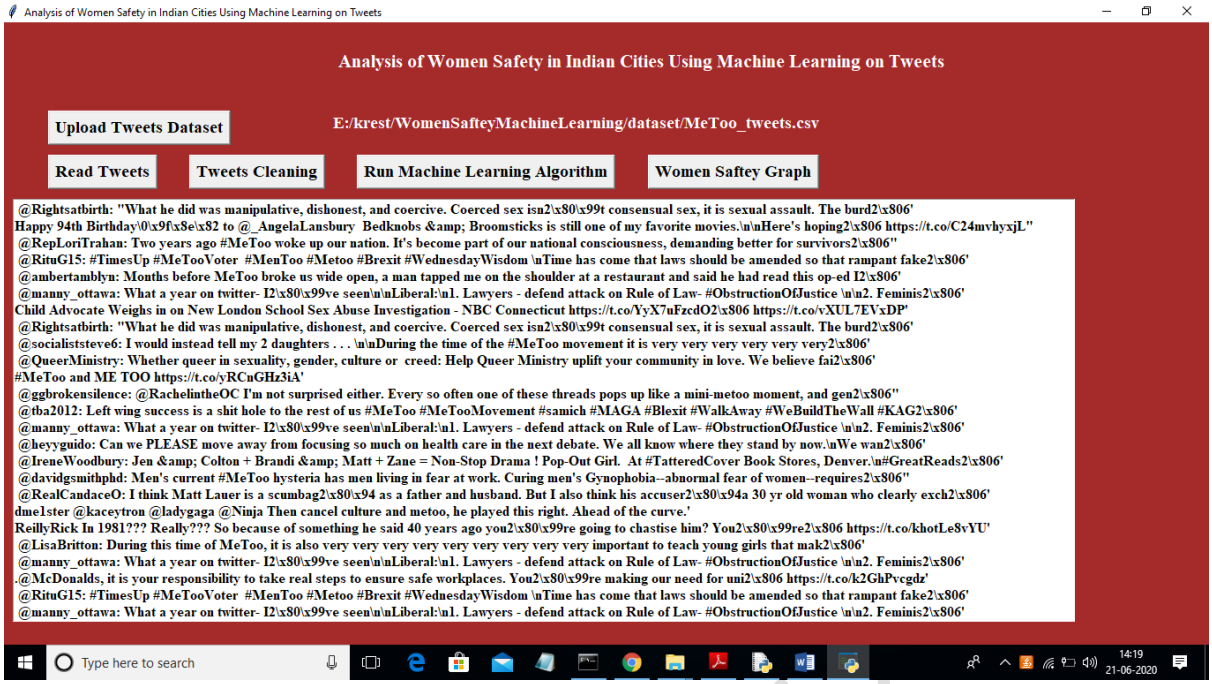

Fig 5.3 In above screen each line represents one tweet and you can scroll down above screen text area to view all tweets. In above screen we can see all tweets contains special symbols and stop words and to clean those tweets click on 'Tweets Cleaning' button

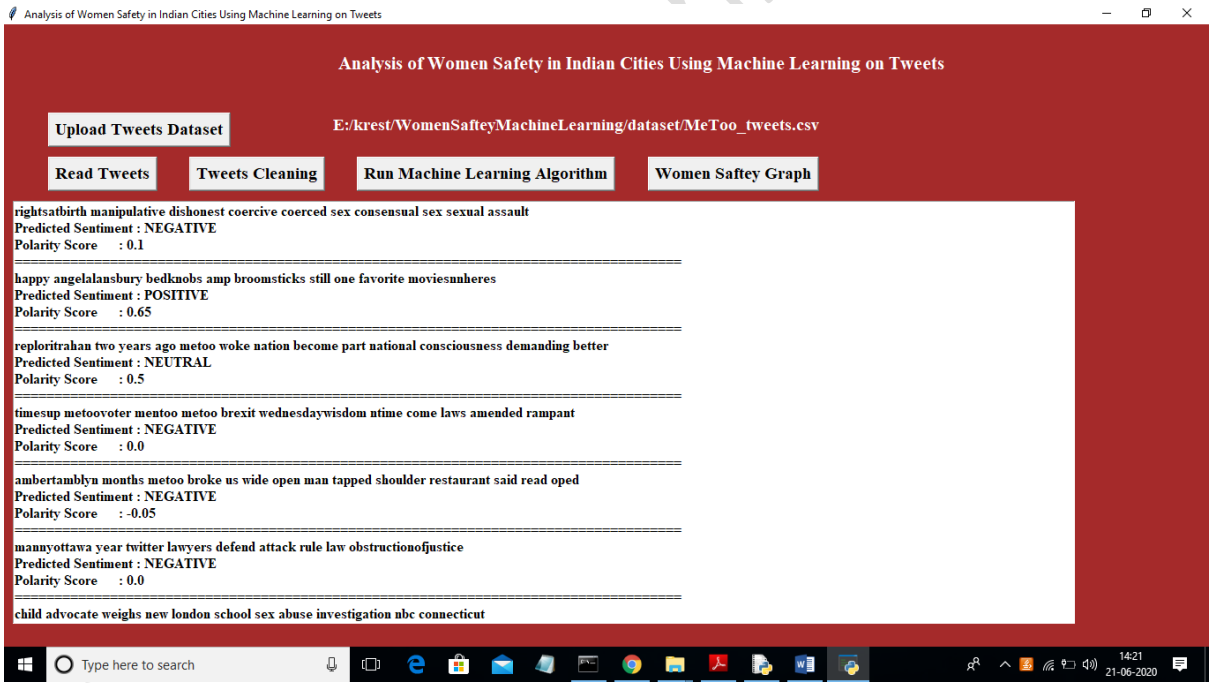

Fig 5.4 In above screen each tweet having tweet text and then displaying tweets sentiments with polarity score. Scroll down above text area to see all tweets. Now click on 'Women Saftey Graph' button to get below results and by seeing that result user can easily understand whether area is safe or not. If area is safe then more peoples will express either positive or neutral tweets and if not safe then more peoples will discuss negative tweets.

Journal of Engineering Sciences

Vol 13, Issue 04, APRIL /2022

ISSN NO:0377-9254

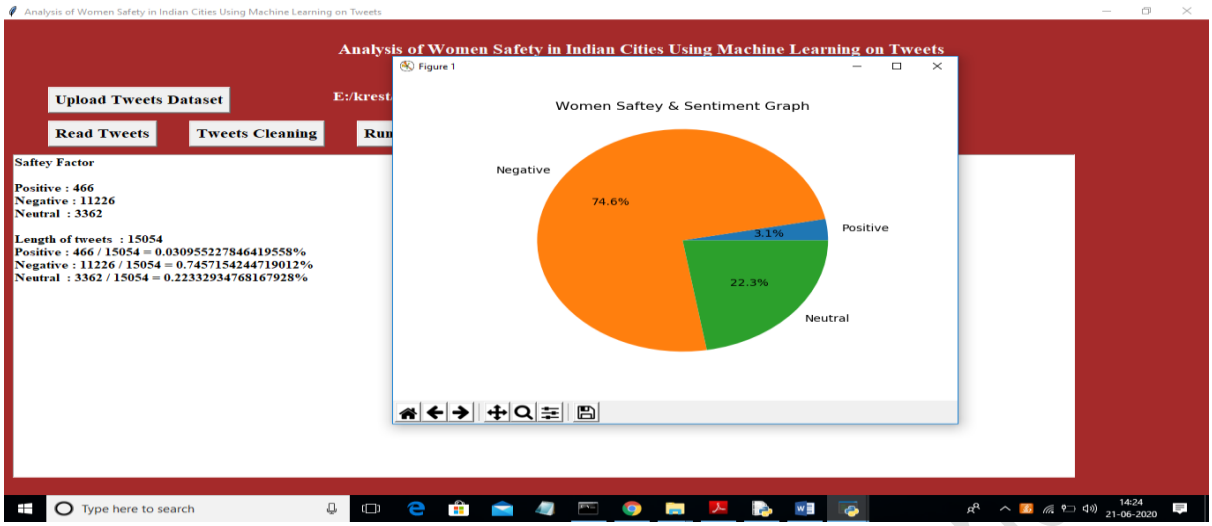

Fig 5.5 In above screen 0.74 multiply by 100 will give 74% which means 74% peoples are talking negative and area is not safe and only 22 and 3% peoples are talking positive and neutral.

6.CONCLUSION

We addressed a number of laptop mastering methods during the lookup paper to assist us organise and analyse the huge volume of Twitter records we have collected, which consists of thousands and thousands of tweets and textual content messages posted each and every day. The SPC technique and linear algebraic Factor Model techniques, which assist to similarly categorise the records into significant groupings, are two laptop mastering algorithms that are specially profitable and beneficial when it comes to evaluating sizable quantities of data. Support vector machines are a kind of laptop mastering approach that is regularly used to extract beneficial data from Twitter and acquire an grasp of the repute of women's security in Indian cities.

REFERENCES

[1]. Agarwal, Apoorv, Fadi Biadsy, and Kathleen R. Mckeown. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.Association for Computational Linguistics, 2009.

[2]. Barbosa, Luciano, and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data." Proceedings of the 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, 2010.

[3]. Bermingham, Adam, and Alan F. Smeaton. "Classifying sentiment in microblogs: is brevity an advantage?."Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[4]. Gamon, Michael. "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[5]. Kim, Soo-Min, and Eduard Hovy. "Determining the sentiment of opinions." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.