SENTIMENTAL ANALYSIS USING TWITTER

B. Muruganantham[1], Srishti Agarwal[2], Aditya Raj Arora[3]

[1,2,3]**Department of Computer Science Engineering, SRM Institute of Science and Technology, Chennai, India**
E-mail: [1]muruganb@srmist.edu.in, [2]srishtiagarwal_govind@srmuniv.edu.in, [3]adityaraj_arora@srmuniv.edu.in

Abstract
With the abrupt growth in technology the use of social media increases day by day, users share and exchange their emotions and opinions on day to day basis on the social websites. Twitter is one of the social websites where daily users post tweets with their emotions and opinions which are necessary for the development. To express the correct sentiment and overall opinion of a large number of tweets the process of sentiment analysis is done. It analyses the sentiment of each tweet and gives the overall output whether the sentiment of the person about the tweet is positive, neutral or negative. It identifies the correct sentiment of each word and analyses the emotions and opinion from the text. It explores how text analysis techniques can be used to dig some data on different trends of tweet languages. The data collected from the tweets is filtered according to the user's requirement. So, content filtering is done to show only interest related posts. The continuous analysis of live and latest tweets is analyzed correctly which gives a highly accurate model with respect to the latest reviews. By gathering a large number of tweets, live data is processed and correct sentiment of the data is analyzed which expresses the opinion of the mass in this huge platform. Usually, words with the same meaning all the variants of the label rather than the exam same label are not classified accurately, therefore the label within the data set is analyzed by Naive Bayes which improves the classification of the tweets considerably.

Keywords-- Naive Bayes classifier, Tokenization Lexicon, Sentiment Analysis, Twitter

INTRODUCTION

Before the web, when a person wanted to make a choice, they usually asked for opinions from their peers in the surroundings. But as the corporation needed to search the opinions and sentiments of the public using their products and services, they started conducting polls, opinion surveys, centers and forum teams. In several cases, the opinion was hidden in a long piece of paper, where it's tough for a person or a reader to search out the source and to extract the sentences connected with the opinions and to read and segregate them, organizing the opinions into distinguishable separate forms.

Over five hundred million tweet area unit are being denote daily, which has created twitter an essential and vital platform for information where moods of social people and their views and opinions are analyzed. Twitter is a vital platform for expanding and enhancing the products of different businesses and it also is a platform to spread awareness and discuss common needs of the people in the medical field where the medical requirements are analyzed and then taken into consideration. These messages help to grasp the desires of individuals and their views regarding product, options and services.

As a result of the analysis all streaming further as offline information on Twitter that area unit on the market within the variety of Tweets area unit analyzed to gather the positive, negative and neutral comments of the users.

Therefore, in our work, our aim is to analyze sentiments, the method of determining the feelings that the author supposed to predict in the messages. These sentiments usually represent the polarity of the messages which tell us whether or not, the opinion in the messages is highly positive, positive, highly negative, negative or neutral. This technique of sentimental analysis constructs the info for extracting the feelings and then classifying it according to the user's requirements from the alternative user generated text content.

LITERATURE SURVEY

[1] Self-constructing bunch algorithmic program. A bunch based mostly approach to rising the potency of cooperative filtering recommendation to cut back the spatiality associated with the amount of product. Similar product area unit classified within the same cluster and dissimilar product area unit sent in several clusters. Recommendation work is then finished the ensuing clusters.

[2] Sum based mostly Sentimental analysis that specialize in utilizing social media, and additional specifically Twitter, for rising and/or gaining insight into a mess of various tasks with within the political domain, organizing the said studies into 3 major classes, particularly providing election forecasts through social media, exploitation sentiment analysis for observation the general public opinion, and eventually, exploitation Twitter for medically charged process scientific discipline tasks.

[3] JEST (Joint-Entity-Sentiment-Topic), this paper offers a short summary of current analysis on medical opinion mining. we've got known 3 main analysis ways to mechanically analyze medical processes in Twitter: exact- precise opinion polarity analyzation, classification, sentiment detection and prediction.

[4] Naïve Bayes exploitation N-gram and POS-tags, featured area taken encloses hash tags, ret weets, link, emojis, punctuation and special symbols in addition with features like POS and preceding polarity of the words to distinguish the tweets as neutral, positive and negative.

[5] K-Nearest Neighbor could be a way to utilize Twitter users defined hash tags in tweets as a way of classification of the sentiments kind exploitation punctuation, ret weets, single words, links, n-grams and completely different patterns that sorts feature, the area unit is then combined into a single feature vector for the classification of the sentiment. The K-Nearest Neighbor strategy is used to assign sentimental labels by

constructing a single featured vector for each example in the coaching and check set.

PROPOSED WORK

The general process of the proposed framework is as follows:
For the training set, we have to prepare a set of unknown tweets that are pre-classified. The sample tweets are training set can be endured in a pre-processing stage as (refer fig. no. 1) follows:

Tokenization: It is a process or a method where every tweet is split into little separate tokens because usually a tweet contains a large sentence with URLs, links, HTTP symbols, emoticons and words. So, we split it into small tokens.

HTTP symbol removal: After the tokenization is done and all the tweets are pre-classified where the address and different HTTP symbols area unit aloof or removal from the training set of tweets. These area unit are of least importance and so the tweet classification is taken into account after the removal of these symbols.

Figure 1. Architecture Diagram

Stop words Removal: These words are typically the articles or the prepositions like "the", "of", "a" etc. that doesn't affect any desiring to the dataset. And additionally, do not place any important part in the classification.

Addition of Lexicons: When the coaching set is pre-processed, then for the corresponding words within the tweets that are pre-classified are going to supplemental. Lexicon area unit supplemental are outlined pre in lexicon domain. The lexicon contains a word with sentiment with its potential alternative or words having same meaning (synonyms). According to the polarity total appointed to the lexicons, they are appointed as highly negative, negative, positive, highly positive or neutral. Now, the dataset got ready for the testing. Unknown tweets are tested with the dataset against the sample unknown tweets as follows: The tweet sample unknown undergoes constant lexicon addition stages and preprocessing within the coaching set as Tokenization: It is a process or a method where every tweet is

split into little separate tokens because usually a tweet contains a large sentence with URLs, links, HTTP symbols, emoticons and words. So, we split it into small tokens. HTTP symbol removal: After the tokenization is done and all the tweets are pre-classified where the address and different HTTP symbols area unit aloof or removal from the training set of tweets. These area unit are of least importance and so the tweet classification is taken into account after the removal of these symbols. Stopwords Removal: These words are typically the articles or the prepositions like "the", "of", "a" etc. that doesn't affect any desiring to the dataset. And additionally, do not place any important part in the classification. Addition of Lexicons: Lexicon area unit supplemental are outlined pre in lexicon domain. The lexicon contains a word with sentiment with its potential alternative or words having same meaning (synonyms). According to the polarity total appointed to the lexicons, they are appointed as highly negative, negative, positive, highly positive or neutral. Now, the dataset got ready for testing.

Classification of the tweets: The most important and ultimate stage when Lexicons are added is that the way the tweets are classified. This classification is done by exploitation of the N-B classifier supported the assigned polarity of the lexicons.

IMPLEMENTATION

Dataset - There doesn't seem to be any giant public knowledge part of Twitter messages. For the classification the dataset is used from the commonplace Sentiment120. From Sentiment120, the extraction of tweets is done which is used in coaching set to coach the classified dataset. Twitter interfere is applied and the programming made exploitation of Sentiment120. To specify a languages the Twitter API's parameter are used. As the parameter is about English thus all the coaching data have tweets solely in English type Sentiment120 to gather the polarity of the tweets. All emojis are used as a question terms within Twitter's API, for instance ":)" and ":-)" both categorical positive feelings. Similarly, ":(" returns highly negative tweet. The massive Sentiment120 data, tweets associated with "Medical BMC" that may be a medical information extracted and then used as a coaching data to coach the given classifier. Dictionary lexicon domain: One vital issue, implementation and the purpose read from the projected system in lexicon domain. Stop words: A group of these plays a vital role while classifying of the tweet. Normally, "Stop words" are those words which helps in builds the meaning out of words or in other words prepositions or articles for example "a", "the" etc. These are the words which have least significance while classification is considered. The tweet is first checked with stop words to remove unwanted traffic from the tweet by then these words are removed and we are left with raw tweet then this tweet is often advanced with the maximum and the sentiment word is drawn out for classification.

RESULT AND DISCUSSION

The experiment contains a set of one twenty tweets from past 6 months that are twitted by the users. This data set is collected for the assembling of the tweets where the tweets from the data are tested. The hundred and twenty tweets are tested for classification. Firstly, the present approach which is existing and Naive Bayes classification the approach based on Lexicons. In proposed lexicon, the classification is first done using the coaching set which is then tested using the test tweet sample.

To, check if the lexicon-based method shows good result than present system, the experiment is done. Table 1.1 shows analysis carried by the experiment showing comparative results. 120 tweets were collected in 6 months. The proposed lexicon method shows improvement of an average of more than 1% over the present approach.

Tweets regarding medical reviews were collected. Total of 120 tweets in 6 months were accessed. The proposed system showed

major improvement in comparison to the present methods and Naive Bayes classifier approach. The proposed combination worked better.

Table 1. Results

Months	No. Of tweets in total	Classified tweets using present	No of tweet classified correctly
Aug	20	12	13
Sep	20	14	18
Oct	20	15	18
Nov	20	12	15
Dec	20	11	13
Jan	20	13	18

CONCLUSION

Naive Bayes is integrated with lexicon approach which is established in proposed system in place of regular one. The existing way of solving problem is having so many problems. First of all the classifier had failed to classify tweets successfully, in given scenario if the sentimental words used in the test tweet sample and in the coaching data set are all not the same then the tweet is not processed in rightful manner even though sample and data set means the same. To tackle this problem, we use lexicon-based approach in dataset. The data set dictionaries contain all possible synonyms with the sentimental words which is used so frequently so by adding these it will give desired output. Adding lexicon to tweets which results in adding synonyms and antonyms to sentiment word in tweet which is used to test which then results in improvement in accuracy of classified data.

One main problem is the liability of using various varieties of sentimental words. These approaches are lexicon based and it provides a solution to this problem as it permits the user to use various sentimental words as dictionary' which contain many synonyms which helps in providing the required output. The prepared approach is capable to provide handful of accuracy and provide many aspects of a tweet given by the user.

ACKNOWLEDGMENT

REFERENCES

1. Daniella Bal, Flavius Frasincar, Alexander Hogenboom, Malissa Bal, Franciska American state Erica Jong, Uzay Kaymak, "Exploiting Emoticons in Sentiment Analysis", 2013.
2. Liu Bing "Sentiment Analysis and Opinion Mining", Morgan Publishers, 2012.
3. Dr.Khalid N.Alhayyan and Dr.Imran Ahmad "Analyzing & discovering the necessary period of time Trends in wheezy Twitter Stream".
4. J. Ramteke, S.Shah, D.Godhia, and A. Shaikh, "Election result prediction victimization Twitter sentiment analysis," in ingenious Computation Technologies(ICICT), International Conference on, 2016, vol. 1, pp. 1–5.
5. ArvindArasu, Hector Garcia- Molina, "Extracting Structured knowledge from net Pages",337-348, 2003
6. P. Burnap, R. Gibson Using Twitter to predict, Vol 41, pp 230- 233, 20.
7. J.Ramteke, S. Shah, D.Godhia, and A. Shaikh, "Election result prediction victimization Twitter sentiment analysis," in ingenious Computation Technologies(ICICT), International Conference on, 2016, vol. 1, pp. 1–5.
8. M. Desai and M .Mehta, "Techniques for sentimental analysis of Twitter data", 2016 International Conference on Computing, Communication and Automation(ICCCA), 2016.
9. Alexander Pak and St.Patrick Paroubek. In Proceedings of the Seventh International Conference Language Resources and analysis (LREC'10), May 2010.
10. Amandeep Kaur, Vishal Gupta, "A Survey on Sentiment Analysis and Opinion Mining Techniques", Journal OF rising TECHNOLOGIES IN net intelligence, Vol.5, 2013.