Sentiment Research on Twitter Data

A. Brahmananda Reddy, D.N.Vasundhara, P. Subhash

Abstract— The scale of social network data that is being generated is increasing exponentially day by day. Public and private opinion of various subjects or issues are expressed in social media. Sentiment analysis is a method of analyzing the sentiment of a statement that it embodies. Twitter is one of the social medias that is gaining popularity nowadays and most people are using this platform to express their opinions. Sentiment analysis on Twitter is an application of analyzing the sentiment of twitter data(tweets) conveyed by the user. The research on this problem statement has grown consistently. The main reason behind this is the challenging format of tweets that are posted, and it makes the processing difficult. The tweet format would be the number of characters, slangs, abbreviations, emojis, http links and so on. In this paper the aim to describe the methodologies adopted, the process and models applied, along with a generalized approach using python. Sentiment analysis aims to determine or measure the attitude of the writer with respect to some topic.

Keywords- Sentiment Analysis, tweet, twitter, sentiment, social media

I. INTRODUCTION

The age of net has remodeled the approach individuals specific their views. it's currently done over web log posts, on-line discussion forums, product review websites etc. Social media platforms (also called microblogging sites) may be a media wherever individuals specific their opinions on one thing or somebody. Twitter is that the most typically used microblogging web site by the individuals to post their opinions [1]. Organizations need a tool that helps in analyzing the feedback given by the individuals on their product or service, and this analysis are often done by police investigation the feeling of the posts. The feeling is painted as positive, negative, that square measure sub-categorized as powerfully positive, sapless positive, powerfully negative sapless negative, and neutral. It will analyze emotions towards entities like product, services, organizations, people, issues, events, topics and their attitudes.

II. EXISTING SYSTEM

The existing system 'Sentiment Analysis' takes the static data which is already extracted from a social media platform. The data extracted is stored in a csv file or Excel file which is the input to the program or application. For each statement the program analyses, the output would be a floating-point number which is termed as polarity. The polarity values range from -1 to +1. Based on the polarity obtained the program determines the emotion of the statement.

- The emotion is classified as positive, negative, neutral.
- If polarity>0 then the emotion is positive.
- If polarity= 0 then the emotion is neutral.
- If polarity<0 then the emotion is negative.

Drawbacks:

- The user who is analyzing the statements has to go through the entire document (csv file) to get overall general report.
- The emotions are classified only into three categories i.e., positive, negative, neutral.
- The data is stored prior to the analysis.

III. PROPOSED SYSTEM

This system deals with performing functions dynamically through an online social media i.e., Twitter. Twitter posts of electronic products creates a dataset. Tweets are short messages with slang words and misspellings. So, the sentence level sentiment analysis is performed. This can be done in seven phases. In the first phase, input data is given. Here the input data refers to a username or a hashtag. Then, the number of tweets to be analyzed are specified. Those tweets are retrieved from the twitter database. Then in the third phase, the retrieved twitter data is stored in a database. In fourth phase, the tweet is processed. This step is performed before feature extraction. Processing steps include removing URLs, removing stop-words, avoiding mis-spellings and slang words. Mis-spellings square measure avoided by commutation continual characters with 2 occurrences. Slang words contribute abundant to the feeling of a tweet. Hence, a slang word lexicon is maintained to switch slang words occurring in tweets with their associated meanings. Next part is feature extraction. A feature vector is formed mistreatment relevant options.

Overcome The Drawbacks Of Existing System:

- It would be better if the result is represented in the form of bar graph or pie chart.
- To get a better understanding on the emotions, the emotions should be classified into seven categories. They are:
  - Strongly Positive
  - Positive
  - Weakly positive
  - Neutral
  - Strongly Negative
  - Negative
  - Weakly Negative

Revised Version Manuscript Received on 16 September, 2019.

Dr. A. Brahmananda Reddy, Associate Professor, Department of Computer Science and Engineering, VNR VJIET, Hyderabad, Telangana, India.
(Email: brahmanandareddy_a@vnrvjiet.in)
D.N.Vasundhara, Associate Professor, Department of Computer Science and Engineering, VNR VJIET, Hyderabad, Telangana, India.
(Email: vasundhara_d@vnrvjiet.in)
Dr. P. Subhash, Associate Professor, Department of Computer Science and Engineering, VNR VJIET, Hyderabad, Telangana, India.
(Email: subhash_p@vnrvjiet.in)

• Instead of storing the data prior to the analysis, it can get real time data from twitter by giving a hashtag or username to analyze the tweets of a person or a specified hashtag

The proposed system overcomes the drawbacks of the prevailing system

Now, it has fell upon our coaching set then, it's required to extract helpful options from it which might be utilized in the method of classification. however, 1st let's discuss some text format techniques which is able to aid America in feature extraction:

• **Tokenization:** it's the method of breaking a stream of text into words, symbols and alternative substantive parts referred to as tokens. they will be separated by whitespace characters and/or punctuation characters. Tokenization is performed so it will examine tokens as individual parts that structure a tweet.

• **URLs** and user references (identified by tokens http and @) are removed if the user is interested in only analyzing the text of the tweet.

• Punctuation marks and digits/numerals may be removed if for instance the user wishes to compare the tweet to a list of English words.

• **Lowercase Conversion:** Tweet may be normalized by

$$P(s|M) = \frac{P(s) \cdot P(M|s)}{P(M)}$$

converting it to lowercase which makes its comparison with an English dictionary easier.

• **Stop-words removal:** Stop words area unit a category of some very common words that embrace no supplementary data once employed in a text and area unit so claimed to be useless. Examples comprise "a", "an", "the", "he", "she", "by", "on", "there", "here" etc. it's typically convenient to get rid of these words as a result of they hold no further data since they're used virtually equally altogether categories of text, as an example, once computing prior-sentiment-polarity of words in a very tweet per their frequency of incidence in several categories and victimization this polarity to calculate the typical sentiment of the tweet over the set of words employed in that tweet.

Finally using different classifiers, tweets are classified into strongly positive, weakly positive, neutral, strongly negative and weakly negative classes. Based on the number of tweets in each class, the final sentiment is derived.

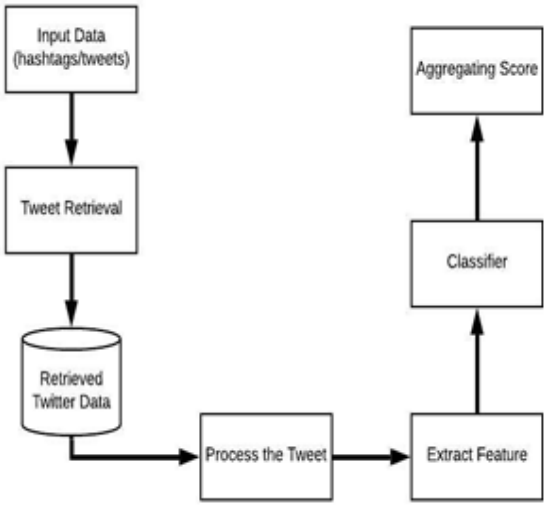

Figure 1: System Architecture

IV. IMPLEMENTATION

Input: Give the input either the username or the hashtags and the no of tweets that you want to analyse the tweets.

Step 1: Tweet cursor from the tweetpy package will retrieve the tweets related to the given search word.

Step 2: The Retrieved tweets will undergo tokenization and then the process of cleaning where the punctuation marks, emoticons, URLs, stop- words will be removed.

Step 3: The output of the above process are the features where they have sent for the analysis to the Naïve Bayes classifier.

Step 4: Now the Classifier will be processing the features and then assigning the polarities to each of them ranging from -1 to 1.

Step 5: Now adding each polarity of the given certain tweet net polarity will be used to classify the given tweet into its respective category.

Output: The output will be the classification of % of tweets in the particular category and the pie chart which will depict all the categories that have been classified.

4.1. Building the Classifier:

The classifier used is Naïve Bayes Classifier. The Bayes Theorem tells that

Where s is Sentiment and M is the Messages.

Here the dataset with videogames reviews is used. Filtering all the reviews by scores and dividing the examples equally between positive class (score = 1) and negative class (score = -1).

4.2. Preparing Data:

Then will be preparing the data by tokenization, cleaning of the data and then move to Bag-of - Words. Generate the feature vector for each of the document. Bag of Words will be counting the frequency of number of times the token has been appeared in each document.

No of columns is tokens that are unique in collection of documents.

No of rows is total documents in whole collection

Now converting the X_train data into a vector called tf_train and X_test data into a vector called tf_test

4.3. Building:

Naive Bayes approach is based on Bayes' theorem which uses probabilistic learning function. Here, Multinomial approach is used.

P is sum of all feature vectors with score-1 P=$\sum tf\_[ytarin = 1] + 1$

Q is sum of all feature vectors with score -1 Q=$\sum tf\_[ytarin = -1] + 1$

Here 1 is added to both P, Q to ensure that

each token has been taken at least once.

log-count ratio r:

r = log $((P/\sum P)/(Q/\sum Q))$

And b:

b = (log $length(P)$) $length(Q)$

Now coefficients calculated then will produce predictions on testing set. A linear classifier is fit, the linear equation is:

y = mx + b

pre_preds= tf_test. T + b

This is a "naïve" method as it assumes that features are independent, that is they will not interact. Also, the assumption made by BOW that order of the tokens does not matter. This method achieves good results.

V. RESULTS

The result is going to be shown within the variety of a pie-chart in Figure a pair of, that constitutes seven major emotions: powerfully positive, sapless positive, positive, neutral, negative, sapless negative and powerfully negative. For neutral, it represents that the tweet / hashtag's aggregate score is that of zero. However, this project will list desired variety of recent tweets as such by the tip user.

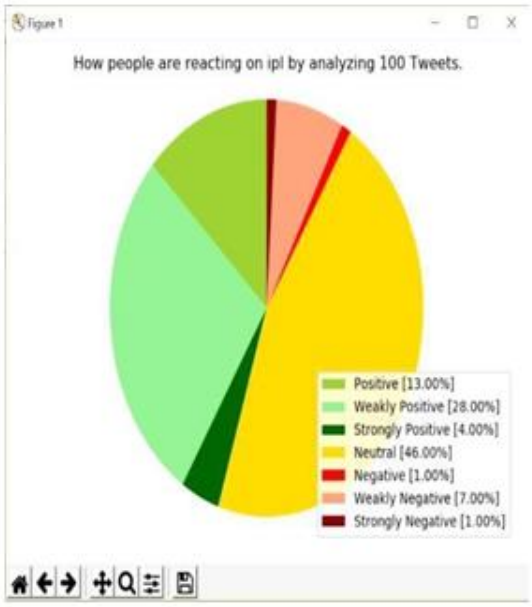

Figure 2: Pie graph of the analyzed tweets

VI. CONCLUSION

Twitter sentiment analysis is developed to investigate public's views towards a tweet / hashtag. Input is given i.e., either the username or a hashtag. Then the tweet is retrieved from twitter information that undergoes feature extraction. Associate in Nursing economical feature vector is formed by doing feature extraction in 2 steps when correct pre-processing. within the start, twitter specific options area unit extracted and additional to the feature vector. After that, these options area unit aloof from tweets and once more feature extraction is finished as if it's done on traditional text. These options are additional to the feature vector. Classification accuracy of the feature vector is tested victimisation Naïve Thomas Bayes classifier. Associate in Nursing accuracy of seventy-eight.38 it had been reached

REFERENCES

1. Hamid Bagheri, Md Johirul Islam,"Sentiment analysis of twitter data", Annual International Conference "Dialogue"(2017) (pp. 14-28)
2. David Zimbra, M. Ghiassi and Sean Lee, "Brand-connected Twitter Sentiment Analysis mistreatment Feature Engineering and therefore the Dynamic design for Artificial Neural Networks", IEEE 1530-1605, 2016.
3. Bhumika Gupta, Monika Negi, Kanika Vishwa, "Study of Twitter Sentiment Analysis mistreatment Machine Learning Algorithms on Python" International Journal of laptop Applications 0975- 8887, 2017
4. Aliza Sarlan, Chayanit Nadam and Shuib Basri, "Twitter Sentiment Analysis", 2014 International Conference on data Technology and Multimedia (ICIMU), Putrajaya, Malaya Gregorian calendar month eighteen – twenty, 2014.
5. Alexander Pak, Apostle Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Proceedings of the International Conference on Language Resources and analysis,
6. LREC 2010, 17-23 could 2010, Valletta, Malta,2010
7. Mining Twitter information with Python (Part half-dozen – Sentiment Analysis Basics), marcobonzanini 2015, https://marcobonzanini.com/2015/05/17/mining-twitter-data- link "https://marcobonzanini.com/2015/05/17/mining-twitter-data-with-python-part-6-sentiment-analysis-basics/" with-python-part-6-sentiment-analysis- basics/
8. Naive Bayes for Sentiment Analysis, Medium Corporation,2018, https://medium.com/@martinpella/naive-link"https://medium.com/@martinpella/naive-bayes-for-sentiment-analysis-49b37db18bf8" bayes-for-sentiment-analysis-49b37db18bf8