

Twitter Sentimental Analysis

Harshada Borade¹, Abhijit Naik², Shraddha Yalgekar³

^{1,2,3}Information Technology, Padmabhushan Vasantdada Patil Pratishthan's College of Engineering, Sion Mumbai

Abstract—Twitter is amongst one of the social networking website where people can post, update, write comments in the form of “tweets”. Twitter basically serve as a mean for an individual to express and share an opinion or view on a specific subject, whether it is politics, market trends, social affairs. As we know twitter contains very large numbers of short messages created by user of this microblogging website. The content of messages or tweets can vary from personal thoughts to public statements. We proposed a method known as analysis of sentiments and classify the sentiments into positive, negative and neutral. In these paper we are using the concept of NER (Named Entity Recognition) and Classification Algorithm to retrieve the data and classify it accordingly. Sentiment Analysis also been used in various e-commerce website such as amazon, flipkart in which user of a product either post a good, bad or neutral comments of a product they purchase in the form of feedback.

Keywords—Classification; Sentiment analysis; Candlestick charts; Longest common subsequence algorithm for numbers; Multi numerical attributes; Nse, Bse stock average.

1. INTRODUCTION

Twitter Sentiment Analysis plays a crucial where in it not only helps various companies, website understands how they are doing with their customers via the tweets, it also gives them a better depiction of they can buildup against the competitors. Result of sentiment analysis can be shown in various form such as using text, using emoticons, emojis. Emojis for the tweets based on either it is a positive, neutral or negative comment. It is technique widely used in text mining so as to analyze sentiments of text which is also known as opinion mining. The collected corpus from twitter can be arbitrarily large.

Twitter audiences indeed varies from regular user to celebrities, politicians, athletes, company representatives, therefore it is possible to collect text post of users from distinct social and interest group, tweets can also be small in length and thus can be equivocal and unbiased in nature. As a part of Natural Language Processing (NLP). Algorithm like naïve bayes along with classification has been used in predicting polarity of text sentiments

Sentiments analysis of twitter data may depend upon the sentence level and document level .which means whether a tweet is a sentence or a document, based on these analysis takes place further

Sentiment analysis has been used in variety of area like not only in social media but in e-commerce website, in companies. Not only various data mining algorithm helps in implementing analysis of tweets but python language plays a major role while implementing this project. We will be importing various library of python textblob tweepy and also methods like clean_tweet to remove links special characters etc also we'll be installing NLTK (Natural Language Processing Toolkit) corpora .Corpora is nothing but large and structured set of texts

2. BACKGROUND

SENTIMENTS ANALYSIS WORKING.

This section introduces the how data is being fetched at backend using a appropriate method that is aim to display the correct output, for this it may require to get twitter api keys like consumer key, consumer secret key, access token key, access token secret key. after getting key we can now have the access to twitter data and tweets to do further processing.

A. Formation and Analysis

1. Firstly the data (tweets) from various social networking sites is fetched and extracted for sentiment analysis through After fetching the user data, preprocessing of the data is done which basically involves three step process: Tokenization, Stemming and Filtering.

In tokenization the punctuation are removed such as “comma”, “apostrophe”, “fullstop”, “exclamation”, “question mark symbol”.

In Stemming the root word is extracted .In filtering the data is filtered so that the resultant output should be relevant to

what user want,ie in filtering step, it filters data that is closely related to the the given input. After applying the all above, machine learning algorithm is used to train the data further as shown in figure 1

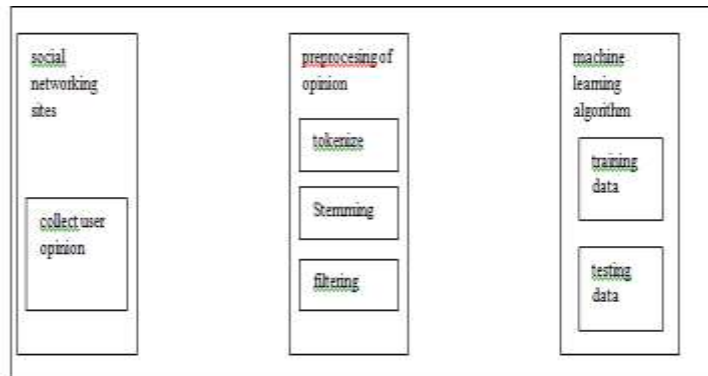


Figure1: Fetching data

B. NLP in twitter sentiments analysis

Natural language processing is being used in the particular implementation, the goal is for computer system to process or understand “natural language” in order to perform various human like task such as language translation, answering question, also it has fast growing collection of useful application obtain from NLP field. therefore they are range from simple to complex.

NLP also used for displaying emotions based on how confidence the tweet is along with the degree of positive, negative, neutral as shown below

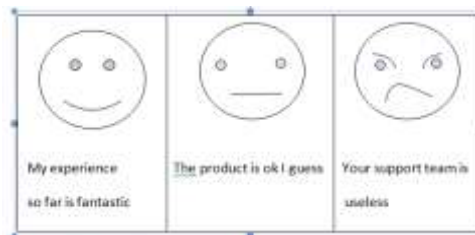


Figure 2: Emotions based on positive, negative and neutral comments

Figure shows the review of a e-commerce product purchased by the customers, numerous customer has distinct review regarding the product based on which we can have clear picture of how the product works for respective customers, the emoji in first column state that the feedback is positive, the on ein the second column belongs to neutral, and the last emoji depict the negative review of a product.

C. Libraries For data processing

Tweepy: It is a python client for official twitter API supports for accessing twitter via basic authentication and newer method like Oauth, however twitter has stopped accepting basic authentication so, Oauth is now the only way to use twitter API.

TextBlob: TextBlob is also one of the popular libraries of python, It is used to process textual data stands on NLTK, textblob comes with some advance features such as spelling correction, sentiment extraction. As we can see in the figure given below, textblob is also used for tagging parts of speech.

```
In [9]: blob = TextBlob("Machine Learning Coding is best thing in the world")
blob.tags

Out[9]: [('Machine', 'NNP'),
('Learning', 'NNP'),
('Coding', 'NNP'),
('is', 'VBZ'),
('best', 'JJS'),
('thing', 'NN'),
('in', 'IN'),
('the', 'DT'),
('world', 'NN')]
```

Figure 3: Textblob output

N-grams with textblob

Here N is basically the number and N-Gram is the chunks of words in group

```
In [10]: blob = TextBlob("Machine Learning Coding is best thing in the world")
blob.ngrams(n=3)

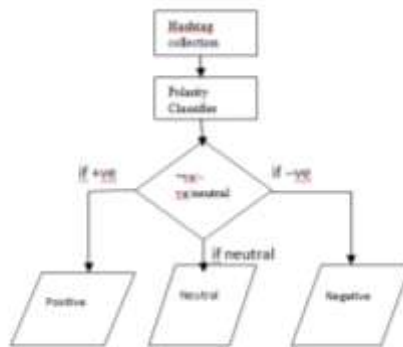
Out[10]: [WordList(['Machine', 'Learning', 'Coding']),
WordList(['Learning', 'Coding', 'is']),
WordList(['Coding', 'is', 'best']),
WordList(['is', 'best', 'thing']),
WordList(['best', 'thing', 'in']),
WordList(['thing', 'in', 'the']),
WordList(['in', 'the', 'world'])]
```

Figure4: Textblob using N-gram

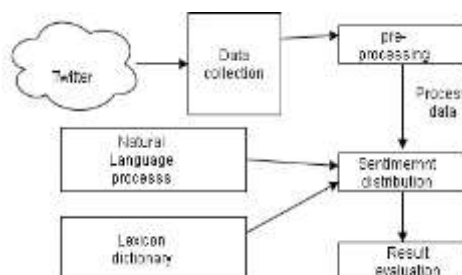
Naïve -Bayes Classifier

This classifier uses a simple approach which is bayes theorem, which describes how the conditional probability of each set of possible causes for a given observed outcomes can be computed from the knowledge of the probability of each causes along with conditional probability of the outcome of each cause

$$p\left(\frac{c}{d}\right) = p\left(\frac{d}{c}\right) \cdot \frac{p(c)}{p(d)}$$



3. Methodology



Twitter

Here twitter act as source of, from where data is fetched and collected which stored in database, it can be tweets, document, links which can be collected as a form of data

Data Collection

Data collection unit collect data and prepare the dataset for further pre-processing of data, here data is collected as a whole unit and is stored

Data Preprocessing

Processing data in simple terms means to manipulate or transform a data to produce meaningful result, common data processing methods are classification, calculation, sorting.

Mining twitter is sometimes a challenging task, because the data which we collect is the raw data, therefore in order to apply classifier, it is mandatory to clean preprocess and clean raw data, task involves removal of punctuation marks, removal of slang words, removal of stop words, removal of emojis if any, compress elongated word such as happyyyyy.compression will remove the extra numbers of y's.

Sentiment Distribution

After classifying the processed data using appropriate algorithm, sentiment distribution accordingly assign a particular sentiments for a tweet whether the tweet says good then it will considered as positive tweets, if a tweets conveys a negative message than it will considered as negative tweet and likewise segregation will done. Lexicon dictionary act as input to sentiment distribution lexicon contains a set of words with their meaning

Result Evaluation

Final evaluation will be done to check whether desired output is getting or not.

4. Conclusion and Future work

Implementing a model which aims to do sentiment analysis of data using different classifier and data mining algorithm, it is easier to classify the tweets and more we can do is we can improve the training dataset so that accurate result will be produced. The task of sentiment analysis, especially in the domain of micro blogging is still in the developing stage and far from complete. So the couple of ideas are proposed by us which we feel are worth exploring in the future and may result in further improved performance

Computation of overall tweet score can be done for a single keyword which can provide overall sentiments of public regarding a topic, moreover we would like to make a web application for users to input keywords for getting analyzed results.

5. Future Research

As a part of this project it is planned to conduct further research into impact of doing analysis of enormous amount of data,also carrying out analysis of tweets which contains hundreds and thousands of words

6. References

[1] Tweet data preprocessing and segmentation to NER.[Mr.Sanket Patil¹, Prof. Varsha Wangikar², Prof. K. Jayamalini³]

[2] Twitter Sentiment Analysis [Aliza Sarlan¹, Chayanit Nadam², Shuib Basri³]

[3] www.towardsdatascience.com

[4]www.wikipedia.com