SENTIMENT ANALYSIS OF GENERALIZED TEXT AND TWEETS

Arpit Upadhyay[1], Nishi Sharma[2], Aryan Chaudhary[3], Divya Jain[4]

[1]*Student, Dept. of Information Technology, Inderprastha Engineering College, Uttar Pradesh, India*
[2]*Student, Dept. of Information Technology, Inderprastha Engineering College, Uttar Pradesh, India*
[3]*Student, Dept. of Information Technology, Inderprastha Engineering College, Uttar Pradesh, India*
[4]*Student, Dept. of Information Technology, Inderprastha Engineering College, Uttar Pradesh, India*

Abstract - *Social media nowadays are the main resources from which we can gather information about people's opinions and sentiments towards various topics as they spend hours daily on social media and yield their view. In this research paper, we explain the functioning of sentimental analysis and how to connect to Twitter and run sentimental analysis queries. We run experiments on different queries from daily world news to show interesting results. At last, we come across that the sentiment for tweets that are neutral are very high which clearly shows the limitations of our present works.*

***Key Words*: Sentiment, Textblob, Naive Bayes, NLP**

1. INTRODUCTION

Sentiment analysis is the computational study of people's reactions, opinions, and emotions toward an entity. The input can describe individuals, events, or topics which most likely are coming from reviews, tweets, and comments. The two expressions sentiment analysis and opinion mining are interchangeable and refer to the same meaning, however, in some resources opinion mining is considered slightly different with sentiment analysis. The point that some researchers differ between opinion mining and sentiment analysis is the opinion mining extract the people's opinion from text data. However, sentiment analysis focuses on the sentiment of the given text data. In this research, we just are looking for the sentiment of data and its values. The overall purpose of the thesis is to study and get a deeper comprehension of how sentiment analysis and opinion mining works and then using the knowledge to make an application by utilizing different tools and frameworks. In simple words, the application aims to read the data from Twitter and then work on the sentiment of collected tweets and show the results to the users. The final product of this application will be used to analyze data related to daily news and it could be attracted to different user groups such as normal users, companies, business holders, politicians, etc.

2. RELATED WORK

The bag-of-words model is one of the most widely used feature models for almost all text classification tasks due to its simplicity coupled with good performance. The project work represents the text to be classified as a bag or collection of individual words with no dependence of one word with the other, i.e. it takes no notice of grammar and order of words within the text. This work is also popular in sentiment analysis and has been used by different researchers in this field. The simplest way to incorporate this model in our classifier is by using n-grams as features. In general terms, the n-gram is a neighboring sequence of "n" words in our text, which is completely unconventional of any other words or grams in the text. So we can conclude that unigram is a group of individual words in the text to be classified, and we assume that the probability of repeating one word will not be changed by the existence or non-attendance of any other word in the text. This is a very simplifying theory but it has been given to provide rather good performance. One simple way to use n-grams as features is to assign them with an unquestionable prior polarity, and take the average of the overall polarity of the text, whereas the overall polarity of the text could only be computed by summing the prior polarities of individual unigrams. The previous polarity of the word would be positive if the word is normally used as a portent of positivity, for example, the word "beautiful"; while it will be negative if the word is generally related to negative implications, for example, "bitter". There can also be degrees of polarity in the model, which means how symptomatic is that word for that particular class. The words like "extraordinary" would probably have great subjectivity along with positive polarity, while the word "ordinary" would although have positive prior polarity but probably with vulnerable subjectivity. There are different ways of using the prior polarity of words as features. The uncomplicated supervised approach is to use publicly available dictionaries present online which maps a word to its polarity. The SentiWordNet 3.0 is a resource that gives the probability of each word belonging to positive, negative, and neutral classes. Another approach is to make a custom prior polarity dictionary from our training data according to the occurrence of each word in each particular class. For instance, if a certain word is repeating more often in the positive labeled phrases in our training dataset then we can evaluate the probability of that word belonging to a positive class to be higher than the probability of occurring in any other class. This approach has been concluded to give better performance since the prior polarity of words is better suited and fitted to a particular type of text and is not very common like in the previous approach. However, this approach is a supervised approach because the training data has to be labeled in the appropriate classes before it is possible to evaluate the relative occurrence of a word in each of the classes. While in python there are so many libraries that can be used to classify the words.

3. LITERATURE SURVEY

Sentiment analysis could be a comparatively new research topic so there's still plenty of space for further research during this domain. In twitter mainly due to the limit of 140 characters per tweet which makes the user by force to specific opinion compressed in a very short text. The best results until now have been reached in sentiment classification using supervised learning techniques such as Naive Bayes and Support Vector Machines, but the manual labeling required for the supervised approach is very costly. Some amount of work has been done on unsupervised and semi-supervised approaches, and there is a lot of scope of improvement. Different researchers testing some new features and classification techniques very often just compare their results to general performance. There is a requirement for proper and formal comparisons between these results arrived through various features and classification techniques to choose the best features and most efficient classification techniques to choose the best features and most efficient classification techniques for some particular applications. It is generally observed that having a larger training data pays off to some better degree, after which the accuracy of the classifier remains almost constant even if we are adding more number of labeled tweets in the training data. Barbosa et al. used many tweets labeled by online resources, instead of labeling them by hand, for training the classifier. But in this case, there is a loss of accuracy of the labeled sample data in doing so (which is modeled as an increase in error) it has been noted that if the accuracy of training labels is greater than 50%, the accuracy of the resulting classifier is much higher. Thus in this way, if there is an extremely large number of tweets, the fact that our labels are noisy and inaccurate can be settled with. By the use of positive or negative emoticons to assign labels to the tweets. Like in the previous case they used a large number of tweets to diminish the effect of noise in their training data.

4. METHODOLOGY

We use two methods for this project one is live tweets analysis and the second is sentiment analysis of the text. TextBlob is a python library and grants a simplistic API to access its methods and do fundamental NLP tasks. NLP tasks. The main aim of the research is to find out the sentiments with the help of supervised classifiers.

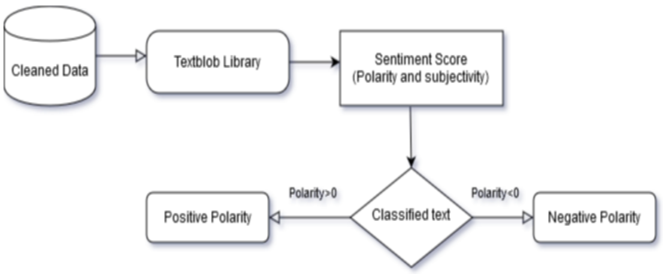


Fig 1: Working of Textblob

The methodology for the live tweet analysis is:-

1. We have collected a large amount of positive, negative, and neutral tweets with the help of twitter API from twitter using tweepy library. The size of our collected data can be enormously large.
2. We then find out the sentiments of the tweets with the help of textblob and save it to our database.
3. Then we read the sentiments from our database. The raw sentiment scores are going to be all over the place and noisy.
4. Then, clean and smooth the sentiments.
5. With this, we create a live tweets sentiment graph with the help of dash.

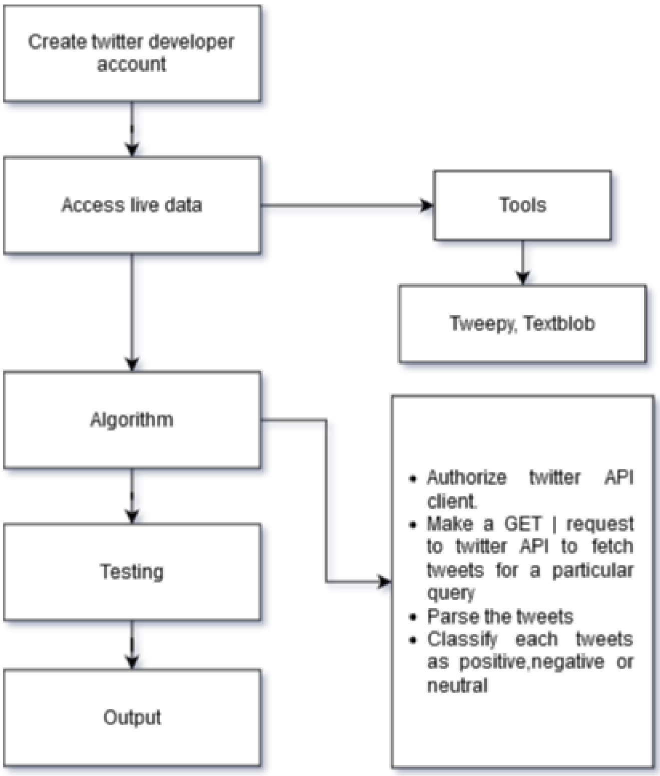

Create twitter developer account
Access live data
Tools
Tweepy, Textblob
Algorithm
Testing
Output
Authorize twitter API client.
Make a GET | request to twitter API to fetch tweets for a particular query
Parse the tweets
Classify each tweets as positive,negative or neutral

Fig 2: Live Tweets Analysis

The methodology for the sentiment analysis of the generalized text is:-

We use some NLP packages such as Textblob and spacy.

1. We used a textblob and spacy python library for the NLP function such as POS tags, tokenization, entities, and sentiments, etc.

Textblob uses two types of analyzers

- Pattern-based (Default)
- Naive Bayes

We use the Naive Bayes analyzer for sentence classification and Pattern-Based analyzer for file processing. The textblob returns two properties, polarity, and subjectivity which are used to find out the sentiment.

5. RESULTS

Dash library is used to design the web-based interface for the live tweet analysis.

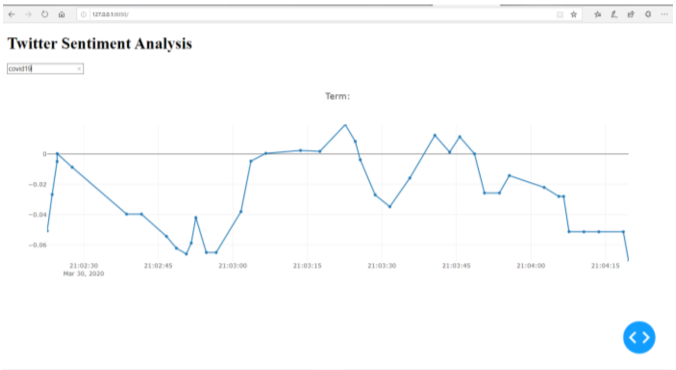
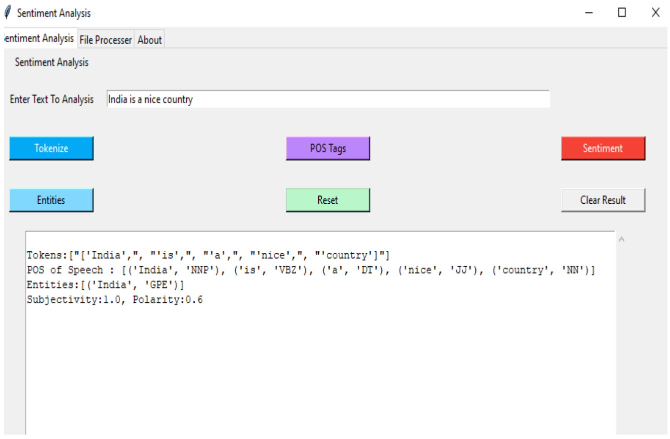


Fig 3: Live tweet Analysis Interface

Tkinter package is used to design the interface for generalized text classification. In this, we design two tabs one is for sentence analysis and the second one is for text file analysis.

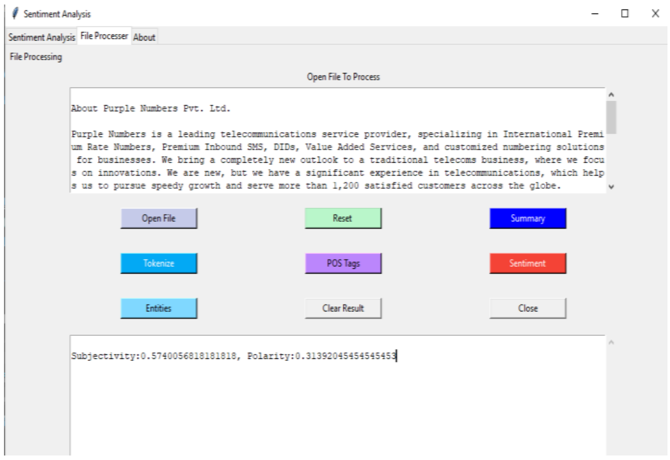


Sentiment Analysis

Enter Text To Analysis: India is a nice country

```
Tokens:["['India',", "'is',", "'a',", "'nice',", "'country']"]
POS of Speech : [('India', 'NNP'), ('is', 'VBZ'), ('a', 'DT'), ('nice', 'JJ'), ('country', 'NN')]
Entities:[('India', 'GPE')]
Subjectivity:1.0, Polarity:0.6
```

Fig 4: Sentence analysis



```
About Purple Numbers Pvt. Ltd.

Purple Numbers is a leading telecommunications service provider, specializing in International Premi
um Rate Numbers, Premium Inbound SMS, DIDs, Value Added Services, and customized numbering solutions
 for businesses. We bring a completely new outlook to a traditional telecoms business, where we focu
s on innovations. We are new, but we have a significant experience in telecommunications, which help
s us to pursue speedy growth and serve more than 1,200 satisfied customers across the globe.
```

```
Subjectivity:0.5740056818181818, Polarity:0.3139204545454545
```

Fig 5: Text file analysis

6. CONCLUSIONS

The main aim of twitter sentiment analysis, especially in the field of micro-blogging, is right now in the development stage and there is still a lot of work from completion. So now we propose a set of ideas which we sense are worth exploring in the future and may result in much more improved performance. Till now we have worked with some of the very simplest unigram models; we can further make and improve those models by adding some extra information like closeness of the word with a negative emotion word. We could also specify a window next to the word (a window could be for example of 2 or 3 words) under consideration and the effect of negation may be consolidated into the model if it rests within that window. For example, if the negation is right next to the word that we have taken, it may just reverse the polarity of that word and then farther the negation is from the word the more diminished its effect should be. Currently, we are exploring Parts of Speech separate from the unigram models, we can try to contain POS information within our unigram models in the future time. So now to say instead of calculating a single probability for each word like P(word | obj) we can instead have multiple probabilities for each according to the Part of Speech the word belongs to. For example, we may have P(word | obj, verb), P(word | obj, noun), and P(word | obj, adjective). Naive Bayes performance is slightly better and SVM is having a marginal decrease in performance, while there is a consequential decrease in accuracy when only adjective unigrams are used as features. But we know that these results are for the classification of reviews and may be used as verification for sentiment analysis on micro blogging websites like Twitter.

ACKNOWLEDGEMENT

The authors would like to extend their esteemed thanks to Mr. Krishan Kumar, Professor, IT Department, Inderprastha Engg. College Special thanks to Dr. Pooja Tripathi, HOD, IT Department, Inderprastha Engg. College.

REFERENCES

[1] Albert Biffet and Eibe Frank. Sentiment Knowledge Discovery in Twitter Streaming Data. Discovery Science, Lecture Notes in Computer Science, 2010, Volume 6332/2010, 1-15, DOI: 10.1007/978-3-642-16184-1_1

[2] Alec Go, Richa Bhayani and Lei Huang. Twitter Sentiment Classification using Distant Supervision. Project Technical Report, Stanford University, 2009.

[3] Tejal Rathod, Mehul Barot, Trend Analysis on Twitter for Predicting Public Opinion on Ongoing Events, International Journal of Computer Applications, 2018

[4] Hamid Bagheri, Md Johirul Islam, Sentiment analysis of twitter data, Iowa State University

[5] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2002.

[6] Chenhao Tan, Lilian Lee, Jie Tang, Long Jiang, Ming Zhou and Ping Li. User Level Sentiment Analysis Incorporating Social Networks. In Proceedings of ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), 2011.

[7] Afroze Ibrahim Baqapuri, Twitter Sentiment Analysis Project Technical Report, National University of Sciences & Technology Islamabad, Pakistan 2012

[8] Badruddin Kamal, Supervisor Abu M. Hammad Ali, Co-supervisor Dr. Mumit Khan, Application of sentimental analysis in adaptive user interfaces, Thesis Report, BRAC University, Dhaka, Bangladesh

[9] Avijit Pal, Argha Ghosh, Bivuti Kumar, Sentiment analysis on Twitter, Project Technical Report

[10] Forum Kapadia, Supervisor Dr. Leon Reznik, Sentiment Analysis Android Application, Project Technical Report, Rochester Institute of Technology Rochester, New York, 2017

[11] Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL), 2002.

[12] Rudy Prabowo and Mike Thelwall. Sentiment Analysis: A Combined Approach.Journal of Informetrics Volume 3, Issue 2, April 2009, Pages 143-157, 20