International Journal of Engineering &Technology, 7 (3.12) (2018) 351-353

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Real Time Opinion Mining and Analysis of Twitter Data

K. Senthil Kumar[1], Mohammad MusabTrumboo[2], Vaibhav[3], SatyajaiAhlawat[4]

[1]Asst. Professor, Department of Information Technology, SRM Institute of Science and Technology
Chennai, India
[2,3,4]Department of Information Technology, SRM Institute Of Science and Technology, Chennai, India
**Corresponding Author Email: [1]senthil.ku@ktr.srmuniv.ac.in, [2]iammusaib@gmail.com,*
[3]vaibhavtripathi1995@gmail.com, [4]satyajaiahlawat9@gmail.com

Abstract

This era, in which we currently stand, is an era of public opinion and mass information. People from all around the globe are joined together through various information junctions to create a global community, where one thing from the far east reaches to the people of the far west within seconds. Nothing is hidden, everything and anything can be scrutinized to its core and through these global criticisms and mass discussions of gigantic magnitude, we have reached to the pinnacle of correct decisions and better choices. These pseudo social groups and data junctions have bombarded our society so much that they now hold the forelock of our opinions and sentiments, ergo, we reach out to these groups to achieve a better outcome. But, all this enormous data and all these opinions cannot be researched by a single person, hence, comes the need of sentiment analysis. In this paper we'll try to accomplish this by creating a system that will enable us to fetch tweets from twitter and use those tweets against a lexical database which will create a training set and then compare it with the pre-fetched tweets. Through this we will be able to assign a polarity to all the tweets by means of which we can address them as negative, positive or neutral and this is the very foundation of sentiment analysis, so subtle yet so magnificent.

Keywords: *Sentimental Analysis, TextBlob, Naive Bayes Classifier, NLP, Tweepy, Twitter API.*

1. Introduction

In today's world, technology has developed to a point where most of the things can be achieved through it.It has spread across every field that a man can think of and it has broadened the applications and functionalities of various fields.

With all the development came a new field into existence, the field of sentiment analysis[2][3][11][12][14][15]. This science is relatively new and a lot has to be covered before it can be perfected.

Various methods and procedures are used in sentiment analysis to accomplish various outcomes. It can be used in smart advertisements in such a way that a particular person gets advertisements according to his liking.It can also be used to rate movies and other public releases such as books, TV shows and even political policies.

With the help of sentiment analysis, we can go through a lot of data from the internet and train it according to our needs and aims. In order to accomplish our goals and aims, some procedures are to be implemented and in this paper we will discuss everything with proper decorum..

We set our procedures by first collecting the tweets from twitter, using its developer API[2][3][4][8][18]. For that we used a library in python[8][9][20] called tweepy[14]. After the tweets are collected, we save them in a list. For the purpose of assignment of polarities to the fetched tweets we need a dataset[1][4][11][14] and a training-set[4][7][13][18]. For the training-set we use a lexical database, which includes a lot of phrases and sentences that are pre-analyzed into it with different polarities embedded in them. We create the training-set and the test-set from this file. Once the sets are completely formed and ready, we then require a classifier[1][4][7][17][19] that compares the data set with the collected tweets. The classifier that is going to be used in our program is called the Naive Bayes classifier.

We can implement the Naive Bayes classifier[5][7][8][9][11] through two methods:

1. We can develop a completely new code for it. and
2. Use another library that has naive bayes incorporated into it.

Here we used Text Blob[20] that is developed over NLTK[8][10][16], it helps us to achieve natural language processing by reading the tweets and removing various unnecessary data from it.

2. Proposed System

In the system that we have developed, we will use various technologies to achieve our desired results. To implement the process of sentiment analysis we have used Natural Language Processing [4][8][9][11][12].

NLP helps in the mining of opinions and expressions from tweets, files and texts.

To enable the incorporation of NLP, we used Machine Learning methods to achieve maximum results. Machine Learning methods [4][6][7][15][17] encourage two types of learning methodologies : the unsupervised method and the supervised method. We are going to use supervised method [8][11] in all our processes because of its accuracy and speed.

Now the next step would be to fetch a database and then implement one of the techniques from the supervised learning methods. To increase our accuracy and speed, even further, we used Naive Bayes classifier to classify the test data-set with the training data-set and then create a trained data-set that can be required further in the process. Naive Bayes classifier will enable us to assign a polarity to the texts or tweets in such a manner that we can distinguish between the negative, positive and neutral. All these processes take place after the fetching of the tweets is done through the uplink.

For our proposed system the process of fetching of tweets will be reiterated more than once, depending upon the user. Once all the data has been collected, it will be then processed so that it can be compared with each other.

Tweets regarding various movies, TV shows and products can be fetched and then compared to find out which product or TV show or movie is most favored by the crowd.

3. System Architecture (Block Diagram)

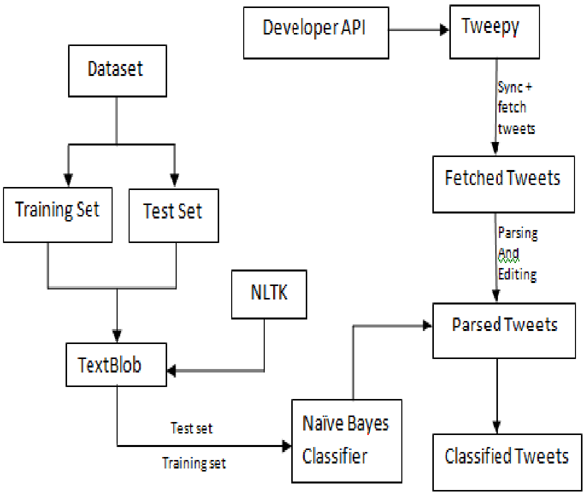


Fig. 1: System Architecture

Technologies used

1.Tweepy
2. TextBlob
3. Naive Bayes Classifier

4. Implementation Process

The implementation of the proposed system will completely take place in python, by using its vast network of libraries. We start the process by first fetching the tweets from twitter using twitter's developer API. The twitter API will be accessed and enabled using the Python library called Tweepy. The number of tweets that can be fetched completely depends upon the user. The user can choose any number of tweets that he wants to get fetched by entering the number manually. Tweepy does not only help us to fetch the tweets but it also allows us to authenticate various required keys with ease.

After the twitter API is synced with the running code, the next step is to fetch tweets based on a particular search string. This search string will enable us to fetch tweets by targeting a particular group of tweets, all related to the search string. The search will be done based on the hashtags and not on the actual words or phrases inside the tweets. This probing of hashtags will be done by an independent and open source python Library called TextBlob. Textblob is built over NLTK library. Hence, it has all the functions of NLTK in it but with extra refinement. After a particular hashtag has been searched and all the tweets that are required are fetched and saved into a list in python, they are processed and certain type of excess data is removed from them, such as special characters, numbers, multiple hashtags etc to make them appear as normal sentences.

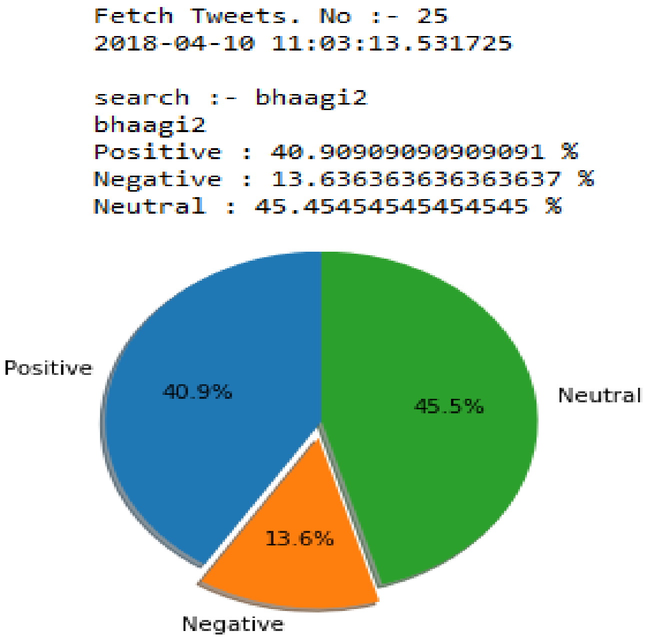
This is also done through the extensive functionality of Textblob Library. While textblob purifies, parses and processes the tweets and preserves them into a new list, it also begins a second procedure to develop a trained data set. The trained dataset will be developed by processing the training-set and the test-set , both of these sets are created from a single database. For the purpose of our code we'll use STS Gold Tweet Corpus[3][7][9]. The mentioned corpus has a lot of text data which is already embodied with their respective polarities. The text data included in this corpus is from various movies, reviews, plays, novels and other literary sources.

This corpus will be called forth using its .csv file through TextBlob. After the file has been accessed it'll be split into two lists one for the training data and other for the test data. Now these two lists will be processed with the help of TextBlob and then a sustainable trained dataset is created which will have polarities assigned to each text item. After this process is done, the processed trained dataset is then saved into a list and then it is compared with the fetched and parsed tweets which are already stored into a separate list. These two lists are processed together to assign polarities to the list which has tweets stored in it. This assignment of the polarities to all the tweets is done through the TextBlob library, as well but this time the process is a bit complicated and an external algorithm is required to ease the process and increase its accuracy and speed.

This process of assigning the polarities to the various tweets is done with the help of a classifier and for our implementation process we'll use the Naive Bayes classifier. The whole list of tweets will be run through it and by comparing it to the trained dataset it will assign a polarity to each of the tweets. These polarities will range from below 0 to 4, where below 0 means negative and above zero means positive and equal to zero stands for neutral.

In our system we'll be searching more than one hashtag so that we can later on compare them and find out which one is the most favorable item amongst the searched hashtags.

5. Results (Graphs and Diagrams)



```
Fetch Tweets. No :- 25
2018-04-10 11:03:13.531725

search :- bhaagi2
bhaagi2
Positive : 40.90909090909091 %
Negative : 13.636363636363637 %
Neutral : 45.45454545454545 %
```

Fig. 2: Pie chart for a movie

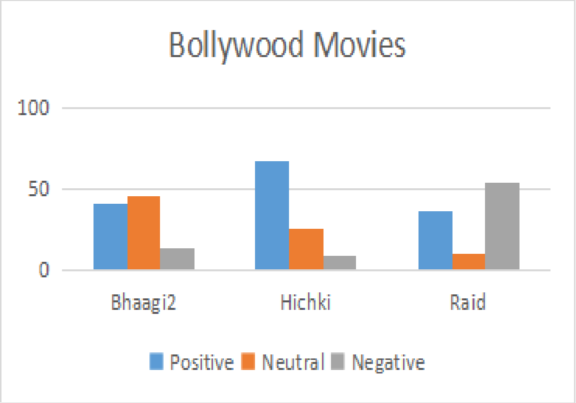


Fig. 3: Comparison of movies using bar chart

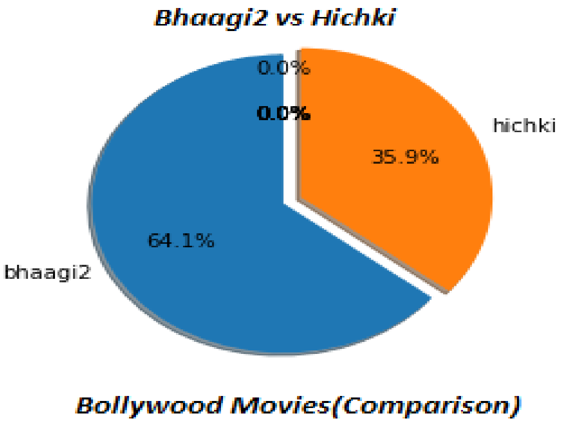


Fig. 3: Comparison of movies using pie chart

With the inclusion of Naive Bayes algorithm into our program the output generated is more accurate than most other algorithms available on the knowledge hubs, not only that, the output achieved is the faster as well.

6. Conclusion

We have taken the most out of the technologies that are currently present in the field of sentiment analysis but this field is still vast and not much has been developed to call it a fully functional field.

It currently works on the principle that the data which is being input is unbiased and that the polarity of the texts that are taken into consideration is proportional to the actual number of texts that are present on a particular query but this principle is not correct and still requires finesse.

More work can be done in the development of techniques that can be used to understand sarcasm and emotional content.

References

[1] Selvan, L.G.S. and Moh, T.S. “A Framework for Fast-Feedback Opinion Mining on Twitter Data Streams.” Published in: Collaboration Technologies and Systems (CTS), 2015 International Conference on 1-5 June 2015.

[2] Rana, S. and Singh, A. “Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques.” 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016.

[3] Trupthi, M., Pabboju, S. and Narasimha, G. “Sentiment Analysis On Twitter Using Streaming Api.” Published in Advance Computing Conference (IACC), 2017 IEEE 7th International 5-7 Jan. 2017.

[4] Soni, A.K. “Multi-Lingual Sentiment Analysis of twitter data by using classification algorithms.” Published in Electrical, Computer and Communication Technologies (ICECCT), 2017 Second International Conference on 22-24 Feb. 2017.

[5] Phand, S.A. and Phand, J.A. “Twitter Sentiment Classification using Stanford NLP” Published in Intelligent Systems and Information Management (ICISIM), 2017 1st International Conference on 5-6 Oct. 2017.

[6] Mehra, R., Bedi, M.K., Singh, G., Arora, R., Bala, T. and Saxena, S. “Sentimental Analysis Using Fuzzy and Naive Bayes.” Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC).

[7] Goel, A., Gautam, J. and Kumar, S. “Real Time Sentiment Analysis of Tweets Using Naive Bayes.” 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016.

[8] Batrinca, B. and Treleaven, P.C. “Social media analytics: a survey of techniques, tools and platforms.” AI and Society. Volume 30. Issue 1(2014). pp 89-116.

[9] Fang, X. and Zhan, J. “Sentiment Analysis Using Product Review Data.” Journal Of Big Data (2015) 2:5

[10] Nielsen, F.A. “A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.” Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages (2011) 93-98.

[11] Kharde, V.A. and Sonawane, S. “Sentiment Analysis of Twitter Data: A Survey of Techniques.” International Journal of Computer Applications 139(11): 5-15, April 2016.

[12] Hutto, C.J. and Gilbert, E. “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.” Proceedings of the Eighth International Association for the Advancement of Artificial Intelligence (AAAI) Conference on Weblogs and Social Media.

[13] Salathé, M. and Khandelwal, S. “Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control.” PLoS Computational Biology 7(10): e1002199

[14] Jain, V. “Prediction of Movie Success using Sentiment Analysis of Tweets.” The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue: The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE’13], San Francisco State University, CA, U.S.A., March 2013.

[15] Gunther, T and Furrer, L. “GU-MLT-LT: Sentiment Analysis of Short Messages using Linguistic Features and Stochastic Gradient Descent.” Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 328–332, Atlanta, Georgia, June 14-15, 2013.

[16] Palanisamy, P., Yadav, V. and Elchuri, H. “Serendio: Simple and Practical lexicon based approach to Sentiment Analysis.” Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 543–548, Atlanta, Georgia, June 14-15, 2013.

[17] Filho, P.P.B. and Pardo, T.A.S. “NILC_USP: A Hybrid System for Sentiment Analysis in Twitter Messages.” Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 568–572, Atlanta, Georgia, June 14-15, 2013.

[18] Rosenthal, S., Farra, N. and Nakov, P. “SemEval-2017 Task 4: Sentiment Analysis in Twitter.” Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 502–518, Vancouver, Canada, August 3 - 4, 2017.

[19] Coppersmith, G., Dredze, M. and Harman, C. “Quantifying Mental Health Signals in Twitter.” Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 51–60, Baltimore, Maryland USA, June 27, 2014.

[20] Anuradha, G. and Varma, D.J. “Fuzzy Based Summarization of Product Reviews for Better Analysis” Indian Journal of Science and Technology, Vol 9(31), August 2016.