Product Based Tweets Detection and Analysis using NLP

Piyush Kumar Singh[1], Ayushi Chourey[2], Tamanna Pancholi[3], Saurabh Vijay[4]
UG Student, Department of Computer Engineering, SVKM's NMIMS, Shirpur
Bhushan Inje[5]
Assistant Professor, Department of Computer Engineering, SVKM's NMIMS, Shirpur

Abstract- In this world of Internet anything and everything is decided by the online activity of a particular individual or body, and this is given by comments, reviews or feedbacks of the users affiliated with it directly or indirectly. These reviews/feedbacks can be the deciding factors for the survival of different individuals in this world. Therefore analyzing these reviews/feedbacks is important, because there can be many spammed reviews or irrelevant comments in the section for feedback and these can affect the individual for whom the comments or reviews matter to survive in the world.

For analysis data from Twitter is taken in the form of tweets which will be in the text format, this data is analyzed to find and determine whether the text provided as input gives a negative or neutral or a positive sentiment, to perform this local polarity of various sentences in the text is identified and evaluation of relationship between them is done, this results in the total text's polarity. This is performed with the help of Natural Language Toolkit of NLP by comparing the words in the data that are meaningful with the dictionary present in NLP. This analysis of results will identify the positivity and negativity of the reviews as per requirement, now this result can be used for the betterment of the particular individual/company.

Keywords- Sentiment Analysis, NLTK, Review analysis NLP, Tweets

I. INTRODUCTION

These days if there is anything happening in this world then it will have its presence on twitter, where many individuals tweet regarding that event from their individual accounts. Now these can be read easily but why not use these tweets[1] in the form of data for determining a positive or negative image of that event happening and use it in favor of the event towards its betterment by processing and analyzing these tweets in the form of data. By doing this and giving an appropriate result at the end of the process can help in ways that is important to the event or the person who has put the event in motion. To perform this method first we need the data on which this process is to be applied, now either we can have someone collect the data manually but this will take a lot of time and this is not beneficiary for anyone.

So to have some benefit from this method we need to feed the system with the data that is live i.e. if someone is tweeting currently than as soon as the tweet is posted we must have the data, this can be done by many methods but we are going to use the tweepy API, which gives reading access to other users for the live data feed on twitter. Now after reading this data through the system this data is then saved into a database where we can access this data at any time use this to process and analyze in different ways that are according to the requirement.[6] If someone needs to analyze the data at current time then it can be done and even if someone needs to analyze this data at any other time than also it can be done since this data is not lost or damaged until and unless done so. What this system focuses on is the positivity and negativity of the content or data provided to it, where positivity is the support that is shown in regarding to that event and negativity is the lack of support to the event.

In section II we have discussed some of the related works, which are somewhat related to our system. In section III we have discussed our proposed system, its architecture and working.

II. RELATED WOKS

Falguni Gupta *et al.* have focused on the classifying the tweets on the basis of location, their work extends the basic classification models by some new features of location. Their results show how these more included features achieve decent improvement in the performance of previous systems. [1]

Gaurav Dubey *et al*. have focused on the data on social media for important facts and information required to make business decisions which are beneficial to a particular company, to tackle this problem they have used text mining techniques in order to perform sentiment analysis for the user generated data on the big faces of Indian politics. [2]

Sushant Kokate *et al.* have discussed an approach based on the behavior of spammers to detect spammed reviews for the manipulation of reviews on some products that are targeted specifically. In order to classify the data they are using J48 Classifier and after collecting this data they are analyzing the behavior of reviews and detecting the spammed reviews. [3]

Hari Bhaskar Sankaranarayanan *et al.* have considered the travelling passengers and their reviews on the travel agencies, this data can be useful to those travel agencies and others if processed properly. To perform this the data is collected from data lakes and analyzed using NLP and machine learning algorithms and tools. [4]

Sahar Jambi *et al.* have focused on a previous system that exists to collect real time data known as EPIC, but had one problem that it cannot analyze that data in real time and not efficiently, so they proposed their own components and used some previously developed products and were able to analyze big data and provide accurate results efficiently in time. [5]

Jingjing Cao *et al.* have designed a framework for taking advantage from review information and the rating for the recommendation performance advancement. They have extracted sentimental feature from reviews done on products. [6]

Xiaojia Pu *et al.* have pointed explicitly towards the overall sentiment in the sentences that play an important and crucial role to determine the level of sentiment. To handle this issue they presented one effective method which is based on Structural SVM for utilizing the overall opinion for analysis of sentiment in the sentence. [7]

Ruxi Yin *et al.* have reviewed various integration algorithms and design methods for the comment spam detection, the comment spam detection is a first starting part for their integrated algorithm. For they have introduced their own classifier through they have proved that detection of spams gives an accurate credibility to the analysis of the reviews on a single or some particular products. [8]

III. PROPOSED SYSTEM

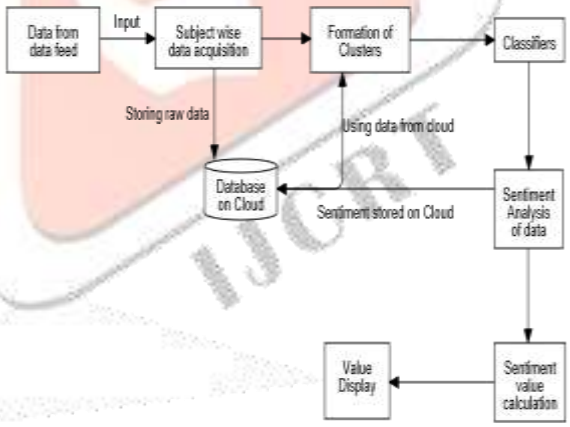

Fig. 1 Architecture of the system

1. *Data Acquiring*

To acquire the raw and real time data the system is making use of tweepyAPI provided by twitter. This API has a single owner who is given a unique id by twitter. Now the API gives the owner two keys where on is consumer key and another is consumer secret key, one sample key is shown in following figure.

Fig. 2 One sample for the consumer key

Now to make any API requests on one's accounts behalf they are provided by Access token and Access token secret, one sample access token is shown in following figure.

Fig.3 One sample for the Access key

Now to collect data from twitter our system makes use of these keys and tokens which are unique for each and every account and user.

2. *Filter the data*

The data if acquired will have lots of tweets from various fields, topics and events. But we need data according to the requirement, the data is filtered by a single most common unit that will be a word (i.e. an event's name, a person's name, etc.), by doing this we will filter out every unnecessary data which is not required. Now if someone requires every tweets regarding IPL then we will select the word as 'IPL' to filter the data and have the required tweets to be analyzed. Now during filtering the data if a tweet is found then the output is shown in figure below.

Fig. 4 Output Screen for found tweet

During filtering if required data is not found then the tweet is skipped, the output is shown in figure below.

Fig. 5 Output screen for a skipped tweet

3. *Storage*

The data needs to be stored in order to perform various analysis on the data. This data can either be stored on a remote storage or on a cloud storage, we prefer cloud since accessing the data on cloud will be easy and analyzing the data on cloud will save time and resources. The output screen for storage of data on the cloud can be seen in the following figure.

Fig. 6 Output screen for the data on Cloud

4. *Analysis of data*

The data that is stored on the cloud is then used for analysis, this analysis is done by the help of Sentiment analysis which uses Natural Language Toolkit aka nltk.

a. Sentiment Analysis

Sentiment analysis is used to perform one very detailed analysis of data in the form of text coming from different sources. The provided text is analyzed to find and determine whether the text provided as input gives a negative or neutral or a positive sentiment, to perform this local polarity of various sentences in the text is identified and evaluation of relationship between

them is done, this results in the total text's polarity. Sentiment analysis also uses advanced NLP techniques to detect the polarity associated with both concepts and entities present in the text.

5. Result

The result to be displayed will be the positive or negative sentiment displayed by the text. This sentiment of the textual data will be calculated for the tweets coming in on the basis of timing and on arrival of new tweets their sentiment will be calculated by summing them in the data and calculating the sentiment for the whole data globally. Further analysis will be done until there are no more tweets or the live feed of data is switched off.

The analyzed result needs to be displayed in a format, now either we display the data on the terminal in textual form or we can use that data to make a graphical representation of that data which updates in real time as real time data feed can never be stopped and since there is always data incoming then this data is going to be analyzed as well and the analyzed result will be displayed. The display of result on the terminal screen is shown in the figure shown below.

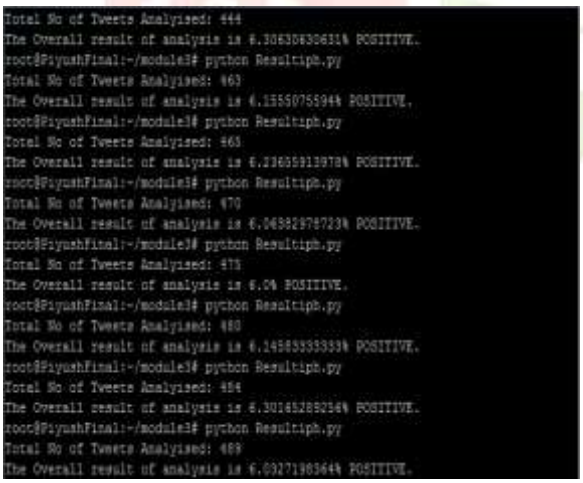

Fig. 7 Output screen for the calculated sentiment

IV. CONCLUSION

There is a lot of data online but to collect this data properly and according to the needs is not possible easily, and if done manually it is time and resource consuming, and after this is done then also it is not easy to perform analysis on that collected data and give a valid output which should not consume much time. We performed this task by using Naive Bayes classifier, K- means for processing the data and used Natural Language Tool Kit for the analysis of data. This experiment proved to be successful and provided accurate and appropriate results at the end which are not time consuming. We performed sentimental analysis to get the output by which we can determine the product growth and the changes that to be done in its future model. This result give us the total outcome of all the tweets in terms of positive, neutral or negative polarities using Naive Bayes classifier and k-means and with the use of natural language processing we filter the content on basis of above polarities, this give us perfect idea of product use in market.

In this proposed system we find polarity for reviews of a single tweet from every user and then give the polarity of that tweet, then the system finds polarity for all tweets similarly. Then mean for polarity of all tweets is calculated which gives the final positivity and negativity, by this the result that we get at last is more accurate than the previous systems which use all reviews at once and provide the sentiment.

V. REFERENCES

1. Falguni Gupta, Swati Singal "Sentiment Analysis of the Demonitization of Economy 2016 India Region wise" *2017 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence*, IEEE 2017.
2. Gaurav Dubey, Shilpi Chawla, Kirandeep Kaur "Social Media Opinion Analysis for Indian Political Diplomats" *2017 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence.*
3. Sushant Kokate, Bharat Tidke "Fake Review and Brand Spam Detection using J48 Classifier" International Journal of Computer Science and Information Technologies, Vol. 6 (4), IEEE 2017.
4. Hari Bhaskar Sankaranarayanan, Jayprakash Lalchandani "Passenger Reviews Reference

Architecture using Big Data Lakes" 2017 *7th International Conference on Cloud Computing, Data Science & Engineering.*

5. Sahar Jambi, Kenneth M. Anderson "Engineering Scalable Distributed Services for Real-Time Big Data Analytics" 2017 IEEE Third International Conference on Big Data Computing Service and Applications.
6. Jingjing Cao, Wenfeng Li "Sentimental Feature based Collaborative Filtering Recommendation" BigComp 2017, IEEE 2017.
7. Xiaojia Pu, Gangshan Wu and Chunfeng Yuan "Sentiment Analysis with the Exploration of Overall Opinion Sentences" National Science Foundation of China under Grant No.61321491 and No.61223003, and Collaborative Innovation Center of Novel Software Technology and Industrialization. 2017 IEEE
8. Ruxi Yin, Hanshi Wang, Lizhen Liu "Research of Integrated Algorithm Etablishment of a Spam Detection System" 2015 4th International Conference on Computer Science and Network Technology, 2015 IEEE.