A Baseline Based Deep Learning Approach of Live Tweets

Anjana Jimmington
M.Tech Scholar , Department of Computer Science and Engineering, Kerala Technical University, Malabar College of Engineering and Technology, Desamangalam-Kerala

Abstract- **In this scenario social media plays a vital role in influencing the life of people. Twitter , Facebook, Instagram etc are the major social media platforms . They act as a platform for users to raise their opinions on things and events around them. Twitter is one such micro blogging site that allows the user to tweet 6000 tweets per day each of 280 characters long. Data analyst rely on this data to reach conclusion on the events happening around and also to rate a product. But due to massive volume of reviews the analysts find it difficult to go through them and reach at conclusions. In order to solve this problem we adopt the method of sentiment analysis. Sentiment analysis is an approach to classify the sentiment of user reviews, documents etc in terms of positive (good), negative (bad), neutral (surprise). I suggest an enhanced twitter sentiment analysis that retrieves data based on a baseline in a particular predefined time span and performs sentiment analysis using Textblob. This scheme differs from the traditional and existing one which performs sentiment analysis on pre saved data by performing sentiment analysis on real time data fetched via Twitter API. Thereby providing a much recent and relevant conclusion.**

Index Terms- micro blogging, twitter, sentiment analysis, Textblob, Twitter API

I. INTRODUCTION

In the past few years, there has been a huge growth in the use of micro blogging platforms such as Twitter. Spurred by that growth, companies and media organizations are increasingly seeking ways to mine Twitter for information about what people think and feel about their products and services. Apart from that data analysts also make use of this data for interpreting about eminent personalities and various events happening.

The online medium has become a significant way for people to express their opinions and with social media, there is an abundance of opinion information available. Using sentiment analysis the polarity of opinion can be found such as positive, negative or neutral by analyzing the text of the opinion. Sentiment analysis has been useful for companies to get their customer's opinions on their products predicting outcomes of elections, and getting opinions from movie reviews. The information gained from sentiment analysis is useful for companies making future decisions.

Many traditional approaches in sentiment analysis uses the bag of words method. The bag of words technique does not consider language morphology, and it could incorrectly classify two phrases of having the same meaning because it could have the same bag of words. The relationship between the collection of words is considered instead of the relationship between individual words. When determining the overall sentiment, the sentiment of each word is determined and combined using a function. Bag of words also ignores word order, which leads to phrases with negation in them to be incorrectly classified. Other techniques discussed in sentiment analysis include Naive Bayes, Maximum Entropy, and Support Vector Machines.

Sentiment analysis refers to the broad area of natural language processing which deals with the computational study of opinions, sentiments and emotions expressed in text. Sentiment Analysis (SA) or Opinion Mining (OM) aims at learning people's opinions, attitudes and emotions towards an entity. The entity can represent individuals, events or topics. An immense amount of research has been performed in the area of sentiment analysis. But most of them focused on classifying formal and larger pieces of text data like reviews.

With the wide popularity of social networking and microblogging websites and an immense amount of

data available from these resources, research projects on sentiment analysis have witnessed a gradual domain shift. The past few years have witnessed a huge growth in the use of microblogging platforms. Popular microblogging websites like Twitter have evolved to become a source of varied information. This diversity in the information owes to such microblogs being elevated as platforms where people post real time messages about their opinions on a wide variety of topics, discuss current affairs and share their experience on products and services they use in daily life. Stimulated by the growth of microblogging platforms, organizations are exploring ways to mine Twitter for information about how people are responding to their products and services. A fair amount of research has been carried out on how sentiments are expressed in formal text patterns such as product or movie reviews and news articles, but how sentiments are expressed given the informal language and message-length constraints of microblogging has been less explored.

Twitter is an innovative microblogging service aired in 2006 with currently more than 550 million users . The user created status messages are termed tweets by this service. The public timeline of twitter service displays tweets of all users worldwide and is an extensive source of real-time information. The original concept behind microblogging was to provide personal status updates. But the current scenario surprisingly witnesses tweets covering everything under the world, ranging from current political affairs to personal experiences. Movie reviews, travel experiences, current events etc. add to the list. Tweets (and microblogs in general) are different from reviews in their basic structure. While reviews are characterized by formal text patterns and are summarized thoughts of authors, tweets are more casual and restricted to 140 characters of text. Tweets offer companies an additional avenue to gather feedback. Sentiment analysis to research products, movie reviews etc. aid customers in decision making before making a purchase or planning for a movie. Enterprises find this area useful to research public opinion of their company and products, or to analyze customer satisfaction. Organizations utilize this information to gather feedback about newly released products which supplements in improving further design. Different approaches which include machine learning(ML) techniques, sentiment lexicons, hybrid approaches etc. have been proved useful for sentiment analysis on formal texts. But their effectiveness for extracting sentiment in microblogging data will have to be explored. A careful investigation of tweets reveals that the 140 character length text restricts the vocabulary which imparts the sentiment. The hyperlinks often present in these tweets in turn restrict the vocabulary size. The varied domains discussed would surely impose hurdles for training. The frequency of misspellings and slang words in tweets (microblogs in general) is much higher than in other language resources which is another hurdle that needs to be overcome. On the other way around the tremendous volume of data available from microblogging websites on varied domains are incomparable with other data resources available. Microblogging language is characterized by expressive punctuations which convey a lot of sentiments. Bold lettered phrases, exclamations, question marks, quoted text etc. leave scope for sentiment extraction. The proposed work attempts a novel approach on twitter data by aggregating an adapted polarity lexicon which has learnt from product reviews of the domains under consideration, the tweet specific features and unigrams to build a classifier model using machine learning techniques.

II. LITERATURE SURVEY

The related work section covers the other aspects of Twitter data usage, with an entirely different approach as discussed in the thesis. An analysis of Big Data technologies Info Sphere Big Insights and Apache Flume [6] was conducted by Birjali et al. Multiple sets of data for various research purposes was first collected from Twitter by Apache Flume, stored in Hadoop, and then displayed with Big Sheets after being ana-lyzed using Info Sphere Big Insights. They chose Twitter as their Big Data source, due to the increasingly large amount of data generated daily by its users. This method uses the Hadoop Distributed File System (HDSF) in order to utilize the Map Reduce feature, enabling the collection of larger data sets (Tweets). Map Reduce counts the number of times a matching data set is iterated and then displays the results. Apaches Flume Next Generation (NG) was used to collect the Tweets used in this case study. Flume NG uses a process that first collects data (Tweets) from multiple sources and

holds them in memory, and then stores them in the HDSF using JAQL script, which is a data processing and query language. After a thorough examination of Info Sphere Big Insights analytics, a separate data collection tool developed from Apache Flume was tested, and the results were analyzed using Info Sphere Big Insights. It was determined that the technique used by the tool developed from Apache Flume was not only superior to older methods, but faster as well. A paper on the Intelligent Mining of Public Social Networks Influence in Society(MISNIS) tool [7] highlights several key limitations on current methods, such as Twitter API restrictions and dependency on hashtags and keywords for categorization, and demonstrates how MISNIS overcomes these limitations, increasing productivity by 80% and 40% respectively. MISNIS uses polarity sentiment analysis, and does not use a language dependent lexicon. While this approach is limiting, it does not negate MISNISs apparent superiority, and is open to further development in future. Joao P. Carvalho and his collaborators [7] demonstrate MISNIS by applying it to track, catalogue, analyze, and trace current events in Portugal; however, MISNIS canbe applied in many other fields with various other research questions. It can collect, store, manage, mine, and display data by using Computational Intelligence, Information Retrieval, Big Data, Topic Detection, User Influence and Sentiment Analysis. This method uses geolocation to collect Tweets within Twitter's API restriction of1% data collection, then traces the collected Tweets back to the users accounts to collect additional Tweets that meet the search criteria from multiple Twitter API accounts. A file of every viable user was created and maintained to facilitate this process. Mongo DB was used for all data storage, and a REST API was used to handle the data once it was collected. In addition, the REST API is also the tool used to collect data from individual users. This method does not make collection limitless, as it is also minimally restricted by Twitter. An insightful exploratory analyzer, demonstrates the capabilities of Tweets Characterization Methodology (TCHARM) [8] to organize collected Tweets based on geographical location, the time of the Tweet, as well as its contents. TCHARM uses the Text And Spatio-TEmporal (TAST) distance measure in order to group similar Tweets based on all three categories. This means that TCHARM is capable of grouping Tweets about the same, or similar subjects, from geographically close, or specified regions, that were Tweeted around the same time. The case study conducted in this paper to demonstrate TCHARMSs performance searched for and categorized Tweets related to the 2014 FIFA world cup. Through this study it was determined that the TAST feature utilized by TCHARM produced a more even distribution of the three factors tested for by TCHARM than did other methods. The authors also address avenues for future work based on TCHARMs limitations. One such limitation is the length of time it takes to set the specifications of TCHARMs features. It is also suggested that the K-means algorithm used by TCHARM may collect too broad a range of Tweets containing the three factors for categorization. While this means that some collections of Tweets are more loosely related than is desirable, it does not affect the overall higher efficiency demonstrated by this method. TCHARM can handle a high number of Tweets in its data collection due to its use of Apache Spark as its platform, and collects Tweets quickly on an hour to hour recurring basis.

III. LIVE TWEET ANALYSIS SYSTEM

In this system we suppose that a user in general searches for tweets related to a particular keyword at current time using his twitter credentials, retrieves tweets and finally performs sentiment analysis on them so as to reach at a conclusion.

A. Architecture

The following system shows the architecture of the proposed scheme. The system consists of four modules.

Creating Twitter API

In order to retrieve live tweets based on baseline, the user should initially request twitter for its authentication credentials.

Tweets Retrieval

Here tweets are retrieved from the twitter API dynamically based on the Keyword name input and given count.

Preprocessing

The tweets are imported to a .csv file from the twitter API, these tweets consist of unnecessary words, whitespaces, hyperlinks and special characters. First we need to do filtering process by removing all unnecessary words.

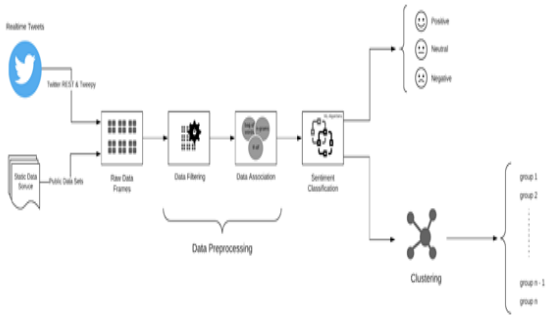

Figure 1. System Architecture

Sentiment Analysis

Sentiment analysis is finally done on the preprocessed data

B. Proposed Scheme

In this method we uses textblob as a method to find the polarity of the text (positive text, negative text or neutral text). The tweets are imported from the Twitter using the (API) provided by the Twitter Developer. From these API various fields like tweets, source, retweets, likes, language, user etc. can be scrapped. After collecting these data, we can analyses the various famous person thoughts on anevent or occasion

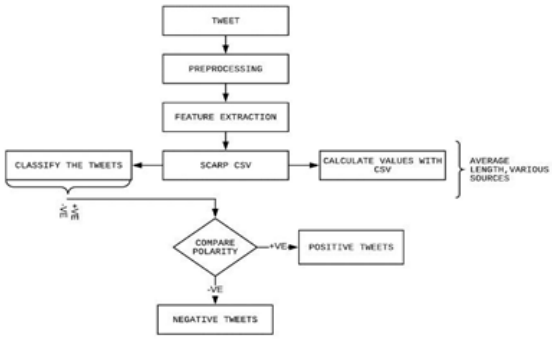

Fig 2 Architectural Flow of Twitter Analysis

The figure 2 explains the extraction of tweets id from twitter through API, then preprocess the data that are extracted. Preprocessing includes exclusion of unwanted fields, segregating the fields important for analysis. Once the fields are extracted and segregated CSV is created. Using this CSV, the length of the message, Likes, retweets for the id is extracted and various results are derived. With the scraped tweets, classify the tweets whether positive or negative or neutral.

C. Dataset Description

In this proposed system, we have used the dataset called result.csv which contains the newly fetched tests Rdata set. csv. It contains the following fields Tweets, Len, ID, Date, Source, Likes, RT's (Retweets), SA (Sentimental Analysis).

D. Software Description

In the system the graphs such as Table, Bar graph, Line graph are generated with the help of Jupyter notebook. The predefined functions are pandas, numpy, matplotlib, pyplot, list, Dictionary. Pandas is used for converting from csv file to dataset. Numpy is one of the essential library for scientific calculating in Python. It delivers a high-performance multidimensional array object, and apparatuses for experimenting with these arrays. Python comprises of numerous built-in container categories: lists, dictionaries, sets, and tuples. A list is the Python equal of an array, but is resizable and can contain elements of different types. A dictionary stores (key, value) pairs, like a Map in Java or an object in JavaScript. Python library such as Text Blobare used for processing the textual data. It provides API for processing natural language processing (NLP) such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. Tweepy isused for accessing Twitter API and it is open sourced.

E. Data Analysis and Visualization

In Twitter users tweets their opinion on an occasion or anything including a commodity or even an personality.. From their thoughts, importance of that occasion and the polarity of their tweet are analysed. Some of the analysis with the dataset as follows.

- Visualize the various source of the tweet.
- Calculate the polarity of the tweets fetched
- Visualize the Polarity of tweet (positive, negative, neutral)
- Calculate the general review of the tweeters
- Calculate the individual review of the tweeters
- Visualise the tweeters opinion in the form of pie graph

F. Advantages of Proposed Scheme

1. The system gives us a review on day to day happenings.
2. Provides impartial reviews.
3. Fast analysis
4. Easily understandable by all.

IV. EVALUATION

Our scheme has a few differences from traditional password based scheme. The first is the adopting live streaming of data. The second is that the output value is tweeters current opinion.

Based on these features, our proposal has advantages as follows:

- Lower computational cost
- High Accuracy
- Supporting privacy of users

The polarity of tweets can be expressed at different levels whether the expressed opinions in a document or sentence is either positive or negative. The subjectivity of tweets is basically finding of subjective words and text that show the presence of opinions. In the result shown inTable2 we can see the polarity of each baseline.

V. RESULT

This is the sample output for the project for the keyword Donald Trump for 1000 tweets.

```
Enter Keyword/Tag to search about: DonaldTrump
Enter how many tweets to search: 1000
```

Fig 3 Searched keyword and number of tweets

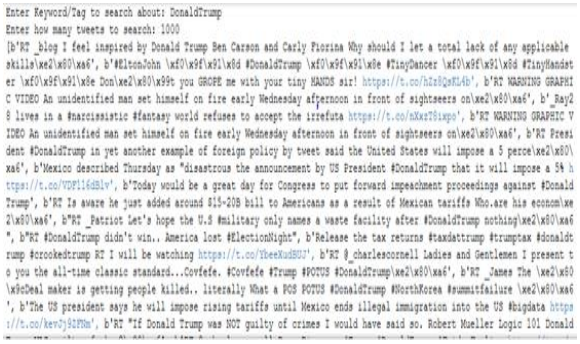

Fig 4 Fetched Tweets for Donald Trump

```
How people are reacting on DonaldTrump by analyzing 1000 tweets.

General Report:
Tweeters donot like DonaldTrump
```

Fig 5 General Tweeters Report

```
How people are reacting on DonaldTrump by analyzing 1000 tweets.

General Report:
Tweeters donot like DonaldTrump

Detailed Report:
3.70% people thought it was positive
28.90% people thought it was weakly positive
3.00% people thought it was strongly positive
1.90% people thought it was negative
5.70% people thought it was weakly negative
0.10% people thought it was strongly negative
41.00% people thought it was neutral
```

Fig 6 Detailed Report

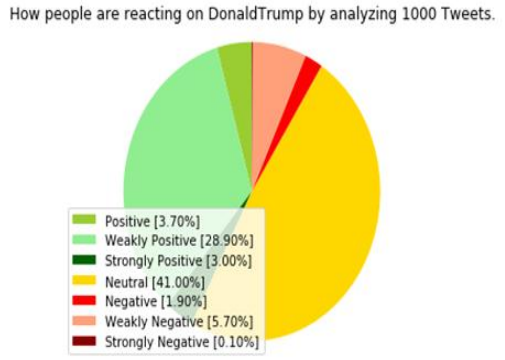

Fig 7 Pie Chart

VI. CONCLUSION

Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on analyzing the sentiments of the tweets and feeding the data to a machine learning model to train it and then check its accuracy, so that we can use this model for future use according to the results. It comprises of steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model. This research topic has evolved during the last decade with models reaching the efficiency of almost 85%-90%. But it still lacks the dimension of diversity in the data. Along with this it has a lot of application issues with the slang used and the short forms of words. Many analyzers don't perform well when the number of classes are increased. Also, it's still not tested that how accurate the model will be for topics other than the one in consideration. Hence sentiment analysis has a very bright scope of development in future.

A. Future Scope

We can perform deep sentiment analysis of text, in different areas of application. It is not adequate to say that a text is an inclusive positive or inclusive negative. Users would like to know which separate topics are talked about in the text, which of the mare

positive and which are negative. So, there will be an inclination towards greater use of NLP techniques (such as syntactic parsing),in addition to machine learning methods.

- A more elaborate web-based application can be made for my work infuture
- By using various classification strategies we further improve the results
- By the use of sentiment analysis, I forecast the future consequence s or at least anticipate them better, when people tweet about present scenario.

REFERENCES

[1] https://www.sas.com/enus/insights/analytics/big-data-analytics.html.

[2] https://www.mckinsey.com/industries/healthcare-systems-and-services/our- insights/the-big-data-revolution-in-us-health-care.

[3] https://www.slideshare.net/BernardMarr/big-data-25-facts.

[4] https://www.lexalytics.com/technology/sentiment.

[5] M. Birjali, A. Beni-Hssane, and M. Erritali, \Analyzing social media through big data using infosphere biginsights and apache ume," Procedia Computer Science, vol. 113, pp. 280-285, 2017.

[6] J. P. Carvalho, H. Rosa, G. Brogueira, and F. Batista, \Misnis: An intelligent platform for twitter topic mining," Expert Systems with Applications, vol. 89, pp.374-388, 2017.

[7] X. Xiao, A. Attanasio, S. Chiusano, and T. Cerquitelli, \Twitter data laid almost bare: An insightful exploratory analyser," Expert Systems with Applications, vol. 90, pp. 501-517, 2017.

[8] A. O. Durahim and M. Coskun, \# iamhappybecause: Gross national happi- ness through twitter analysis and big data," Technological Forecasting and Social Change, vol. 99, pp. 92-105, 2015.

[9] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, \Detecting suicidality on twitter," Internet Interventions, vol. 2, no. 2, pp. 183-188, 2015.

[10] Y. Yu and X. Wang, \World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans tweets," Computers in Human Behavior, vol. 48, pp.392-400, 2015.

[11] F. Corea, \Can twitter proxy the investors' sentiment? the case for the technol- ogy sector," Big Data Research, vol. 4, pp. 70-74, 2016.

[12] https://klout.com/corp/score.

[13] M. Daniel, R. F. Neves, and N. Horta, \Company event popularity for financial markets using twitter and sentiment analysis," Expert Systems with Applications, vol. 71, pp. 111-124, 2017.

[14] A. C. Pandey, D. S. Rajpoot, and M. Saraswat, \Twitter sentiment analysis using hybrid cuckoo search method," Information Processing & Management, vol. 53, no. 4, pp. 764-779, 2017.

[15] N. Oliveira, P. Cortez, and N. Areal, \The impact of microblogging data for stock market prediction: using twitter to predict returns, volatility, trading volume and survey sentiment indices," Expert Systems with Applications, vol. 73, pp. 125-144, 2017.

[16] Y. Huang, D. Guo, A. Kasako, and J. Grieve, \Understanding us regional linguistic variation with twitter data analysis," Computers, Environment and Urban Systems, vol. 59, pp. 244-255, 2016.

[17] E. Sulis, D. I. H. Faras, P. Rosso, V. Patti, and G. Ruo, \Figurative messages and in twitter: Differences between# irony,# sarcasm and# not," Knowledge-Based Systems, vol. 108, pp. 132-143, 2016.

[18] M. Oussalah, B. Escallier, and D. Daher, \An automated system for grammatical analysis of twitter messages. a learning task application," Knowledge-Based Systems, vol. 101, pp. 31-47, 2016.

[19] https://statistics.laerd.com/spsstutorials/wilcoxon-signed-rank-test-using- spss-statistics.php.

[20] M. A. Moreno, A. Arseniev-Koehler, D. Litt, and D. Christakis, \Evaluating college students' displayed alcohol references on facebook and twitter," Journal of Adolescent Health, vol. 58, no. 5, pp. 527-532, 2016.

[21] X. Lin, K. A. Lachlan, and P. R. Spence, \Exploring extreme events on social me- dia: A comparison of user reposting/retweeting behaviors on twitter and weibo," Computers in Human Behavior, vol. 65, pp. 576-581, 2016.

[22] C. S. Park and B. K. Kaye, \The tweet goes on: Interconnection of twitter opinion leadership,

network size, and civic engagement," Computers in Human Behavior, vol. 69, pp. 174-180, 2017.
[23] A. Acar and Y. Muraki, \Twitter for crisis communication: lessons learned from japan's tsunami disaster," International Journal of Web Based Communities, vol. 7, no. 3, pp. 392{402, 2011.
[24] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, \Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," Computers in Human Behavior, vol. 63, pp. 433{443, 2016.
[25] NaiveBayesian.http://www.statsoft.com/Textbook/Naive-Bayes-Classier.
[26] http://scikit-learn.org/stable/modules/svm.html.
[27] RandomForest.http://www.stat.berkeley.edu/breiman/RandomForest/ cc home.htm.
[28] R. Daniulaityte, R. W. Nahhas, S. Wijeratne, R. G. Carlson, F. R. Lamy, S. S. Martins, E. W. Boyer, G. A. Smith, and A. Sheth, \time for dabs: Analyzing twitter data on marijuana concentrates across the us," Drug & Alcohol Depen- dence, vol. 155, pp. 307{311, 2015.
[29] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, \Predicting u trends using twitter data," in Computer Communications Workshops (INFOCOM WK- SHPS), 2011 IEEE Conference on. IEEE, 2011, pp. 702{707.
[30] V. Kayser and A. Bierwisch, \Using twitter for foresight: An opportunity?" Futures, vol. 84, pp. 50{63, 2016.
[31] A. Nakhasi, R. Passarella, S. G. Bell, M. J. Paul, M. Dredze, and P. Pronovost, \Malpractice and malcontent: Analyzing medical complaints in twitter," in 2012 AAAI Fall Symposium Series, 2012.
[32] S. Gaglio, G. L. Re, and M. Morana, \A framework for real-time twitter data analysis," Computer Communications, vol. 73, pp. 236{242, 2016.
[33] https://www.djangoproject.com/.
[34] https://aws.amazon.com/.
[35] http://textblob.readthedocs.io/en/dev/.
[36] http://scikit-learn.org/stable/autoexamples/model selection/plot precision recall.html.
[37] http://www.pewinternet.org/2016/11/11/social-media-update-2016/.