International Journal of Innovative Research in Computerand Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488|

||Volume 8, Issue 5, May 2020||

Sentiment Analysis of Election Related Tweets Using Twitter API

Prof. Balaji Bodkhe[1], Lakshita Patel[2], Amey Mahakhode[3], Rahul Rane[4], Prathamesh Metkar[5]

Assistant Professor, Modern Education Society's College of Engineering Pune, India[1]

UG Students, Modern Education Society's College of Engineering Pune, India[2,3,4,5]

ABSTRACT: In the context of text classification for sentiment analysis, the instances are naturally fuzzy and therefore to get a clear-cut outcome by extracting the opinion in a new innovative and fuzzy combination way and assign a relevant sentiment to the tweets on Twitter, usually either positive or negative is taken into consideration. Due to this new approach we get an advancement for extracting the opinions of people more deeply even if instance in a sentence are manipulated more complexly along with better fault tolerance and also, not to forget, it can be used for Cyberhate speech detection. Using Cloud Architecture for deploying the system gives us even more flexibility to access it from anywhere in the world

I.INTRODUCTION

From the last few decades analyzing the twitter data using different analysis techniques has became popular. Twitter has become the point of fascination for several researchers to predict the outcome of various events like elections, cricket matches, customer-brands, stock-market, movie review, influence of particular politician. The meaning of word sentiment is the opinion or view of a person towards the particular area. In this paper we focus on sentiment analysis of election tweet using twitter API.

With the increased use of twitter, it is easy to find the contents on any topic and also the people's view on the particular topic. As people are free to express their views openly on twitter related to any topic or event. A domain that is particularly discussed on twitter is politics. As more and more people express their views and opinions related to politics, twitter has become the most valuable platform of people's opinion. Political predication is very interesting topic to analyze and find the people's opinion about the different political parties, reviews of the people on different political leaders and that can help to predict whether the person or the party can win the elections. In our project we mainly focus on predicting the result of elections, and predict the winning chances of particular political leader based on the reviews he/she has on twitter. We are extracting the tweets using twitter API, using twitter API we will extract the real time tweets and the extracted tweets are stored in the .csv file. This tweets are extracted using hashtags and using twitter access token Preprocessing is done on the collected tweets i.e. removal of noise and useless data. After the preprocessing phase the features or the reviews are extracted andthen compared.Based on the reviews tweets are classified as good, average and bad.The tweets are than compared on the basis of the keywords. This keywords are than passed to the classification algorithm which will classify the tweets in good, average or bad tweets.

This will generate huge amount of data which will be stored and processed in Apache Hadoop. Apache Hadoop is an architecture which works efficiently and perform computation on the large datasets. Cloud computing functionality such as hadoop is very efficient on the distributed data. Hadoop has an internal framework which is known as MapReduce, which has two task Map task and Reduce task. Map task takes the set of data and converts it into another set of data in which individual elements are broken down into different tuples. The Reduce task takes the output of previously executed map task as input and combines those data tuples into smaller set of tuples.

As soon as data is available on hadoop analytical techniques are applied on the tweets which are collected. Apache Mahout provides different clustering, classification and categorization algorithm which is then applied on the dataset. The next step is to train our data model using mahout, which consists of different algorithm. The next step is to use our classifier in order to check if our classifier is working well or not the Apache Mahout splits the dataset into two parts i.e. training model and the testing model. The training model is used to train the classifier and the testing model is used to test the training model. Using the testing model the accuracy of the training model and classifier is checked. From the collected tweets we will give 75% of the tweets to build the training model and remaining 25% is used to build the testing model which are selected randomly. After the analysis we will be able to find the opinion of the people on different candidates which are standing for the elections and can find whether they are good, average or bad according to the people's poll.

This whole data is stored on the cloud as cloud provides flexible storage system we can scale services to meet our needs and we can access the cloud services even without internet services. Cloud also provides high security for the data. Cloud architecture is best suited for changing workloads . So for storing the tremendous amounts of tweets cloud is the best option for storage.

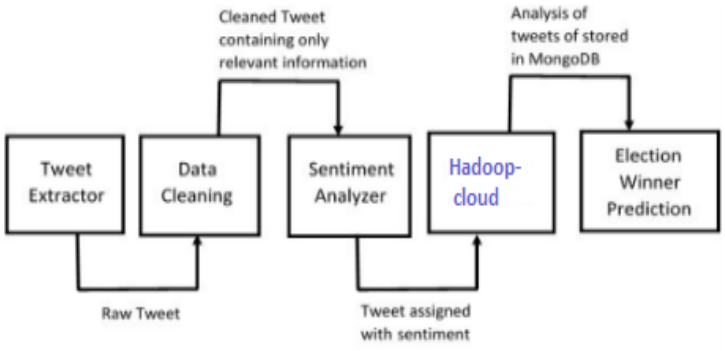

Fig. 1. Architecture of the proposed model

Figure 1 : System flow

II.LITERATURE SURVEY

Most of the authors, expertise and every other people in today's world use the platform of social media to discuss their views and opinions on world's day to day happenings. Almost 6 billion people in today's world use social networks to discuss their views and share data among each other. One of such social media platforms provided for today's world is twitter. Almost 140 million tweets per day are tweeted every day. This creates a large amount of data if considered for analysis and storage of such big data. Here are some techniques used by authors to store this large amount of twitter data and analyse it.

Traditionally the classification of Spastic Hemiplegia (SH) was done by specialized physicians only, but with the introduction of Support Vector Machine (SVM) as a technique for classification, classification of SH disease has become possible. The problem of recognizing which Type of SH a patient suffers is a multi-class pattern recognition task which is done by the svm easily. The result analyses shown by the authors A. J. Salazar, O. C. De Castro, R. J. Bravo in this paper indicates that the use of SVM is a model for adequate Gait pattern recognition. While still further research is necessary. According to the reference from above paper we can conclude that SVM also have some disadvantages. A common one is the lack of transparency in results. SVM is mostly used in the field of medical.[1]

In support vector machine we rank the input factors by the weight vector w in the decreasing order. The rank of the element decides its priority. According to Jing-Rong Chang, Mu-Yen Chen, Long-Sheng Chen, And Shu-Cih Tseng suggests that SVM algorithm is mostly used in medical field such as medicine classification, disease identification and so on to diagnose the disease in a short period of time. Here also we can see that it is used for medical field.[2]

In this paper the authors Simon D. Duque Anton, Sapna Sinha, Hans Dieter Schotten, two data sets captured in industrial environments are analysed for attack-based anomalies. SVMs are used to detect attacks of seven different categories and 35 different subtypes. According to the reference from above paper we can conclude that SVM also have some disadvantages. A common one is the lack of transparency in results. Since the dimensions might be very high, SVM may not be able to predict the accuracy of the events properly.[3]

From the above reference we can say that Support Vector Machine is mostly used in medical fields where the ordering or the ranking is necessary, but in our project, we mainly focus on predicting the outcomes of the particular events. So, the SVM algorithm would not be the accurate algorithm for our project. The most accurate algorithm for our project is Naive Bayes Classifier.

There are many people who are disabled and are unable to perform and use the complex computers and lack knowledge for using different complex devices. Computer technology enables people with severe disability to perform and operate any

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488|

||Volume 8, Issue 5, May 2020||

computer related tasks. Authors has proposed a decision tree is a model used for making predictions through a classification process. Through a decision tree, experts can describe questions and figure out ways to solve problems.[4]
In this paper authors Yi Hou, Praveen Edara, and Carlos Sun proposed a lane changing assistance system that would help the driver to change the lane using different parameters. They have used combination of two algorithms i.e. Bayes classifier and decision tree. When both the algorithms are combined gives good result which is more accurate and more precise.[5]
The problem with decision tree is that if there is small change in the data it can cause a large change in the structure of the decision tree causing instability and in our proposed system data is changing constantly. Also training of decision tree is relatively expensive as complexity and time taken is more.

Objective Quality Assurance plays a very important and vital role in the quality of the image in this paper authors Soo-Chang Pei, Life Fellow, and Li-Heng Chen has proposed various ways to improve the quality of the image by the use of different classification algorithm like random forest. With the rapid development in image processing quality of the image is the main aspect author suggests various measures to improve the quality using random forest.[6]
Target detection is aimed at detecting and identifying target pixels based on specific spectral signatures, and is of great interest in hyperspectral image (HSI) processing. Target detection can be considered as essentially a binary classification. Random forests have been effectively applied to the classification of HSI data. However, random forests need a huge amount of labelled data to achieve a good performance, which can be difficult to obtain in target detection. In this paper, authors Yanni Dong, Student Member, Bo Du, and Liangpei Zhang propose an efficient metric learning detector based on random forests, named the random forest metric learning algorithm, which combines semi multiple metrics with random forests to better separate the desired targets and background.[7]
Random forest has some disadvantages like anensemble model is inherently less interpretable than an individual decision tree. So, we used Naïve-Bayes classifier algorithm.

In this paper authors Alberto Sanchis, Alfons Juan, and Enrique Vidal proposed confidence estimation which is majorly used in speech recognition to detect words in the recognized sentence that have been likely to be misrecognized. Confidence estimation can be seen as a pattern for classification problem in which a set of words are obtained for eachof the presumed word in order to find whether it correct or incorrect. Authors has proposed a smoothed naïve Bayes classification model to which benignant combines these words. The model itself is a combination of word-dependent and word-independent Naïve-Bayes models. As in statistical language modelling, the purpose of the generalized model is to smooth the estimates given by the specific models. This classification model than compared with confidence estimation. Thisresults shows that the good performance of word graph-based probabilities can be improved by using the naïve Bayes classification algorithm.[8]

Authors in this system tells us about clinical decision support system, which makes the use of advanced data mining techniques to help clinician make proper decisions, has received good attention in recent years. In this paper authors proposed with large amounts of clinical data generated every day, Naïve-Bayesian classification can be used to dig the valuable information which can be used to improve clinical decision support system. Even though clinical decision support system is very promising but it can have security issues and privacy problems. [9]

In this paper authors has proposed a new technique for signal classification and jamming detection in wide-band radios. Theory of compressed sensing is uncovered to recover the population which is distributed through WB spectrum from sub-Nyquist samples, this reduces the high rate of sampling requirements at the receiver. From the recovered features of each narrow-band signal are extracted. These features are then used to train a simple and powerful classifier, that is Naïve-Bayes classifier. The training model of naive bayes classifier is used classify Naive bayes signals into their respective modulations and it also detect the jamming on different Naive Bayes signals, which are the main contributions of this paper. The proposed algorithm is then evaluated with different heuristic setups and is shown to perform better when compared to a recently proposed feature-based jamming detection algorithm.[10]

We are using Naive Bayes algorithm as Naive Bayes requires a small amount of training data to estimate the test data. So, the training period is less. It works well on larger data sets. It is easy to implement. It works properly in any situation where decision tree, random forest and SVM does not provide accurate results. The accuracy of Naïve-Bayes algorithm is better than the accuracy of other algorithms.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488|

||Volume 8, Issue 5, May 2020||

In paper [11], author makes use of twitter authentication service providing tokens for the twitter data for every transaction with twitter server. Using these tokens, tweets are mined using *hashtags.* These mined tweets are then obtained in a .csv file which are later broken into various forms. Punctuation is removed which is a disadvantage as sometimes punctuations can point to a neutral or calm mindset of the user who tweeted. Upper cases are converted to lower which hides the emotions of the person in tweet as upper case words can denote anger or a feeling of excitement from the user. Removal of control characters also degrades the analysis of data as it displays the feeling of the person who tweeted regarding that particular event. This storage of twitter data into .csv file gives access to only limited number of tweets which is very least accurate and least sufficient if compared to today's highly generated data.

In paper [12], author Paramita Ray and AmlanChakrabarti have although used lexicon analysis, they still have overcome the emoticon analysis which has been described in paper [11]. Still this paper [12] removes punctuation which hides the sentiments of the user which must have been described by the user in the tweets by using symbols such as exclamations. This symbols after removal will not provide proper analysis of that particular tweet and hence will not provide accurate analysis. The removal of digit is another issue as in today's world, young generation mostly use short word slangs which contain digits in them instead of the proper spelling. On removal of such digits, it might cause loss of words which describe the sentiment of the user who tweeted the tweet. Alphanumeric characters also depict the emotions and important points tweeted in the tweet of the user and removal of this alphanumeric characters will damage the sentiment analysis accuracy. Removing URLs will remove some important URLs tweeted by user in order to provide a specific data through the means of video which will indirectly depict the emotions of the person who tweeted the tweet regarding that particular event. Thus, whenever the tweets are supposed to be stored on the cloud, they would be stored as a whole, which gives additional advantage for profane storage.

In paper [13], author mines the tweets which are geo-tagged. Therefore, tweets of only a specific location are taken into consideration for analysis which is not sufficient if compared to overall tweets which are being tweeted every second from the whole world. Also, the miner uses a single system to perform the mining and analysis of tweets which can be very slow and inaccurate at times. On the other hand, if the system is deployed in cloud, tweets from multiple locations can be accessed through multiple servers deployed on various locations in the world without any delay. This alternative also provides replication and fault-tolerance in case one system fails, the other one still keeps working absolutely fine.
The author in paper [14] uses retweet package because it allows you to extract up to 20,000 tweets and in a suitable format. But looking into it statistically, there are 7000 tweets every second being tweeted. Therefore, 20,000 tweets technically are a data of just 3 seconds which is not at all worth to consider for analysis. Whereas, the flow data to be stored on cloud is continuous and also can consider comparatively large number of tweets with the help of Hadoop-cloud. Thus, we cannot consider the above-mentioned technique for handling large number of tweets for analysis.

In paper [15], author discusses about analysis of data from twitter API by storing them in MySQL. In today's world more than 140 million tweets tweeted every day. This large amount of twitter data of today's world is not possible to store in MySQL as it is not much scalable. MySQL is not much scalable and does not provide the ability to run in heterogeneous environment. MySQL will not provide parallel computing for the analysis of large data which is a big requirement of today's heavy data generating world. Therefore, the twitter API stored using MySQL eventually falls insufficient in making proper analysis of today's twitter data.

The author from paper [16] uses NoSQL techniques like MongoDB for storage of Twitter API data and analysis data. MongoDB is considered better than MySQL as it is more scalable and provides storage of unstructured data in flexible JSON like document format. Although MongoDB provides high scalability along with horizontal scalability, it is still not enough to store the day to day's high generating data every minute. The processing of this every minute generating data needs parallel computing which is not supported in MongoDB. This is one of the reason MongoDB also falls insufficient in the analysis of today's high generated of twitter data.

As we have seen many users use MySQL, NoSQL or Hadoop to store a large amount of twitter data. Also, these users use various processing techniques like MapReduce of Hadoop which requires high end machines which are generally expensive and require a lot of maintenance. This processing also degrades the performance of machines gradually leading to possibilities of loss of data and server failures eventually. Therefore, Hadoop-cloud can be considered as one of the best options for the storage of this large amount of data of today's high data generating world. Using Hadoop-cloud, user need not to worry about the maintenance of the machines used for processing as all the maintenance related tasks are done by the

cloud service provider. Hadoop-cloud is highly scalable which is very much required. The provision of storage of varied data sources also plays an important role which can be very much needed at times. It is more of easier to use and is very cost-effective providing a very high throughput and performance. It is highly fault-tolerant too along with the feature of open source. Moreover, Hadoop-cloud supports multiple languages and most importantly, parallel computing can be done easily using MapReduce. As a result, Hadoop-cloud is completely compatible for storage of today's large amount of data and analysis of the same.

REFERENCES

1. Jing-Rong Chang, Mu-Yen Chen, Long-Sheng Chen, and Shu-Cih Tseng, "Novel approach for spastic hemiplegia classification through the use of SVM"
2. A. J. Salazar, O. C. De Castro, R. J. Bravo, "Why customers don't revisit in tourism and hospitality industry?" In the Proceedings of the 26th Annual International Conference of the IEEE EMBS San Francisco, CA, USA.
3. Simon D. Duque Anton, Sapna Sinha, Hans Dieter Schotten, "Anomaly- based Intrusion Detection in industrial data with SVM and random forest".
4. IEEE transactions on neural systems and rehabilitation engineering, vol.20, no.4, July 2012, "Pruning a Decision Tree for Selecting Computer-Related Assistive Devices for People with Disabilities".
5. Yi Hou, Praveen Edara, Carlos Sun, "Modelling Mandatory Lane Changing Using Bayes Classifier and Decision Trees" In the IEEE transactions on intelligent transportation systems.
6. Soo-Chang Pei, Life Fellow, Li-Heng Chen, "Image Quality Assessment Using Human Visual DOG Model Fused With Random Forest" In the IEEE transactions on image processing, Vol. 24, No. 11, November 2015.
7. Yanni Dong, Bo Du, Liangpei Zhang, "Target Detection Based on Random Forest Metric Learning" In the IEEE journal of selected topics in applied earth observations and remote sensing, Vol. 8, No. 4, April 2015.
8. Alberto Sanchis, Alfons Juan, Enrique Vidal, "A Word-Based Naïve Bayes Classifier for Confidence Estimation in Speech Recognition" In the IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 2, February 2012.
9. Ximeng Liu, Rongxing Lu, Jianfeng Ma, "Privacy-Preserving Patient-Centric Clinical Decision Support System on Naive Bayesian Classification." In the IEEE Journal of Biomedical and Health Informatics, December 2014.
10. M. O. Mughal, S. Kim, "Signal Classification and Jamming Detection in Wide-band Radios Using Naive Bayes Classifier" In the Journal of Latex Class Files, Vol. 14, No. 8, August 2015.
11. Dr. Nirmala C R, Roopa G M, Naveen Kumar K R, "Twitter Data Analysis for Unemployment Crisis" In the 2015 International Conference on Applied and Theoretical Computing and Communication Technology.
12. Paramita Ray, AmlanChakrabarti, "Twitter Sentiment Analysis for Product Review Using Lexicon Method" In the 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI) Zeal Education Society, Pune, India
13. Yutaka Arakawa, ShigeakiTagashira and Akira Fukuda, "Relationship Analysis between User's Contexts and Real Input Words through Twitter". In the IEEE Globecom 2010 Workshop on Ubiquitous Computing and Networks.
14. Sahar A. El_Rahman, FeddahAlhumaidiAlOtaibi, Wejdan Abdullah AlShehri, "Sentiment Analysis of Twitter Data" In the 2019 International Conference on Computer and Information Sciences (ICCIS).
15. J. Jayadharshini, R. Sivapriya, S. Abirami, "Trend square: An Android Application for Extracting Twitter Trends Based on Location" In the Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India.
16. WiesławWolny, "Knowledge Gained from Twitter Data" In the Proceedings of the Federated Conference on Computer Science and Information Systems, 2016.