International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 4 • 2017

Navie Bayes and K-Means Hybrid Analysis for Extracting Extremist Tweets

Himanshu Nagar[1], Chetna Dabas*[2] and J.P. Gupta[3]

[1] *Department of Computer Science and Engineering Jaypee Institute of Information Technology, JIIT Noida, India*
[2] *Department of Computer Science and Engineering Jaypee Institute of Information Technology, JIIT Noida, India*
[3] *Ex Chancellor Lingaya's University Faridabad, India*

***Abstract*: Tremendous growth of data has affected the organizations way of handling data. In order to handle this big data in an effective manner, softwares are being used to gather and pre-process the same. The prime challenge associated with big data in organizations is that of analyzing and extracting semantic information for future actions in minimum time. Predicative analysis is a study related to predication of a future trends and corresponding outcomes. Machine learning and regression technique are commonly used to carry out predictive analysis. Due to the outstanding performance in handling large- scale datasets with noisy data machine learning techniques have become quite popular in conducting predictive analysis. The aim of this paper is to identify the impact of preprocessing and sentiment analysis on the performance of Naïve Bayes classifier and K-Means clustering considering the fact the Naïve Bayes classifier is an effective method for carrying out sentiment analysis and K-Means clustering is an effective method for clustering the positive and negative tweets. The proposed work presents the preprocessing on ISIS Twitter dataset and then applies Naïve Bayes and K-Means clustering. The result of the proposed work is visualized using a confusion matrix. It was observed a part of this work that even with high level of noise; K-Means method gave better result in extracting extremist tweets from non-extremist tweets.

Keywords: RStudio, Naïve Bayes, K-Means, World Cloud**.**

1. INTRODUCTION

Social media is new platform of communication among people all over the world. People express their reviews and opinion on social media site such as WhatsApp, Facebook, Instagram and Twitter. This reviews and opinion will be good or bad to the affected side. In the last few year, there is increase the interest to extract sentiment from the text. In today's world people use many social networking sites to share the ideas and their views by comments and sharing on all type of subjects on daily basis, the sites being used are like Youtube and Twitter and it is the easiest way to do it. The data and information obtained from the data of these social networking sites is in a huge amount and can be of good use for the businesses for decision making so business are very much interested in the data to extract knowledge for their future decisions. For example, predictive models derived from social media for successful product may facilitate mall mart owner making more profitable decisions.

Natural Language Processing (NLP) research, there is quite a challenge to train computer to handle huge data that are easily differentiated by human. The step that must undergo to make the opinion can be analyzed by automated system is it need to be quantified, turning the fuzzy of emotion into something that are measurable, then it can be calculated, classified and clustering . Now days, there are many different active terrorist group in the world now a days and almost every day twitter, Facebook and Newspapers report about terrorist attacks in different kind of places. Many terrorist organizations use the Internet to spread propaganda. Propaganda usually includes virtual messages, presentations, audio and video files that contain explanations, justifications or promotion of terrorist activities. Objective of terrorist propaganda is to generate an anxiety and fear in a population by releasing violent videos like killing people who fight against terrorist groups. One of the most common approaches to stop these groups is to suspend accounts that spread propaganda when they are discovered.

The most important process in data mining is the data preprocessing, which prepare the raw dataset for cleaning and then for the further processing. After data preprocessing we apply sentiment package on dataset. This sentiment package gives the polarity of tweets. Tweets polarities are positive, negative, neutral. We create train and test data sets then we are applying the classification algorithm classifying the tweets and clustering the negative words and positive words. The purpose of this study is addressing the problem of classifying tweets as supporting ISIS analyzes and implements the most appropriate method for sentiment analysis ISIS tweets dataset. This work may be to helpful analysts to detect twitter users that promote radical views.

The whole study is organized in the following manner: literature study related to Naïve Bayes and K-Means techniques us given in Section II, existing methods and techniques are discussed in Section III, Section IV presents the experimental setup used. The proposed method and work done for the method is explained in Section V in which the description of data set used is also given. In Section VI results obtained and analyzed are given in detail and finally the entire study is concluded in section VII.

2. LITERATURE REVIEW

Author describes the data preprocessing step, which prepare the raw dataset for the further steps. Aim of this study identify the impact of preprocessing step apply of a dataset and performance of Naïve Bayes classifier, which identify the spam emails. Author compares the result of both non-preprocessed dataset result and preprocessed dataset result [1].

Author compares the Fuzzy C-Means and K-Means clustering technique a noisy and huge datasets. They used free cloud computing solution Apache Mahout/Hadoop. Author use the English Wikipedia's articles dataset which is 30GB dataset and files format are XML. After applying preprocessing step of dataset size reduces, the file is 11GB. They found Fuzzy C-Means can lead to worse cluster quality than K-Means [2].

Author describes stock market predication using sentiment analysis. They use stock dataset and text dataset. Text dataset is microblog text, which is use for sentiment analysis. Stock data is use for predict the stock market. They use 3 modules which are: Microblock Filter, Sentiment Analysis, Stock Predication. Microblock Filter module is based on Latent Dirichlet Allocation to get the financial microblogs. Sentiment Analysis module first sets-up a financial lexicon method, and then got the sentiments of the microblogs obtained from the Microblock Filter Module. The Stock Predication Module proposes a user-group model, which modified the significance of different people and merges it with stock historical dataset to predict the move of the Shanghai Composite Index [3].

Author describes a personality of a human behavior. People are required to take a personality test to find their personality. Social site is the place where person express themselves to the world. Posts made by users of social site can be analyzed to obtain their personal information. Author use the text classification to predict their personality based on text written by Twitter user. Author use two language English and

Indonesian. Classification method use is Naïve bayes and Support Vector Machine. Naive Bayes slightly outperformed the other methods with 60% [4].

Author present to classified tweets sentiments used Naïve Bayes classification algorithm based on trainers' perception, which divided into three categories: negative, positive or neutral. In this paper Author use 25 trainers participated. Each trainer classifies the sentiments of 25 tweets of each keyword, and then result from the classification training was then used for as an input for naïve Bayes. The accuracy of this naïve bayes is 90% ± 14% measured by total number of correct tweets per total classified tweets. Author archive high of accuracy using Naïve Bayes technique. Naïve Bayes method is suitable to train data and classify sentiment from twitter, Facebook and other social network data. Naïve Bayes technique is good to classify large number of tweets data [5]

Author explained sentiments mining from social networking site for Arabic language. They use two approached first is naïve bayes and second approached was traditional probabilistic classifier algorithm. They use social networking site facebook as a source of data. They determine the polarity of tweets. They consider their results to be good because half of the corpus was classified. Moreover, the accuracy is 85%. Accuracy of classified tweets was high [6].

Author apply data mining tool to predicting box office performance of movies. They collected dataset from social networking sites and web sources like: YouTube, IMDb movies and Twitter dataset. Author label three class: Hits, Neutral and Flop. They use weka tool for K-means clustering algorithm implementation. They apply the sentiment analysis on comment extracted form YouTube. They apply three steps; first step author normalized the train data set. Second step applying K-Means clustering algorithm and third step they apply predicative modeling. They predict popularity of top actress is deciding to the success of movies [7].

Author is explaining Sentiment analysis using Naïve Bayes classifier. They use Indonesian language feedback from OSNs of "XYZ Online Distro pages''. They predict Indonesia small medium enterprise organization market product. They have a problem on citizen talked in Bahasa and local language so build a sentiment analysis model is not easy task. They proposed approached use feature extraction and selection to selected words from learning dataset of Indonesian corpus and then classifying dataset to the classes of target objects and sentiments. The level of accuracy proposed by author is 97.27% and recall 86.86%. This study help to small enterprise industry and medium enterprise industry in Indonesia to have a better understanding of their target market needs and wants [8]

In this journal Author Erik Cambria explain the affective computing and sentiment analysis. Affective computing and sentiment analysis is most important of artificial intelligent and all research area. It has great capability of customer relationship management and recommendation system. Business intelligence is also an important factor behind corporate interest in the area of sentiment analysis and affective computing. Basic task of this paper is polarity detection and emotion recognition. Polarity classification more advanced analyses. It can be apply on sentiment and find pro and con expression. Another task is multimodal fusion that means integrate all single modalities into single representation. Two type of fusion technique feature level and decision level. It is used to improve the emotion recognition. This technique identify the polarized words, smile, voice pitch, pauses, gaze as relevant feature. Three main categories for sentiment analysis: knowledge base, statistical method, hybrid approaches [9]

3. EXISTING METHODS AND TECHNIQUES

3.1. Naïve Bayes:

Naive Bayes is a probabilistic classifier algorithm based on applying Bayes Theorem. Naive Bayes computes the probability p of a document x being of class y : p (y|x). Given a document x to be classified, represented by a vector x = (x_1, ...,x_n) the conditional probability can be written using Bayes' theorem as:

$$p(y|x)=\frac{p(y)\,p(x|y)}{p(x)}$$

Naïve Bayes model is robust and stable, easy to interpret and computation efficient because independent variables are assumed.

3.2. K-MEANS

Clustering is a partitioning base technique. In partitioning base technique all the object is consider initial a single cluster. Object is divided into number of partition by iteratively locating the point between the partitions. K-Means algorithm has long development history. K-Means algorithm is popular clustering algorithm. K-Means clustering select K initial cluster center randomly. K is parameter, which value assigned by the user. K is a number of clusters. Each data object in a dataset will be assigned nearest cluster.

Suppose Z ={z_1, z_2, z_3, …….., z_n} N is data object in dataset Z.

K is number of cluster user assumption.

The distance between the two vectors i.e. data points and centroid is measured using many measures which are Cosine similarity distance measure, Euclidean distance measure (EDM), Squared EDM. Here we have used the Euclidean distance measure for calculating the distances.

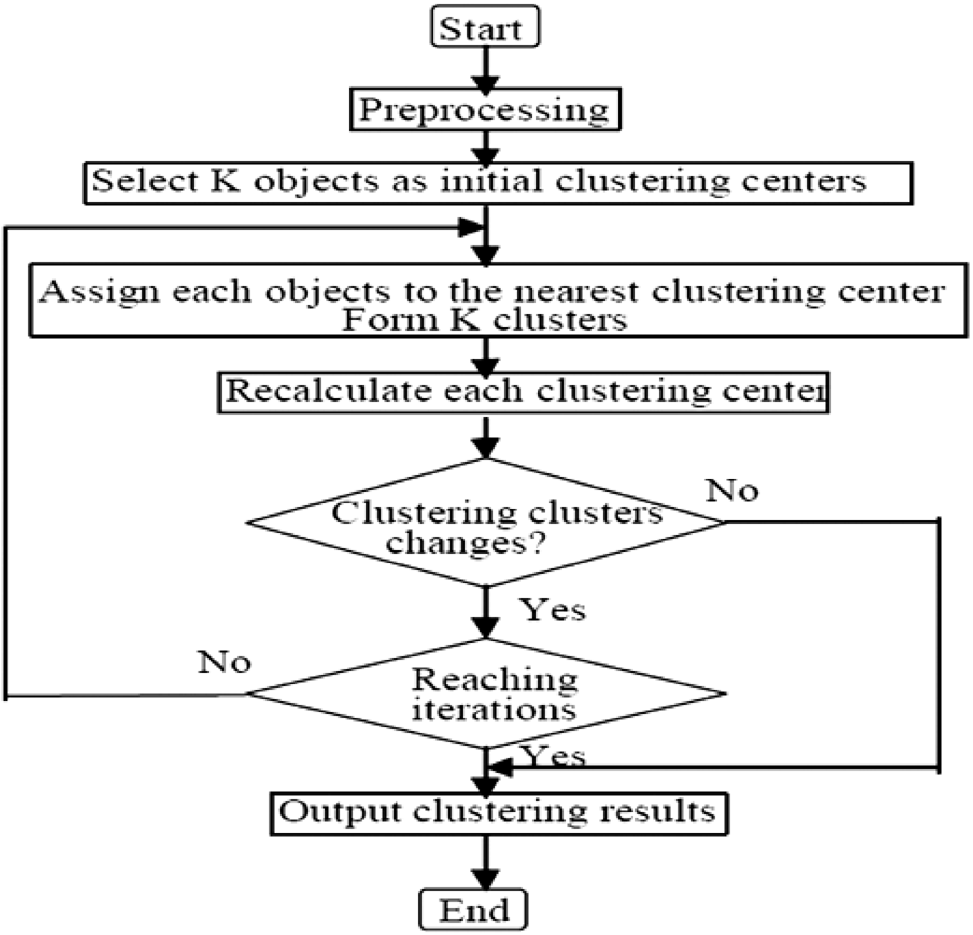

Start
Preprocessing
Select K objects as initial clustering centers
Assign each objects to the nearest clustering center
Form K clusters
Recalculate each clustering center
Clustering clusters changes?
No
Yes
No
Reaching iterations
Yes
Output clustering results
End

Figure 1: K-Means Clustering [21]

Euclidean distance between two data object

$$D_{\left(z_i z_j\right)} = \sqrt{\left(zi1 - zj1\right)2 + \left(zi2 - zj2\right)2 + \ldots + \left(zix - zjx\right)2}$$

Where is $z_i = (z_{i1,} z_{i2,\ldots\ldots} z_{ix})$ and $z_j = (z_{j1,} z_{j2,\ldots\ldots} z_{jx})$ .z_i and z_j are the data object with x dimension.

Step Procedure of K-Mean algorithm:

Input: Dataset Z is an input data.

Output: K cluster C_1, C_2,……C_k.

- Chooses K data objects as the initial clusters C_1, C_2,……C_k
- According to distance formula calculate the distance between each data object and each cluster center and assign each data object to the smallest cluster center.
- Again calculate clustering center of each cluster.
- If the cluster centers do not change or the iterations has reached assigned iterations value the algorithm end, if not go to step 2 and go on [21].

4. EXPERIMENTAL SETUP

The experimental were conducted using a tool called RStudio. This tool supported the R language (version 3.3.1). It is suitable of the machine learning technique. In our experiment we use two different algorithms Naïve Bayes classifier and K-Means Clustering. Experiment performs on following hardware configuration: RAM 4GB, Processor Intel Core i3.

5. PROPOSED METHOD AND WORK DONE

A. Social media is not communicates with families and friends but also promote some negative tweets and radical tweets. First step is collecting proper data sets for our model.

B. In a raw data format data can consist of variables. Only the Tweets in ISIS Twitter data are selected.

C. When the data is collected the data is not clean due to several reasons like: misspelled words, retweet, strange symbols, etc. Preprocessing step is the most important step. Preprocessing step is where all the selected tweets get clean up. We apply the some steps like: Removing Retweet Tag, All URL, Some Symbols, Spelling Correction and lexical analysis (means splits the text into words).

D. The tweet dataset is divided into trained data and test dataset. The dataset used to construct a model is called the training dataset. A training sample dataset is a collection of instances {Zi}ni = 1 = {Z1, Z2, ..., Zn}, that apply as a input in a learning algorithm for a statistical model. To examine a performance of the test dataset is used which has same feature and characteristics as the train dataset.

E. The sentiment analysis applied on tweets data. Each tweet is classified into three categories: positive, negative, neutral. Positive tweets show the user are happy, joy, optimist other positive feeling. Negative tweets show the users are angry, aggressive, sad and other feeling. Neutral posts are not expressing any feeling. So negative tweets not contain any opinionated words.

F. Apply the Naïve Bayes algorithm on the trained data set. Trained data is being as features the Naïve Bayes to calculating the remaining tweets for classification purpose. Than we being calculate the accuracy of Naïve Bayes in our experiment.

G. Final step we applied the K-Means algorithm on Positive and Negative tweets. K-Means separated the positive words and negative words.

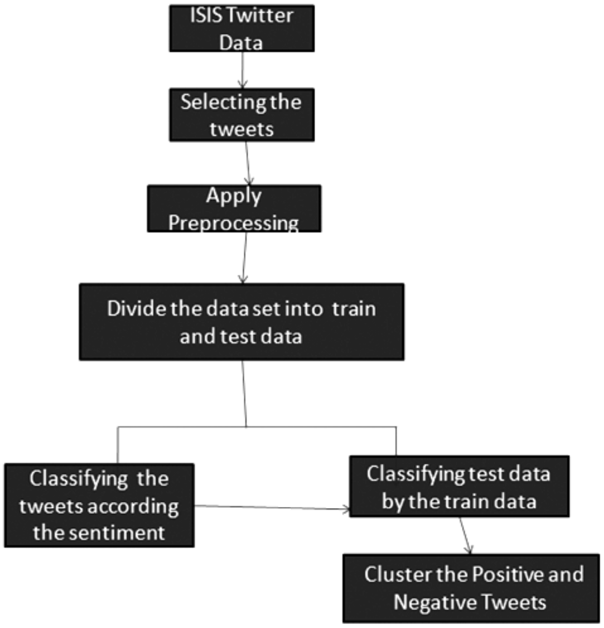

ISIS Twitter Data
Selecting the tweets
Apply Preprocessing
Divide the data set into train and test data
Classifying the tweets according the sentiment
Classifying test data by the train data
Cluster the Positive and Negative Tweets

Figure 2: Overall Process

5.1. Data Collection

In our research we use ISIS tweets dataset which are available on www.kaggle.com site. Data set consists of 17410 observation and 8 variables. Variables are like Name, username, location, followers, numberstatues, time, tweets. After applying preprocessing step and sentiment function on ISIS dataset then we select in positive and negative tweets. In this tweets data there are 2830 observation. We create train data in train data 1980 observation that are 70% of the tweets data. Test data 850 observations that are 30% of tweets data.

Table 1
The datasets used for the experiments

Data Set	*Description*
ISIS	17410 observations
Tweets_data	2830 observations [it contains only positive and negative tweets]
Train	1980 observation randomly select 70% tweets
Test	850 observations

Table 2
Information that is available about ISIS Tweets dataset

Name	*Name of person who tweets*
Username	Name of user account
Description	The description that the users provide about himself/herself
Location	Location of user
Followers	Number of followers
Numberstatuses	Number of user status
Time	The time zone
Tweets	Tweets statements

5.2. Pre-Processing Text

When a data is collected form twitter account data was not clean. Data contained noise that mean data contained error, outlier, misspelled words, retweets and strange symbols like #, @, !. Before using the dataset it was cleaned the dataset to deal with a missing values and blank values etc. Such noise and outliers can effect on accuracy of the result therefore preprocessing the data sets is necessary. To eliminating noise, outliers blank values , missing values, streaming words, URL, HTML character, English words and prepare the corpus for building the model the following preprocessing step are done:-

- Convert the all the Tweets to lower or upper case.
- All URL are removed in ISIS data set. Tweets like http://t.co/EPaPRlph5W http://t.co/4VUYszairt,http://t.co/bJ24uNFNIT,http://t.co/99nsW0Nzxa,http://t.co/1hLwtP3XOd http://t.co/jcviGUdedv, http://t.co/SzOgGMvPMI.
- Removed all # hash tag and @ in data set. Tweets like @AbdirahmanBash2, @KhalidMaghrebi, @IbnNabih1, @Polder_Mujahid
- Removed the punctuation in data set like comma (,), dot (.) etc.
- Removed a digit or number.
- Removed RT (Retweet Tag).
- Stop words which are consider to be common words in English language and this words do not contributes in a sentiment analysis.
- Many a times, people do misspell words which may take away the meaning of the sentence therefore, these words have been correct.

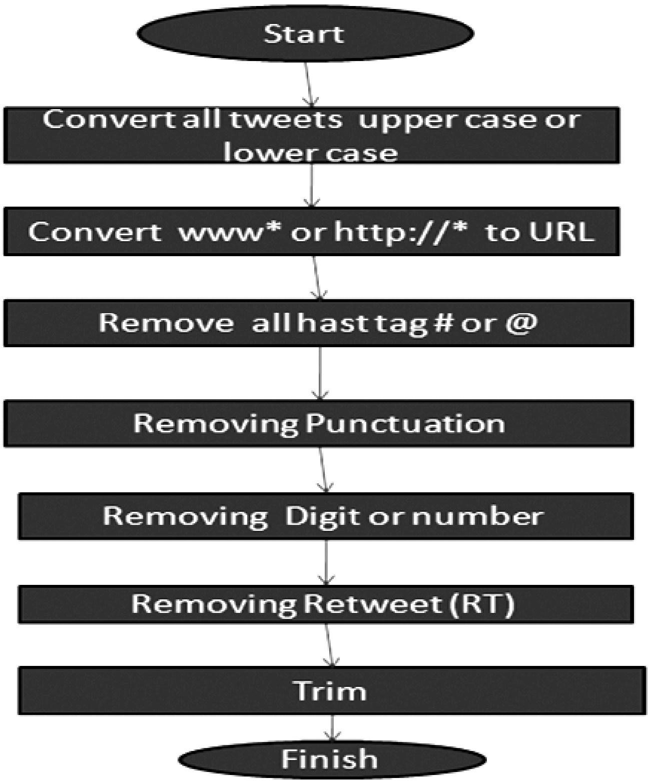

Start
Convert all tweets upper case or lower case
Convert www* or http://* to URL
Remove all hast tag # or @
Removing Punctuation
Removing Digit or number
Removing Retweet (RT)
Trim
Finish

Figure 3: Pre-processing Step

6. RESULT ANALYSIS

We performed the classification and clustering techniques on dataset. We performed the naïve bayes classifier techniques on the ISIS Twitter data sets and features to evaluate the ability these techniques have to classifies extremist from non-extremist tweets. We apply the sentiment function on a tweets data then we found probability tweets. Positive tweets probability is 0.05967835, Negative tweets probability is 0.10287191 and Neutral tweets probability is 0.83744. We separate only positive and negative tweets that are 2830. Naïve Bayes classifies the tweets according the sentiment in Table 1.

Table 3
Positive and Negative Tweets

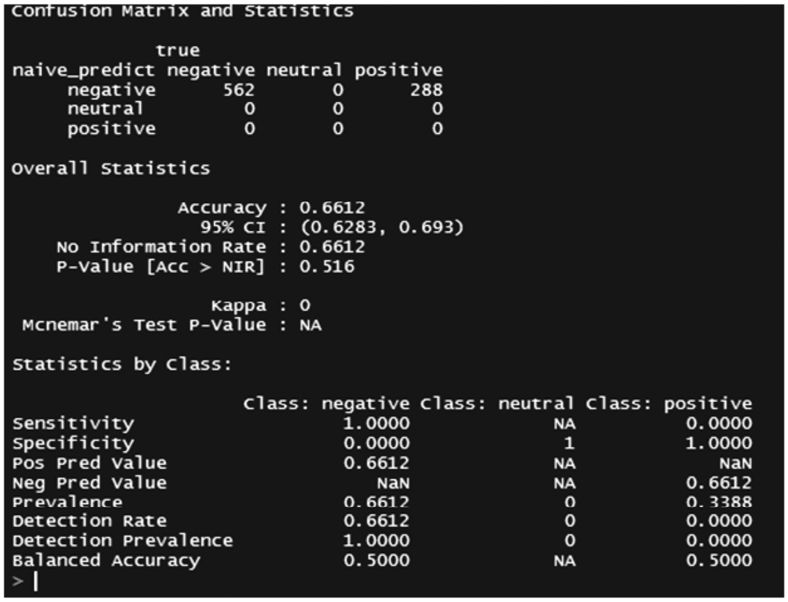
Negative Tweets	Positive tweets
562	288

Accuracy is the proportion of the sum of true results and the total number of instances. It shows the percentage of total instances that were correctly classified

$$\text{Accuracy} = \frac{(\text{True Positive} + \text{True Negative})}{(\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative})}$$

The results of our experiments are visualized using a confusion matrix. A confusion matrix, known also as error matrix or contingency table is used in machine learning to visualize the results of a naïve bayes classifier algorithm. Accuracy of classifying of tweets is 66.12% means that the Naïve Bayes classification was able to precisely estimate the classes of the post as positive, negative, neutral.

Apply the K-Mean technique on separately positive tweets and negative tweets. Since k=3, the 3 clusters. Negative tweets cluster snapshot result is in figure 5. We got cluster, after applying the K-Means algorithm on negative tweets. Negative words like: Dead, ISIS, Sriya, Iraqi, Kill and so on, Negative cluster are:



```
Confusion Matrix and Statistics

             true
naive_predict negative neutral positive
     negative      562       0      288
     neutral         0       0        0
     positive        0       0        0

Overall Statistics

               Accuracy : 0.6612
                 95% CI : (0.6283, 0.693)
    No Information Rate : 0.6612
    P-Value [Acc > NIR] : 0.516

                  Kappa : 0
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: negative Class: neutral Class: positive
Sensitivity                   1.0000             NA          0.0000
Specificity                   0.0000              1          1.0000
Pos Pred Value                0.6612             NA             NaN
Neg Pred Value                   NaN             NA          0.6612
Prevalence                    0.6612              0          0.3388
Detection Rate                0.6612              0          0.0000
Detection Prevalence          1.0000              0          0.0000
Balanced Accuracy             0.5000             NA          0.5000
> |
```

Figure 4: Result Snapshot for confusion matrix using Naïve Bayes Classifier

Negative Cluster$_1$= {arm, fight, dead, soldier, attack,…}

Cluster$_2$ = {Kill}

Cluster$_3$ = {isis, sriya}

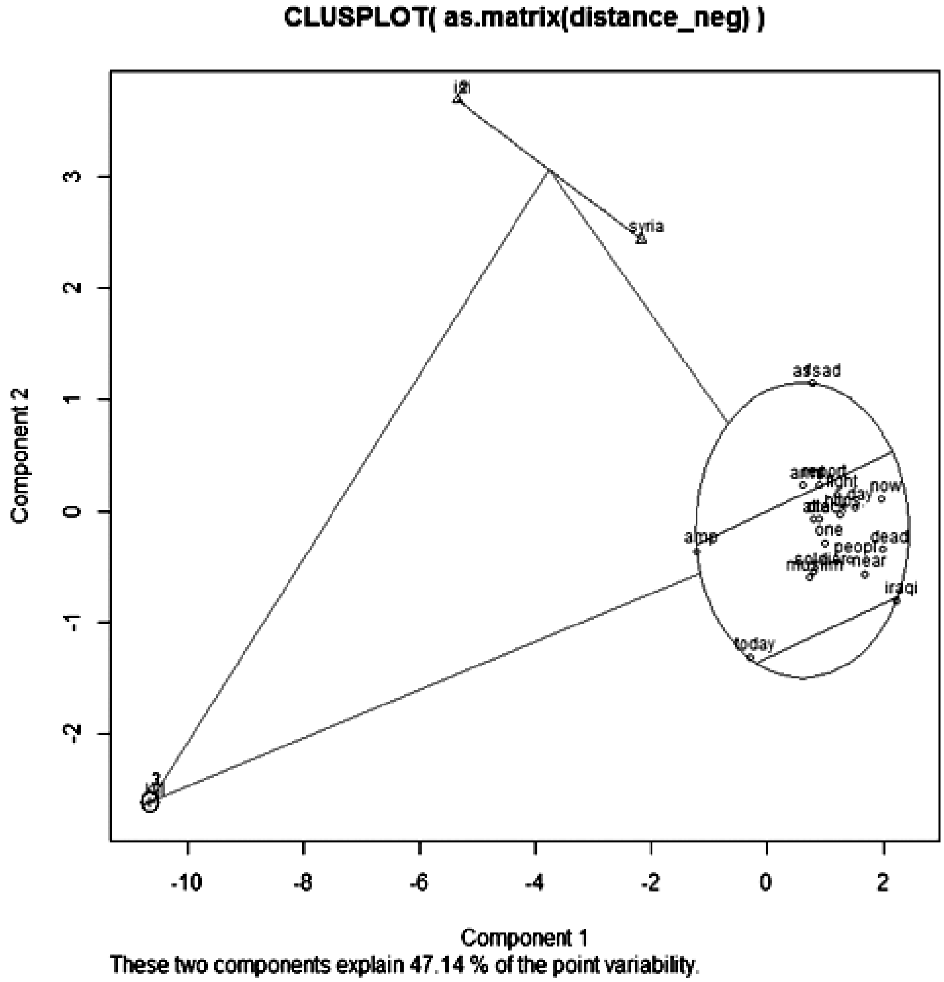


Figure 5: Negative Tweets Result Snapshot

Positive tweets cluster snapshot result is in figure 6. We got the positive cluster, after applying the K-Means algorithm on positive tweets. Positive words like: allah, amp, good, support,love make and so on. Positive cluster are:

Positive Cluster$_1$ = {allah, amp,…}

Cluster$_2$ = {good, support, just, see, new,…..}

Cluster$_3$ = {follow, love, make,….}

Word cloud is a text mining method, which allows us to highlight the most frequent use word in dataset. World cloud is a visual represent of text data. In figure 7(a) show the negative tweets person word cloud, which are ishfaqahmad, melvynlion, fidaeefulaani, warnews, whitecat7, abuhumayra4 and so. on. Figure 7(b) show the positive tweets person word cloud, which are warreporter1, warrepoter2, ibnkashmir, mobiayubi, and so on.

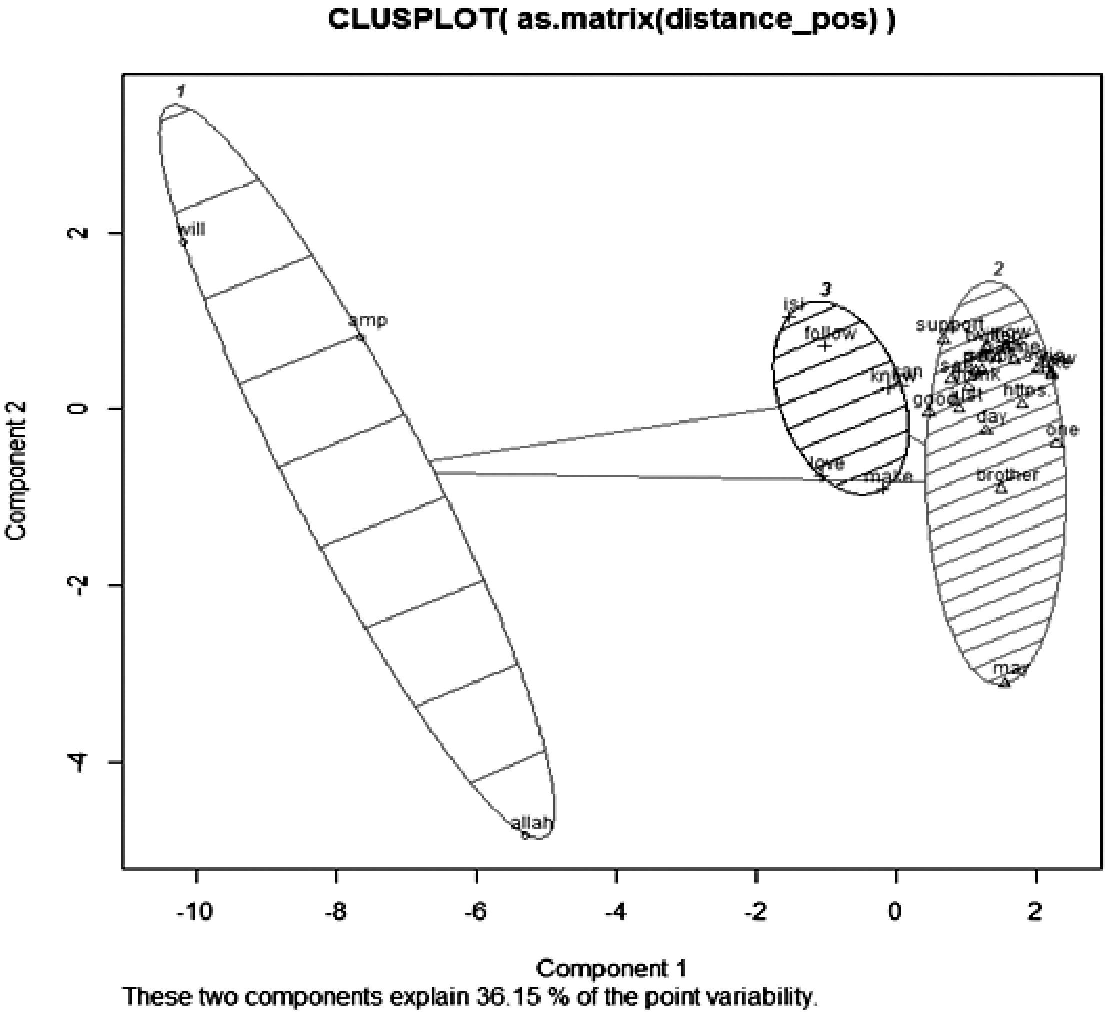


Figure 6: Positive Tweet Result Snapshot

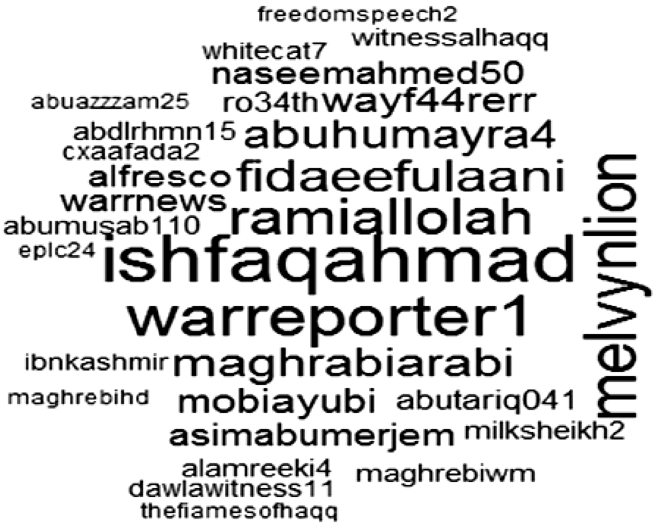


Figure 7: (a) Negative word cloud Snapshot

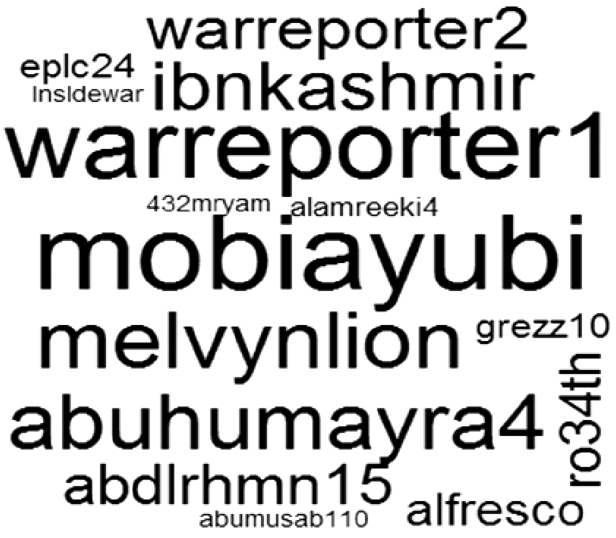


Figure 7: (b) Positive word cloud Snapshot

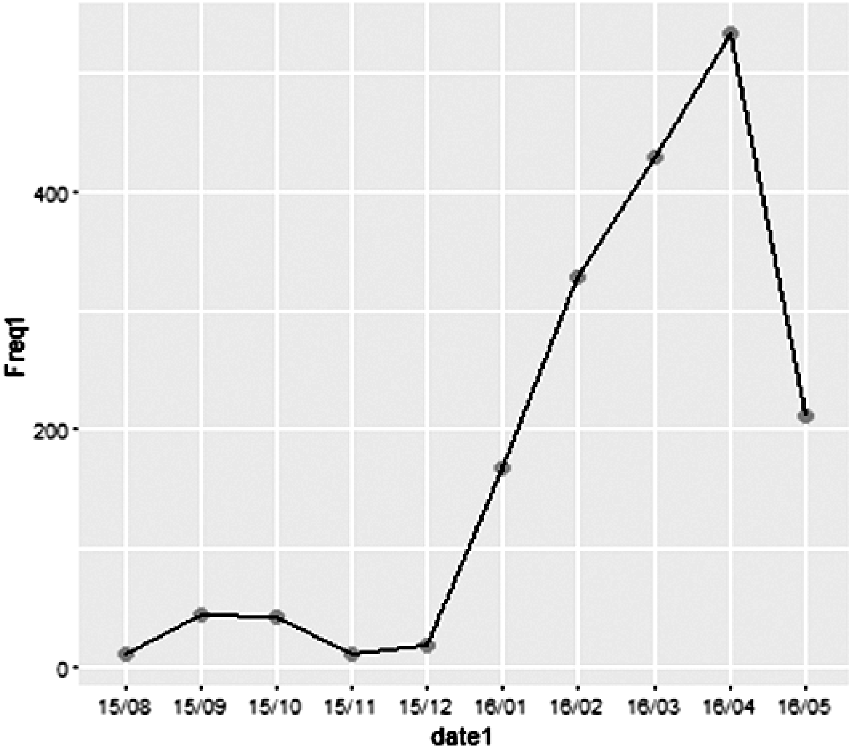


Figure 8: (a) Negative tweets graph Snapshot

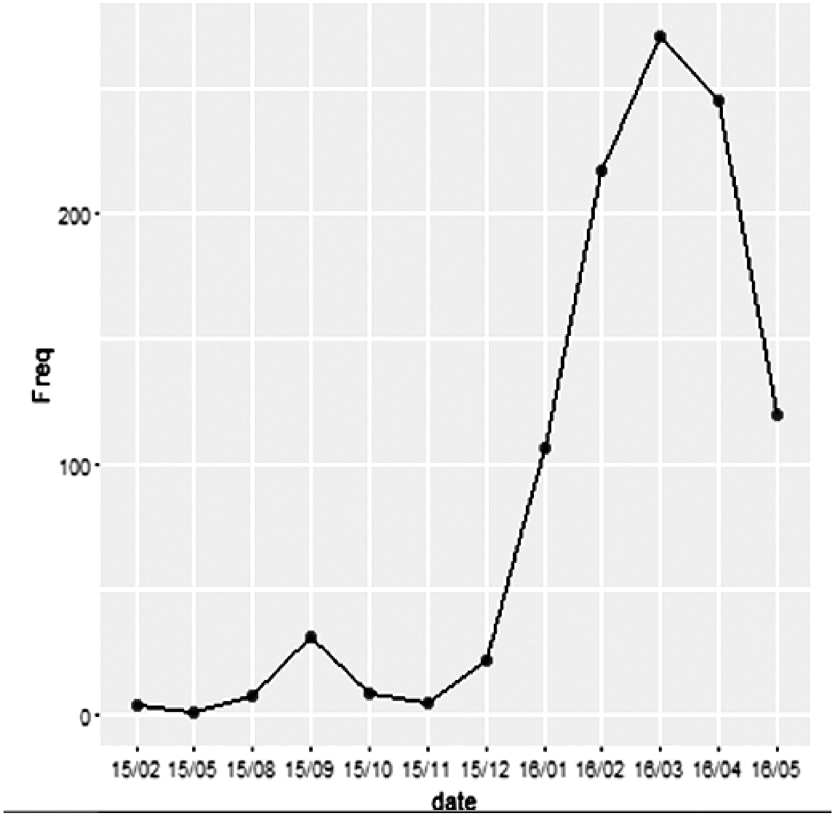


Figure 8: (b) Positive Tweets graph Snapshot

Figure 8(a) Show the Negative tweets graph Snapshot, which is represent that in December 2015 to April 2016 negative tweets increase after that May 2016 negative tweets decrease. In this graph X axis has Month and Y axis has Frequency of negative tweets. Figure 8(b) Show the Positive Tweets graph snapshot which is represented that in August 2015 positive tweets are increase after that September 2015 positive tweets are decrease than December 2015 to March 2016 positive tweets are increase than April 2016 and May 2016 positive tweets are decrease

7. CONCLUSION

From our study, we archive high degree accuracy using naïve bayes classifier. Naïve Bayes classifier is suitable to train and classify sentiment form a tweeter data. This method is classifying the large number of tweets. This technique suitable for the business sentiment classifies. Sentiment analysis is an effective way of classifying the Opinions. The classification decisions made by the Naïve Bayes classifier give a good accuracy because every time a decision with the higher probability is being made. We concluded that in this dataset with high level of noise, but K-means give the better result.

In this work we covered many things regarding the attempt to classify tweets on twitter. Still there are many ways in which this may be improves. In Future we improve the accuracy of naïve bayes classifier. We also go beyond positive and negative sentiment polarity detection and extract our work emotional knowledge text. We can improve our work is to make other classifier work not only on English tweets but also on tweets written in other local languages tweets

REFERENCES

[1] Priyanga Chandrasekar and Kai Qian," The Impact of Data Preprocessing On the Performance of Naïve Bayes Classifier", 2016 IEEE 40th Annual Computer Software and Applications Conference, 2016.

[2] Rui Máximo Esteves and Chunming Rong," Using Mahout for clustering Wikipedia's latest articles", Third IEEE International Conference on Cloud Computing Technology and Science, 2011.

[3] Bo Zhao, Yongji He, Chunfeng Yuan, and Yihua Huang,"Stock Market Prediction Exploiting Microblog Sentiment Analysis", International Joint Conference on Neural Networks (IJCNN), 2016.

[4] Bayu Yudha Pratama and Riyanarto Sarno," Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM", International Conference on Data and Software Engineering, 2015.

[5] Mohd Naim Mohd Ibrahim and Mohd Zaliman Mohd Yusoff," Twitter Sentiment Classification Using Naïve Bayes Based on Trainer Perception", IEEE Conference on e-Learning, e-Management and e-Services, 2015.

[6] Maher M. Itani, Lama Hamandi, Rached N. Zantout and Islam Elkabani, "Classifying Sentiment in Arabic Social Networks: Naïve Search versus Naïve Bayes", 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA),2012.

[7] Krushikanth R. Apala, Merin Jose, Supreme Motnam, C.-C. Chan, Kathy J. Liszka, and Federico de Gregorio1,"Prediction of Movies Box Office Performance Using Social Media", IEEE International Conference on Advances in Social Networks Analysis and Mining, 2013.

[8] Cut Fiarni1, Herastia Maharani1, Rino Pratama1," Sentiment Analysis System for Indonesia Online Retail Shop Review Using Hierarchy Naive Bayes Technique", Fourth International Conference on Information and Communication Technologies (ICoICT), 2016.

[9] Erik Cambria," Affective Computing and Sentiment Analysis", IEEE Intelligent Systems, 2016.

[10] Wenping Zhang, Chunping Li and Yunming Ye," Dynamic Business Network Analysis for Correlated Stock Price Movement Prediction", IEEE Intelligent System, 2015.

[11] Erik Cambria and Bebo White," Jumping NLP Curves: A Review of Natural Language Processing Research", IEEE Computational Intelligence Magazine, 2014.

[12] Erik Cambria, Bjorn Schuller, Yunqing Xia and Catherine Havasi," New Avenues in Opinion Mining and Sentiment Analysis", IEEE Computer Society, 2013

[13] Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang, and Aijun An," Mining Online Reviews for Predicting Sales Performance: "A Case Study in the Movie Domain, IEEE Transactions on knowledge and data engineering, 2012.

[14] Chihli Hung and Hao-Kai Lin," Using Objective Words in SentiWordNet to Improve Word-of- Mouth Sentiment Classification", IEEE Intelligent Systems, 2013.

[15] Vadim Kagan and Andrew Stevens and V.S. Subrahmanian," Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election," IEEE Intelligent Systems, 2015

[16] Dragomir Radev,"Natural Language Processing and the Web", IEEE Intelligent Systems, 2008.

[17] Desheng Dash Wu, Lijuan Zheng, and David L. Olson," A Decision Support Approach for Online Stock Forum Sentiment Analysis", IEEE Transactions on systems, man, and cybernetics: systems, 2014.

[18] ZHANG Lumin, JIA Yan, ZHU Xiang, ZHOU Bin, HAN Yi," User-Level Sentiment Evolution Analysis in Microblog, IEEE Intelligent Systems, 2014.

[19] Soujanya Poria, Erik Cambria, Erik Cambria, Federica Bisio and Amir Hussain,"Sentiment data flow Analysis of Dynamic Linguistic Patterns", IEEE Computational intelligence magazine, 2015.

[20] Nehal Mamgain1, Ekta Mehta2, Ankush Mittal4 and Gaurav Bhatt3," Sentiment Analysis of Top Colleges in India Using Twitter Data", 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 2016.

[21] Yeh Hanmin, Lv Hao and Sun Qianting,"An Improved Semi- Supervised K-Means Clustering Algorithm", Information Technology, Networking, Electronic and Automation Control Conference, IEEE 2016.