TEXT ANALYSIS ON TWITTER DATASET USING NEURAL NETWORK AND BAYESIAN CLASSIFIER

[1]V.Abinaya , [2]V.Jayashree, [3]K.Kalaivani
[1,2,3]Assistant Professor ,
Department of Commerce with Computer Application,
Dr.N.G.P Arts and science college, Coimbatore, India

Abstract: The growth of social media has caught attention in recent years with respect to the information that can be utilized for research purposes across various industries. The information inside social media can be used for various purposes from sentiment analysis, news updates, to users opinion mining, recommendation systems and organization gain more insights on their customers, products, promotion using social media data to take competitive advantages. Utilizing data mining technique is the best and the right way to extract information but the problem lies in choosing the right technique. Every technique is unique in nature and has its own limitation and advantages based on the goal of the researchers. It is noted that different techniques yields different results and the success of data analysis depends on the right choice of technique selection and the type data used.

Twitter being the most widely used social media platform to update the current events to the others through short text messages called tweets. Using twitter large companies gain knowledge on their customers and products to use it on their promotional and marketing purposes and even companies have started to provide updates and services through twitter media. In this study a classification technique is employed on twitter data set that has tweets (short text messages) on Iphone7. Using Naïve Bayes algorithm classification of tweets into positive and negative classes is carried out and the results are compared to neural network algorithm.

The main goal of such a text analysis is to discover how the audience (tweets)reacts to IPhone7. The Twitter data that is collected will be classified into two categories; positive or negative. An analysis will then be executed on the classified data to analyze what percentage of the audience (tweets) falls into each category.

Index Terms -: Text analysis, neural network, Naïve Bayes classifier, twitter

1. INTRODUCTION

Information technology has enabled users to interact, share and stay connected. Social media is a platform that offers to view, create and share information on interests, ideas and opinions. Social media sites differ with respect to content, frequency and usability and it operates in such a way that a user can receive and send information to many others who are connected. The various social media sites that are available today are WhatsApp, Tumblr, Instagram, Twitter, BaiduTieba, Pinterest, LinkedIn, Google+, YouTube etc, each differs uniquely from other in terms of content, for example Face book uses pictures and text, whereas YouTube uses only videos. The growth of social media becomes popular with respect to development of mobile technology, enabling users to access via their phones.

Twitter is a popular social media site where users post, search and share news, events, trends on different categories with respect to their interest. Twitter uses a text message as a primary content with a limitation of 140 characters in length. These text contents are called as tweets. Upon posting a tweet, a follower can read the tweets and get informed. It is estimated that around 6,000 tweets are tweeted (posted) every second and the number amounts to 200 billion tweets per day. Twitter being a real-time information sharing platform, it is often used by companies for campaigns and promotions. For the purpose of promotions and campaigns, a company must understand the type of mood or the trend that interest the twitter community.

Presently, research on text classification within Twitter has indicated that people use Twitter for different reasons. (Java A, Song X, Finin T, 2007) discovered four main user purposes on Twitter:

- Daily chitchat: Most posts on Twitter talk about daily course or what people are presently executing and this is the largest and most common user of Twitter
- Conversations: About one-eighth of all posts hold a conversation and this form of communication was used by almost 21 % of users
- Sharing information: About 13 % of posts contained a URL (i.e., website address), directing readers to another information source; and
- Reporting news—many Twitter users report latest news or comment about current events on Twitter.

The process of reading data and generating newer information is called data mining. Data mining involves various methods such as classification, clustering, prediction, association rule mining etc. On the other hand text mining involves information retrieval, sentiment analysis, pattern recognition, categorization, clustering and summarization. With huge volume of data from twitter, organization started utilizing twitter data to get the latest trends, opinions, moods, topic, interest and comments to understand their customers on their product and services. Text mining is commonly used in the following areas,

News filtering: Organizing news articles manually is difficult and text categorization can be used to filter news categorically.

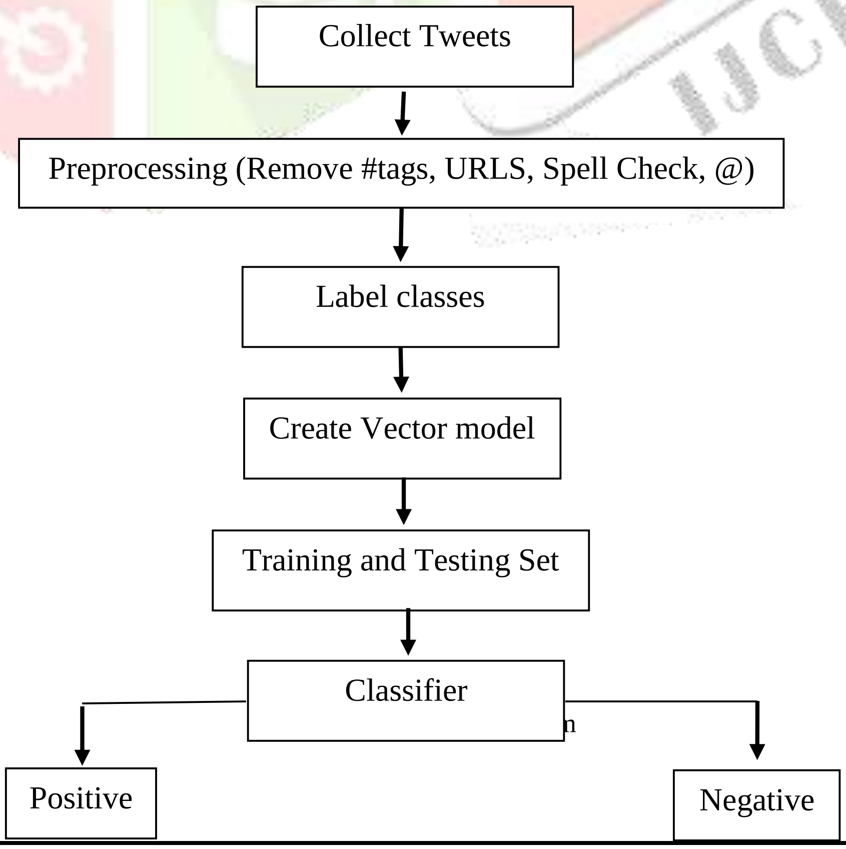
Document retrieval and organization: Digital libraries and search engines uses documents to retrieve and organize large volumes of documents, literatures, books, etc

Opinion mining: Customer reviews can be mined to get their opinion and reviews mostly contain different forms of expression and classification is the best technique to separate different opinions

Filtering: Classifying and filtering of documents, information with respect to textual occurrences, example includes, spam filtering, email filtering etc.

1.1 METHODOLOGY DIAGRAM

,



2. EXPERIMENT ANALYSIS

R is a programming language and an environment for statistical computing and graphics. It is free software, released under the GNU General Public License (GPL) and runs on all common operating systems. It is the leading open-source system for statistical computing. R consists of a base distribution and add-on packages, contributed by members of its open-source community. The base distribution contains R's basic functionality, such as the plotting functions and statistical models. Add-on packages extend R with diverse functionalities, such as graph handling, machine learning algorithms, and advanced plotting capabilities. Both the base distribution and the add-on packages are distributed through the Comprehensive R Archive Network (CRAN). Currently, there are more than 5000 packages on CRAN.

2.1 Building the classifier

The next step of the experiment involves building the classifier using Naïve Bayes and Neural network. There are several packages available for the classifiers. For Naïve bayes, we build the classifier using the packages, e1071, tm, KlaR, and RTextTools. For Neural Network, we use library neuralnet packages. Once the classifiers are built, the next step is to train the classifier. For the training and testing purpose we use two data sets separately with 50 numbers of tweets on each data set.

The testing datasets is composed of both positive and negative tweets. The tweets in the data set are already labeled with the classes 'Positive' and 'Negative'. The main aim of the experiment was to find out how accurately the classifiers classify the testing data sets. For references the algorithm developed for two classifiers are attached in the appendix section.

2.2 Evaluation Metrics

Classifiers are commonly evaluated using either a numeric metric, such as accuracy, or a graphical representation of performance, such as a receiver operating characteristic (ROC) curve. Metrics help us understand how a classifier performs; many are available, some with numerous tunable parameters.

Classification metrics are calculated from true positives (TPs), false positives (FPs), false negatives (FNs) and true negatives (TNs), all of which are tabulated in the so-called confusion matrix (Fig. 1). The relevance of each of these four quantities will depend on the purpose of the classifier and motivate the choice of metric.

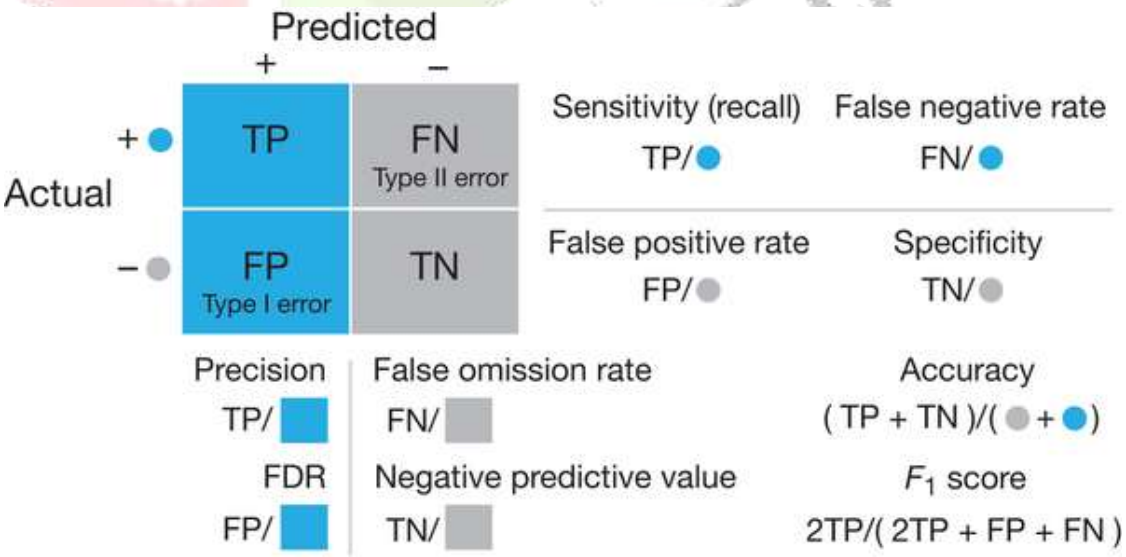


Fig 2.1 Confusion matrix

There are four different metrics: accuracy, sensitivity, precision and F1. Accuracy is the fraction of predictions that are true. Although this metric is easy to interpret, high accuracy does not necessarily characterize a good classifier. For instance, it tells us nothing about whether FNs or FPs are more common. A useful measure for understanding FNs is sensitivity (also called recall or the true positive rate), which is the proportion of known positives that are predicted correctly. However, neither TNs nor FPs affect this metric, and a classifier that simply predicts that all data points are positive has high sensitivity.

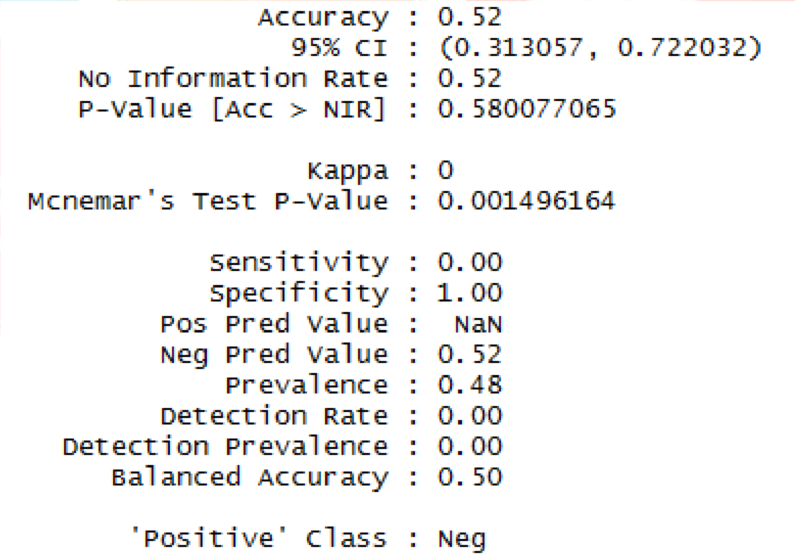
Specificity, which measures the fraction of actual negatives that are correctly predicted, suffers from a similar weakness: not accounting for FNs or TPs. Both TPs and FPs are captured by precision (also called the positive predictive value), which is the proportion of predicted positives that are correct. However, precision captures neither TNs nor FNs.

2.3. Naïve Bayes Classifier

The process of creating the classifier involves creation of term document matrix and applying the naïve bayes algorithm. The amount of time took for the algorithm to build the model is 0.02 sec.

NEG / POS	NEG	POS
NEG	0	0
POS	12	13

Table 2.1 Confusion matrix



```
              Accuracy : 0.52
                95% CI : (0.313057, 0.722032)
   No Information Rate : 0.52
   P-Value [Acc > NIR] : 0.580077065

                 Kappa : 0
Mcnemar's Test P-Value : 0.001496164

           Sensitivity : 0.00
           Specificity : 1.00
        Pos Pred Value :  NaN
        Neg Pred Value : 0.52
            Prevalence : 0.48
        Detection Rate : 0.00
  Detection Prevalence : 0.00
     Balanced Accuracy : 0.50

      'Positive' Class : Neg
```

Fig 2.2 Accuracy

Table 2.2 Accuracy details

Classifier	Accuracy	Sensitivity	Specificity
Naïve Bayes	0.52	0.00	1.00

2.4 Neural Net Classifier

The neural net classifier was applied to our twitter data set and the classifier took 0.04 sec to build the model. The misclassification error was found to be 0. Meaning all the 50 instances is correctly classified. Although there were 0.017% of error rate reached with the hidden layer 3.

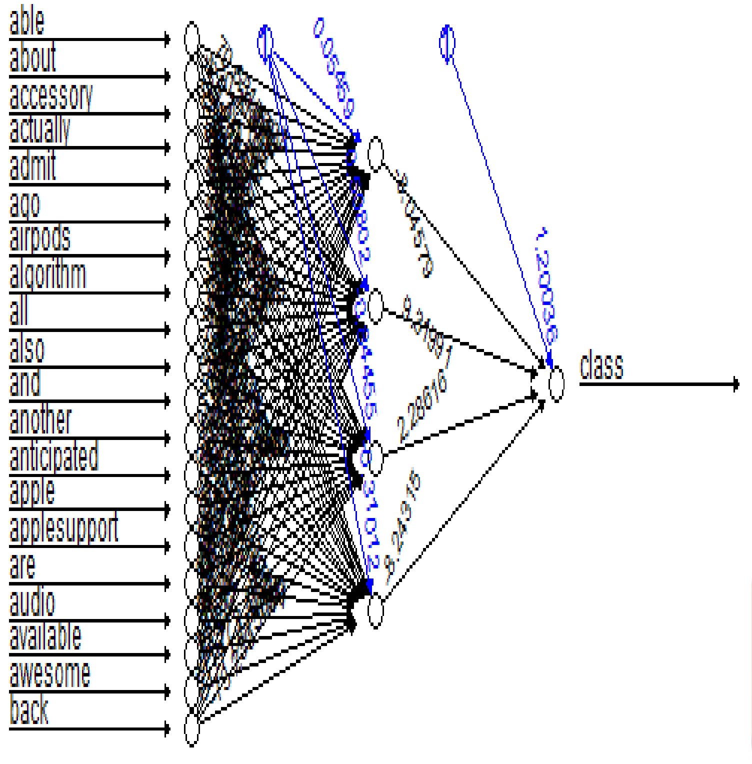
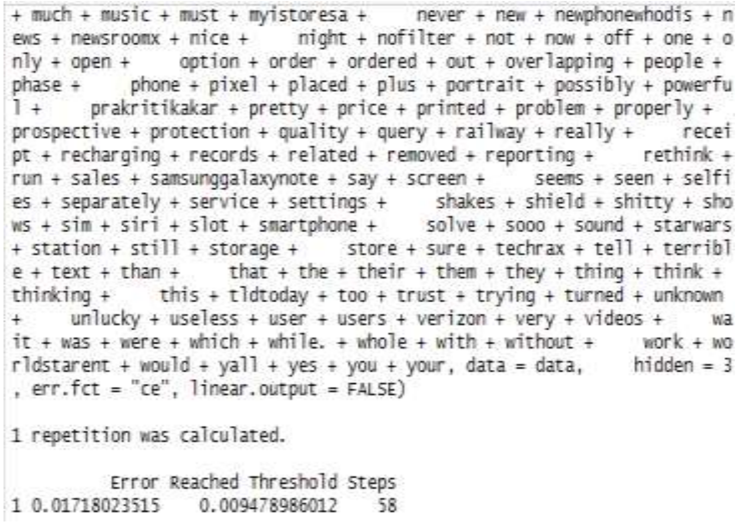


Fig 2.3 Neural network with 3 hidden layers

Table 2.3 Confusion matrix

	NEG	POS
NEG	27	0
POS	0	23



```
+ much + music + must + myistoresa +     never + new + newphonewhodis + n
ews + newsroomx + nice +     night + nofilter + not + now + off + one + o
nly + open +     option + order + ordered + out + overlapping + people +
phase +     phone + pixel + placed + plus + portrait + possibly + powerfu
l +     prakritikakar + pretty + price + printed + problem + properly +
prospective + protection + quality + query + railway + really +     recei
pt + recharging + records + related + removed + reporting +     rethink +
run + sales + samsunggalaxynote + say + screen +     seems + seen + selfi
es + separately + service + settings +     shakes + shield + shitty + sho
ws + sim + siri + slot + smartphone +     solve + sooo + sound + starwars
+ station + still + storage +     store + sure + techrax + tell + terribl
e + text + than +     that + the + their + them + they + thing + think +
thinking +     this + tldtoday + too + trust + trying + turned + unknown
+     unlucky + useless + user + users + verizon + very + videos +     wa
it + was + were + which + while. + whole + with + without +     work + wo
rldstarent + would + yall + yes + you + your, data = data,     hidden = 3
, err.fct = "ce", linear.output = FALSE)

1 repetition was calculated.

         Error Reached Threshold Steps
1 0.01718023515    0.009478986012    58
```

Fig 2.4. Error Rate reached

Table 2.3 Accuracy for Neural Network

Classifier	Accuracy	Sensitivity	Specificity
Neural Network	1.00	1.00	1.00

3.DISCUSSION

According to the table 2.3, Neural network's Accuracy was found to be 1, Senstivity was found to be 1 and specificity is also 1. Whereas for Naïve Bayes, the Accuracy was found to be 0.52, Sensitivity is 0.00, and specificity is 1.00.

Table 3.1 Comparitive Table of Classifier

Classifier	Accuracy	Sensitivity	Specificity	Time
Neural Network	1.00	1.00	1.00	0.04
Naïve Bayes	0.52	0.00	1.00	0.02

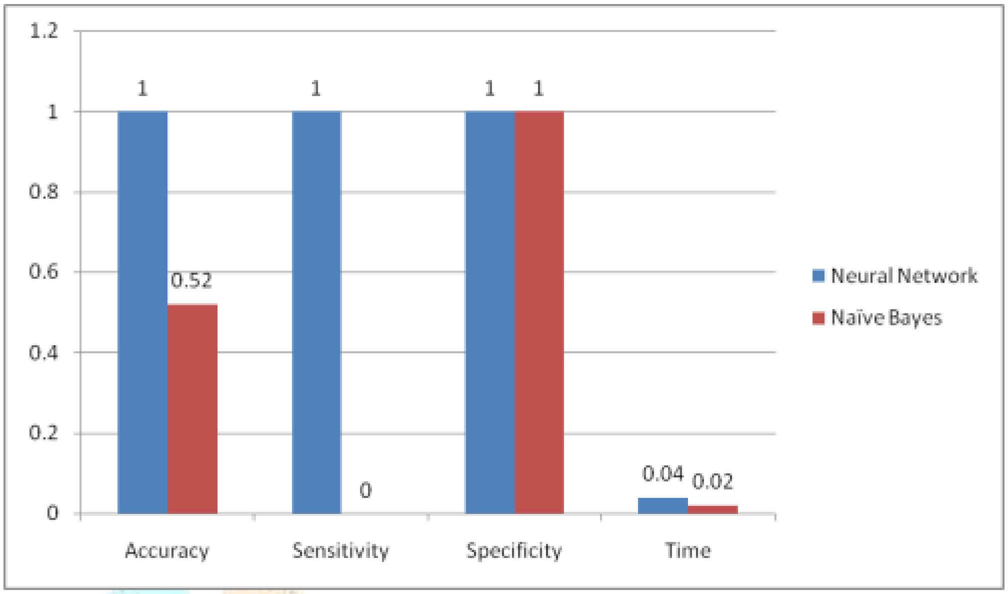


Fig 2.5 Performance Comparison

From the table we conclude that neural network has higher accuracy rate than Naïve bayes, but comparative with time, the model building time was found to bit higher for Neural networks than Naïve bayes. The amount of time varies according to the hidden layers of the perceptrons.

4. CONCLUSION

To conclude, it has illustrated that an effective text classification analysis can be performed on a Iphone7 product by collecting a sample users opinions from Twitter. Throughout the duration of this project many different data analysis tools were employed to collect, clean and mine sentiment from the dataset. Such an analysis could provide valuable feedback to providers and help them to spot a negative turn in user's comments. Discovering negative trends early on can allow them to make decisions on how to target specific aspects of their products and features in order to increase its customer satisfaction.

It is apparent from this study that the machine learning classifier used has a major effect on the overall accuracy of the analysis. Commonly used algorithms for text classification were examined such as Naïve Bayes, Neural Network. Through the evaluation of different algorithms it was found that out of the models examined the Neural Network algorithm had the highest performance on this iphone7 dataset.With machine learning algorithms constantly being developed and improved, massive amounts of computational power becoming readily available both locally and on the cloud, and large amounts of data being uploaded to social media sites every day, sentiment analysis will become standard practice for marketing and product feedback.

5.REFERENCES

[1] A. Althubaity, A. Almuhareb, S. Alharbi, A. Al-Rajeh, and M. Khorsheed, "KACST Arabic text classification project: overview and preliminary results," in Proceedings of the 9th IBIMA Conference on Information Management in Modern Organizations, Marrakech, Morocco, January 2008.

[2] A. Go, R. Bhayani and L. Huang, "Twitter sentiment classification using distant supervision ", CS224N Project Report, Stanford, (2009), pp. 1-12

[3] A. L. Churchill, E. G. Liodakis, and S. H. Ye, "Twitter relevance filtering via joint bayes classifiers from user clustering," Journal of University of Stanford, 2010.

[4] A. Zubiaga, D. Spina, V. Fresno and R. Martínez, "Classifying trending topics: a typology of conversation triggers on twitter", Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, (2011) October, pp. 2461-2464

[5] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu and M. Demirbas, "Short text classification in twitter to improve information filtering", Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval, ACM, (2010)
[6] A El-Hajj, L Safatly, M Bkassiny, M Husseini, Cognitive radio transceivers: RF, spectrum sensing, and learning algorithms review. Int J Antenn Propag 11(5), 479–482 (2014)
[7] Bird, Steven, Edward Loper, Ewan Klein, 2009. Natural Language Processing with Python.
[8] C Ghosh, C Cordeiro, DP Agrawal, M Bhaskara Rao, Markov chain existence and hidden Markov models in spectrum sensing, in Proceedings of the IEEE International Conference on Pervasive Computing & Communications (PERCOM) (Galveston, 2009), pp. 1–6
[9] C. Thongsuk, C. Haruechaiyasak and S. Saelee, "Multi-classification of business types on twitter based on topic model", Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2011 8th International Conference,IEEE, (2011)May, pp. 508-511
[10] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: conversational aspects of retweeting on twitter," in Proceedings of the 43rd Annual Hawaii International Conference on System Sciences (HICSS '43), pp. 1530–1605, Honolulu, Hawaii, USA, January 2010.
[11] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents (SMUC '10), pp. 37–44, ACM, October 2010.
Eui-Hong (Sam) Han, George Karypis, Vipin Kumar; "Text Categorization Using Weighted Adjusted k-Nearest Neighbor Classification", Department of Computer Science and Engineering. Army HPC Research Centre, University of Minnesota, Minneapolis, USA. 1999.
[12] F. Thabtah, M. Eljinini, M. Zamzeer, and W. Hadi, "Naïve Bayesian based on Chi Square to categorize Arabic data," in Proceedings of the 11th International Business Information Management Association (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Citeseer, Cairo, Egypt, January 2009.
[13] H. Becker, M. Naaman and L. Gravano, "Beyond trending topics: Real-world event identification on twitter", Proc.of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11), (2011)
[14] Heide Brücher, Gerhard Knolmayer, Marc-André Mittermayer; "Document Classification Methods for Organizing Explicit Knowledge", Research Group Information Engineering, Institute of Information Systems, University of Bern, Engehaldenstrasse 8, CH -3012 Bern, Switzerland, 2002.
[14] Hunt, E.B., Marin. and Stone,P.J. (1966). Experiments in induction, Academic Press, New York.
[15] J Zheng, F Shen, H Fan, J Zhao, An online incremental learning support vector machine for large-scale data. Neural Comput Appl 22(5), 1023–1035 (2013)
[16] J. Benhardus and J. Kalita, "Streaming trend detection in Twitter," International Journal of Web Based Communities, vol. 9, no. 1, pp. 122–139, 2013.
[17] J. Benhardus, "Streaming trend detection in twitter", National Science Foundation REU for Artificial Intelligence, Natural Language Processing and Information Retrieval, University of Colarado, (2010)
[18] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," Journal of Computational Science, vol. 2, no. 1, pp. 1–8, 2011.
[19] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on Twitter," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309, Association for Computational Linguistics, Edinburgh, UK, July 2011.
[20] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman and J. Sperling, "Twitterstand: news in tweets ", Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, (2009) November, pp. 42-51
[21] Java A, Song X, Finin T. Why we twitter: understanding microblogging usage and communities. Joint 9th WEBKDD and 1st SNA-KDD Workshop '07. 2007. July, pp. 841-842
[22] K Yue, Q Fang, X Wang, J Li, W Weiy, A parallel and incremental approach for data-intensive learning of Bayesian networks. IEEE Trans Cybern 99, 1–15 (2015)

[23] K. H. Lim and A. Datta, "Interest classification of Twitter users using Wikipedia," in Proceedings of the 9th International Symposium on Open Collaboration, ACM, Hong Kong, August 2013.

[24] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal and A. Choudhary, "Twitter trending topic classification", Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference, IEEE, (2011)December, pp. 251-258

Kumar, E., 2011. Natural Language Processing. New Delhi: I.K International Publishing house