RESEARCH ARTICLE OPEN ACCESS

Application Research of Text Classification Based on Random Forest Algorithm

Mr D.Purushothaman MCA.,M.E.,[1], Kotipaiah Gari Yamuna [2]

[1] Asst. Professor, Department of Computer Applications
[2] Student, Department of Computer Applications
[1], [2] Chadalawada Ramanamma Engineering College (Autonomous)

ABSTRACT

Various ensemble classification methods have been proposed in recent years. These methods have been proven to improve classification accuracy considerably. One of the most widely used ensemble methods is Random Forests, an ensemble of CART, it uses bagging or bootstrap aggregating. In the paper, the use of the Random Forests classifier for text classification is explored. We compare the accuracy of the Random Forest classifier to other pre-existing and freely available methods on Reuters-21578, the standard text test collection. The results showed that the model can be applied to text classification; The text classification model based on random forest had the best effect, compared with the results of a text classification model based on CART, REPTree and J48 and F1-Measure reached 0.777; The text classification model based on random forest is convenient, intuitive and effective, and the evaluation results are reliable. It can provide a new idea for the research of text classification.

Keywords: - Various ensemble classification methods have been proposed in recent years. These methods have been proven to improve classification accuracy considerably.

I. INTRODUCTION

Due to the increasing amount of text information available on the Internet, interest in automatic analysis of the information has also increased. There are a number of techniques to search, organize, and process this information and one of these techniques is text classification, by which an unknown text document is assigned to one of several classes. Text classification has been an active topic in computer science for over forty years. There are many different algorithms and techniques used to perform the classification[1]. And, some have proved to be better at certain tasks than others. Naïve Bayes (NB), K-nearest neighbor (k-NN), Support Vector machines (SVM), the Recchio method and Neural Network (NN) have demonstrated good text classification as classic methods. The classification methods described above are the most commonly used ones in automatic text classification. Each of them has some unique advantages. Some of them are easier to implement, such as k-NN. Others are more complex, but they are more robust and adaptive, such as SVM. There also suffer from additional shortcomings. The linear classifiers, such as Rocchio, may have the cancroids of a class falling outside the clustered documents. To further enhance the performance of automatic text classification, we developed a new method, called RF-Miner, based on Random forests algorithm. Random forests (also called RF) was introduced by Leo Breiman in 2001[2]. Random Forest is "competitive in accuracy with the best classification algorithms that are out there now". It is a decision-tree-based ensemble classifier that can achieve classification accuracy. Random Forests is wildly applied to many fields ranging from clinical research to financial decision-making[3]. Random Forests does not over fit, runs fast and efficiently on large datasets such. It does not require assumptions on the distribution of the data, which is interesting when different types or scales of input features are used. These outstanding properties make it suitable for text classification.

II. LITERATURE SURVEY

Ensemble of feature sets and classification algorithms for sentiment classification

The ensemble framework is applied to sentiment classification tasks, with the aim of efficiently integrating different feature sets and classification algorithms to synthesize a more accurate classification procedure. First, two types of feature sets are designed for sentiment classification, namely the part-of-speech based feature sets and the word-relation based feature sets. Second, three well-known text classification algorithms, namely naive Bayes, maximum entropy and support vector machines, are employed as base-classifiers for each of the feature sets. Third, three types of ensemble methods are used to draw sentiment

Deep convolutional neural networks for sentiment analysis of short texts

Sentiment analysis of short texts such as single sentences and Twitter messages is challenging because of the limited contextual information that they normally contain. In this work they propose a new deep convolutional neural network that exploits sentiment analysis of short texts. They apply our on Stanford Twitter Sentiment corpus (STS), which contains Twitter messages. For the SSTb corpus, our approach achieves state-of-the-art results for single sentence sentiment prediction in both binary positive/negative classification

Short text classification in Twitter to improve information filtering

In microblogging services such as Twitter, the users may become overwhelmed by the raw data. One solution to this problem is the classification of short text messages. As short texts do not provide sufficient word occurrences. To address this problem, we propose to use a small set of domain-specific features extracted from the author's profile and text. The proposed approach effectively classifies the text to a predefined set of generic classes such as News, Events, Opinions, Deals, and Private Messages

III. EXISTING SYSTEM

For the study of text classification, many predecessors have done a lot of excellent work. In Existing system, proposed an improved KNN algorithm based on clustering, improved the feature selection function by connecting the accurate coefficients of several feature selection functions to form a new feature selection function, and finally used SVM to classify and proposed an improved algorithm based on bagging's Chinese text classifier. Based on the increase of text information and the development of text processing technology, the application of text classification is more and more. For example, public opinion monitoring, emotional analysis, commodity classification, news classification, etc.

Disadvantages

- Inaccurate
- Low Quality of text extraction
- Poor Classification

IV. PROPOSED SYSTEM

In the traditional random forest algorithm, the number and quality of feature selection are not prominent. But for books and other large capacity text classification, the more the number and quality of text features (classification decision tree attribute), the better the classification effect will be. Therefore, this project proposes a tr-k method which Tr-k method for the extraction of text features, first of all, preprocess the input text set, remove stop words. Then, feature set K1 is extracted by TF-IDF and feature set K2 is extracted by textrank. The intersection of feature set K1 and K2 is taken as the clustering center of K-means, and feature set K3 is obtained. By using three text feature extraction methods, we can get more diversified and high-quality text feature sets.

Advantages

- Accurate
- High Quality of text extraction
- Better Classification

V. RANDOM FOREST

A Random Forests is a classifier consisting of a collection of tree-structured classifiers {h(x,Θλ k),k=1,...} where the {Θk} are independent identically distributed random vectors and each casts a unit vote for the most popular class at input X. Random Forests is a multi-classifiers system. It constructs the numerous trees as sub-classifiers (or called internal classifiers). It combines CART's tree generation idea and the bagging predictor to create tree forests. Using the vote from multiple trees, it can yield more accurate, in general, classification on the test data set. Its accuracy is determined by the correlation are weak and the trees are strong, the final vote will be more accurate. The RF is acclaimed as one of the most accuracy classifiers developed to date. Random Forests algorithm is based on Bagging Sampling and CART. And through a voting system, it becomes a very accurate predictor.

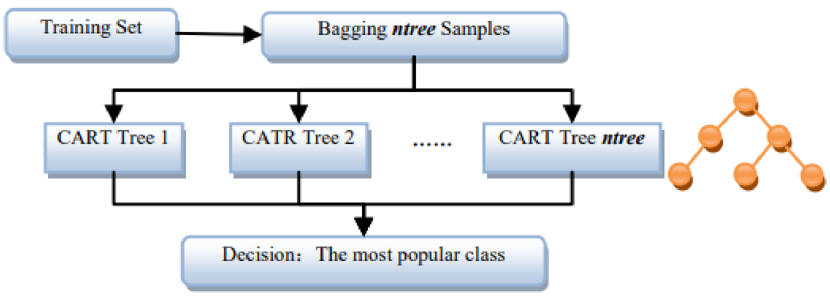


Fig.1, Random Forest work mechanism

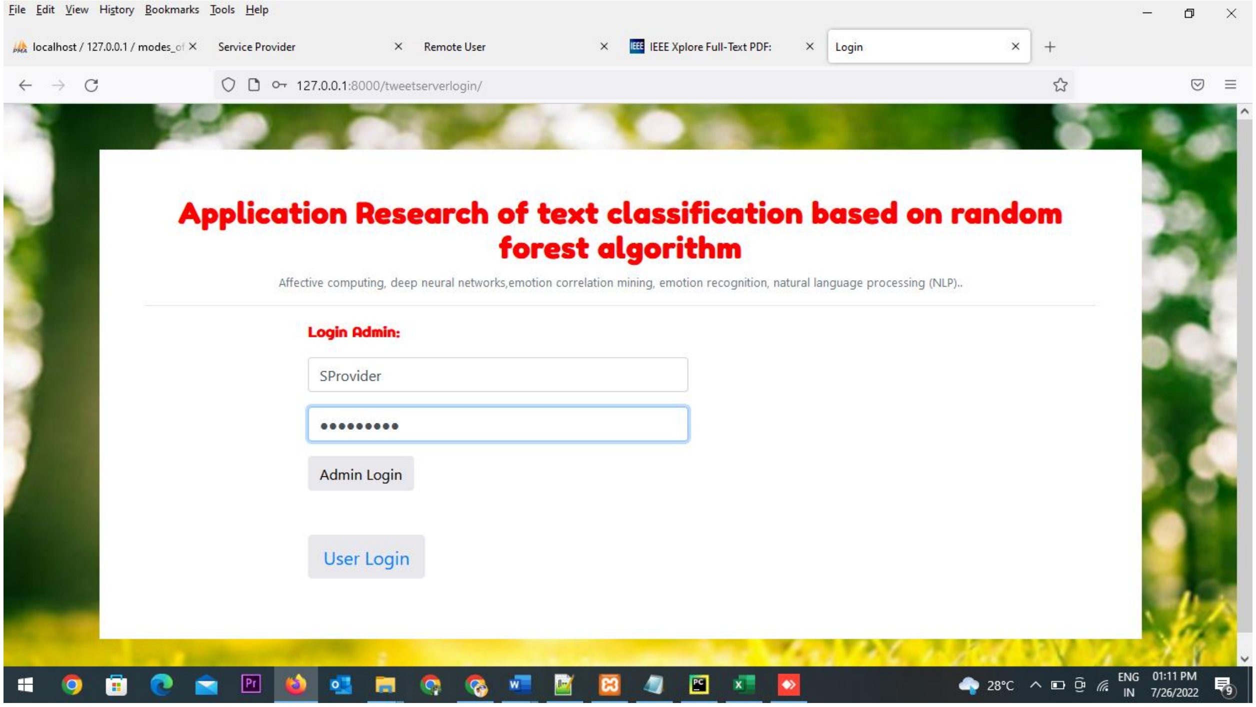
Bagging Sampling

Breiman introduces bagging Sampling in 1996. In recent years, it becomes quite popular as other famous sampling methods: boosting (including Adaboosting), v-fold cross-validation, leaf-one cross-validation, randomization, etc. A detailed study and applications of this method are given in [4][5] etc. Bagging is the acronym of bootstrap aggregating. It has following two phrases, sampling and voting. In the above two phrases, the sampling is the kernel and the first randomness in Random Forests.

CART

Classification and Regression tree (also called CART) was introduced by Leo Breiman, et al. In 1984. It creates a tree based on the training data set. At each tree node, it will find the locally best split rule by using some split method. In Random Forests, each classification tree is a no pruned CART tree. The advantage of using the no pruned tree over using original CART, a pruned tree, is the resulting decrease in the correlation among tress. Even the no pruned tree will affect the strength if each tree but the reduced correlation will improve the final accuracy after combination of all trees. Without pruning, each tree generation will be much simpler and quicker. Therefore, a side benefit of this approach is its assistant on smaller time consumption when generating hundreds of trees.

VI. RESULT


Application Research of text classification based on random forest algorithm

Affective computing, deep neural networks,emotion correlation mining, emotion recognition, natural language processing (NLP)..

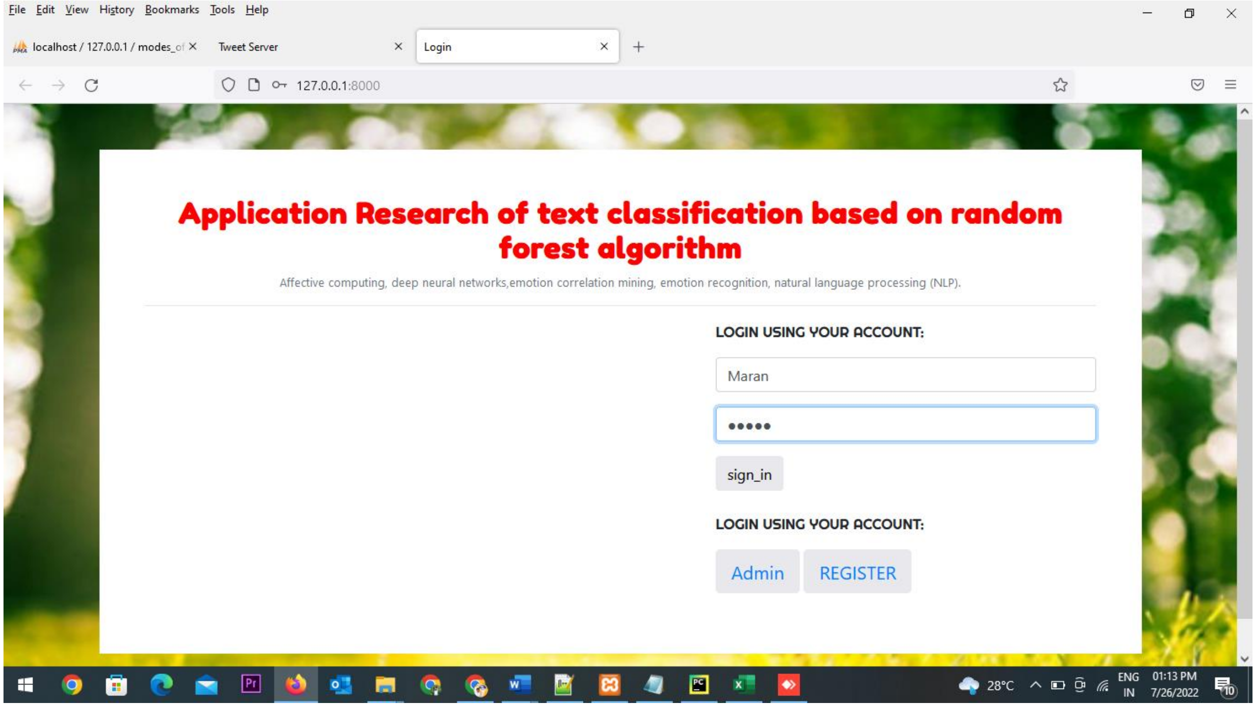
Login Admin:

SProvider

•••••••••

File Edit View History Bookmarks Tools Help
localhost / 127.0.0.1 / modes_of
Service Provider
Remote User
IEEE Xplore Full-Text PDF:
Login
127.0.0.1:8000/tweetserverlogin/
Admin Login
User Login
28°C
ENG IN
01:11 PM
7/26/2022

Fig 2. Admin Login


Application Research of text classification based on random forest algorithm

Affective computing, deep neural networks,emotion correlation mining, emotion recognition, natural language processing (NLP).

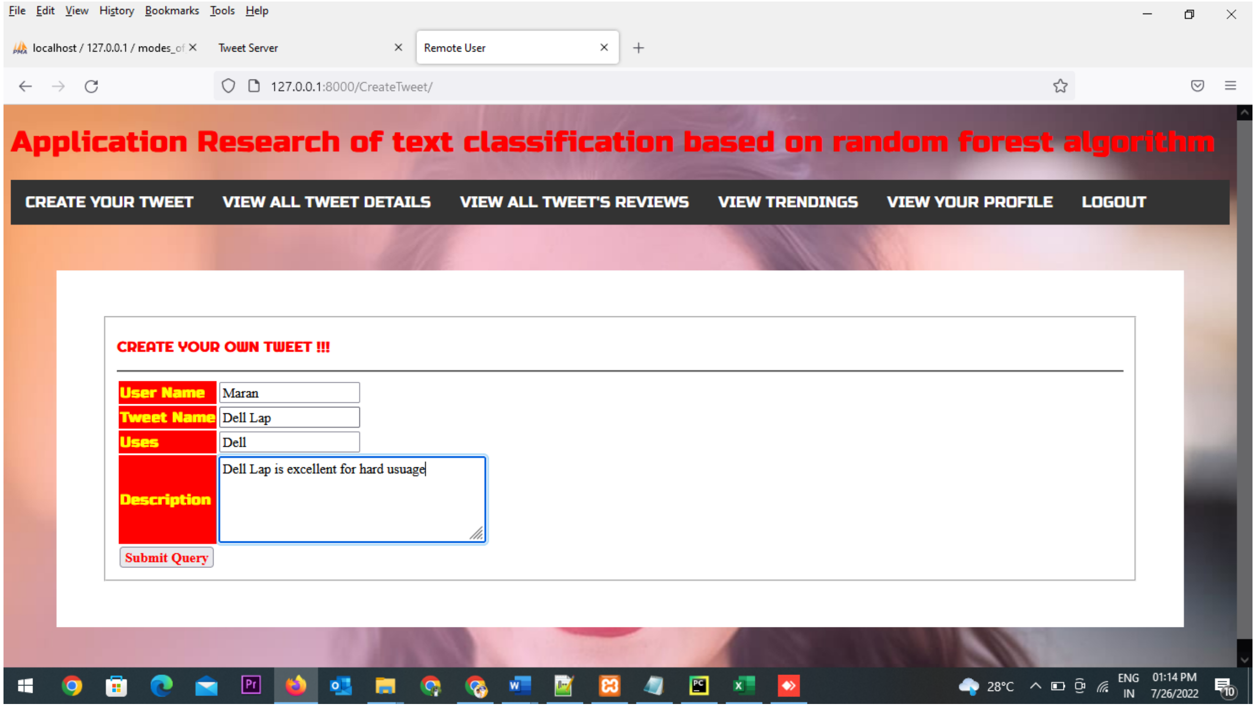
LOGIN USING YOUR ACCOUNT:

Maran

•••••

LOGIN USING YOUR ACCOUNT:

Fig 3. User Login


File Edit View History Bookmarks Tools Help
localhost / 127.0.0.1 / modes_of
Tweet Server
Remote User
127.0.0.1:8000/CreateTweet/
28°C
ENG IN
01:14 PM 7/26/2022

Application Research of text classification based on random forest algorithm

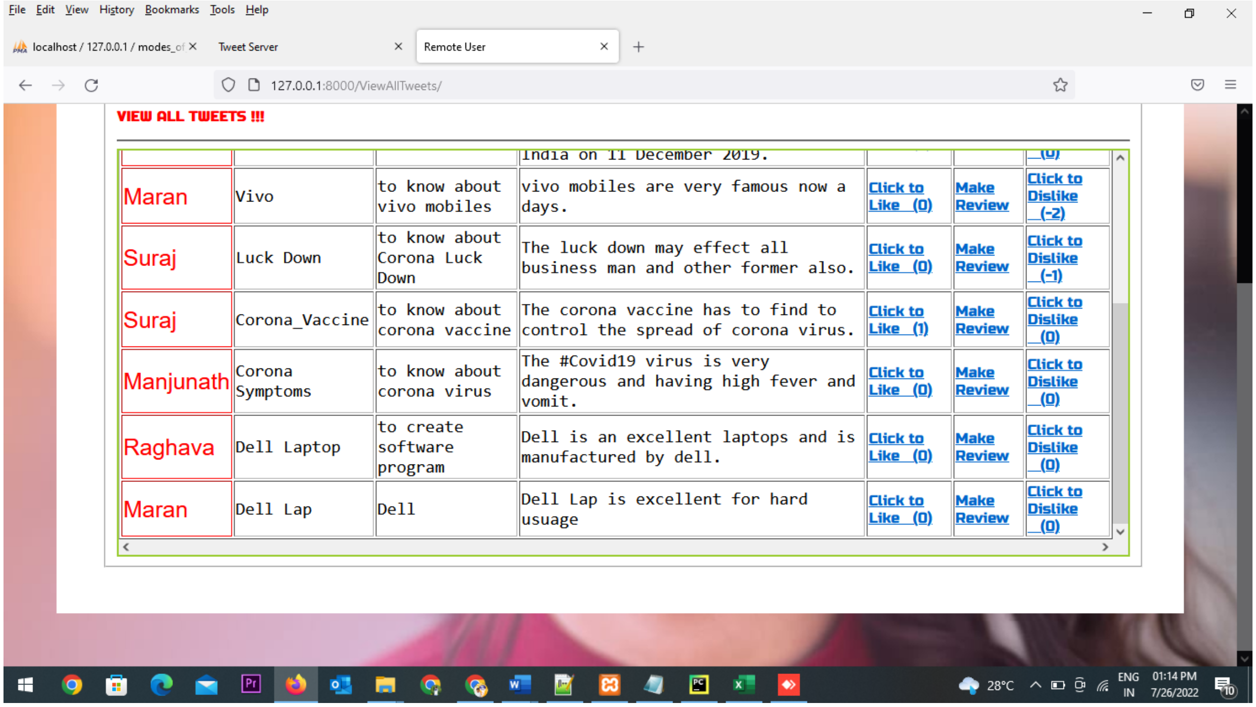
CREATE YOUR TWEET | VIEW ALL TWEET DETAILS | VIEW ALL TWEET'S REVIEWS | VIEW TRENDINGS | VIEW YOUR PROFILE | LOGOUT

CREATE YOUR OWN TWEET !!!

User Name	Maran
Tweet Name	Dell Lap
Uses	Dell
Description	Dell Lap is excellent for hard usage

Submit Query

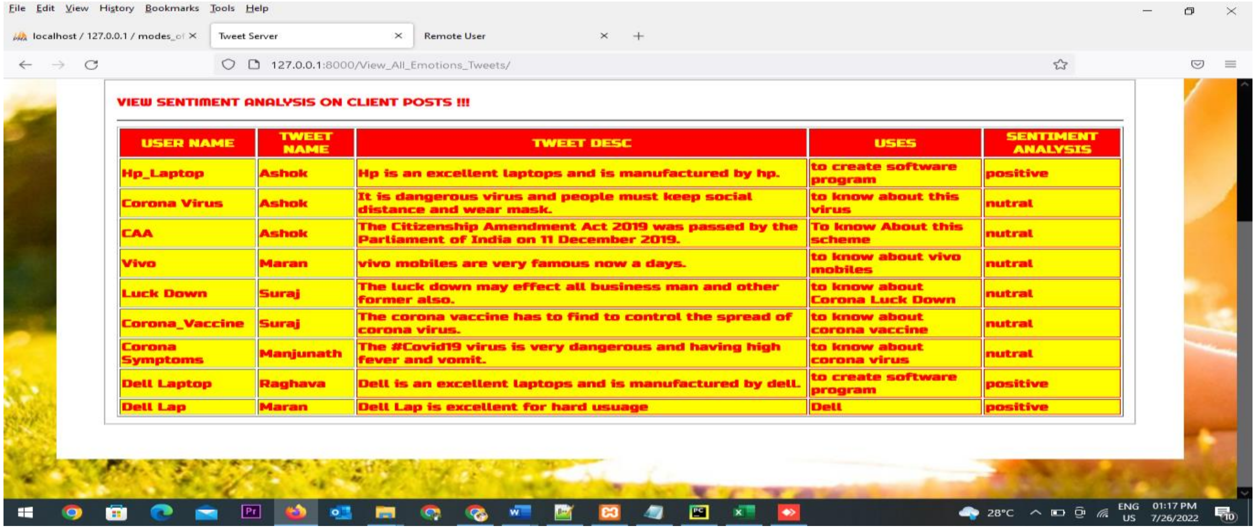
Fig 4. Creating Tweet


File Edit View History Bookmarks Tools Help
localhost / 127.0.0.1 / modes_of
Tweet Server
Remote User
127.0.0.1:8000/ViewAllTweets/
28°C
ENG IN
01:14 PM 7/26/2022

VIEW ALL TWEETS !!!

			India on 11 December 2019.			(0)
Maran	Vivo	to know about vivo mobiles	vivo mobiles are very famous now a days.	Click to Like (0)	Make Review	Click to Dislike (-2)
Suraj	Luck Down	to know about Corona Luck Down	The luck down may effect all business man and other former also.	Click to Like (0)	Make Review	Click to Dislike (-1)
Suraj	Corona_Vaccine	to know about corona vaccine	The corona vaccine has to find to control the spread of corona virus.	Click to Like (1)	Make Review	Click to Dislike (0)
Manjunath	Corona Symptoms	to know about corona virus	The #Covid19 virus is very dangerous and having high fever and vomit.	Click to Like (0)	Make Review	Click to Dislike (0)
Raghava	Dell Laptop	to create software program	Dell is an excellent laptops and is manufactured by dell.	Click to Like (0)	Make Review	Click to Dislike (0)
Maran	Dell Lap	Dell	Dell Lap is excellent for hard usage	Click to Like (0)	Make Review	Click to Dislike (0)

Fig 5. All Tweets


File Edit View History Bookmarks Tools Help
localhost / 127.0.0.1 / modes_of
Tweet Server
Remote User
127.0.0.1:8000/View_All_Emotions_Tweets/
28°C
ENG US
01:17 PM 7/26/2022

VIEW SENTIMENT ANALYSIS ON CLIENT POSTS !!!

USER NAME	TWEET NAME	TWEET DESC	USES	SENTIMENT ANALYSIS
Hp_Laptop	Ashok	Hp is an excellent laptops and is manufactured by hp.	to create software program	positive
Corona Virus	Ashok	It is dangerous virus and people must keep social distance and wear mask.	to know about this virus	nutral
CAA	Ashok	The Citizenship Amendment Act 2019 was passed by the Parliament of India on 11 December 2019.	To know About this scheme	nutral
Vivo	Maran	vivo mobiles are very famous now a days.	to know about vivo mobiles	nutral
Luck Down	Suraj	The luck down may effect all business man and other former also.	to know about Corona Luck Down	nutral
Corona_Vaccine	Suraj	The corona vaccine has to find to control the spread of corona virus.	to know about corona vaccine	nutral
Corona Symptoms	Manjunath	The #Covid19 virus is very dangerous and having high fever and vomit.	to know about corona virus	nutral
Dell Laptop	Raghava	Dell is an excellent laptops and is manufactured by dell.	to create software program	positive
Dell Lap	Maran	Dell Lap is excellent for hard usuage	Dell	positive

Fig 6. View Text analysis on tweets

CONCLUSION

Experiments above shows that when nodesize=1, mtry=14, ntree=35,the RF classifier achieved the best results, F1-Measure get 0.777. We introduce the use of Random Forests as a new and useful modeling technique in the field of Text Classification. This method is easy to use and fast, requiring only the tuning of three parameters (but their value is not critical).Free software implementations are available. It shows similaror better discriminate capability than CART, J48 and REPTree on the typical datasets, reuters-21578, analyzed in this work. The random forests algorithm provides not only a comparable best classification method, but also some techniques on other statistics, such as out-of-bag estimation, outlier detection, variable importance rank, etc. All these have been implemented in the software for research purpose. Ongoing research includes refinements of the selection method in order to deal appropriately with strongly correlate attributes and the application of RF to other problems in the text classification field.

REFERENCES

[1] Niusha Shafiabadya, L.H. Leeb, R. Rajkumarc, V.P. Kallimanid, Nik Ahmad Akramc, Dino Isac: Using unsupervised clustering approach to train the Support Vector Machine for text classification. Neurocomputing. Vol.211(2016),p.4-10.

[2] Breiman L:Random Forests.Machine Learning.Vol.45(2001),p.5-32.

[3] Fallon NG , Fielding S, Fernandes PG: Classification of Southern Ocean krill and icefish echoes using random forests. ICES JOURNAL OF MARINE SCIENCE.Vol.72(2016),p.1998-2008.

[4] Breiman L. Bagging Preditors [J]. Machine Learning, 1996, 24(2):123-140.

[5] Luo J, Meng B, Quan CQ, Tu XH: Exploiting salient semantic analysis for information retrieval. ENTERPRISE INFORMATION SYSTEMS.Vol.10(2016),p.959-969.

[6] Z. Zhao and X. Ma, "Text emotion distribution learning from small sample: A meta-learning approach," in *Proc. Conf. Empirical Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLPIJCNLP)*, 2019, pp. 3948–3958.

[7] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

[8] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, Mar./Apr. 2013.

[9] X. Lin, M. Zhang, Y. Liu, and S. Ma, "A neural network model for social-aware recommendation," in *Proc. Asia Inf. Retrieval Symp.*, 2017, pp. 125–137.