



## TEXT CLASSIFICATION USING THE RANDOM FOREST ALGORITHM: AN APPLICATION STUDY

G Mahesh<sup>1</sup>, A. Ramya Sri<sup>2</sup>, T. Sai Madhuri<sup>3</sup>, B. Swathi<sup>4</sup>, K. Sivananda Reddy<sup>5</sup>

<sup>1</sup> *Asst. Professor, Krishna Chaitanya Institute of Technology & Sciences, Markapur, A.P, India*  
<sup>2,3,4,5</sup> *Scholar, Krishna Chaitanya Institute of Technology & Sciences, Markapur, India*

---

### ABSTRACT:

In view of the poor classification effect of traditional random forest algorithms due to the low quality of text feature extraction, a random forest method for text information is proposed. In view of the difficulty in controlling the quality of traditional random forest decision trees, a weighted voting mechanism is proposed to improve the quality of decision trees. This algorithm uses tr-k method based on text feature extraction to improve the quality and diversity of text features, and uses the latest Bert word vector generation model to represent the text. Experimental data in the Python environment show that this method can achieve better results in text classification than IDF based random.

**Keywords :** Text Classification, Random forest algorithm, vector generation mode, tr-k method

---

### [1] INTRODUCTION

With the rapid development of science and technology, since the 1990s, more and more data information has been generated, 80% of which is stored in text. Therefore, people can't use the traditional manual filtering for huge amounts of text information. Text processing based on natural language processing emerges as the times require. In recent years, there is more and more research on text classification, mainly focusing on Naive Bayes, K-means clustering, SVM and other algorithms. Random forest algorithm is widely used in all walks of life due to its advantages of fast training speed, easy parallel computing in the era of big data, strong anti-interference ability and excellent anti over fitting ability, and has achieved the effect of traditional methods.

For the study of text classification, many predecessors have done a lot of excellent work. For example, Zhou Qingping proposed an improved KNN algorithm based on clustering; Yang, 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; Zhang Xiang 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.

Many advantages of random forest algorithm (RFA) make experts and scholars carry out many improved application research on RFA. In 1995, tin Kam ho first proposed the concept of random forest. Later, Leo boeiman proposed that RFA is a classification and prediction model. M p Perrone, In coope and others proposed that in the classification stage, RF class labels are synthesized from the classification results of all decision trees, and are the most commonly used methods in voting and probability average. In terms of application, EI atta proposed a method to predict the activity of cannabinoid receptor (CB2) agonist using RF in bioinformatics; in ecology, eruan et al. Studied air prediction using RFA; in genetics, retroria used RFA in gene recognition. Moreover, RFA has achieved good results in biochip, information extraction.

## **[2] LITERATURE SURVEY**

Parul Kalra, Deepti Mehrotra et. al studied , As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge .Text classification which classifies the documents according to predefined categories .In this paper we are tried to give the introduction of text classification, process of text classification as well as the overview of the classifiers and tried to compare the some existing classifier on basis of few criteria like time complexity, principal and performance.

Mikhail V.KotsaMikhail A.Ryabinin et. al. presented A weighted random survival forest in this paper. It can be regarded as a modification of the random forest improving its performance. The main idea underlying the proposed model is to replace the standard procedure of averaging used for estimation of the random survival forest hazard function by weighted averaging where the weights are assigned to every tree and can be viewed as training parameters which are computed in an optimal way by solving a standard quadratic optimization problem maximizing Harrell's C-index. Numerical examples with real data illustrate the outperformance of the proposed model in comparison with the original random survival forest.

Zhang et. al proposed several context-based methods for text categorization. One method, a small modification to the PPM compression-based model which is known to significantly degrade compression performance, counter-intuitively has the opposite effect on categorization performance. Another method, called C measure, simply counts the presence of higher order character contexts, and outperforms all other approaches investigated.

Timor Kadir et. al studied, Decision trees are attractive classifiers due to their high execution speed. But trees derived with traditional methods often cannot be grown to arbitrary complexity for possible loss of generalization accuracy on unseen data. The limitation on complexity usually means suboptimal accuracy on training data. Following the principles of stochastic modeling, we propose a method to construct tree-based classifiers whose capacity can be arbitrarily expanded for increases in accuracy for both training and unseen data. The essence of the method is to build multiple trees in randomly selected subspaces of the feature space. Trees in, different subspaces generalize their classification in complementary ways, and their combined classification can be monotonically improved. The validity of the method is demonstrated through experiments on the recognition of handwritten digits.

Jiajia Liu Yudong Ye et. al. presented an approach to predict the activity of analogues of 2,4,6-trisubstituted 1,3,5-triazines as cannabinoid receptor (CB2) agonists using random forest technique. We compute twenty molecular descriptors for a data set of 58 analogues for the component, and depending on values of these descriptors we train random forest to find a relation between biological activity and molecular structure of analogues. The results obtained by random forest were compared with the decision tree and support vector machine classifiers and the random forest has 100% overall predicting accuracy and for decision tree and support vector machines were 93% and 67% respectively.

In the Existing system used Naive Bayes. In Naive Bayes, texts are classified based on posterior probabilities generated based on the presence of different classes of words in texts. This assumption makes the computations resources needed for a naïve bayes classifier far more efficient than non-naïve bayes approaches which are exponential in complexity. Moreover, it is found that Naive Bayes is the Less accurate model for text classification.

The Disadvantages of Existing System are

The main limitation of Naive Bayes is the assumption of independent predictor features. Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it's almost impossible that we get a set of predictors that are completely independent or one another.

- Less quality text classification by using naive bayes.
- we haven't implemented tf-idf concept for classification

In this Paper, the proposed method is based on the Random forest and is proposed to Perform text classification. In the traditional random forest algorithm, the number and quality of feature selection are 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 paper proposes a tr-k method which combines TF-IDF, textrank and K-means to improve the effect of text classification. The full name of the TF-IDF method is term frequency inverse document frequency.

The Advantages of Proposed System are Random forests overcome several problems with decision trees, including:

- Reduction in over fitting: by averaging several trees, there is a significantly lower risk of over fitting.
- Less variance: By using multiple trees, you reduce the chance of stumbling across a classifier that doesn't perform well because of the relationship between the train and test data. tr-k method which combines TF-IDF, textrank and K-means to improve the effect of text classification.
- Random forest has achieved good results in biochip, information extraction and other fields.

### [3] SYSTEM ARCHITECTURE

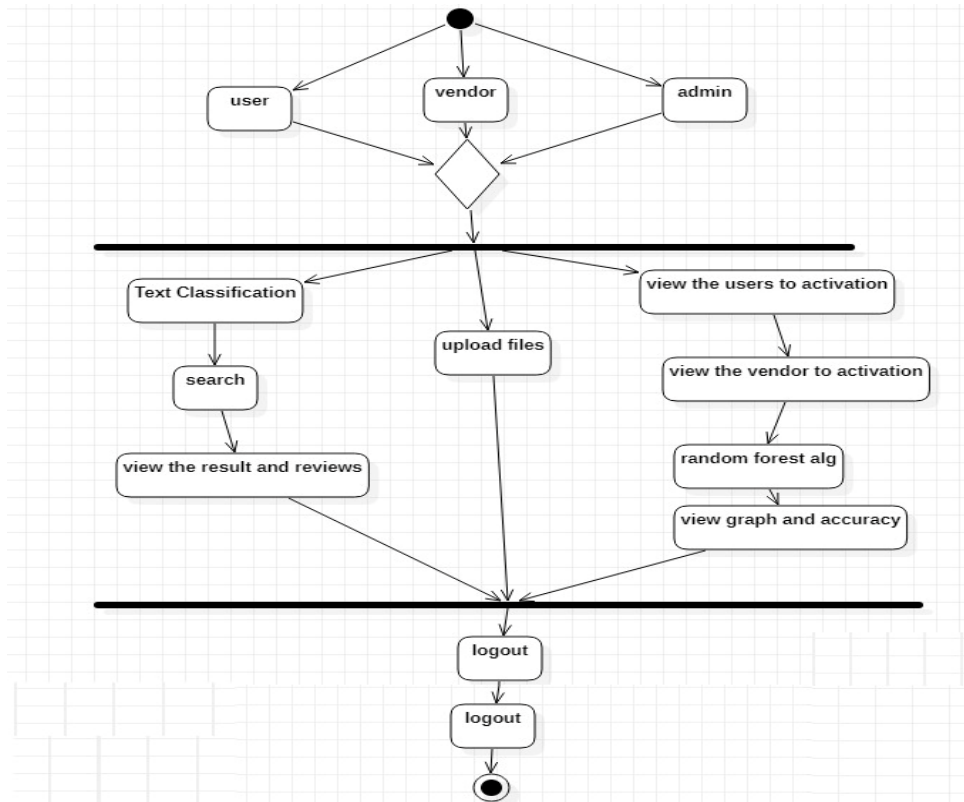


Fig. 1 Activity Diagram For Text Classification Using the Random Forest Algorithm

### 3.1 SOFTWARE ENVIRONMENT

In this paper we have implemented source code in Python Programming language and Django. Python is a general - purpose interpreted, interactive, object -oriented, and high-level programming language. An interpreted language, Python has a design philosophy that emphasise the code readability (not ably using whitespace indentation to delimit code blocks rat her than curly bracket s or keywords), and a syntax that allows programmers to express concept s in fewer lines of code than might be used in languages such as C++ or Java. It provides construct s that enable clear programming on both small and large scales. Python interpreters are available for many operating systems. CPython, the reference implement at i on of Python, is open source soft ware and has a community- based development model, as do nearly all of its variant implementations. CPython is managed by the non- profit Python Soft ware Foundation. Python features a dynamic type system and automatic memory management . It support s multiple programming paradigms, including object - oriented, imperative, functional and procedural , and has a largeand comprehensive standard library.

Django is a high-level Python Web framework that encourages rapid development andclean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. It's free and open source. Django's primary goal is to ease the creation of complex, database-driven websites. Django emphasizes reusability and "plug ability" of components, rapid development, and the principle of don't repeat yourself. Python is used throughout, even for settings files and data models.

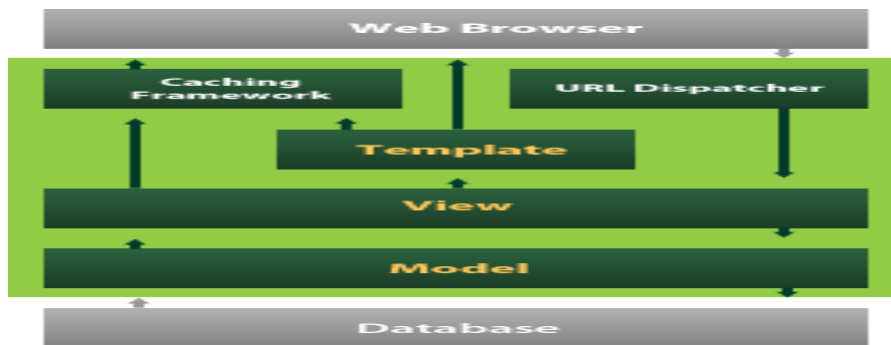


Fig 2. Design Pattern Used in Django-Model-View-Template-Controller Architectural Pattern

Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models

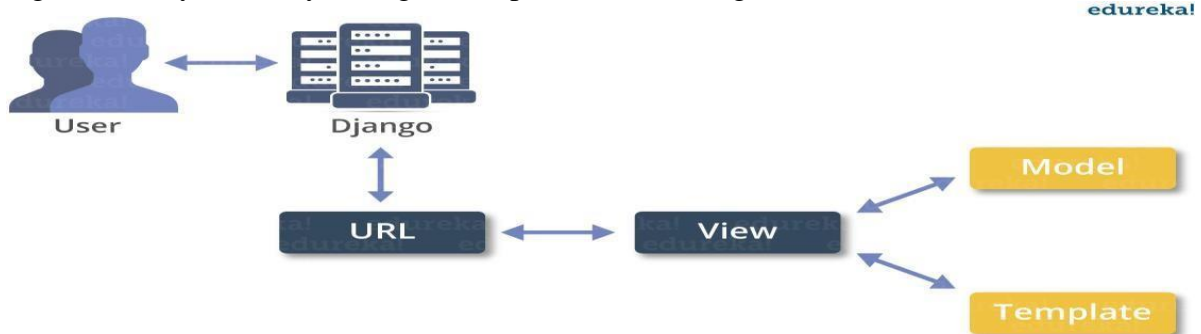


Fig. 3 :Model-View-Template(MVT) Architecture

## [4] IMPLEMENTATION

### 4.1 Modules Description

**i) User:** The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user registers, then the admin can activate the customer. Once the admin activates the customer then the customer can login into our system. user will search for music product then he will get list of music products. He will purchase products and will give the review and ratings to product. whenever user click on classification link he will get sentimental analysis of music products.

**ii) Uploader:** The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user registers, then the admin can activate the customer. Once the admin activates the customer then the customer can login into our system. up loader will upload different type of music products.

**iii) Admin:** Admin can login with his credentials. Once he login he can activate the users. The activated user only login in our applications.. Once he login he can activate the vendor. The activated vendor only login in our applications. after that admin will perform random forest classification and will get the accuracy and precision reports.

**iv) Machine learning:** Machine learning refers to the computer's acquisition of a kind of ability to make predictive judgments and make the best decisions by analyzing and learning a large number of existing data. The representation algorithms include deep learning, artificial neural network, decision tree, enhancement algorithm and so on. The key way for computers to acquire artificial intelligence is machine learning.

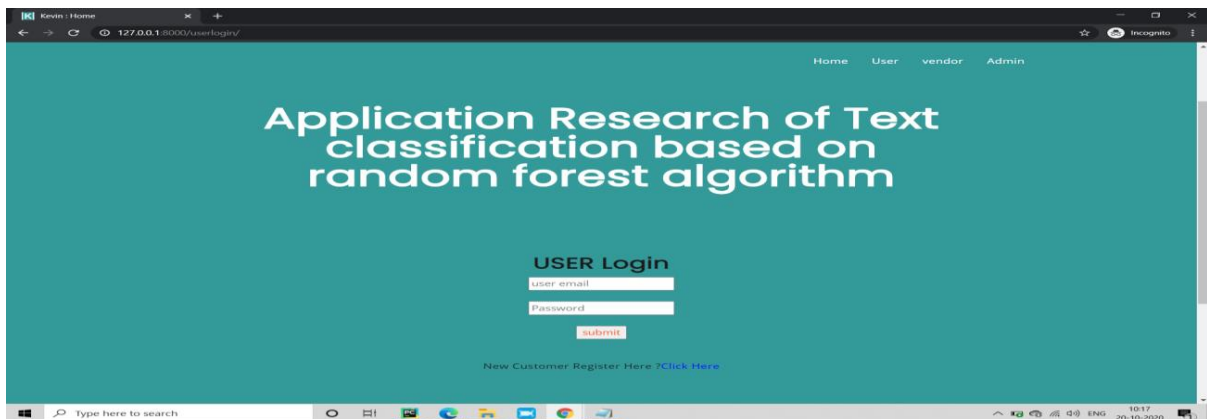
Nowadays, machine learning plays an important role in various fields of artificial intelligence. Whether in aspects of internet search, biometric identification, auto driving, Mars robot, or in American

presidential election, military decision assistants and so on, basically, as long as there is a need for data analysis, machine learning can be used to play a role.

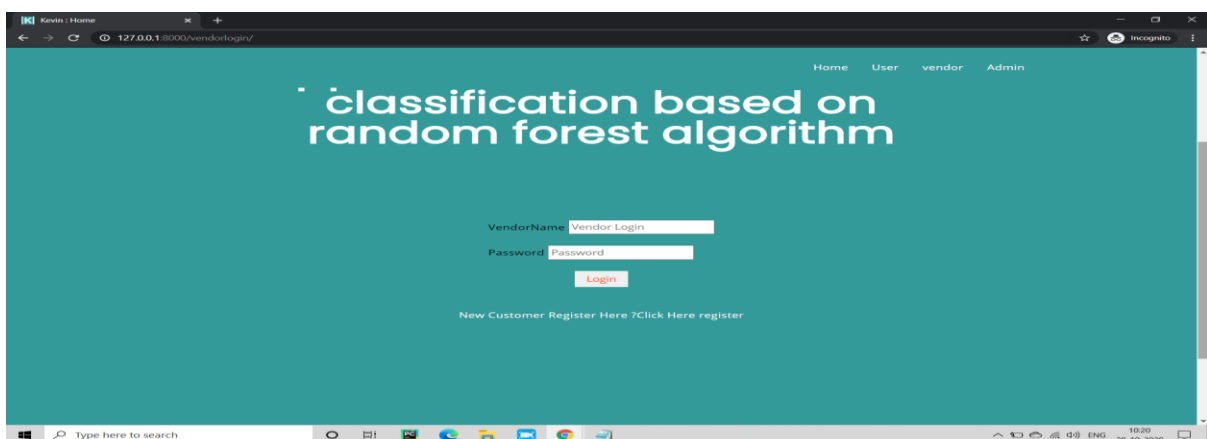
## 4.2 SCREEN SHOTS



**Fig.4 : Home Page of Using Text Classification Using the Random Forest Algorithm**

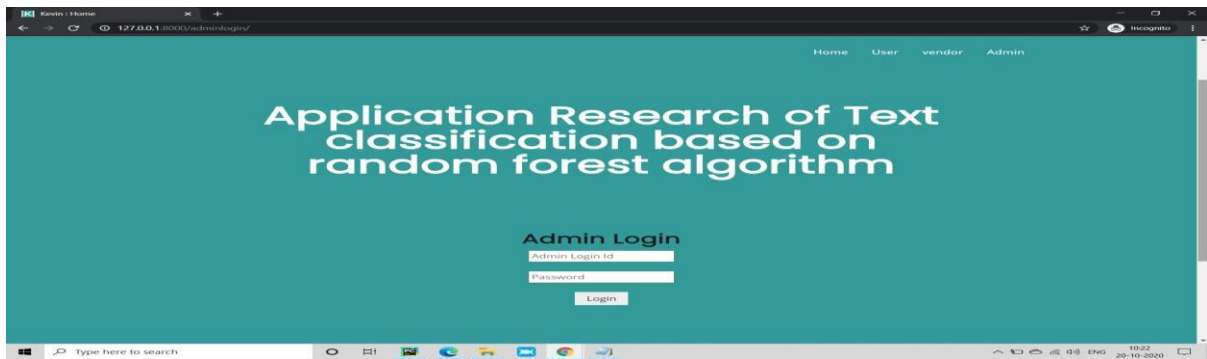


**Fig.5: User Login form of Text Classification Using the Random Forest Algorithm**

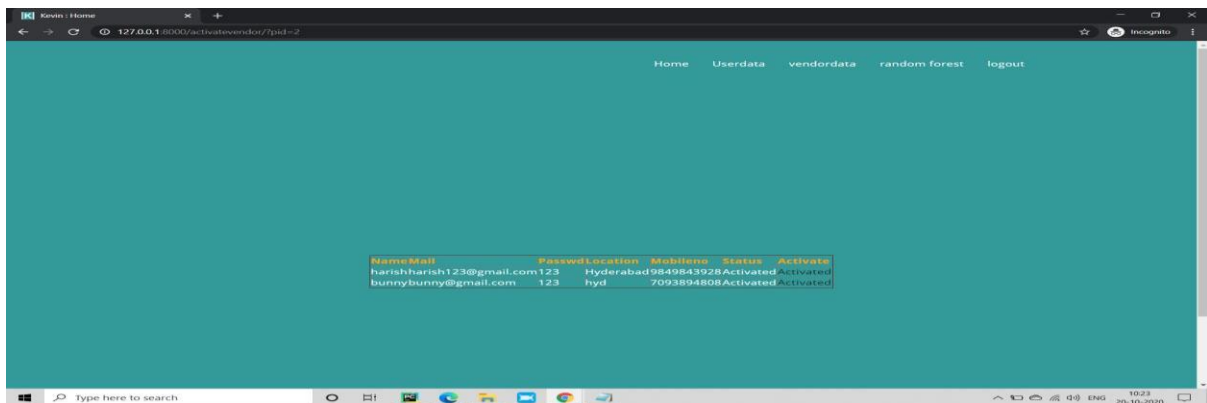


**Fig. 6 :Vendor Login Page of Text Classification Using the Random Forest Algorithm**

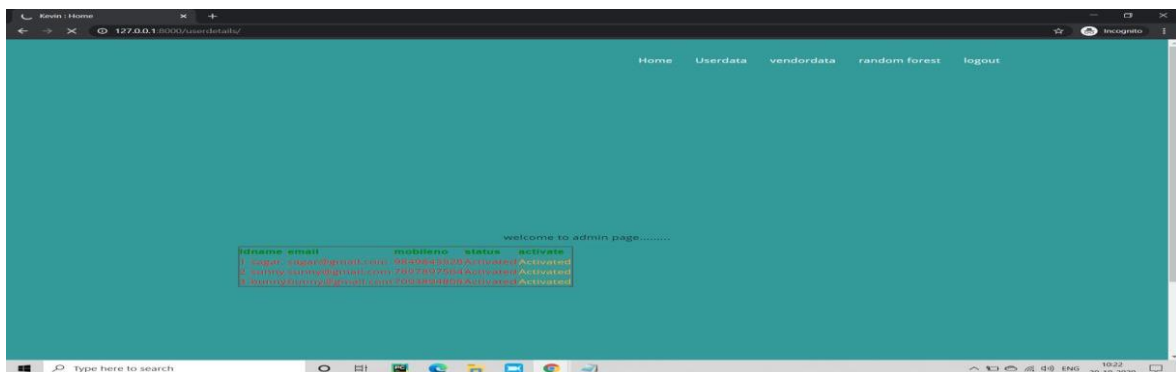




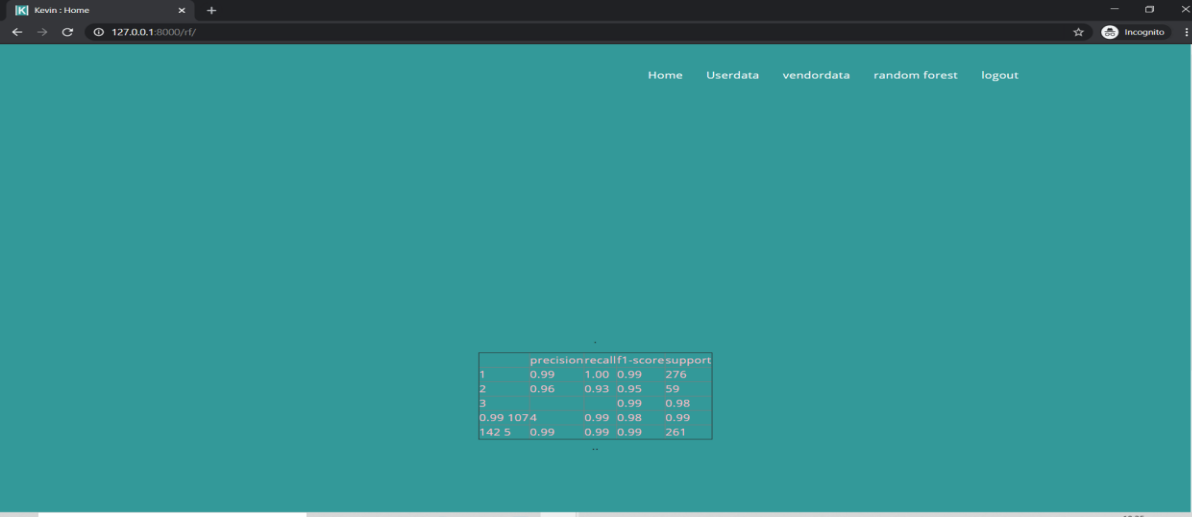
**Fig. 7 :Admin Login Page of Text Classification Using the Random Forest Algorithm**



**Fig. 8 :Admin Approve the User to activate the account of Text Classification Using the Random Forest Algorithm**



**Fig. 9 : Admin Approve the Vendor Account to activate of Text Classification Using the Random Forest Algorithm**



	precision	recall	f1-score	support
1	0.99	1.00	0.99	276
2	0.96	0.93	0.95	59
3	0.99	0.99	0.98	0.99
1425	0.99	0.99	0.99	261

**Fig. 10: This page Shows the Final Output of Text Classification Using the Random Forest Algorithm**

## [5] CONCLUSION AND FUTURE WORK

In this paper, the input text data set of a random forest algorithm is processed to improve the classification effect. At the same time, we use the Bert word vector model to improve the quality of text representation, and then improve the classification accuracy of the final random forest. Experiments show that the model can effectively improve the classification accuracy and F1 value. In future research, improvement in operational efficiency and overall prediction accuracy could be analyzed. In addition, this algorithm is designed for unbalanced data and not suitable for industries with high turnover rate. How to increase the universality of the algorithm still needs to be studied further. In future we can also do text classifications for videos and images.

## REFERENCES

- [1] Korde V, Mahender C N. Text classification and classifiers: A survey [J]. International Journal of Artificial Intelligence Applications, 2012, 3 (2) :85.
- [2] Utkin L V , Konstantinov A V , Chukanov V S , et al. A weighted random survival forest [J]. Knowledge-Based Systems, 2019, 177(AUG.1):136-144.
- [3] Yang Y. An examination of text categorization methods [C] // International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1999: 42-49.
- [4] Mantas C J , Castellano J G , Serafín Moral-García, et al. A comparison of random forest based algorithms: random credal random forest versus oblique random forest [J]. 2019.
- [5] T.K.Ho. Random Decision Forest [J]. In Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montreal, Canada, 1995, 8:278-282.
- [6] Breiman L. Random forests [J]. Machine Learning, 2001, 45 (1) :5~32.
- [7] L K Hansen, P Salamon. Neural network ensembles [J]. Pattern Analysis and Machine Intelligence, 1990, 12 (10) :993~1001.
- [8] M P Perrone, L N Cooper. When networks disagree: Ensemble method for neural networks [A]. Artificial Neural Networks for Speech and Vision [C]. New York: Chapman & Hall, 1993. 126~142.
- [9] El-Atta A H A, Moussa M I, Hassanien A E. Predicting Biological Activity of 2,4,6 trisubstituted 1,3,5-triazines Using Random Forest [J]. 2014, 303: 101-110.
- [10] Erwan Scornet, Gérard Biau, Jean Philippe Vert. Consistency of random forests [J]. Eprint Arxiv, 2015, 9(1):2015--2033.
- [11] Petralia F, Wang P, Yang J, et al. Integrative random forest for gene regulatory network inference [J]. Bioinformatics, 2015, 31(12):i197.



- [12] Kimura S, Tokuhisa M, Okada-Hatakeyama M. Inference of genetic networks from time-series of gene expression levels using random forests[C]. Computational Intelligence in Bioinformatics and Computational Biology. IEEE, 2017:1-6.
- [13] Janitzka, Silke, Tutz, Gerhard, Boulesteix, Anne-Laure. Random forest for ordinal responses: Prediction and variable selection[J]. Computational Statistics & Data Analysis, 96:57-73.
- [14] Lee J , Yu I , Park J , et al. Memetic feature selection for multilabel text categorization using label frequency difference[J]. Information Sciences, 2019, 485:263-280.
- [15] Tang X , Dai Y , Xiang Y . Feature selection based on feature interactions with application to text categorization[J]. Expert Systems with Applications, 2018.
- [16] Ayesha Mariyam, SK Althaf Hussain Basha, and S Vishwanadha Raju "Applications of Multi-Label Classification", International Journal of Innovative Technology and Exploring Engineering(IJITEE),pp.86-89,ISSN:2278-3075, Volume-9 Issue-4S2, March 2020,Blue Eyes Intelligence Engineering & Sciences Publication.
- [17] P M Yohan, Sk Althaf Hussain Basha, B Sasidhar, A.Govardhan , “ Automatic Named Entity Identification and Classification using Heuristic Based Approach for Telugu”, IJCSI(International Journal of Computer Science Issues), Volume 10, Issue 6,November 2013,ISSN: 1694-0814
- [18] Donapati Srikanth SK Althaf Hussain Basha, T Naveen Kumar , V. Anand , “Categorization of Academic Student Performance using Hybrid Techniques” International Conference on Advanced Computing Methodologies (ICACM-2013), Hyderabad, pp.325- 330,2013.
- [19] Ch. Prakash, Sk Althaf Hussain Basha, , D. Mounika, G. Maheetha, “An Approach for Multi Instance Clustering of Student Academic Performance in Education Domain”, IJJDWM Journal, Volume 3,Issue 1,pp.1-9,Feb.2013,ISSN: 2249-7161
- [20] SK Althaf Hussain Basha, A.Govardhan, “MICR: Multiple Instance Cluster Regression for Student Academic Performance in Higher Education”, International Journal of Computer Applications(IJCA), Volume 14– No.4,2011,pp.23-29, ISSN: 0975-8887
- [21] Sreedhar Jinka Sk. Althaf Hussain Basha, Suresh Dara, Baijnath Kaushik, “Sequence Labelling for Three Word Disambiguation in Telugu Language Sentences”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology ( IJSRCSEIT) Volume 2 , Issue 7 , pp.311-315, 2017, ISSN : 2456-3307.
- [22] Baijnath Kaushik , Sk. Althaf Hussain Basha, Sreedhar Jinka, D Praveen Kumar, “ Sequence Labelling for Two Word Disambiguation in Telugu Language Sentences”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology ( IJSRCSEIT) Volume 2 , Issue 7 , pp.321-327, 2017, ISSN : 2456-3307.
- [23] D. Praveen Kumar, Sk. Althaf Hussain Basha, Sreedhar Jinka, Baijnath Kaushik, A. Jagan, “Empirical Analysis of Context Sensitive Grammars and Parse Trees for Disambiguating Telugu Language Sentences”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology ( IJSRCSEIT) Volume 2 , Issue 7 , pp.328-331, 2017, ISSN : 2456-3307.
- [24] A Jagan, Sk. Althaf Hussain Basha, Sreedhar Jinka, Baijnath Kaushik, D. Praveen Kumar, “NLP: Context Free Grammars and Parse Trees for Disambiguating Telugu Language Sentences”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology ( IJSRCSEIT) Volume 2 , Issue 7 , pp.332-337, 2017, ISSN : 2456-3307.
- [25] Sd.Muneer , Sk Althaf Hussain Basha, A.Govardhan, V.Uday Kumar “ Generate Eligible Students using Decision Trees-A Frame work for Employee Ability” International Journal of Advanced Computing(IJAC), Volume 4,Issue 2,2012,pp.68-76, ISSN: 0975-7686.
- [26] SK Althaf Hussain Basha, Pammi Pavan Kumar, Jinka Sreedhar, “Innovative Techniques and Technologies in Translation in a Multilingual Context -2012”, Third International Conference on Translation, Technology and Globalization in Multilingual Context, ITA, New Delhi.
- [27] Y.Sri Lalitha, Sk. Althaf Hussain Basha, N.Sandhya, Dr. A.Govardhan "Web Usage Mining- A Survey", IEEE International Advance Computing Conference (IACC 2009) , Thapar University Patiala,1684-1689,2009, ISSN: 978-981-08-2465.
- [28] SK Althaf Hussain Basha, A.Govardhan “A Comparative Analysis of Prediction Techniques for Predicting Graduate Rate of University”, European Journal of Scientific Research (EJSR) ,Vol.46 No.2,2010, pp.186-193, ISSN No:1450-216X