ISSN: 0974-5823 Vol. 7 No. 6 June, 2022

International Journal of Mechanical Engineering

Text Classification Method Utilizing a Hybrid Learning Algorithm

[1]Mr. Syed Ahmeduddin, [2]Mohd Abdul Arshad, [3]Mohammed Aseeb Ahmed, [4]Mohd Adnan Ali, [5]Mohammed Faiz Qureshi

[1]Assistant Professor, Department of Computer Science and Engineering, Lords Institute of Engineering and Technology, Hyderabad.

[2,3,4,5]Research Scholar, Department of Computer Science and Engineering, Lords Institute of Engineering and Technology, Hyderabad.

ABSTRACT— In text categorization, files are organised into specified groups depending on the material they contain. Suggestions for new supervised learning methods text that can be properly classified is more than enough Appropriate information for learning. In-depth look at a innovative text categorization technique that relies on smaller training materials to be used as an alternative to letters and the laws of pattern recognition Pre-classified documents are utilised to generate features extracted. Conceptualizing naive Bayes is then applied to the resulting features and classification is made. Genetic Algorithms [2] are ultimately reduced to a single concept was inserted as a last-minute addition to sustain and enhance. Experimental using this method, findings demonstrate that the classification is much more accurate. More precise than the currently used [3] text classification techniques

Index Terms— Text classification, Feature extraction, Genetic Algorithm, Naive Bayes.

I.INTRODUCTION

Text categorization having risen to prominence within field of information extraction like an essential tool. Records must be classified manually within predetermined categories depending on actual contents. Automated text categorization was the subject of numerous methods [3]. Automated text categorization procedures are being developed with the help of existing techniques.

The concept of an apriory algorithm was shown to be widely understood in the context of text categorization. There are many intriguing correlations and associations that can be discovered via association rules [1]. It can be helpful to uncover those connections between large volumes of transactional data. Instead of using a greatest a probability prediction, Naive bayes utilizes the maximum likelihood method. [3] It presupposes that every letter in a paper is independently of all of the other keywords in that document based on its class. For learning and practice, the Naive Bayes takes a massive amount of records even if various studies [7] have shown its effectiveness. To begin, the genetic algorithm utilizes a procedurally chosen population of rules to establish an initial population. A string of bits can be used to indicate every rule. The fittest members of the appropriate sample, and also their descendants, make up a new number of individuals based on the principle of natural selection and survival. Generally, a rule's ability to correctly classify a collection of training samples is used as a measure of its overall quality.

A novel algorithm for text categorization is introduced in this report. To extract feature sets from pre-classified textual information, word relationship i.e. association rules [1] are utilised. Finally, the notion of Genetic Algorithm was included for final classification using Naive Bayes Classifier. Algorithm implementation and testing have been completed. This system is an excellent textual classification, according to the findings of the experiments.

II.RELATED WORK

Nave Bayes Classification with the Association Rule

Employing the classification rules of data analysis, the Nave Bayes classification is used for text categorization indicated the classification model of text the Naive Bayes classifier's reliability The rules of collaboration [4] However, this strategy does not take into account the fact that it does not take into account the how to calculate an instance of a negative outcome correctness of any classification decision might to the ground in certain instances. It's all you need to categorise a text. Various classes' probabilities are calculated by using the value estimates of the complete set the haphazard assortments.

Genetic algorithms are used to choose text categorization features

In classification tasks, feature selection is very important for enhancing classification performance. This research investigates a genetic algorithm-based [5] feature search strategy for various text representations. In one hand, this optimization algorithm can identify a segmentation results that is ideal for classification performance; from the other contrary, it could identify any feature extraction method with the least dimensional that ensures the highest generalization ability. Three of the finest classifications have been chosen to test this strategy, including Naïve Bayes classifier (NB), Nearest Neighbours (KNN), and Support Vector Machine (SVM). Our goal is to see if utilising F-measure, [8] we can increase textual classification accuracy by employing evolutionary techniques rely feature extraction. 20 Newsgroups and Reuters-21578 were used as test datasets for the experiments. In addition, the findings were fascinating.

Text classification using a parallel learning algorithm

Text classification describes the process of organising papers into a set of predetermined experiences and interactions by the contents of the records themselves. In order to learn effectively, supervised machine learning techniques currently in use to properly categorise text require a substantial amount of labelled documents. An additional method that leverages sizable pools of unidentified materials in addition to the goal of identifying that are already accessible is the application of the Expectation-Maximization (EM) technique to this problem. This is an alternative method. Furthermore, the amount of time required to learn with any of these extensive materials that are not labelled is very long. This work presents an original parallel learning technique for the job of text classification. The Estimation technique as well as the classification Algorithm [3] has been combined in the parallel method. In order to provide the desired results. Our objective is to shorten the time spent in computing throughout the learning and classification processes. On a sizable ensemble of Windows computers that we referred to as the PIRUN Cluster, we analysed how well our parallel method performed. We publish the outcomes both in terms of consistency plus precision. Based on these findings, it appears that the parallel method that was developed is able to deal with big collection of documents.

Keep the space's border under constant surveillance for new patterns as they emerge. Mining and discovering new knowledge from data

The term emerging patterns (EPs) refers to a class of useful and widely applicable knowledge patterns. Biomedical profiling data has been used in recent studies to address complex cancer detection challenges and provide higher classification accuracy than other methods. However, finding EPs is a difficult and time-consuming task.

In this research, we investigate how tiny changes in the data can be used to incrementally adjust and preserve the short border definitions of all developing patterns. Maintaining the boundaries ensures that no preferred features are lost because EP spaces are convex. New data insertion, old data deletion, growth of new characteristics, and redundancy of existing attributes are all handled by the techniques that we present here. On six sets of benchmark data, we compared these progressive methods to a more effective method that starts from scratch. According to the data, it is clear that incremental algorithms outperform the traditional "From-Scratch" approach.

Comparison of Two Text Categorization Algorithms

When it comes to categorising natural language documents, inductive learning can be a useful tool. Natural language processing providers are rapidly relying on character recognition. Research on text classification using machine learning and information extraction techniques had previously been uneven, prevented from

making assumptions about just the efficacy of certain techniques. A Probabilistic classification and a recursive partitioning technique were tested on two text classification databases in this research. We found that both algorithms performed reasonably well and allowed for essential part of the business among false negatives, which we believe, is a good sign. Because of its ability to handle enormous selected features, the decision tree [7] method's sequential feature selection technique is especially useful for document classification. Although this approach relies on a preliminary which was before of characteristics, the findings are nevertheless confirmed.

III.METHODOLOGY

An association rule-based text classification method is described in this project, which uses less training words or documents to improve classification accuracy. In present text classification algorithms, we need to extract every keyword from the texts and then construct features/TF-IDF levels and afterwards training these attributes with ml algorithms such as Naive Bayes or Decision Trees [7]. There can be a delay in execution and a decrease in classification results due to a large amount of data.

To solve the aforesaid issue, the author is using the Association Rule on the trained dataset to identify the most frequently occurring words, which can then be used to training machine learning techniques as well as to categorize text. This technique reduces the length of texts since it removes all less frequent terms and trains ML with frequent words, resulting in faster execution time and higher accuracy.

Positive and negative examples have been spotted in the following code. The word sets that do not fit our class are considered negative sets, while those that do are deemed positive.

Proposed Algorithm:

n = number of class, m = number of associated sets

1. For each class i = 1 to n do

2. Set pval = 0, nval = 0, p = 0, n = 0

3. For each set s = 1 to m do

4. If the probability of the class (i) for the set (s) is maximum then increment pval else increment nval

5. If 50% of the associated set s is matched with the keywords set do step 6 else do step7

6. If maximum probability matches the class i then increment p

7. If maximum probability does not match the class i increment n

8. If (s<=m) go to step 3

9. Calculate the percentage of matching in positive sets for the class i

10. Calculate the percentage of not matching in negative sets for the class i

11. Calculate the total probability as the summation of the results obtained from step 9 and 10 and also the prior probability of the class i in set s

12. If (i<=n) go to step 1

13. Set the class having the maximum probability value as the result.

Nave Bayes Classifier and Association Rule:

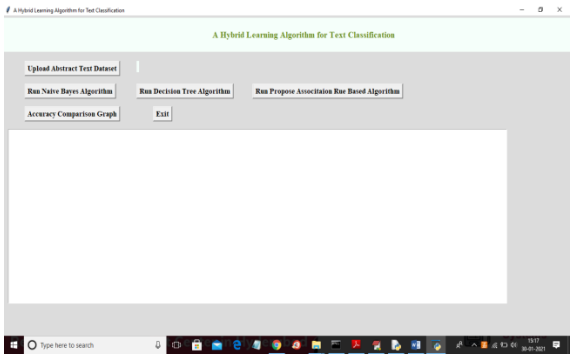
Data sets used in both methods are identical, thus the following results may be drawn from them [4]. Using only half of the Training data, the proposed technique worked well.

A Decision Tree Based on Association Rules: Only 76 percent of the entire 33 data sets were used to train in character recognition using an association rule-based tree structure. It is possible to categorise text with 78% accuracy utilising only 50% of the data as training for the suggested method. One of the main issues with decision tree-based classification is that it fails miserably to identify classes. Even with three times the size of data sets, our method outperforms the competition.

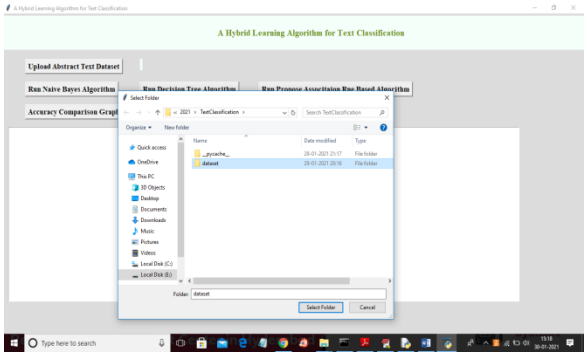
Involvement in Genetic Algorithm: Using 69 percent of the training data, genetic algorithm-based text classification performed well, however the technique is time-consuming [2] [5].

IV.RESULT AND DISCUSSION

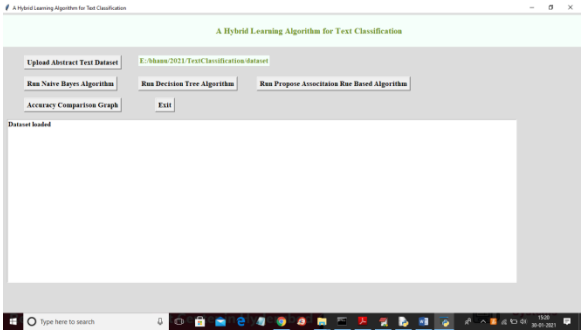
Run project to get below result



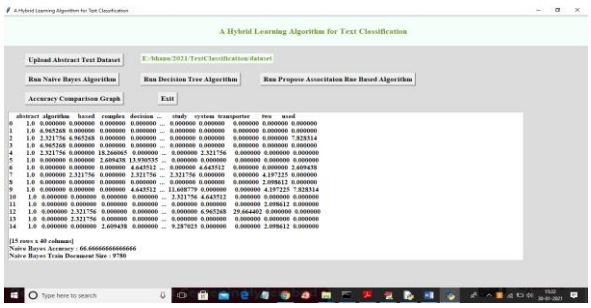
To upload the dataset, go to the 'Upload Abstract Text Dataset' section of the results page shown above and click the button.



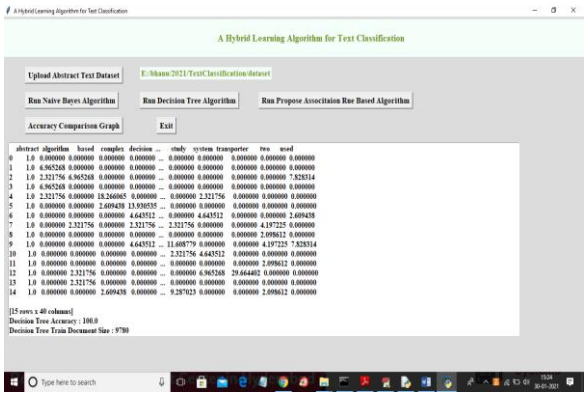
Once the 'dataset' folder has been selected and uploaded, click the 'Select Folder'' button to begin loading the dataset and obtain the results shown below.



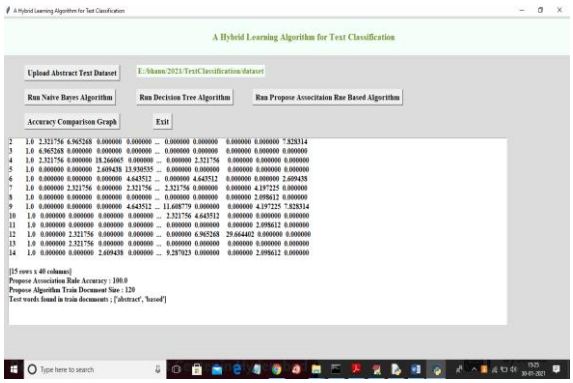
To train the abstracts dataset with the naive bayes algorithm, click on the 'Run Nave Bayes Algorithm' button in the outcome dataset.



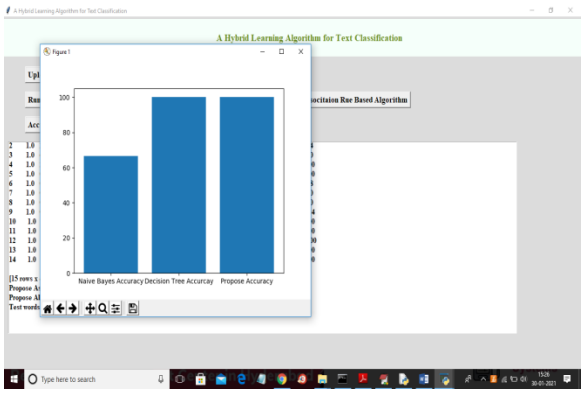
The TF-IDF idea is used to turn all words in the dataset into features, which are then put to the Nave Bayes method to calculate its correctness. The naive bayes accuracy in the above screen is 66%, and this algorithm was training on 9780 total characteristics. In order to train an ML algorithm, the application will divide the entire dataset into two halves, one for training and the other for testing. When you're ready, click on the 'Run Decision Tree Algorithm' tab.



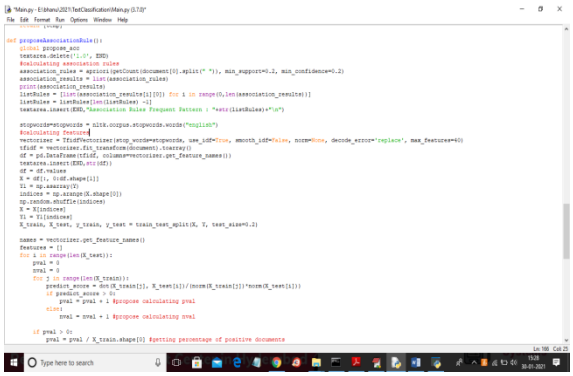
To see the decision tree's accuracy, click on the 'Run Propose Association Rule Based Algorithm' button and enter the 9780 features that were used to train the tree in the above results.



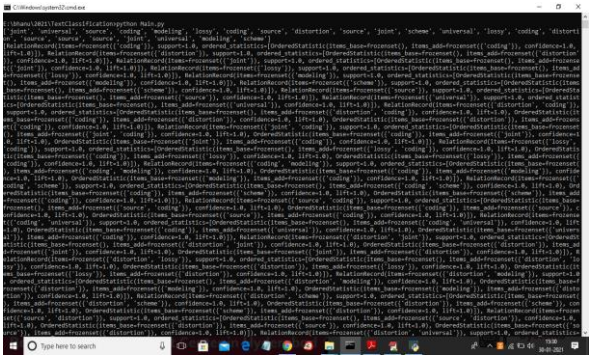
In the aforementioned results, the proposed approach achieved 100% accuracy, however it trained with only 100 attributes because this technique removes infrequent objects from the training set. Here, both the proposal and the decision tree achieved the same levels of accuracy, although the number of features utilised to train them differed considerably. To see the following graph, click on the 'Accuracy Comparison Graph' tab.



The x-axis shows the name of the algorithm, and the y-axis shows the accuracy of that algorithm. The code for implementing the algorithm proposed in the study is shown in the following result.



The association rule calculation is shown in the black console below.



V.CONCLUSION

In this study, a novel hybrid text categorization methodology was introduced. Previous techniques demand more data to be trained and take longer to compute as a result of both. The suggested hybrid technique uses less training dataset and takes less time to compute than the current alternatives. To our surprise, we were only able to get 78% accuracy out of 50% of the training data. A class couldn't have been classified, thus we deleted this position and enlarged the labelled training data in order to get a more acceptable outcome. If we create a Frequently Patterns (FP) growing trees, it will take less time. A bigger sample size will be preferable even if the results of the experiment were favourable. Text classification [3] can be improved by incorporating further genetic algorithm features. We plan to use this technique for developing pattern detection in the near future [6].

VI.REFERENCES

[1] Agarwal, R., Mannila H., Srikant R., Toivonan H., Verkamo A., "A Fast Discovery of Association Rules," Advances in Knowledge Discovery and Data Mining, 1996.

[2] Anwar M. Hossain, Mamunur M. Rashid, Chowdhury Mofizur Rahman, "A New Genetic Algorithm Based Text Classifier," In Proceedings of International Conference on Computer and Information Technology, NSU, pp. 135-139, 2001.

[3] Canasai Kruengkrai, Chuleerat Jaruskulchai, "A Parallel Learning Algorithm for Text Classification," The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2002), Canada, July 2002.

[4] Chowdhury Mofizur Rahman, Ferdous Ahmed Sohel, Parvez Naushad, Kamruzzaman S. M, "Text Classification Using the Concept of Association Rule of Data Mining," In Proceedings of International Conference on Information Technology, Kathmandu, Nepal, pp 234-241, May 23-26, 2003.

[5] Eshita Sharmin, Ayesha Akhter, Chowdhury Mofizur Rahman, "Genetic Algorithm for Text Categorization," In Proceedings of International Conference on Computer and Information Technology, BUET, pp. 80-85, December, 1998.

[6] Jinyan Li, Thomas Manoukian, Guozhu Dong, and Kotagiri Ramamohanarao. Incremental Maintenance of the Border of the Space of Emerging Patterns. Data Mining and Knowledge Discovery, 9 (1): 89- 116, July 2004

[7] Lewis, D., and Ringuette, M., "A Comparison of Two Learning Algorithms for Text Categorization," In Third Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93, 1994.

[8] Masud M. Hassan, Chowdhury Mofizur Rahman, "Text Categorization Using Association Rule Based Decision Tree," Proceedings of 6th International Conference on Computer and Information Technology, JU, pp. 453-456, 2003.