International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 03, ISSN No. 2455-2143, Pages 19-24
Published Online July 2023 in IJEAST (http://www.ijeast.com)

IMPROVING TEXT CATEGORIZATION IN INFORMATION RETRIEVAL: AN ENHANCED KERNEL RIDGE APPROACH

Dr. S.Vijayarani, Dr. R. Janani, R.Nandhini
Assistant Professor,
Department of Computer Science,
School of Computer Science and Engineering,
Bharathiar University, Coimbatore.

Abstract - **Information Retrieval (IR) systems play a crucial role in identifying relevant documents from a collection based on user queries. Popular examples of IR systems include search engines like Google, which locate web documents that match given words. In library settings, IR systems are employed to search digital records containing information about books rather than the books themselves. The primary objective of this research is to automatically classify information within a table. Categorization is a fundamental requirement for text retrieval systems. This research encompasses three key steps: Pre-processing, Searching, and Classification. For the searching phase, two existing algorithms, Boyer Moore and Brute Force, are utilized, and Naïve Bayes based searching algorithm is proposed. In the classification phase, existing algorithms, Linear Regression, Random Forest Regression and Kernal Ridge Regression are employed and enhanced algorithm called Enhanced Kernal Ridge for Text Categorization (EKRTC) is used. Experimental results demonstrate that EKRTC achieves the highest accuracy.**

Keywords: **Information Retrieval, Text Classification, Preprocessing, Searching, Regression**

I. INTRODUCTION

Text mining, also known as text data mining or text analytics, is the practice of extracting valuable insights and meaningful information from a large collection of documents. Text mining, discovering useful knowledge from unstructured or semi-structured text, is attracting increasing attention. Text mining usually involves the process of unstructured text and deriving patterns within the unstructured data and finally evaluates and interprets the output (Janani, R., & Vijayarani, S. 2016). There are various tasks related to text mining, including text categorization, text clustering, concept/entity extraction, fine-grained taxonomy creation, sentiment analysis, document summarization, and entity relation modeling. Information retrieval, lexical analysis to examine word frequency distributions, pattern recognition, annotation and tagging, information extraction, link and association analysis, visualization, and predictive analytics are all part of text analysis. The goal of text categorization is to develop algorithms and models that can accurately and efficiently classify documents into predefined categories, allowing for effective organization, retrieval, and analysis of large collections of text data (Sebastiani, F. 2002). This task has numerous applications in various domains, including document organization; email filtering, sentiment analysis, spam detection, news categorization, and recommendation systems.

The remaining sections are structured as: Section 2 illustrates related works. Section 3explains the methodology of this research work. Results and discussion are given in Section 4. The conclusion of this research work is given in Section 5.

II. RELATED WORK

In a study by (Bhope, V. D., & Deshmukh, S. N. 2015), was proposed a pattern-based techniques and pattern evolving techniques. The results demonstrated that this model improved the accuracy of knowledge retrieval from textual data. (Zhang, K., Xu, H., Tang, J., & Li, J. 2006) explored the use of support vector machines (SVM) for text retrieval, considering both local and global context. Various techniques leveraging local and global context have been developed for text retrieval. Another class of techniques employs semantic analysis concepts, such as ontology-based similarity measures. (Sriurai, W. 2011) conducted a comparison between the feature processing techniques of Bag of Words (BOW) and topic modeling. Text classification algorithms, including Naive Bayes (NB), Support Vector Machines (SVM), and Decision Trees, were utilized for experimentation. The results demonstrated that the topic-modeling approach for representing documents yielded the best performance, with an improvement of 11.1% in F1 measure compared to the BOW model.

In a study conducted by (Charjan, D. S., & Pund, M. A. 2013), the focus was on developing efficient mining

International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 03, ISSN No. 2455-2143, Pages 19-24
Published Online July 2023 in IJEAST (http://www.ijeast.com)

algorithms for discovering patterns from large data collections and searching for useful and interesting patterns. In the field of text mining, pattern mining techniques are utilized to identify various text patterns, including frequent item sets, closed frequent item sets, and co-occurring terms. (Gupta, V., & Lehal, G. S. 2009) analyzed the performance and effectiveness of stemmers in applications like spelling checkers, highlighting the variations across languages. Simple stemmer algorithms typically involve the removal of suffixes using a list of common suffixes, while more complex ones leverage morphological knowledge to derive stems from words. The paper provides a comprehensive overview of common stemming techniques and existing stemmers for Indian languages.

III. METHODOLOGY

The main objective of this research work is to automatically categorize the information which is stored on the table. Categorizing the information is one of the primary requirements of the text retrieval systems. This research work has three important steps, they are Pre-processing, Searching and Classification. The system architecture of proposed system is depicted in Figure 1.

A. Dataset

The Annexure II dataset is used for this experiment, and it was obtained from www.annauniv.edu/research website. This data set consists of a list of journal names from various disciplines. The number of instances of this dataset is 22780 and five attributes, namely serial number, source title, ISSN number and country. The source title describes the name of the journal's, the ISSN number refers to the International Standard Serial Number of a journal and the country refers to the place where the journal is being published. In this research work, the source title attribute is considered for automatic classification.

B. Pre-Processing

It is a preliminary task of text data in order to prepare it for the primary processing or for further analysis. They are used to extract the structured text information from the raw text data. In this research work, stop word removal is used.

Stop word removal- Stop words are language specific functional words which carry no information. It may be of the following types such as pronouns, prepositions and conjunctions. These kinds of word carry less meaning, so these words are filtered out in preprocessing technique.

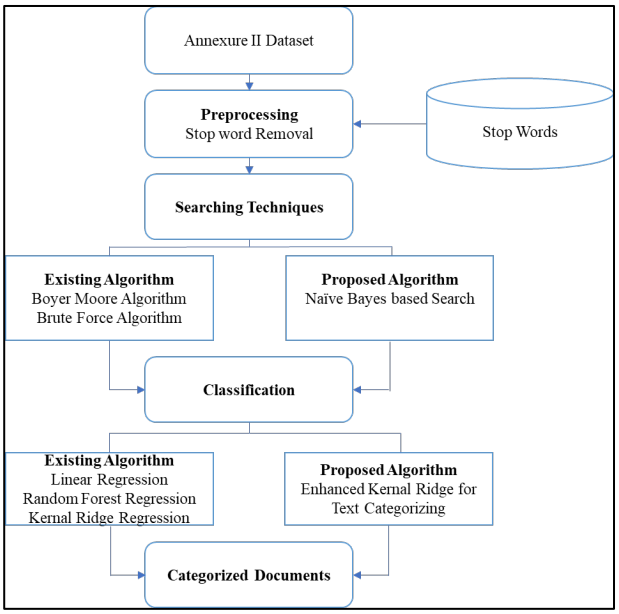

Fig1.System Architecture

C. Searching Techniques

Keyword searching is a technique used to retrieve relevant information from a collection of documents or a database based on specific keywords or terms. It involves searching for documents or records that contain the specified keywords or closely related variations of those keywords. The keyword searching process typically involves defining the keywords, formulating the search query, submitting the query to the search engine, and retrieving the keywords. In this research, two existing search techniques such as, Boyer Moore (Lecroq, T. 1992) and Brute Force (Charras, C., & Lecroq, T. 2004) and Naïve Bayes based Search algorithm is proposed.

1.Naïve Bayes based Search Algorithm

In this research work the proposed Naïve Bayes based Search algorithm using dictionary refers to the different alphabetical terms of different disciplines. First, the Naïve Bayes algorithm isused on labeled data to predict the class labels for the preprocessed documents. This approach allows for more accurate and flexible classification based on the learned patterns and features from the training data (Rish, I. 2001, August). Typically, a simple function is applied to the key to determine its place in the dictionary. The output of the pre-process technique Pi is compared with Dndictionaries. The string Pi goes to the storage buffer where the dictionaries $D1, D2, D3, \ldots . Dn$ are stored are searches for the keywordKi.

$IfPi = Ki \in Dn$ Then the particular Pi is labeled with the relevant class. If the keyword is not found, then the instance is ignored.

There are two types of cases namely for classifying the document which are stored in table format using the dictionary method they are.

International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 03, ISSN No. 2455-2143, Pages 19-24
Published Online July 2023 in IJEAST (http://www.ijeast.com)

- Best case: The best case occurs when the search term is found in any one of the dictionaries from $D1, D2, \ldots . Dn$.
- Worst case: The worst case occurs when the search term is not found in any kind of dictionaries from $D1, D2, \ldots . Dn$.

Algorithm 1: Naïve Bayes based Search

Input: Preprocessed Document: $P1, P2, \ldots, Pn$
Dictionary: $D1, D2, \ldots, Dn$
Output: Classified Keywords
Step 1: Train a supervised learning algorithm using a labeled documents where each document Piis associated with a class label based on its content.
Step 2: Initialize variables: $K = 0, i, j$.
Step 3: for $(i = 1 \text{ ton})$ then
Step 4: check the preprocessed document Pi.
Step 5: Use the trained classifier to predict the class label for Pi.
Step 6: Assign the predicted class label to Pi.
Step 7: If the prediction confidence is below a certain threshold, consider the classification uncertain or assign it to a special class.
Step 8: Increment the key value.
Step 9: Stop the process.

D. Classification

Text classification, also known as text categorization, is the process of automatically assigning predefined categories or labels to textual documents based on their content. It involves training a machine learning model to recognize patterns and relationships in the text data, allowing it to classify new, unseen documents into the appropriate categories. In this research work, three existing algorithms such as Linear Regression (Su, X., Yan, X., & Tsai, C. L. 2012), Radom Forest Regression (Segal, M. R. 2004) and Kernal Ridge Regression (Vovk, V. 2013) are used and Enhanced Kernal Ridge for Text Categorization is proposed.

1. Enhanced Kernel Ridge for Text Categorization (EKRTC)

Enhanced Kernel Ridge for Text Categorization (EKRTC) is an improved approach for text categorization tasks. The traditional Kernel Ridge algorithm, which is a kernel-based machine learning method, is enhanced with additional techniques to achieve better performance in text categorization (Cortes, C., Mohri, M., & Rostamizadeh, A. 2012). The main steps involved in the EKRTC algorithm for text categorization is as follows,

- Text Preprocessing: The text documents are preprocessed such as stop-word removal, to prepare the documents for further analysis.
- Feature/Keyword Extraction: Relevant features or keywords are extracted from the preprocessed text documents.
- Kernel Matrix Construction: A kernel matrix is constructed using the extracted features. The kernel matrix captures the similarity or dissimilarity between pairs of documents in the feature space.
- Training: The EKRTC model is trained using the kernel matrix and the corresponding class labels of the training documents. The goal is to learn a classification function that can accurately predict the class labels of unseen documents.
- Enhanced Regularization: In EKRTC, additional regularization techniques are incorporated to improve the model's generalization performance. This can include techniques such as L1 or L2 regularization, kernel parameter tuning, or model selection using cross-validation.
- Testing and Prediction: The trained EKRTC model is applied to new, unseen documents for prediction. The model assigns class labels to these documents based on their similarity to the training documents in the feature space.

Algorithm 2: Enhanced Kernal Ridge for Text Categorization

Input: Training documents with class labels
Testing documents
Output: Predicted class labels for testing documents
Step 1: Preprocess the testing documents.
Step 2: Extract keywords from the preprocessed documents
Step 3: Construct a kernel matrix using the extracted keywords
Step 4: Train the EKRTC model:
Apply enhanced regularization techniques to improve generalization performance
Solve the kernel ridge regression problem with regularization
Step 5: Test the EKRTC model:
for each testing document do:

International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 03, ISSN No. 2455-2143, Pages 19-24
Published Online July 2023 in IJEAST (http://www.ijeast.com)

Compute the similarity with training documents using the kernel matrix
Predict the class label based on the learned classification function
Step 6: Return the predicted class labels for testing documents

IV. RESULT AND DISCUSSION

All the experiments are carried out on a 2.00 GHz Intel CPU with 4 GB of memory and running on windows 10. The proposed algorithm was experimented with Anna University Annexure II dataset. The performance factors of searching algorithm are,

- Search Time: It refers to the time taken for searching the pattern within the input text. It can be estimated by comparison of each character in pattern with the input text.
- Relevancy: It refers to the accuracy of the algorithm which is correctly classified instance and incorrectly classified instances.

The performance metrics of classification is,

- Correctly classified instance are also called true positive rate or the recall rate, which measures the proportion of actual positives which are correctly identified instances.

True positive = correctly identified

$$TPR = \frac{TP}{P} = \frac{TP}{(TP+FN)} \quad (1)$$

- Incorrectly classified Instances is the false positive rate of error in the predicted data for the test set.

False positive = incorrectly identified

$$FPR = \frac{FP}{N} = \frac{FP}{(FP+TN)} \quad (2)$$

Where P is positive instances and N is negative instances

- Search time: The time complexity of an algorithm quantifies the amount of time taken by an algorithm to run as a function of the length of the string representing the input. Search time measures the amount of required time for classifying the information.

Table 1: Search Accuracy

Measures	Existing Algorithms		Naïve Bayes based Search
	Boyer Moore Algorithm	Brute Force Algorithm	
Correctly Classified Instances (%)	69.97	74.10	89.74
In Correctly Classified Instances (%)	29.17	23.14	10.01

From **Table 1**, the existing search algorithms searches the keyword with lower accuracy when compared to the Naïve Bayes based Search algorithm. The proposed algorithm searches the keyword with higher accuracy. **Fig 2** shows the search accuracy.

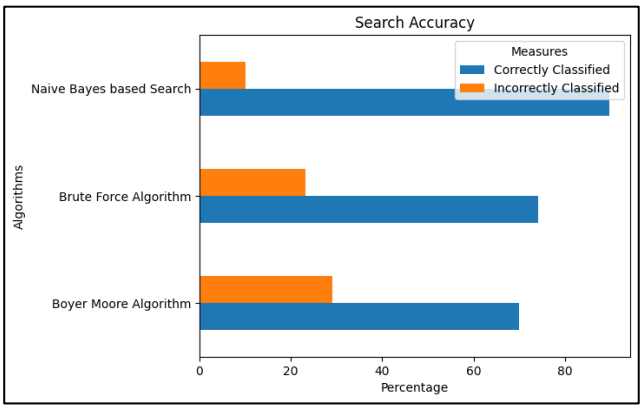

Fig 2: Search Accuracy

Table 2 shows the time taken for keyword searching using existing and proposed algorithms. From this analysis, the proposed algorithm takes less time to search the keywords. The search time of existing and proposed algorithms are depicted in **Fig 3**.

Table 2: Search Time

Algorithm	Time taken (sec)
Boyer Moore Algorithm	2.76
Brute Force Algorithm	2.39
Naïve Bayes based Search Algorithm	2.01

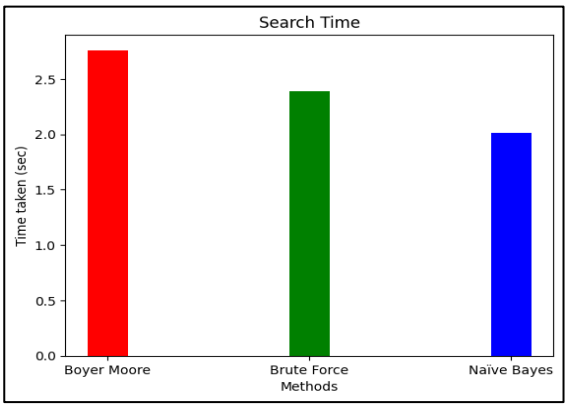

Fig 3: Search time for Keyword Searching

International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 03, ISSN No. 2455-2143, Pages 19-24
Published Online July 2023 in IJEAST (http://www.ijeast.com)

In **Table 3,** the classification accuracy of existing and proposed algorithm is shown. From this, the enhanced algorithm yields better accuracy when compared to the other existing algorithms. The classification accuracy is shown in **Fig 4**.

Table 3: Classification Accuracy

Measures	Existing Algorithms			Enhanced Kernal Ridge for Text Categorization (EKRTC)
	Linear Regression	Random Forest Regression	Kernal Ridge Regression	
Correctly Classified Instances (%)	59.23	62.84	69.47	79.94
In Correctly Classified Instances (%)	39.14	38.96	27.10	17.11

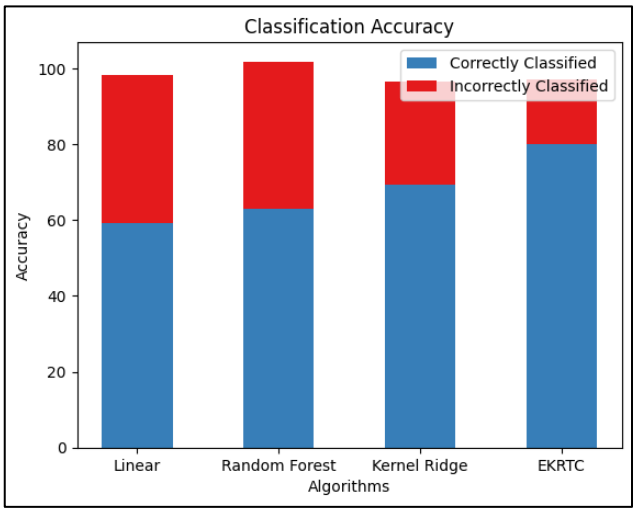

Fig 4: Classification Accuracy

Table 4 shows the classification time of existing and proposed algorithm. From this, it is found that the enhanced algorithm takes minimum time for classification. The classification time of existing and enhanced algorithm is shown in **Fig 5.**

Table 4: Classification Time

Method	Time taken (sec)
Linear Regression	3.95
Random Forest Regression	3.01
Kernal Ridge Regression	2.46
EKRTC	2.11

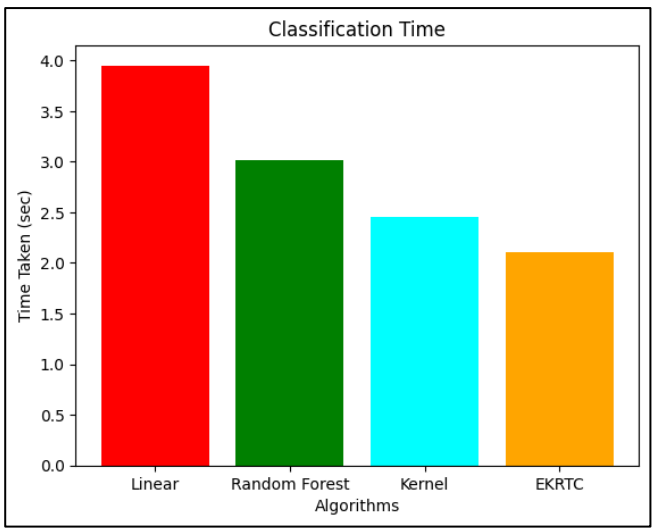

Fig 5: Classification Time

V. CONCLUSION

Categorizing information is a crucial aspect of text retrieval systems. This research work focuses on three essential steps: Pre-processing, Searching, and Classification. The initial step in automatic information categorization involves pre-processing. Pre-processing is a text mining technique that simplifies complex text and enhances the information content for better interpretation by machines. In this study, the information is pre-processed using techniques such as stemming and stop word removal. In the searching phase, two existing search methods is implemented for searching the information, they are namely ,Boyer Moore Algorithm and Brute Force Algorithm and a new algorithm is proposed

namely, Naïve Bayes based Search algorithm. existing algorithms, Linear Regression, Random Forest Regression and Kernal Ridge Regression are employed and enhanced algorithm called Enhanced Kernal Ridge for Text Categorization (EKRTC) is used. Experimental results demonstrate that EKRTC achieves the highest accuracy.

VI. REFERENCES

[1]. Bhope, V. D., & Deshmukh, S. N. (2015). Information Retrieval using Pattern Deploying and Pattern Evolving Method for Text Mining. International Journal of Computer Science and Information Technologies, 6(4), 3625-3629.

[2]. Charjan, D. S., & Pund, M. A. (2013). Pattern Discovery For Text Mining Using Pattern Taxonomy. International Journal.

[3]. Charras, C., &Lecroq, T. (2004). Handbook of exact string-matching algorithms (pp. 1-17). King's College.

[4]. Cortes, C., Mohri, M., &Rostamizadeh, A. (2012). L2 regularization for learning kernels. arXiv preprint arXiv:1205.2653.

[5]. Gupta, V., &Lehal, G. S. (2009). A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), 60-76.

[6]. Janani, R., & Vijayarani, S. (2016). Text mining research: A survey. Int. J. Innov. Res. Comput. Commun. Eng, 4(4), 6564-6571.

[7]. Lecroq, T. (1992). A variation on the Boyer-Moore algorithm. Theoretical Computer Science, 92(1), 119-144.

[8]. Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

[9]. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

[10]. Segal, M. R. (2004). Machine learning benchmarks and random forest regression.

[11]. Sriurai, W. (2011). Improving text categorization by using a topic model. Advanced Computing, 2(6), 21.

[12]. Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3), 275-294.

[13]. Vovk, V. (2013). Kernel ridge regression. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, 105-116.

[14]. Zhang, K., Xu, H., Tang, J., & Li, J. (2006). Keyword extraction using support vector machine. In Advances in Web-Age Information Management: 7th International Conference, WAIM 2006, Hong Kong, China, June 17-19, 2006. Proceedings 7 (pp. 85-96). Springer Berlin Heidelberg.