www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2534-2539

e-ISSN : 2583-1062
Impact Factor : 7.001

COMPARATIVE ANALYSIS OF TEXT CLASSIFICATION PERFORMANCE: EVALUATING ARTIFICIAL NEURAL NETWORKS, NAIVE BAYES, AND SUPPORT VECTOR MACHINES

Sakarapalli Jayalakshmi[1]

[1]GMR Institute of Technology

ABSTRACT

Text classification is one of the most important tasks of machine learning the appropriate choice of algorithm can make all the difference. Artificial Neural Networks, Support Vector Machine, and Naïve Bayes have been tested to prove which one works best for text classification. ANN simulates the structure of a natural neural network and thus is robust across a wide array of classification tasks. Due to its ease and performance, Naïve Bayes is still very popular, especially in text categorization, while SVM is popular due to its high accuracy in binary and multi-class classifications. Comparison emphasizes, the emphasis will be on evaluating these algorithms using the f1-macro metric for single-label and multi-label datasets. ANN has competitive results in single-label classification, staying close to the accuracy performance of SVM. In the more complex tasks of multi-label classification, ANN outperformed both SVM and Naive Bayes with a much simpler network architecture. The findings thus underline the strong points and weaknesses of every algorithm, spelling important lessons on their appropriateness in different scenarios for text classification. ANNs' ability to handle these complex, multi-label tasks definitely makes them a potential versatile tool in machine learning and gives SVM and Naïve Bayes strong alternatives depending on the classification context.

Keywords: Text Classification, Artificial Neural Networks (ANN), Naïve Bayes, Support Vector Machines (SVM), Classification Algorithm.

1. INTRODUCTION

Text classification is the backbone of machine learning, and the basis for sorting, analyzing, and giving meaning to a great deal of textual information which may be semi- or unstructured. In this world of ever-growing digitalization, text automatic categorization into predefined classes has reached importance in several related applications spanning from spam control, document organization, sentiment analysis, and information retrieval. This is achieved by the fact that text classification success is determined by the selected algorithms, and each of them bears its own strengths and weaknesses. The most commonly implemented algorithms for text classification are Artificial Neural Networks, Support Vector Machine, and Naïve Bayes. ANN is among the most prominent ones because it successfully implements the idea of modeling intricate patterns and interrelations in data just the way a human brain works. Naïve Bayes, being of simple form and computationally efficient, found a tremendous number of applications in text classification where the issues of interpretability and speed are among the priorities. Support Vector Machines, on the contrary, having solid theoretic basics and quite sophisticated functionality in high-dimensional space, proves to be pretty effective in reaching high accuracy for binary and multi-class classification problems. This paper surveys and compares these three algorithms on the performance of classification tasks for single- and multi-label problems, respectively. It should give a worthy insight into the choice of the most appropriate one under specific requirements and constraints through comparison of their strengths, weaknesses, and appropriateness for different text classification challenges. We also discuss how each algorithm performs in various contexts, pays attention to the increasing importance of multilabeling and classification of big data.

2. RELATED WORK

This paper compares ANNs with Naïve Bayes and SVMs for single-label and multi-label text classification. The methods are compared using key metrics such as accuracy, precision, recall, and F1-score. The study highlights the advantages of ANN over traditional techniques and throws light on its application in natural language processing and information retrieval. [1].

This paper evaluates several machine learning algorithms for document classification across various datasets. It describes training, testing, and performance metrics such as accuracy, precision, recall, and F1-score. The study identifies the best-performing models for specific tasks, aiding researchers in selecting suitable methods based on their data requirements [2].

The paper suggests an improved approach to machine learning on text classification with a focus on accuracy, speed, and efficiency. The paper tests various algorithms, and the results provide insights into applications in information retrieval and automated document management. [3].

The research compares Naïve Bayes, SVM, and ANN over text classification based on accuracy, precision, recall, and F1-score. SVM gave a better performance than both Naïve Bayes and ANN, especially on the bigger datasets, in terms of the accuracy of classification. [4].

This paper evaluates the Naïve Bayes classifier for classifying Arabic short texts. Despite the complexity of the Arabic morphology and sparsity, Naïve Bayes achieved acceptable accuracy, and the efficiency of Naïve Bayes in the Arabic text classification task is proven by its combination with feature selection techniques. [5].

Performance evaluation of Naïve Bayes for Arabic short text classification: although the Arab morphology and sparsity of feature are problematic, Naïve Bayes achieved fair accuracy, showing its competence, especially if the feature selection techniques are used. [6].

The research compares SVM, Naïve Bayes, and Neural Networks for the purpose of categorizing Arabic text. Due to the richness in morphology and sparse data representation, classification is challenging. In this regard, SVM was found to outperform Naïve Bayes and Neural Networks in terms of accuracy for the task of Arabic text categorization. [7].

It combines Naïve Bayes vectorization with SVM for an improvement in multi-class document classification. The hybrid model allows a better extraction of features in comparison to the traditional methods and, therefore, a more accurate and performing output [8].

The next research presentation is ANNs applied to text classification. ANNs learn very well for high-dimensional data and vast datasets with complex patterns. This model obtained competitive accuracy compared to traditional classifiers, indicating that ANNs are suited for text classification [9].

Here we are going to evaluate how Naïve Bayes performs with regard to the task of document classification. Naïve Bayes is computationally efficient and does very well on large datasets where features are independent. Though Naïve Bayes is a good baseline classifier, in more complex dependencies between features, its performance degrades [10].

The authors compare SVM, Naïve Bayes, and Decision Tree algorithms on performance in text document classification. It happens that SVM has the highest accuracy score at 89%, Naïve Bayes at 85%, and Decision Tree at 83%. The researchers applied a standard dataset of the text documents to evaluate how these machine learning techniques may perform in classifying them [11].

The study compares Naïve Bayes, SVM, and Deep Learning models. Deep learning models were able to attain the highest accuracy above 90%, while SVM achieved 85% and Naïve Bayes reached 80%. On the whole, deep learning models outperform SVM and Naïve Bayes in text classification tasks. [12].

Techniques Investigated. This paper investigates four: SVM, Naïve Bayes, Decision Trees, and Neural Networks. SVM: Known to work very accurately in high-dimensional space. Naïve Bayes: Currently works well over large datasets having independent features but generally is not as good as that of SVM. Neural Networks: Allow robust performance with complex patterns but consume more computational resources [13].

Comparative studies of SVM, Naïve Bayes, and Decision Tree for the purpose of classification of English texts. Feature selection also acts as an aid to enhanced efficiency and accuracy in terms of computation. Optimized models compare well in relation to both accuracy and computational speed in comparison with existing methods [14].

It reviews several supervised learning techniques used for automatic text classification: SVM, Naïve Bayes, and Neural Networks. Techniques are compared on accuracy, computational efficiency, and scalability across different text classification tasks. The SVM and Neural Networks are highlighted to have strong performance, with the Naïve Bayes being efficient but relatively less effective on large datasets [15].

3. METHODOLOGY

The paper explores text classification using ANN and compares its performance with Naïve Bayes and SVM models.

1. Data Preprocessing:

Data Cleaning: As a first step, the dataset needs to be cleaned for an analysis.

This involves:

- Case Folding: All text is converted to lower case for uniformity.
- Example: “The dog is jumping over the fence” → “the dog is leaping over the wall.”
- Stopword Removal: Removing common words that carry no meaning. (Examples of stopwords are from, and, too, is, and).
- Example: “The dog is jumping over the fence.” → “dog jumping fence.”
- Tokenization: Division of the text into individual words or tokens.
- Example:
- "The cat runs." → ["The", "cat", "runs"]

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2534-2539
e-ISSN : 2583-1062
Impact Factor : 7.001

- Stemming: reducing words into their base form or its root.

Example: “The dog is jumping over the fence.” → “the dog is jump over the fenc.”

The datasets that the authors in the document "Single-Label and Multi-Label Text Classification using ANN and Comparison with Naïve Bayes and SVM" by Mahfi et al. (2023) have employed are:

1.20 Newsgroups Dataset (Single-label)

2. BBC Full-Text Dataset (Single-label)

3. Elsevier OA CC-BY Corpus (Multi-label)

2. Feature Extraction:

Transforming Text Data: Cleaned text data is converted into numerical form for processing by machine learning algorithms. This step allows models to understand the text. TF-IDF assigns weighted values to terms based on relevance, highlighting important terms. However, in large vocabularies, high dimensionality can increase computational costs.

3. Deployed Models:

Artificial Neural Network (ANN): This contains dense and batch normalization layers and has been trained on single-label as well as multi-label datasets.

Naïve Bayes: A probabilistic model assuming feature independence, implemented using the Scikit-learn library.

SVM: SVM is one of the popular binary and multi-class classifiers, which is basically separated by the hyperplane.

4. Evaluation:

F1-Score: Measures precision and recall simultaneously, derived from the harmonic mean of precision and recall, highlighting its importance in the precision-recall F1 framework.

F1-Macro: Average F1 scores across classes, which is useful for multi-class or multi-label classification models.

Cross-validation: In K-fold cross-validation, it takes one-fold for training, tests on another fold and validates with the remaining one.

5. System Flow:

The methodology is shown in a flowchart

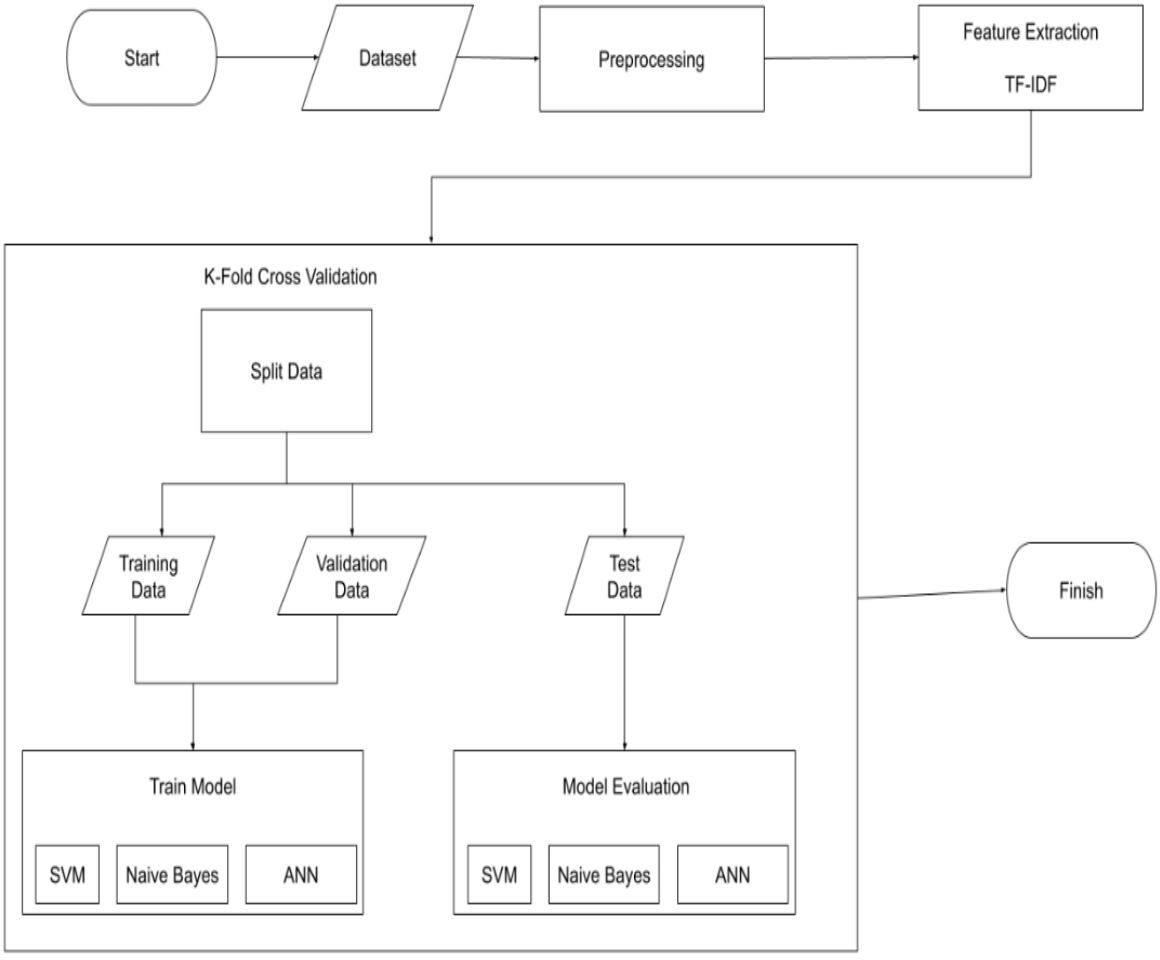


Fig 1: Framework for Machine Learning-Based Text Classification

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2534-2539

e-ISSN : 2583-1062
Impact Factor : 7.001

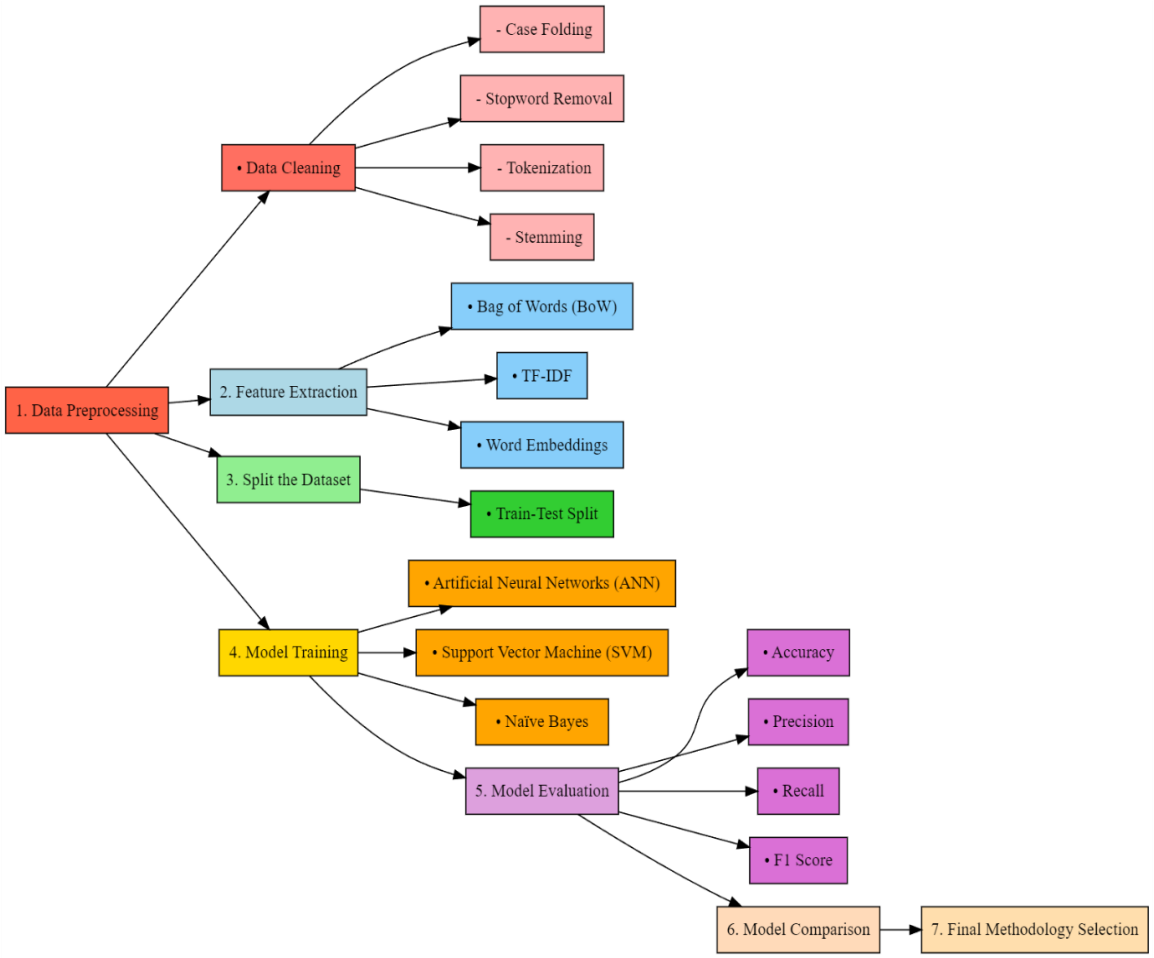


Fig 1: Workflow for Text Classification Methodology

COMPARISON

SVM shows good accuracy and F1-score, particularly in binary and single-label classification, while not performing so well with unbalanced datasets. Naïve Bayes outperforms when the datasets are smaller, but simple datasets; the precision and recall are not as good as it is in complex datasets. ANN works best in multi-label classification, showing higher recall, precision, and F1-score due to its capability to capture intricate patterns, but it consumes more computations and needs huge datasets along with fine-tuning. In general, ANN outperforms Naïve Bayes and SVM in complex tasks, SVM performs well for high-dimensional datasets, and Naïve Bayes works pretty well as a baseline for simpler problems.

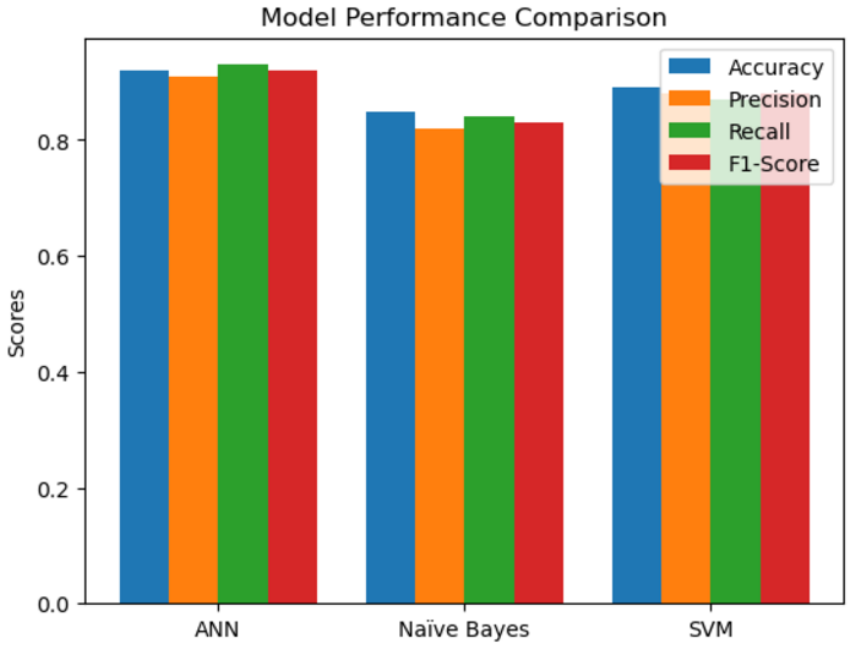
COMPARISION TABLE:

Model	Dataset	Precision	Recall	F1-Macro Score
ANN	20 Newsgroups (Single)	0.79	0.78	0.79
Naïve Bayes	20 Newsgroups (Single)	0.8	0.78	0.78
SVM	20 Newsgroups (Single)	0.83	0.82	0.82
ANN	BBC (Single)	0.96	0.96	0.96
Naïve Bayes	BBC (Single)	0.97	0.97	0.97
SVM	BBC (Single)	0.97	0.97	0.97
ANN	Elsevier OA CC-BY (Multi)	0.55	0.42	0.48
Naïve Bayes	Elsevier OA CC-BY (Multi)	0.49	0.29	0.34
SVM	Elsevier OA CC-BY (Multi)	0.33	0.72	0.44

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2534-2539

e-ISSN : 2583-1062
Impact Factor : 7.001



4. RESULT

The results of comparison of SVM, Naïve Bayes, and ANN are given below.

Single-Label Classification:

• F1-macro for SVM was 0.82 on the 20 Newsgroups while ANN is close at 0.79, and Naïve Bayes was at 0.78.

• On the BBC dataset, Naïve Bayes and SVM are at 0.97 while ANN is at 0.96.

Multi-Label Classification:

• On the Elsevier OA CC-BY dataset, ANN has outperformed SVM at 0.44 and Naïve Bayes at 0.34 with F1-macro at 0.48.

• SVM had high recall of 0.72 but very poor precision of 0.33, while Naïve Bayes and ANN showed higher precision and recall.

5. CONCLUSION

This research provides a structured approach for the application of ANN in text classification, from preprocessing to feature extraction, model training, and evaluation. A comparison with Naïve Bayes and SVM reveals that ANN can be applied to both single-label and multi-label tasks. ANN performs well on multi-class, multi-label classification, where it maps complex patterns, but is computationally expensive. SVM performs well on single-class classification, especially on high-dimensional datasets, making clear class boundaries. Naïve Bayes performs well for the less complex, single-class problem with a lower computational cost, but fails to perform in the multi-class setting due to interdependence between features. ANN is preferred for more complex multi-label tasks, and SVM or Naïve Bayes are preferred for single-label tasks. The future research might be the efficiency improvement of ANN in the multi-label task and balancing precision-recall trade-off in SVM for complex case.

6. REFERENCES

[1] Mahfi, M., Karsana, N., Muslim, K., & Astuti, W. (2023). Single-Label and Multi-Label Text Classification using ANN and Comparison with Naïve Bayes and SVM. JURNAL MEDIA INFORMATIKA BUDIDARMA.

[2] Ali, S. I. M., Nihad, M., Sharaf, H. M., & Farouk, H. (2024). Machine Learning for Text Document Classification-Efficient Classification Approach. IAES International Journal of Artificial Intelligence (IJ-AI), 13, 703-10.

[3] Jain, R. (2023). Comparative Study of Machine Learning Algorithms for Document Classification. International Journal of Computer Sciences and Engineering IJCSE.

[4] Kumar, R. R., Reddy, M. B., & Praveen, P. (2019). Text classification performance analysis on machine learning. International Journal of Advanced Science and Technology, 28(20), 691-697.

[5] Ibrahim, M. F., Alhakeem, M. A., & Fadhil, N. A. (2021). Evaluation of Naïve Bayes classification in Arabic short text classification. Al-Mustansiriyah Journal of Science, 32(4), 42-50.

[6] Sarkar, A., Chatterjee, S., Das, W., & Datta, D. (2015). Text classification using support vector machine. International Journal of Engineering Science Invention, 4(11), 33-37.

[7] Mohammad, A. H., Alwada 'n, T., & Al-Momani, O. (2016). Arabic text categorization using support vector machine, Naïve Bayes and neural network. GSTF Journal on Computing (JoC), 5, 1-8.

[8] Sueno, H. T., Gerardo, B. D., & Medina, R. P. (2020). Multi-class document classification using support vector machine (SVM) based on improved Naïve bayes vectorization technique. International Journal of Advanced Trends in Computer Science and Engineering, 9(3).

[9] Prasanna, P. L., & Rao, D. R. (2018). Text classification using artificial neural networks. International Journal of Engineering & Technology, 7(1.1), 603-606.

[10] Ting, S. L., Ip, W. H., & Tsang, A. H. (2011). Is Naive Bayes a good classifier for document classification. International Journal of Software Engineering and Its Applications, 5(3), 37-46.

[11] Ting, S. L., Ip, W. H., & Tsang, A. H. (2011). Is Naive Bayes a good classifier for document classification. International Journal of Software Engineering and Its Applications, 5(3), 37-46.

[12] Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics of machine learning-based algorithms for text claTing, S. L., Ip, W. H., & Tsang, A. H. (2011). Is Naive Bayes a good classifier for document classification. International Journal of Software Engineering and Its Applications, 5(3), 37-46ssification. Sustainable Operations and Computers, 3, 238-248.

[13] Chavan, G. S., Manjare, S., Hegde, P., & Sankhe, A. (2014). A survey of various machine learning techniques for text classification. International Journal of Engineering Trends and Technology, 15(6), 288-292.

[14] Luo, X. (2021). Efficient English text classification using selected machine learning techniques. Alexandria Engineering Journal, 60(3), 3401-3409.

[15] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. Artificial intelligence review, 52(1), 273-292