ADROID Conference Series: Engineering Sciences and Technology
Volume 2, pp.73-78, 2026

A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR LEGAL TEXT CLASSIFICATION

Aman Kumar, Hridoy Roy, Togon Chakma, Kanderp Narayan Mishra

Department of Computer Science and Engineering, Sharda School of Computer Science and Engineering, Sharda University, Greater Noida, Uttar Pradesh

amankrbup04@gmail.com, hridoyr75@gmail.com, togonchakma100@gmail.com, kanderp.mishra@sharda.ac.in

Corresponding author: amankrbup04@gmail.com

Abstract. This paper provides a comparative analysis of machine learning algorithms for text classification in law. The swelling of legal documents and the growing case backlog have led to a demand for automated systems to aid efficient legal analysis. In this article, text data is first preprocessed and converted to numerical features using TF-IDF. Naive Bayes, Support Vector Machine, and Logistic Regression are three machine learning algorithms that are applied and tested on a real-life legal dataset. Evaluation metrics for such models include accuracy, precision, recall, and F1-score. The experimental results indicate that the Support Vector Machine model achieves the highest accuracy among the considered approaches. The results not only confirm that machine learning methods can be applied to large-scale legal text classification but also serve as a reference point for improving the methods in the future through more advanced models.

Keywords: Legal Text Classification, Machine Learning, TF-IDF, Naive Bayes, Logistic Regression, Support Vector Machine, Natural Language Processing.

1. INTRODUCTION

The acceleration in the number of legal papers and the growing number of cases awaiting hearing have become the key issues in contemporary judicial practices. Professionals in law have to process extensive amounts of textual data, which can be time-consuming and is usually subject to human error. Such an unstructured legal text, manually processed, not only slows down the process of making a decision, but also lowers legal proceeding efficiency. As natural language processing (NLP) and machine learning approaches have developed, text analysis and classification could be automated. These technologies can facilitate the effective management of massive documents and help perform such actions as document classification, information search, and text processing, which also means that fewer legal experts will be required. This research aims to conduct a comparative analysis of various machine learning algorithms in text law classification. This paper preprocesses textual data and transforms it into numerical features through TF-IDF and runs three machine learning models, including Naive Bayes, Logistic Regression, and Support Vector Machine.
The contributions of the paper are as follows:

1. Application of various machine learning algorithms to text classification in law.
2. Comparative analysis of models based on performance measures, i.e., accuracy, precision, recall, and F1- score.
3. Performance of the models on a realistic legal dataset.

Nonetheless, in spite of having multiple machine learning methods at disposal, it is still quite difficult to choose the most adequate model to classify legal texts because legal language is complex and will differ depending on the case. This paper will be structured as follows: the problem statement comes in Section II, the related work in Section III, the methodology in Section IV, the algorithms in the experimental results in Section VI, and the conclusion of the study in Section VII, where limitations and future work are mentioned.

2. PROBLEM STATEMENT

The rising number of legal papers and backlog cases have posed a significant challenge to the judicial system. Lawyers will need to manually examine and categorise a large volume of textual data, which is both time-intensive and error-prone. This paperwork system normally means that there is a delay in the court processes, and the system is not as efficient.

Besides, legal texts are generally unstructured, complex, and thus hard to process using conventional means. The diversity of language, terms, and circumstances makes the correct categorisation of legal documents difficult. Despite the many machine learning methods proposed for text recognition, determining which algorithm to use for legal text recognition remains a challenge. Thus, a procedural and comparative approach to assessing various machine learning models for effective classification of legal texts is required.

3. LITERATURE REVIEW

Table 1: Literature Review of the available Legal Text Classification Methods

S. No	Reference	Contributions	Method/Model	Dataset Used	Limitations
1	[1]	Automated legal document classification system	Naive Bayes	Legal case dataset	Poor performance with complex texts
2	[2]	Comparative study of ML models	Logistic Regression, SVM	Court judgments dataset	Limited feature engineering
3	[3]	Text classification using NLP Methods	TF-IDF + Naive Bayes	Indian legal dataset	Weak contextual sensitivity
4	[4]	Deep learning for legal texts	LSTM	Legal texts collection	Expensive computation
5	[5]	Transformer-based classification	BERT	Large-scale legal dataset	It requires the use of GPU resources
6	[6]	ML-based prediction of legal cases	SVM	Case law dataset	Small-sized dataset
7	[7]	Feature extraction improvement	TF-IDF, N-grams	Legal text dataset	Ignored the semantic meaning
8	[8]	Classification using ensemble methods	Random Forest	Legal dataset	Problem of Overfitting
9	[9]	Comparative NLP approaches	NB, LR	Text dataset	Poor performance on large data
10	[10]	Legal document analysis	CNN	Legal corpus	Complex model tuning
11	[11]	ML for document categorisation	Logistic Regression	Mixed legal dataset	Poor generalization
12	[12]	Hybrid NLP approach	TF-IDF + SVM	Court case dataset	Issue related to Feature dependency

Based on the literature, it can be seen that there are multiple machine learning and deep learning methods that have been used in legal text classification. Nevertheless, a number of studies are associated with limitations, including high computational cost, small sample sizes, and the absence of comparative analysis. Thus, this paper aims to compare various machine learning models to define the most successful one in legal text classification, as per Table 1.

3.1 Dataset Description

The study's sample consists of legal cases gathered by the investigator from publicly available sources. It contains textual data from legal cases, which is used for classification. The dataset is real-world legal data, which is appropriate for the assessment of machine learning models. The dataset consists of several records; each record contains the text of a legal case and its outcome label. Unstructured textual data is in text form and needs to be preprocessed before machine learning.
Table 2 presents the key characteristics of the data:

Table 2: Dataset Description

Feature Name	Description
case_text	Textual content of the legal case document
case_ outcome	The name of the classification label of the case outcome

The dataset is preprocessed prior to implementing machine learning models to eliminate noise, handle missing data, and standardise text data. This ensures that the data entered is appropriate for feature extraction and model training.

4. METHODOLOGY

The present research is systematic in its approach to law text classification using machine learning. The overall workflow consists of data preprocessing, feature extraction, model training, and evaluation. First, information on legal cases, in the form of documents, is gathered and prepared for analysis. The data is unstructured, so pre-processing is performed to clean the text. This involves the transformation of text to lower case, the use of special characters and the elimination of unnecessary spaces that create noise and enhance the quality of data. After preprocessing, feature extraction is performed using the TF-IDF technique, which converts textual data into numerical vectors. This enables machine learning models to process and learn from the textual information successfully. The processed data can then be split into training and test sets at 80:20. The models are developed using the training data, and their performance is tested on the test data. Classification in three machine learning algorithms, which include Naive Bayes, Logistic Regression, and Support Vector Machine, has been implemented. These models are trained using the extracted features to acquire patterns of legal text data. Lastly, the models are measured based on the metrics of accuracy, precision, recall and F1-score. The results of the classification are also analysed in detail by a confusion matrix, as per Fig. 1

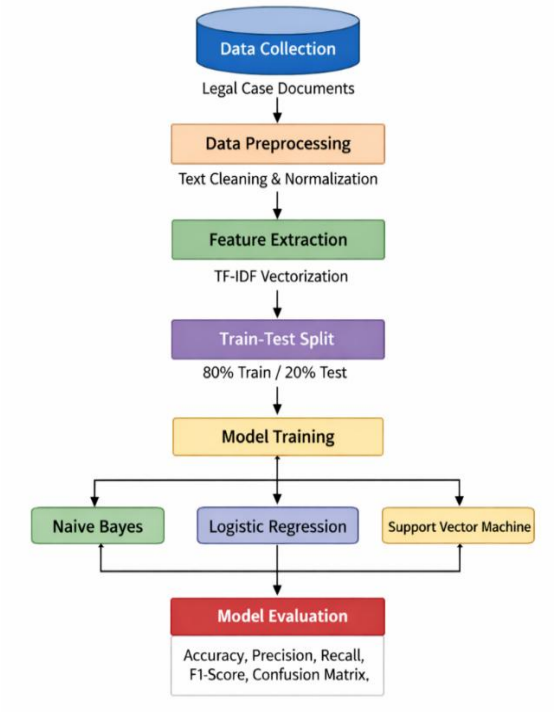

Fig 1 Recommended Approach to Classifying Text in Law.

5. ALGORITHMS USED

A study is conducted on three machine learning algorithms, namely Naive Bayes, Logistic Regression, and Support Vector Machine, in legal text classification. Naive Bayes is a probabilistic classification algorithm that relies on the Bayes theorem. It further assumes that the features are independent and computes the likelihood of each of the classes based on the input text. It is simple and efficient; thus, it can be used on small datasets, but it can also fail to identify complex relationships between words. Logistic Regression is a linear classification method based on the probability of a class label with the help of a sigmoidal form of prediction. It is applicable to numerical characteristics like TF-IDF and yields results that can be interpreted. Non-linear data patterns may, however, be a problem for it. Support Vector Machine (SVM) is an effective classification algorithm that divides data with a hyperplane that has the greatest margin. It works well, especially in high-dimensional data like text, and in many cases, it offers better accuracy than other traditional models. Although it has high performance, it uses more computational resources and training time. Such algorithms are trained on the extracted features and tested to assess how well they classify legal texts.

6. RESULTS AND DISCUSSIONS

This research paper tested the three machine learning models, namely, Naive Bayes, Logistic Regression, and Support Vector Machine, to classify legal text. Table 3 shows the accuracy of every model.

Table 3: Model Performance Comparison of the models

Model	**Accuracy**
Naive Bayes	0.498
Logistic Regression	0.531
Support Vector Machine	0.540

The findings indicate that the Support Vector Machine model is the most accurate of the three models, followed by the Logistic Regression and the Naive Bayes. The comparison of model accuracies is in the form of a graph, as shown in Figures 2 and 3. Besides accuracy, the accuracy of the models was measured by precision, recall, and F1-score because I was interested in the overall analysis of the models. The classification results were also examined using a confusion matrix, which was used to determine the misclassifications. Based on the findings, it can be seen that the Support Vector Machine is more effective as it can operate effectively in high-dimensional feature spaces, and therefore can be used in the task of text classification. The performance of Logistic Regression is also competitive, and Naive Bayes is rather low because of the fact that it assumes that features are independent, and this assumption is not always true with legal texts, which are complex. Generally, the comparative analysis indicates that the Support Vector Machine is the best model to be used in legal text classification in this case study.

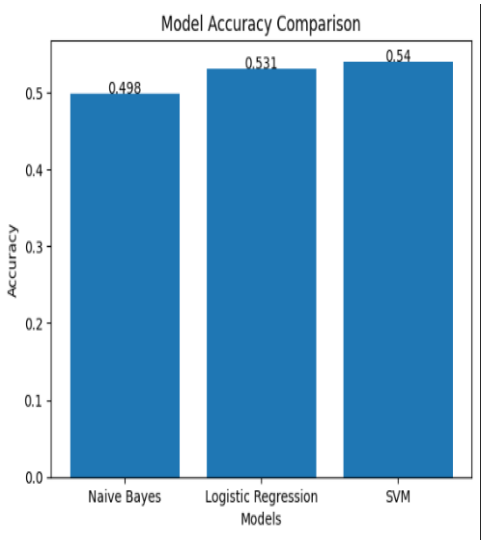

Fig 2: Comparison of the Accuracy of ML Models

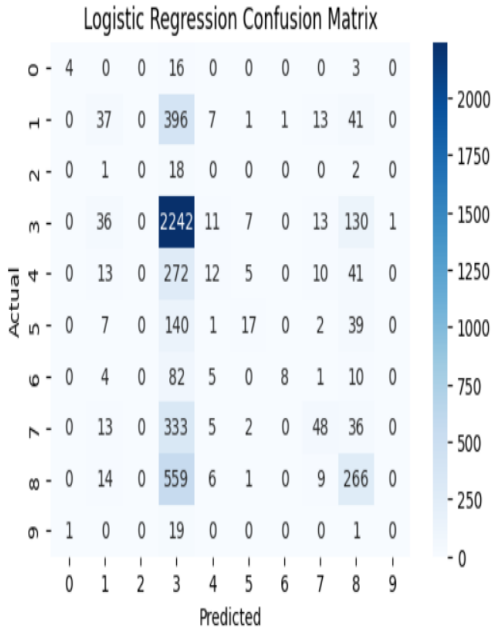

Fig 3: Confusion Matrix of Logistic Regression Model

7. CONCLUSION AND FUTURE WORK

This paper compared machine learning algorithms in legal text classification. The TF-IDF was used to preprocess the dataset and convert it to numerical features. Naive Bayes, Logistic Regression, and Support Vector Machine were three machine learning models implemented and tested. The findings reveal that the Support Vector Machine was the most successful model in this task, achieving the highest accuracy. Logistic Regression also showed competitive performance, whereas Naive Bayes exhibited a rather low level of accuracy given its simplifying assumptions. Even though it has obtained a satisfactory performance, this research has limitations. The sample is not very large and might not be representative of the variety of actual legal texts. Moreover, only simple machine learning models were considered, and the overall accuracy is only satisfactory.
In future work, it is possible to consider more sophisticated methods, such as deep learning and transformer models like BERT, to enhance performance. In addition, the models' generalisation ability can be improved by using larger, more varied datasets.

Conflict of Interest

The authors declare no conflicts of interest regarding the current research.

References

[1] A. Dhar, H. Mukherjee, N. S. Dash, and K. Roy, “Text categorisation: Past and present,” *Artificial Intelligence Review*, vol. 54, no. 4, pp. 3007–3054, 2021.
[2] L. Wan, G. Papageorgiou, M. Seddon, and M. Bernardoni, “Long-length legal document classification,” *arXiv preprint arXiv:1912.06905*, 2019.
[3] J. S. T. Howe, L. H. Khang, and I. E. Chai, “Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments,” *arXiv preprint arXiv:1904.06470*, 2019.
[4] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” *Information*, vol. 10, no. 4, p. 150, 2019.
[5] S. Xu, “Bayesian Naïve Bayes classifiers to text classification,” *Journal of Information Science*, vol. 44, no. 1, pp. 48–59, 2018.
[6] Z. Li, “A classification retrieval approach for English legal texts,” in *Proc. 2019 Int. Conf. Intelligent Transportation, Big Data & Smart City (ICITBS)*, pp. 220–223, 2019.
[7] R. S. Wagh and D. Anand, “A novel approach of augmenting training data for legal text segmentation by leveraging domain knowledge,” in *Intelligent Systems, Technologies and Applications*, Singapore: Springer, pp. 53–63, 2020.

[8] F. A. Braz, N. C. da Silva, T. E. de Campos, F. B. S. Chaves, M. H. Ferreira, P. H. Inazawa, and F. H. Peixoto, “Document classification using a Bi-LSTM to unclog Brazil's Supreme Court,” *arXiv preprint arXiv:1811.11569*, 2018.
[9] N. C. da Silva *et al.*, “Document type classification for Brazil’s Supreme Court using a convolutional neural network,” in *Proc. 10th Int. Conf. Forensic Computer Science and Cyber Law (ICoFCS)*, Sao Paulo, Brazil, pp. 29–30, 2018.
[10] A. Iftikhar, S. W. U. Q. Jaffry, and M. K. Malik, “Information mining from criminal judgments of Lahore High Court,” *IEEE Access*, vol. 7, pp. 59539–59547, 2019.
[11] L. Zhang and D. Moldovan, “Chinese relation classification using long short-term memory networks,” in *Proc. 11th Int. Conf. Language Resources and Evaluation (LREC)*, 2018.
[12] Janani, R., &Vijayarani, S. (2021). Automatic text classification using machine learning and optimisation algorithms. Soft Computing, 25(2), 1129-1145.