LEGAL DOCUMENT CLASSIFICATION (JUDGMENTS, PETITIONS, ORDERS)

# 

**Table of Contents**

[1. Introduction 3](#_Toc203502183)

[2. Problem statement 3](#_Toc203502184)

[3. Research question 3](#_Toc203502185)

[4. Methodology 3](#_Toc203502186)

[5. Results and findings 4](#_Toc203502187)

[6. Conclusion 14](#_Toc203502188)

[References 16](#_Toc203502189)

# 1. Introduction

Classification of legal documents is an important procedure in handling and recovering a large amount of text related to cases in the judicial system. The high rates of increasing digital legal documentation require automated methods that ensure increased efficiency as well as consistency. The analysis demonstrates the use of a conventional machine learning process to classify documents, including judgments, orders, and petitions, as per case determinants. TF-IDF vectorization is used in converting text content into numerical forms, facilitating good feature extraction. The models, such as Support Vector Machine and Random Forest, have been used to address the issue of complexity and variety of legal language, leading to correct classification and better access to legal information.

# 2. Problem statement

Judgments, fresh petitions, and orders are the unstructured textual data produced in large quantities by courts, law firms, as well as legal research platforms. These documents are manually classified, which is time-consuming and lacks consistency, and human error creates delays and inefficiencies in legal processes (Sil and Roy, 2021). The consistency of legal language, differences in length of the document, and terms present within the specific domain all compound the challenge of proper classification. The current retrieval methods frequently involve a keyword search or a rule-based classification that is unable to capture the contextual meaning and cannot scaleably handle an increasing dataset.

# 3. Research question

*Q: How accurately can traditional machine learning models classify legal documents into their respective outcome categories using TF-IDF-based text representations?*

# 4. Methodology

A corpus of thousands of legal texts, namely judgments, orders, and petitions, are experimented with. EDA is performed preliminarily to assess the distribution of document categories and observe text lengths, term frequencies, and the balance of the entire dataset (Chalkidis *et al.* 2021). This step determines the existence of data problems and pre-processing strategies. Text pre-processing is then used to apply undesirable text like punctuation symbols, special symbols, and stop words, and at the same time, leave out the meaningful legal terms (Chen *et al*. 2022). Lemmatization helps to normalize words and limit inflectional forms, which helps to increase the quality of text input consistency. After the pre-processing, the documents are transformed to a numerical form with the help of Term Frequency-Inverse Document Frequency (TF‑IDF). The inclusion of weights on the terms relates to the significance of the words within the whole text and allows such models to recognize meaningful patterns within the body of the law.

Two classical machine-learning algorithms are realized. The former is a linear Support Vector Machine (SVM) and forms a hyperplane that is used to classify documents into their classes based on the outcomes. The optimization of a model is done using grid search methods, which involve hyper parameter tuning. Accuracy, precision, recall, and F1-score are measured to give a descriptive measure of each model. A confusion matrix can be constructed to illustrate how well the classification is successful on discrete categories (Nie *et al.* 2022). There is also plotting of comparative visualizations, such as bar charts, to show the difference in model accuracy and other performance metrics. This approach to the methodological framework guarantees the robustness of classification results, their interpretability, and adherence to the peculiarities of legal text analysis.

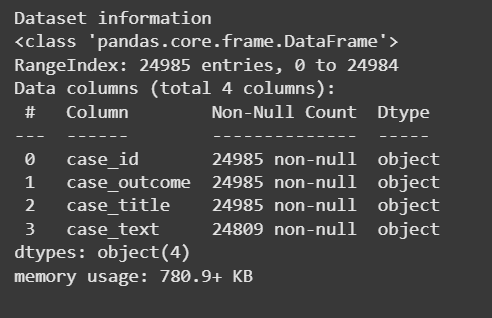
# 5. Results and findings

Chosen dataset contain information of legal documents that have case\_id, case\_outcome, case\_title, and case\_text columns. These documents contain various legal rules where the references are to cases like Black v Liponac [1998] FCA 699 and Colgate Palmolive Co v Cussons Pty Ltd (1993).



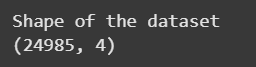
**Figure 1: Legal test classification dataset**

Document collection includes 24,985 records and 4 columns: case\_id, case\_outcome, case\_title, and case\_text. The value of 24,985 is non-null in all columns except the case\_text, which has 24,809 records, and this implies 176 missing text entries.



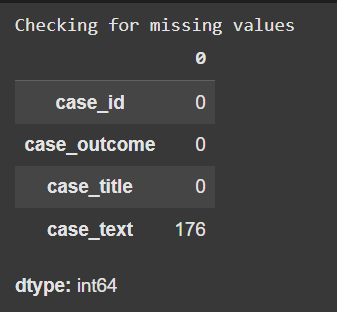
**Figure 2: Dataset information**

Following dataset contains object type data with 24985 entries and 4 columns.



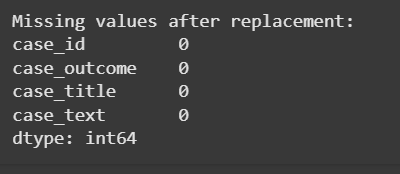
**Figure 3: Shape of the dataset**

Shape of the dataset indicates presence of 24,985 rows and 4 columns, proving the extensive volume of legal documents categorization.



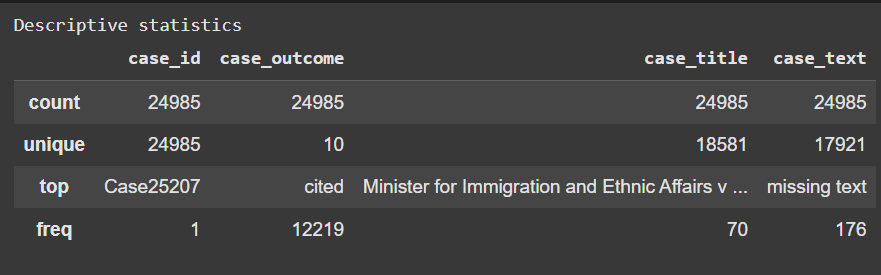
**Figure 4: Checking for missing values**

Missing value analysis shows that the case\_id, case\_outcome, and case\_title columns have no missing values, ensuring full identification and classification information. Nevertheless, in case\_text, there are 176 missing entries out of 24,985 records, which makes around 0.7 percent of the total data.



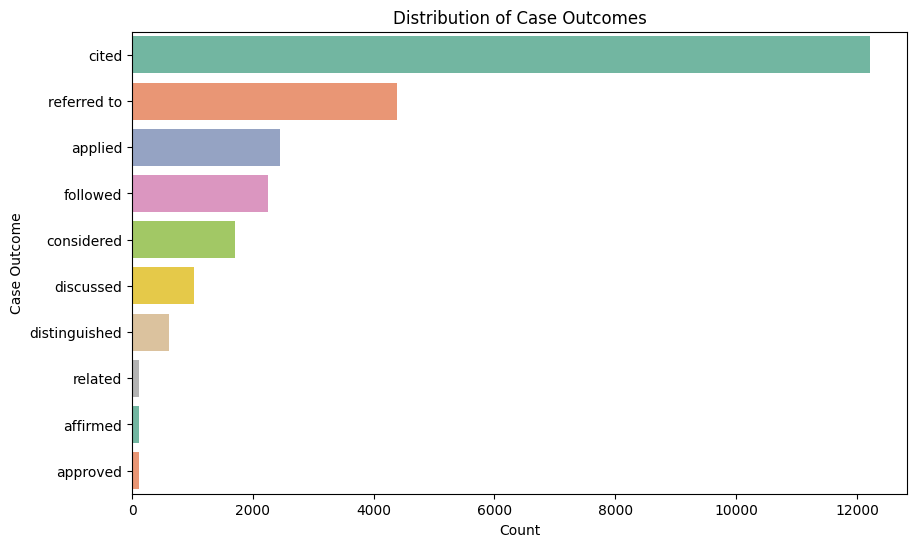
**Figure 5: Missing values after replacement**

Data cleaning processes was used, and all the missing data points are also cleaned now, where each column displays 0 missing data points. The replacement strategies have been performed on the case\_text column, which had 176 missing values.



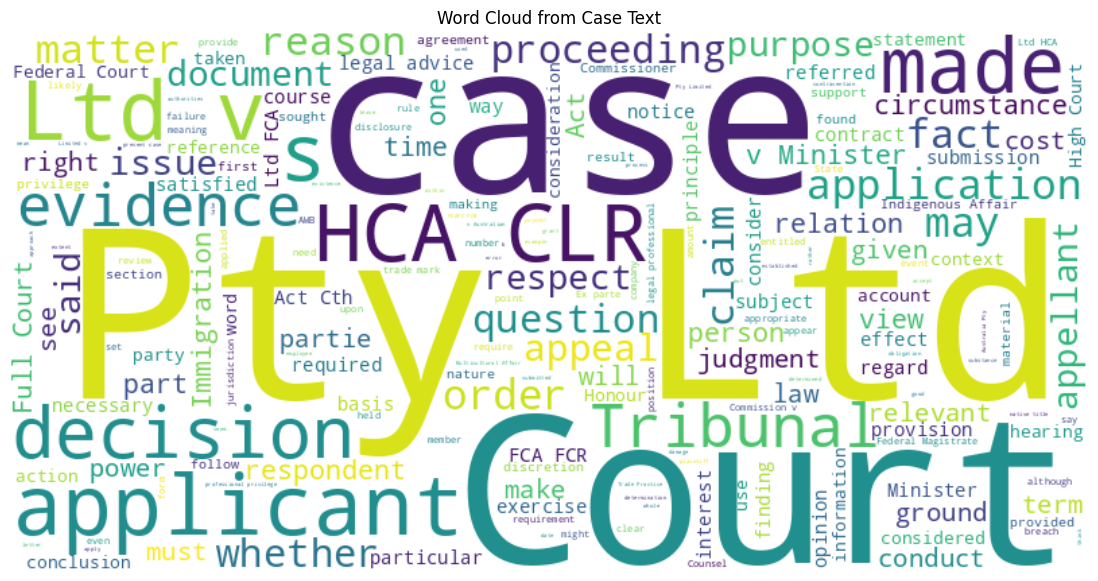
**Figure 6: Descriptive statistics**

Descriptive statistics indicate that there were 24,985 total entries, 10 different case outcomes, and 18581 different case titles. The most common result is the word cited used 12, 219 times, whereas Minister for Immigration and Ethnic Affairs is the most recurring title of the case.



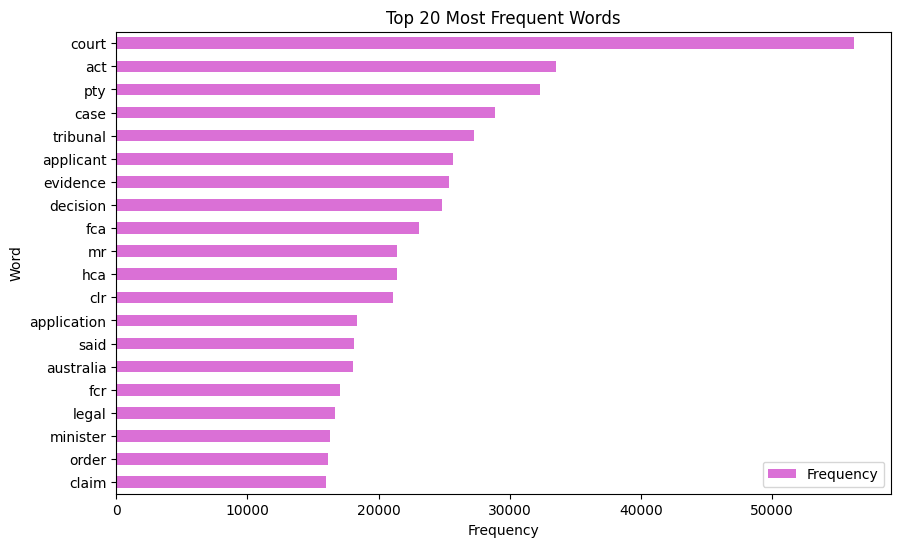
**Figure 7: Distribution of Case Outcomes**

Distribution of the case outcomes shows that there is an imbalance in the classes. The most common category according to citation is with statistics of about 12,000 cases, followed by those referred to as amounting to about 4000 cases.



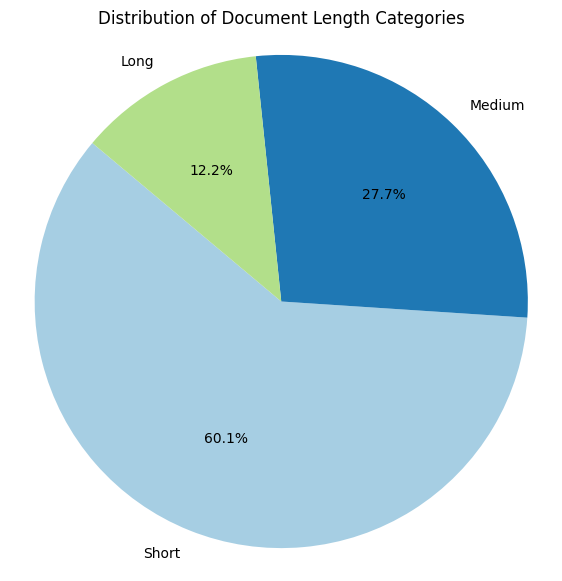
**Figure 8: Word Cloud from Case Text**

Word cloud representation shows the most significant terms in texts of the legal cases, and the words that feature as most significant terms include the use of the words case, court, made, applicant, decision, and tribunal. There are a lot of legal terms like evidence”, “HCA”, “FCLR”, “judgment”, “Minister”, “Ltd” that show the jargon and persons that are usually used in the legal proceedings of Australia.



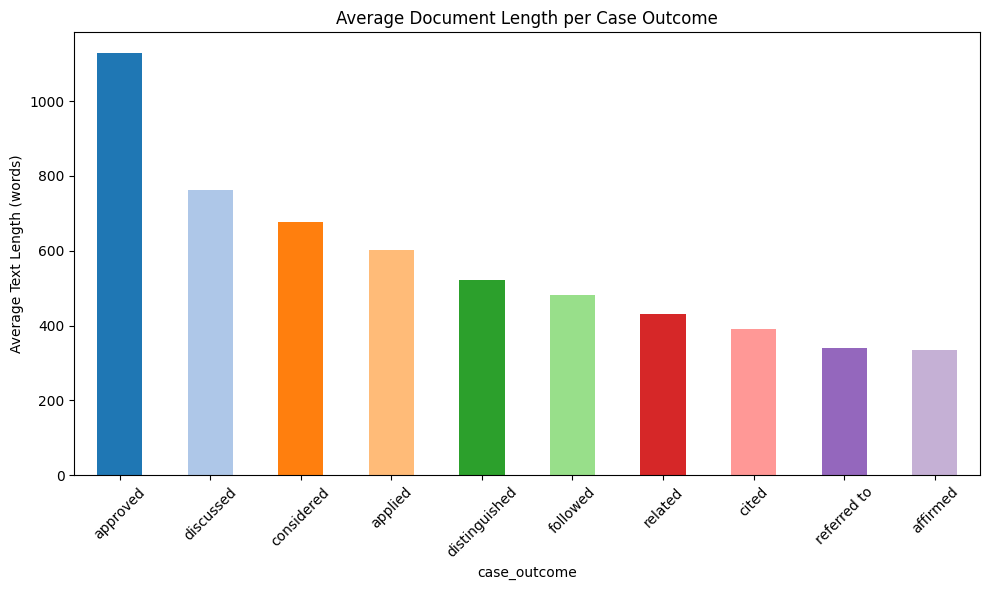
**Figure 9: Top 20 Most Frequent Words**

Bar plot in figure 9 shows the presence of fca, fcr, hca, clr and the use of application, the presence of words, such as said, australia, minister, order, and claim show that the dataset is indeed a legal document.



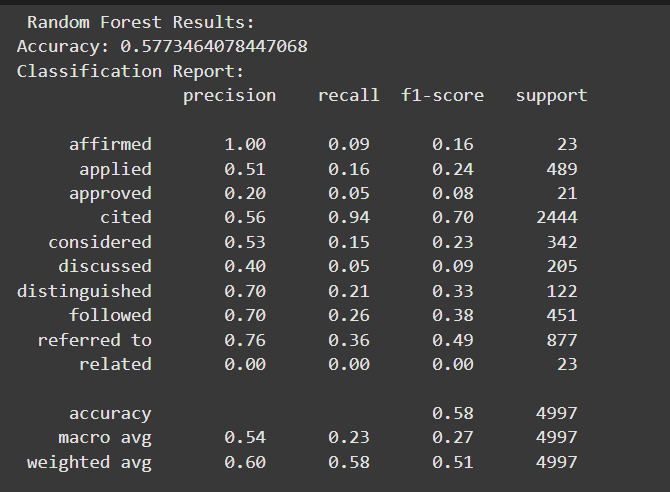
**Figure 10: Distribution of Document Length Categories**

Pie chart in figure 10 shows the length distribution of the documents constitutes 59.1 % documents of the size of short documents, 27.7 % documents of middle-sized documents, and 12.2 percent documents of the size of long documents.



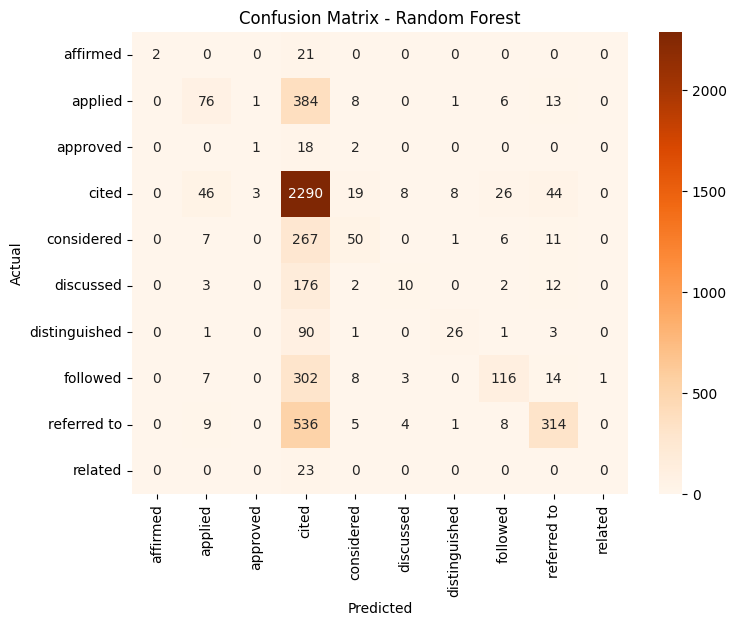
**Figure 11: Average Document Length per Case Outcome**

Following visualisation shows a significant difference between document length distributions, the longest document average length being observed in cases with the outcomes of approved case length at about 1200 words, followed by the length of a discussed case.



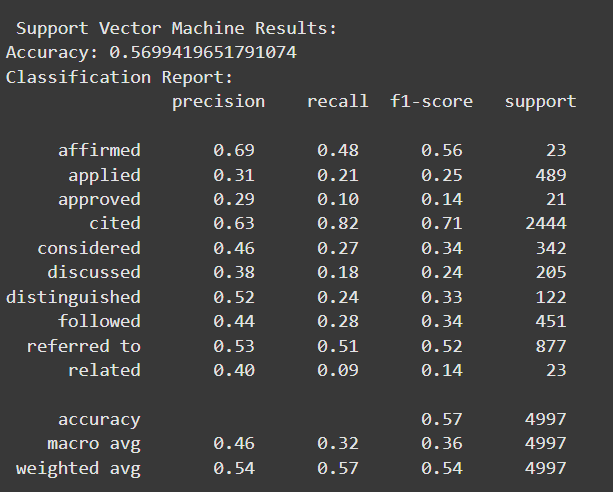
**Figure 12: Classification report for random forest**

Random Forest classifier had an overall accuracy of 57.73 percent and different performances in the categories. The macro average precision was 0.54, and the weighted average F1-score is 0.51, which describes a moderate level of classification in terms of the unbalanced data distribution.



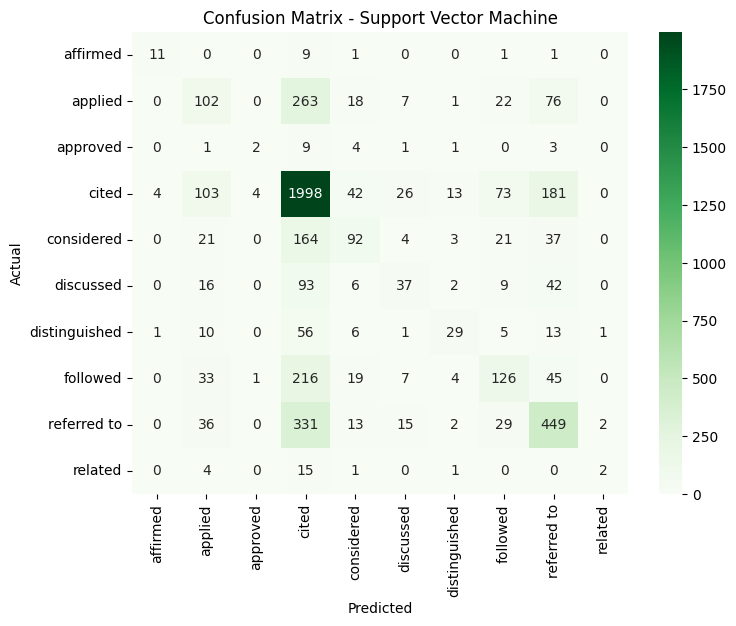
**Figure 13: Confusion matrix for random forest**

Confusion matrix in figure 13 demonstrates the classification trends of the Random Forest, where 2229 cases were determined as correct, whereas a major number of cases are misclassified in the other categories. True positive values are found in the diagonal elements, and the values outside the diagonal show the error of the classification.



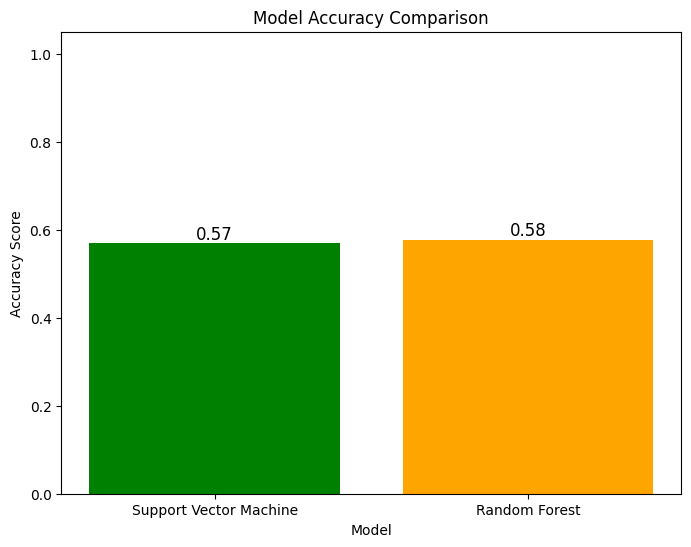
**Figure 14: Classification report for Support Vector Machine**

Support Vector Machine got a total of 56.59 percent accuracy with unique performance features. The precision in the “Affirmed” was 0.69, compared to 0.63 precision and 0.82 recall of the “Cited”. The macro mean precision stood at 0.46 with a weighted average F1-score of 0.54.



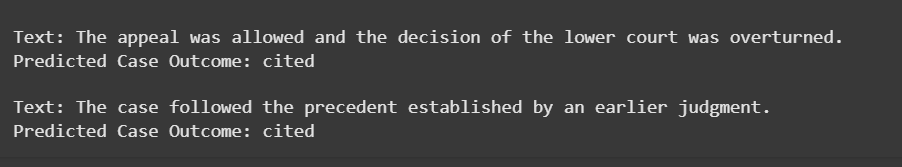
**Figure 15: Confusion matrix for Support Vector Machine**

SVM confusion matrix reveals the classification performance of all the outcome classes, and in this case, the cited category has had the correct prediction of 1998 out of 2444 cases. The significant misclassifications were the case of applied (102 correct of 478) and cross-category errors that involved similar outcomes.



**Figure 16: Model accuracy comparison**

Comparative analysis reveals that Random Forest obtained an accuracy of 58% as compared to 57 percent accuracy obtained by Support Vector Machine, creating a minimal 1 percent performance variation. The two traditional supervised machine learning models indicate competitive performance in the classification of legal documents based on TF-IDF vectorization.

****

**Figure 17: Text classification**

Text classification has been implemented using a Random Forest model, that case is effective, with the input text. The appeal was granted and the lower court's decision is reversed, effectively classifying the outcome of the case as cited.

# 6. Conclusion

Findings from the classification process shows that classic machine learning models, combined with adequate text pre-processing and TF-IDF feature selection, are capable of classifying legal documents with a reasonable level of accuracy. The methodology effectively desegregates unstructured lawyers' text into numerical representations that allow the algorithm to differentiate between judgments, petitions, and orders with some degree of precision. Random Forest performed better with 58 percent accuracy against 57 percent accuracy of Support Vector Machine, which indicates the usefulness of ensemble techniques in the classification of legal text. The solution greatly contributes to operational efficiency in legal research and decision support systems in terms of reducing manual effort and increasing speed and consistency in retrieval, regardless of the type of legal documents involved.

# References

Chalkidis, I., Fergadiotis, M. and Androutsopoulos, I., 2021. MultiEURLEX--A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv preprint arXiv:2109.00904*.

Chen, H., Wu, L., Chen, J., Lu, W. and Ding, J., 2022. A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management*, *59*(2), p.102798.

Nie, Z., Bai, X., Nie, L. and Wu, J., 2022. Optimization of the economic and trade management legal model based on the support vector machine algorithm and logistic regression algorithm. *Mathematical Problems in Engineering*, *2022*(1), p.4364295.

Sil, R. and Roy, A., 2021. Machine learning approach for automated legal text classification. *International Journal of Computer Information Systems and Industrial Management Applications*, *13*, pp.10-10.