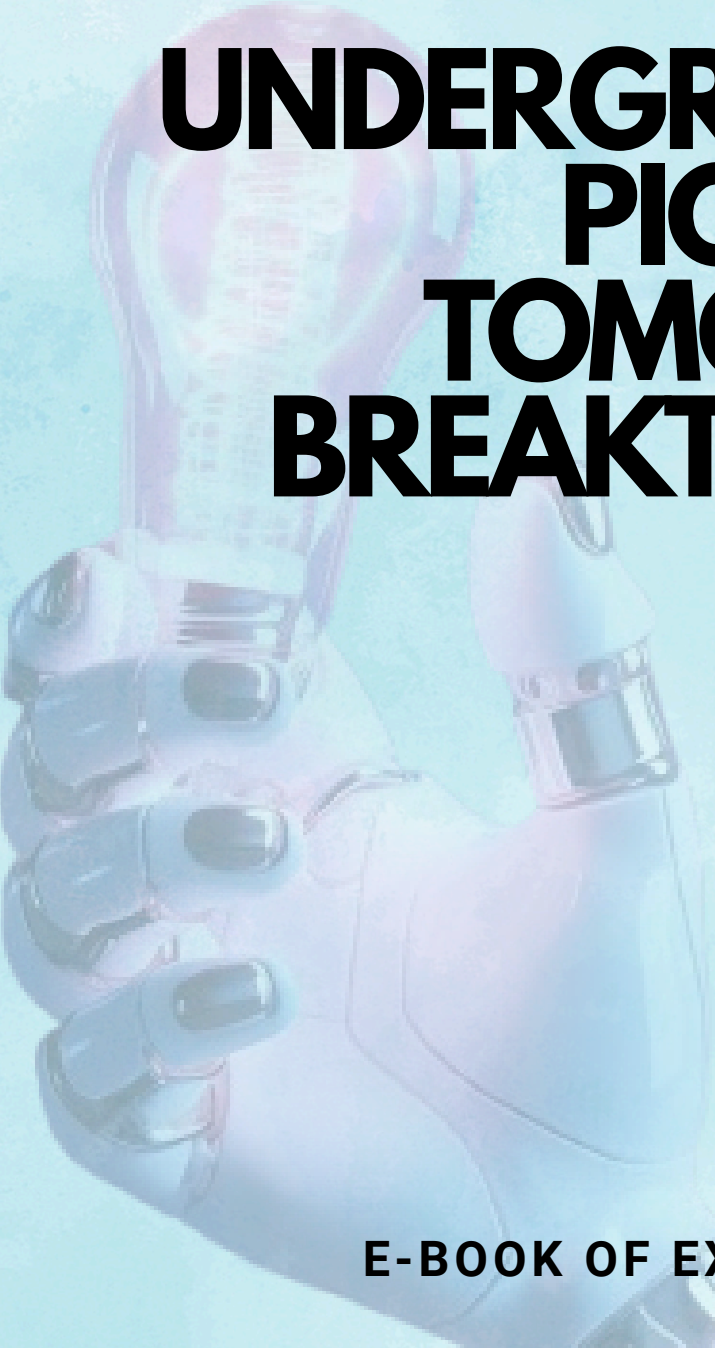




I-URIP 2025

INTERNATIONAL UNDERGRADUATE RESEARCH
INNOVATION, INVENTION AND DESIGN



UNDERGRADUATES PIONEERING TOMORROW'S BREAKTHROUGH

E-BOOK OF EXTENDED ABSTRACTS

CLASSIFYING PARSED LEGAL DOCUMENTS USING A VOTING-BASED ENSEMBLE OF NAÏVE BAYES AND SUPPORT VECTOR MACHINE WITH HIERARCHICAL KNOWLEDGE GRAPH VISUALIZATION

Anis Suraya Latip*, Khyrina Airin Fariza Abu Samah, Huda Liyana Mohd Rapini, and Muhammad Afiq Md Disa

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Melaka
Branch, Jasin Campus, Melaka, Malaysia

*2022862526@student.uitm.edu.my

ABSTRACT

Legal documents (LDs) are often too complex for non-experts, leading to the acceptance of unfavorable terms. This study addresses this by developing a web application to analyze and classify legal contracts. The methodology involves generating a synthetic dataset of financial clauses based on Product Disclosure Sheets from major Malaysian banks. A voting-based ensemble model combining Naïve Bayes (NB) and Support Vector Machine (SVM) is used to classify the extracted clauses, and the findings are presented in an interactive Hierarchical Knowledge Graph (HKG) to visualise the document's structure. The model categorises clauses into six key aspects and the final soft voting ensemble model achieved a classification accuracy of 80.36%. The system's effectiveness was further validated through reliability testing, which demonstrated a significant average time reduction of 71.08% for users analysing documents compared to manual methods ($p < 0.001$). The system's interface, supplemented by AI-generated summaries and a searchable jargon dictionary, proves to be a practical tool for enhancing user comprehension. This project offers a useful tool that empowers the public to make more informed financial decisions by deconstruct complex legal texts.

Keywords: *Legal Documents, Naive Bayes, Support Vector Machine, Ensemble Model, Knowledge Graph*

INTRODUCTION

Legal documents (LDs) exist in many forms, including contracts, loan agreements, policies, and wills. These documents play significant roles in ensuring conformity, upholding legal rights, and guiding decision-making in various sectors. LDs, such as loan agreements, are fundamental in formalising obligations but are often characterized by complex terminology and dense structures, making them difficult for individuals without a legal background to comprehend (Sancheti et al., 2023). Traditionally, reviewing these documents manually is time-consuming due to their length and specialized language (Begum & Goyal, 2021; Smith, 2022). Consequently, insufficient time for thorough reviews can lead to overlooked critical aspects, such as payment terms, termination rights, or hidden charges, resulting in the acceptance of unfavourable conditions (Chalkidis et al., 2020).

To address this problem, this research proposes the development of a web application designed to analyse and classify clauses from parsed legal documents. The core innovation of this system is a voting-based ensemble model that leverages the combined predictive power of two established machine learning algorithms: Naïve Bayes (NB) and Support Vector Machine (SVM). The system's output is not merely a classification result but also an interactive Hierarchical Knowledge Graph (HKG) that visually represents the document's structure and key clauses. The system also integrates on-demand explanations of legal jargon and AI-powered summaries, making LDs more accessible and comprehensible to the public. This project's primary objectives were to design a robust classification system, develop it into a functional web

application, and validate its functionality and reliability. The system's scope was defined by focusing on loan agreements from several major Malaysian banks, using a dataset from January 1, 2024, to December 31, 2024, to ensure its practical relevance.

RESULTS AND DISCUSSION

The project has successfully met its objectives with the development of a web application that classify clauses from parsed legal documents (LDs) using a voting-based ensemble of Naive Bayes (NB) and Support Vector Machine (SVM) models and visualise it in the form of Hierarchical Knowledge Graph (HKG). Functionality testing with 6 test cases was conducted successfully, showing that all of the system features are functioning exactly as intended. The voting-based ensemble of NB and SVM models demonstrated a high performance with 80.36% accuracy. This has proven the system to be effective and functional in assisting users to analyze legal documents more efficiently and with improved comprehension.

Upon the classification of the parsed clauses, the results are then visualized in the form of HKG which visually represents the document's structure and semantics as shown in Figure 1. This central network model is composed of a nested hierarchy of nodes, beginning with a Root Node representing the document. From this, Aspect Nodes (colored boxes) represent high-level categories, while Topic Nodes (hexagons) group semantically related clauses. The individual clauses are represented by color-coded Clause Summary Nodes, which, when expanded, reveal Phrase Nodes that highlight important terms.

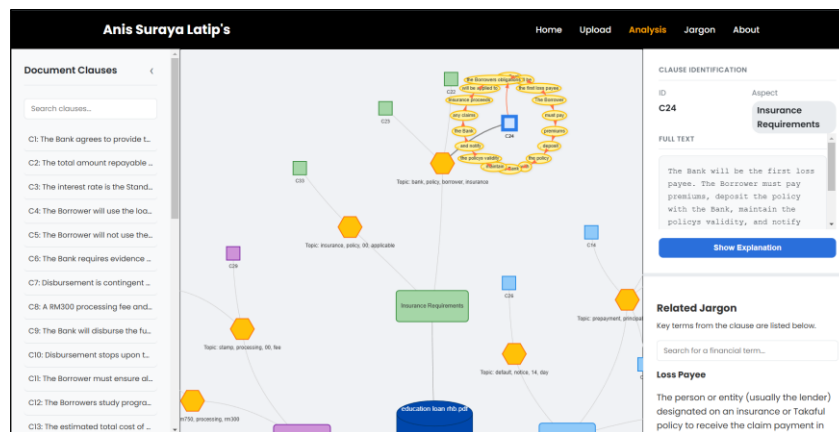


Figure 1: Interactivity of KG

To complement this visual model, the interface includes two interactive sidebars. The left-hand Clause List Sidebar offers a familiar, linear view of the document's content and is synchronised with the graph for seamless navigation. On the right, the Details Sidebar provides in-depth analysis. It features a "flipper card" interface where the front displays information about a selected clause and the back offers an "In Plain English" summary generated by the Gemini 1.5 Flash model displayed in Figure 4. A bottom panel, the "Jargon Explorer," automatically defines relevant terms, functioning as a comprehensive dictionary. Together, these components create an intuitive, multi-modal system that allows users to explore complex documents graphically, textually, and with detailed contextual information.

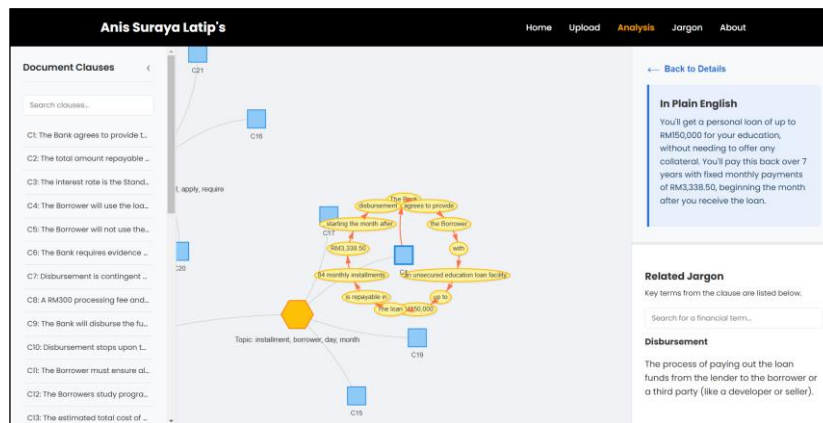


Figure 2: Flipper Card with Clause Summary

Reliability testing was conducted to determine if the system could produce consistent and reproducible results under similar conditions. The test analysed the system's ability to correctly classify clauses, generate meaningful KG structures, and provide jargon definitions. Ten participants, all of whom were members of the public without legal backgrounds, were randomly selected to perform a set of five tasks. Each participant first completed the tasks manually using a paper copy of a loan agreement. They then performed the same tasks using the project's web-based system, which automatically labeled the document, visualised it as a KG, and provided interactive jargon definitions. The time taken to complete each method was independently measured with a stopwatch. The findings showed an average time saving of 71.08% for participants using the system, which demonstrates its reliability and efficiency in reducing the time required to examine a LD.

The reliability test results were statistically analyzed using the Mann-Whitney U Test as illustrated in Table 1. This non-parametric test was chosen due to the small sample size of ten participants, which made it difficult to assume the data was normally distributed. The research question was whether the time taken for manual evaluation was significantly different from the time taken for the web-based system evaluation. The null hypothesis (H_0) stated there was no significant difference in the time taken between the two methods, while the alternative hypothesis (H_1) suggested there was a significant difference. The Mann-Whitney U Test yielded a p-value of <0.001 , which is less than the conventional significance level of $p < 0.05$. Since the p-value was lower than the significance level, the null hypothesis (H_0) was rejected. This led to the conclusion that a statistically significant difference exists between the time spent on manual evaluation and the time spent on the web-based system evaluation. The asterisk symbol (*) next to the p-value in the table indicates that the result is significant, emphasising that the reduction in time when using the system is a real performance improvement and not due to random chance.

Table 1: Mann-Whitney U Test for Manual and Web-based Evaluation

Mann-Whitney U Test of Two Samples		
	Manual	System
Minimum	9.36	2.58
Maximum	15.32	4.3
Median	12.3	3.31
Mean	12.25	3.52
<i>p-value</i>		$< 0.001^*$

CONCLUSION

In conclusion, this project successfully developed a web application that classifies clauses from parsed legal documents (LDs) using a voting-based ensemble of Naive Bayes (NB) and Support Vector Machine (SVM) models and visualises the results as a Hierarchical Knowledge Graph (HKG). The system proved to be effective and reliable, achieving a high predictive accuracy of 80.36% and a statistically significant average time reduction of 71.08% compared to manual methods ($p < 0.001$). While successful, the study was limited by its use of a synthetically generated dataset, fixed categories, and the inability to handle advanced functions like document comparison or tracking changes. Future work should focus on expanding the dataset with real-world documents, adding advanced features like document comparison, and creating a dynamic knowledge base to keep the model's insights current.

ACKNOWLEDGEMENT

The authors would like to express their deepest gratitude to our supervisor, Prof. Madya Ts. Dr. Khyrina Airin Fariza binti Abu Samah, for her invaluable guidance, patience, and the time she sacrificed to mentor us. We are profoundly grateful to Universiti Teknologi MARA for generously funding this research under the Young Researchers Grant Scheme (Project Code: ST-50).

REFERENCES

- Begum, N., & Goyal, A. (2021). Analysis of legal case document automated summarizer. Proceedings of IEEE International Conference on Signal Processing, Computing and Control, 2021-Octob, 533–538. <https://doi.org/10.1109/ISPC53510.2021.9609442>
- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT : The muppets straight out of law school. In T. Cohn, Y. He, & Y. Liu (Eds.), Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020 (Issue i, pp. 2898–2904). Association for Computational Linguistics (ACL). <https://doi.org/10.18653/v1/2020.findings-emnlp.261>
- Sancheti, A., Garimella, A., Srinivasan, B. V., & Rudinger, R. (2023). What to read in a contract? Party-specific summarization of legal obligations, entitlements, and prohibitions. EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings, 14708–14725. <https://doi.org/10.18653/v1/2023.emnlp-main.909>
- Smith, M. (2022). Challenges automating the law: AI in the legal sector. Oxford Fintech and Legaltech Society. <https://www.oxfordfls.org/blog/Challenges-Automating-the-Law-AI-in-the-Legal-Sector/>