**Domain classification of summarized judgments with fast text word embeddings**

Sunkara Sravan Kumar, K Rishi Sai Reddy, Sadineni Sai Kumar, Priyanka Prabhakar Peeta Basa Pati\*  
*Department of Computer Science and Engineering,*  
*Amrita School of Computing* Bengaluru,   
Amrita Vishwa Vidyapeetham, India

\*Corresponding Author: [bp\_peeta@blr.amrita.edu](mailto:bp_peeta@blr.amrita.edu), ORC ID: 0000-0003-2376-4591

***Abstract*— This study uses summarized English judgments to present a machine learning method for cross-lingual classification of Telugu legal text, employing Fast Text word embeddings. This study applies six machine learning methods KNN, Random Forest, Decision tree, MLP, SVM and Logistic Regression to classify Telugu legal text cross-lingually, revealing that random forest achieves the highest accuracy. The model exhibits high accuracy in classifying legal domains, confirmed through extensive comparative studies. With a rigorous model selection process, it provides an effective tool for swift analysis of Telugu legal documents. In essence, this research contributes to advancing cross-lingual natural language processing, assisting scholars and legal professionals in managing linguistic diversity in legal content.**

***Keywords: Cross-Lingual, Domain Classification, Telugu Text, Fast Text, Legal Judgments.***

# Introduction

In an age characterized by the widespread exchange of global information, the pursuit of cross-lingual natural language processing (NLP) emerges as a crucial endeavour. This study hones in on the specific challenge of discerning Telugu legal text using condensed English judgments. The intricacies of legal language in cross-lingual tasks, particularly within the legal domain, necessitate a keen understanding of linguistic subtleties. Legal judgments, serving as foundational elements in global legal systems, harbor a wealth of information. The capacity to categorize Telugu text from English legal summaries becomes indispensable for the streamlined retrieval of information and decision-making within legal circles.

Traditional approaches often grapple with the complexities of legal language, prompting the search for innovative solutions. This research adopts Fast Text word embeddings to enhance semantic understanding and contextual relevance, integrating supervised learning with annotated class labels. The primary aim is to cultivate a robust machine learning model proficient in accurately categorizing Telugu legal text based on English summaries. The paper delves into scalability, scrutinizing the model's performance across diverse dataset sizes. The classifications are promising for legal professionals, researchers, and practitioners in search of effective tools for cross-lingual domain classification. Subsequent sections meticulously outline the methodology, experimental framework, results, and implications, offering a thorough examination of our proposed approach and its impact on the evolving panorama of cross-lingual NLP.

Moreover, the project's implementation of domain classification aligns with the principles of Sustainable Development Goal 16 (SDG 16). The accurate and inclusive translation enhances access to justice and fosters transparent legal systems. Proficient Telugu speakers benefit from improved access to crucial legal information, promoting equal protection under the law. This initiative directly supports SDG 16's goal of establishing effective, accountable, and inclusive institutions by overcoming language barriers in legal contexts. The enhanced comprehension of legal texts in local languages, like Telugu, empowers individuals to engage actively in legal proceedings, reducing the potential for injustice and contributing to the broader objective of constructing fair and inclusive societies. By presenting a machine learning method for categorizing Telugu legal text, this work significantly advances cross-lingual natural language processing. The flexibility of the approach is demonstrated by the use of some classification models in conjunction with Fast Text word embeddings. This research provides scholars and legal professionals with a useful tool for the quick and accurate analysis of Telugu legal documents, addressing linguistic diversity within legal content through a rigorous model selection process and extensive comparative studies.

# Related work

Katsuhito Sudoh et al. [1] proposed a new method for handling lengthy sentences by breaking them into smaller clauses. The original sentence was systematically divided, and these clauses were translated using statistical machine translation (SMT). This approach may introduce challenges in preserving the original sentence's coherence and may lead to potential loss of context during the division and reassembly process. Koehn et al. [2] brought to light an open-source SMT tool named Moses, which opened up new avenues for research and experimentation in the field of machine translation. Notably, they introduced the concept of confusion network decoding, a valuable technique for addressing the translation of ambiguous texts. Alkatheery et al. [3] undertook an extensive evaluation of the quality of Google Translate, which leverages Neural Machine Translation for translating legal texts. Their assessment covered lexical, syntactic, and register-related errors, with lexical errors being the most prominent.

Elnaggar et al. [4] tackled the challenge of multitasking within the legal domain using Deep Learning Models. Their approach encompassed translation, summarization, and multi-label classification. One of their notable accomplishments was mitigating the data scarcity issue through the application of transfer learning. Mouratidis et al. [5] introduced NoDeeLe, a pioneering deep-learning schema designed for assessing NMT systems. Their evaluation encompassed two NMT systems across multiple text genres, including medical, legal, marketing, and literary documents within the English-Greek language pair. Wiesmann et al. [6] conducted comprehensive tests on DeepL Translator and MateCat, a CAT system incorporating machine translation. Their evaluation criteria included comprehensibility, meaningfulness, and the alignment between source and target text, all while considering the specific translation context.

Hung et al. [7] opted to categorize legal texts based on their logical structure and subsequently employed a standard SMT model for translation. They reported notable improvements in translation quality, particularly for English-Japanese legal text pairs, as evidenced by metrics such as BLEU, NIST, and TER scores.

Moneus et al. [8] undertook a meticulous comparison between AI and human translation, focusing on the translation of legal documents. Their findings highlighted that human translation generally yields superior quality, yet both approaches possess their respective advantages and limitations. Lagarda et al. [9] introduced an innovative application of automatic post-editing within an online learning framework. This dynamic post-editing module acquired knowledge and improved its corrections in real time through user feedback, employing cutting-edge online learning techniques. Gao et al. [10] introduced the concept of conceptual referentially transparent inputs (CRTI) and developed a metamorphic testing approach, which utilized back-translation as a benchmark for testing machine translation. Their method successfully identified a notable number of suspicious issues, demonstrating a precision of 74% in Chinese translation and 82% in Vietnamese translation across a dataset of 200 sentences.

Sankaravelayuthan et al. [11] introduced an innovative dataset tailored for evaluating word similarity, diverging from traditional association or relatedness metrics. The assessment framework mirrors cognitive phenomena and maintains dataset consistency. Importantly, the research underscores the dataset's user-friendliness for non-native speakers with minimal exertion. The document's limitation lies in its exclusive focus on the low-resourced language Tamil, lacking exploration of transferability to other languages, and not addressing factors like bilingual dictionary size, embedding algorithm choice, and corpus domain. Menon et al. [12] address the difficulty of assessing sentence semantics through the introduction of a scoring scheme based on local alignment using word embeddings. The paper incorporates a theoretical examination of metrics, reinforced by a separability argument demonstrated through t-SNE plots. But the paper does not provide any quantitative or qualitative results to compare its method with existing approaches or baselines. Priyanka et al. [13] address the difficulty of identifying deceptive information on social media, acknowledging the constraints of conventional approaches. By utilizing word embedding features and datasets like LIAR and ISOT, the investigation utilizes methods such as cosine similarity to recognize correlated news data and categorize domains based on central themes. However, it does not compare their methods with existing approaches or baselines for fake news detection. Jameer et al. [14] confronts the task of enhancing web search outcomes by suggesting a method that utilizes word embeddings on Stack Overflow inquiries. Through assigning vectors to each word within a sentence, the strategy strives to augment the significance of search results. However, it does not explain the details of the word embedding technique, such as the choice of parameters, the size of the vocabulary, and the source of the training data.

Priyanka et al. [15] report the effectiveness of a method using T5 to generate an extractive summary of Indian legal judgments, which vary in length and structure. The paper creates a dataset manually with a lawyer’s help and evaluates the result using the ROUGE score. The drawbacks of the paper are that does not address the ethical or legal implications of using automatic summarization for legal documents.

# Dataset

The dataset, consisting of 311 collected documents of summarized judgments in English & Telugu. Each such document also contained a label corresponding to the respective legal domain that is Civil, Criminal and Writ. In our dataset we have 170 documents in Civil domain, 111 documents in Criminal domain and 29 documents in Writ domain. All these judgement documents are taken from Indiankanoon.com [16] and summarized with a summarization tool [15]. These summaries are translated to Telugu using a translation tool [11]. The Synthetic Minority Oversampling Technique (SMOTE) method was used for balancing these classes.

The dataset consisting of summaries are converted to embedded vectors using FastText embedding algorithm.

# Methodology

The proposed approach for executing the process is shown in Fig 1. where we have done Data collection, Data Preprocessing, Feature Selection, Classification and Model Evaluation which have been briefly explained below.

Data Collection

Data Preprocessing

Feature selection

Classification

Model Evaluation

Fig 1. Flow Schematic block diagram for the methodology.

1. *Data Preprocessing*

The dataset was meticulously cleaned during preprocessing by eliminating null values from the word embeddings dataset and closely examining any anomalies. By removing the outliers, we have made the dataset clean.

1. *Feature Selection*

We have used decision trees where recursive feature elimination method is used to choose the best features by comparing all the features. Where the method selects the best 200 features out of 300.

1. *Classification*

We utilized six different methods, namely, K-Nearest Neighbors (K-NN), Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), Logistic Regression, Decision Trees, and Random Forest, during the classification phase.

1. *Model Evaluation*

We have taken the four evaluation matrices (Accuracy, Precision, Recall, F1score) for model evaluation in our project.

# Results And Analysis

The following table presents the latest results of the experiments.

*Before SMOTE:*

Before executing SMOTE, the machine learning models were put into practice, and each model's F1-score and accuracy were determined as shown in Table 1. At 76.88%, the Random Forest classifier exhibits the highest accuracy and F1 score for the train data. At 61.29%, the Random Forest classifier exhibits the highest accuracy for the test data.

Fig 2 shows the accuracy of each classifier before SMOTE in graph plot where Random Forest classifier shows the highest accuracy as 76.88% for train data. And the Random Forest classifier displaying the highest accuracy at 61.29% for test data.

*After SMOTE:*

After executing SMOTE, the machine learning models were put into practice, and each model's F1-score, accuracy, precision, and recall were determined as shown in Table 2. At 77.42%, the Random Forest classifier exhibits the highest accuracy and F1 score. At 58.06%, the Random Forest classifier exhibits the highest accuracy.

Table 1. Comparison between models before SMOTE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Train Data** | | **Test Data** | |
| **Model** | **Accuracy** | **F1 Score** | **Accuracy** | **F1 Score** |
| KNN | 62.90% | 44.59% | 47.58% | 33.37% |
| Random Forest | 76.88% | 76.60% | 61.29% | 38.98% |
| Decision Tree | 76.88% | 77.06% | 50.81% | 31.05% |
| MLP | 76.88% | 76.21% | 53.23% | 31.39% |
| SVM | 62.90% | 62.61% | 57.26% | 30.05% |
| Logistic Regression | 58.06% | 43.46% | 55.65% | 27.85% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Train Data** | | **Test Data** | |
| **Model** | **Accuracy** | **F1 Score** | **Accuracy** | **F1 Score** |
| KNN | 70.05% | 66.80% | 55.91% | 34.88% |
| Random Forest | 77.42% | 74.87% | 58.06% | 45.25% |
| Decision Tree | 77.42% | 74.87% | 54.84% | 42.34% |
| MLP | 53.92% | 23.35% | 56.99% | 24.20% |
| SVM | 54.38% | 40.76% | 56.99% | 24.20% |
| Logistic regression | 70.51% | 66.32% | 53.76% | 32.87% |

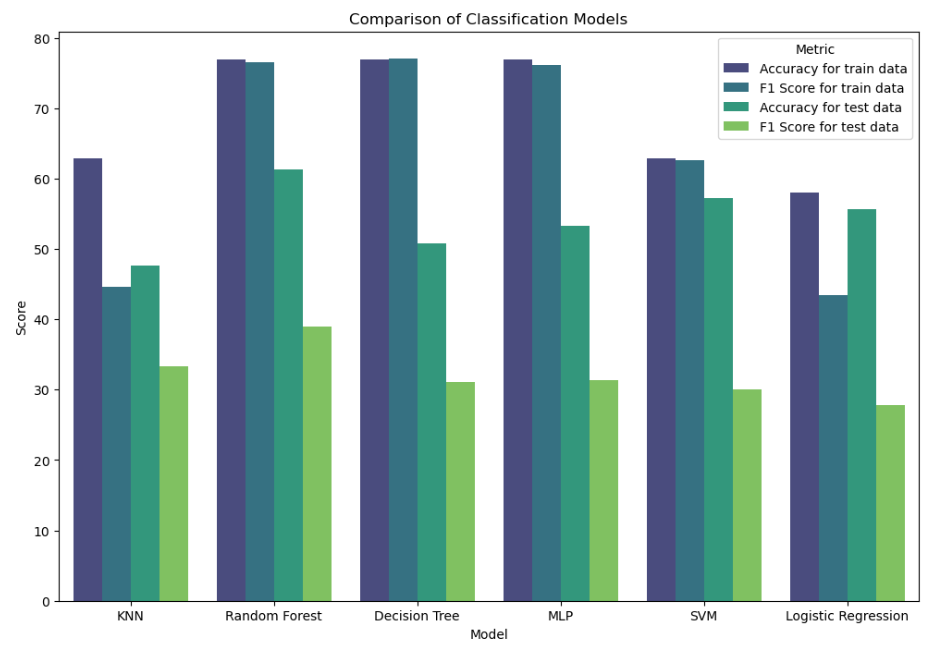


Fig 2. Cross-validation accuracy before smote

Fig 3 shows the accuracy of each classifier before SMOTE in graph plot where Random Forest classifier shows the highest accuracy as 77.42% for train data. with the Random Forest classifier displaying the highest accuracy at 58.06% for test data.

*After Feature Selection:*

After Feature Selection, the machine learning models were put into practice, and each model's F1-score, accuracy, precision, and recall were determined as shown in Table 3. At 77.42%, the Random Forest classifier exhibits the highest accuracy for the train data. At 58.06%, the Random Forest classifier exhibits the highest accuracy for the test data.

The accuracy of each classifier is displayed in Fig4. prior to the graph plot, with the Random Forest classifier displaying the highest accuracy at 77.42% for train data. with the Random Forest classifier displaying the highest accuracy at 58.06% for test data.

Table2. Comparison between models after SMOTE

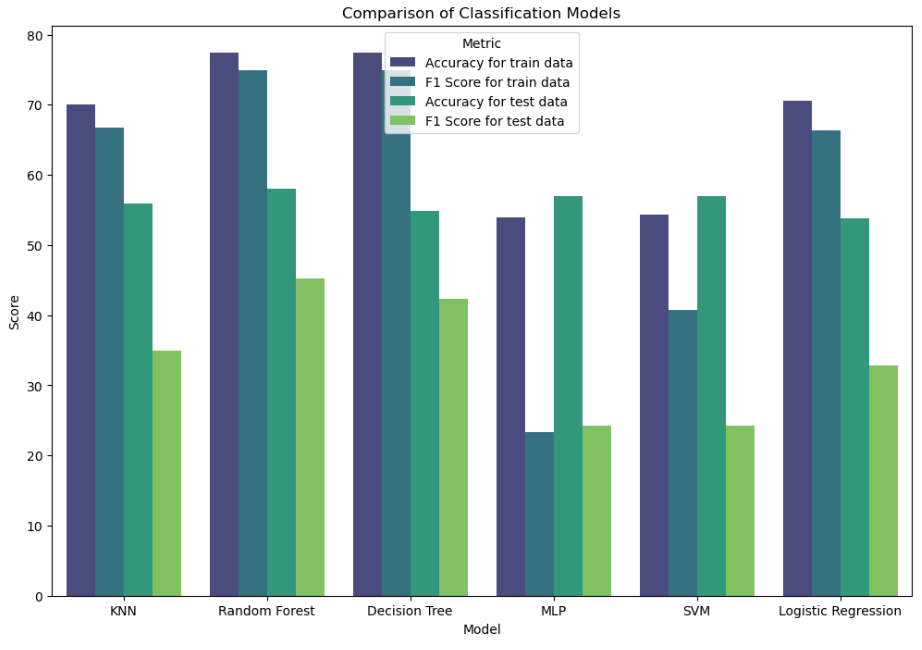


Fig 3. Cross-validation accuracy after smote

###### *Table 3. Comparison between models after feature selection.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Train Data** | | **Test Data** | |
| **Model** | **Accuracy** | **F1 Score** | **Accuracy** | **F1 Score** |
| KNN | 70.05% | 6.80% | 55.91% | 34.88% |
| Random Forest | 77.42% | 74.75% | 58.06% | 45.25% |
| Decision Tree | 77.42% | 74.87% | 54.84% | 42.34% |
| MLP | 77.42% | 74.87% | 56.99% | 24.20% |
| SVM | 54.38% | 40.76% | 56.99% | 24.20% |
| Logistic Regression | 70.51% | 66.32% | 53.76% | 32.87% |

# 

Fig 4. Cross-validation accuracy after SMOTE model

# Conclusion

Random Forest consistently outperforms other models in classifying translated text domains, excelling in generalization across training and test datasets. While KNN shows moderate accuracy but struggles to generalize, Decision Tree and MLP models exhibit overfitting issues. SVM and Logistic Regression perform relatively weaker across metrics. The Random Forest model stands out as the most reliable choice, delivering commendable recall, accuracy, and F1-score on unseen data, making it the preferred model for domain classification of translated text. When we compare the results of ours with the existing models, we are getting scores mostly in the same range from our research of various research papers.

Potential future for development involves broadening the model's language coverage, refining its focus on specific legal domains, and incorporating advanced language models for more nuanced classification. Ongoing improvements include dynamic model training, user interface enhancements, and collaborative efforts with legal professionals for practical assessments. Expanding the dataset and continual benchmarking against evolving legal standards contribute to the system's adaptability and sustained efficacy.

# Acknowledgement

We extend our sincere appreciation to the authors of the prior studies for their invaluable contributions to the classification of legal text documents. We acknowledge the contributions of Amrita Vishwa Vidyapeetham for the infrastructure and support in carrying out this work.

# References

1. Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Tsutomu Hirao, and Masaaki Nagata. “Divide and Translate: Improving Long Distance Reordering in Statistical Machine Translation.” Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR,2010.
2. Koehn Philipp, Hoang Hieu, Alexandra Birch, Callison-Burch Chris, Federico Marcello, Bertoldi Nicola, Cowan Brooke, Shen Wade, Moran Christine, Zens Richard, Dyer Chris, and Bojar Ondrej. "Moses: Open-Source Toolkit for Statistical Automated Translation." Proceedings of the ACL 2007 Demo and Poster Sessions, 2007.
3. Alkatheery Eman Rashed. "Google Translate Errors in Legal Texts: Automated Translation Quality Assessment." AWEJ for Translation and Literary Studies, 2023.
4. Elnaggar Ahmed, Gebendorfer Christoph, Glaser Ingo, and Matthes Florian. "Multi-Task Deep Learning for Legal Document Translation, Summarization and Multi-Label Classification." 2018 Association for Computing Machinery.
5. Mouratidis Despoina, Stasimioti Maria, Sosoni Vilelmini, and Kermanidis Katia Lida. "NoDeeLe: A New Deep Learning Schema for Evaluating Neural Automated Translation Systems." 2018 Association for Computing Machinery.
6. Wiesmann Eva. "Automated Translation in the Field of Law: A Study of the Translation of Italian Legal Texts into German."
7. Bui Thanh Hung, Nguyen Le Minh and Akira Shimazu,"Translating legal sentence by Segmentation and rule selection",International Journal on Natural Language Computing (IJNLC) Vol. 2, No.4, August 2013.
8. Moneus Ahmed Mohammed, and Sahari Yousef Mohammed. "Artificial Intelligence and Human Translation: A Contrastive Study Based on Legal Texts."
9. Lagarda Antonio L., Ortiz-Martinez Daniel, Alabau Vicent, and Casacuberta Francisco. "Translating Without in-Domain Corpus: Machine Translation Post-Editing With Online Learning Techniques."
10. Gao Wentao, He Jiayuan, and Pham Van-Thuan. "Metamorphic Testing of Automated Translation Models Using Back Translation." 2023 IEEE/ACM International Workshop on Deep Learning for Testing and Testing for Deep Learning (Deep Test).
11. S. Jp, Menon, V., Sankaravelayuthan, R., Dr. Soman K. P., and Kumar, M., “Intrinsic Evaluation for English–Tamil Bilingual Word Embeddings: Proceedings of ISTA 2018”, in Advances in Intelligent Systems and Computing, 2020.
12. V. Krishna Menon, M., S., K., H., and Dr. Soman K. P., “Semantic Analysis Using Pairwise Sentence Comparison with Word Embeddings”, in Intelligent Systems Technologies and Applications, Cham, 2018.
13. Priyanka VT, Sanjanasri J.P, Vijay Krishna Menon and Soman KP “Exploring Fake News Identification Using Word and Sentence Embeddings” accepted in the Journal of Intelligent and Fuzzy Systems, IOS Press, Netherlands (ISSN print 1064-1246, ISSN online 1875-8967). - SCIE Indexed.
14. Babu, Jameer, and S. Thara. "Finding the Duplicate Questions in Stack Overflow using Word Embeddings." Procedia Computer Science 171 (2020): 2729-2733, DOI: 10.1016/j.procs.2020.04.296.
15. P. Prabhakar, D. Gupta and P. B. Pati, "Abstractive Summarization of Indian Legal Judgments," 2022 OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, 2022, pp. 256-261, doi: 10.1109/OCIT56763.2022.00056.Indian court judgements. Retrieved from <http://www.indiankanoon.com>