Engineering and Scientific International Journal (ESIJ)
Volume 13, Issue 2, April – June 2026
ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

Automated IPC Section Prediction using Hybrid Deep Learning and NLP Techniques

Dr.N.Sevugapandi[*1], K.Ponraj[2]

[1]*Guest Lecturer,* [2]*II[nd] Year M.Sc(Computer Science)*

[1,2]*Department of Computer Science (UG & PG), Government Arts and Science College, Kovilpatti - 628 503, Tamilnadu, India,*

Email-id: [*]sevugapandi1985@gmail.com, rajpon5916@gmail.com

Abstract — This paper presents an intelligent system for automatic prediction of Indian Penal Code (IPC) sections based on FIR (First Information Report) text. The system uses Natural Language Processing (NLP) techniques combined with a hybrid deep learning model consisting of Bi-LSTM and RoBERTa. The model analyzes case descriptions and predicts the most relevant IPC section along with confidence score. Additionally, the system provides automated legal explanation for better understanding. The proposed system improves accuracy compared to traditional machine learning models and helps in faster legal decision support. This solution can assist police departments, legal professionals, and judicial systems in reducing manual effort and improving efficiency. The legal domain involves complex analysis of textual data such as FIR (First Information Reports), which requires expertise and time to determine the appropriate IPC sections. This paper proposes an intelligent system that automates the prediction of IPC sections using Natural Language Processing (NLP) and a hybrid deep learning model combining Bi-LSTM and RoBERTa. The system processes unstructured legal text, extracts meaningful features, and predicts the most relevant IPC sections with a confidence score. Additionally, the system provides automatic IPC explanations from a structured database, improving interpretability. Experimental results demonstrate that the hybrid approach significantly improves prediction accuracy compared to traditional machine learning methods. The proposed system can assist law enforcement agencies, legal practitioners, and judicial systems in enhancing efficiency and reducing manual effort.

Keywords *— Indian Penal Code (IPC); First Information Report (FIR); Natural Language Processing (NLP); Bi-LSTM; RoBERTa; Hybrid Deep Learning; Text Classification; Legal Text Analysis; IPC Section Prediction; Transformer Model; Sequential Learning; Contextual Embeddings.*

1. Introduction

The increasing volume of criminal cases has made it challenging for law enforcement agencies to manually analyze FIR (First Information Report) documents and accurately assign appropriate IPC sections [9]. FIRs are typically written in unstructured textual format, which makes them difficult to process using traditional rule-based or manual methods. As a result, human analysis becomes time-consuming and is often prone to errors and inconsistencies. With the rapid advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) [1], [8], it has become possible to automate text classification tasks with high accuracy and efficiency. In this research, we propose an intelligent system that utilizes a hybrid deep learning model combining Bi-LSTM [4] and RoBERTa [3]. The Bi-LSTM model effectively captures sequential patterns in the text, while RoBERTa enhances contextual understanding of the legal language [11].

The primary objective of this system is to automatically predict the relevant IPC sections based on FIR descriptions and provide corresponding legal explanations. This approach helps reduce manual workload, improve prediction consistency, and support faster and more reliable decision-making in the legal system.

2. Methodology

The proposed system follows a structured and systematic pipeline for predicting IPC sections from FIR text. Initially, the dataset is collected through manual entry and admin uploads, and all records are stored in a centralized database. The input FIR text is then preprocessed using Natural Language Processing (NLP) techniques such as lowercasing, removal of special characters, tokenization, stopword removal, and text cleaning to ensure uniformity and improve model performance [7], [8].

After preprocessing, feature extraction is performed where token sequences are generated and contextual embeddings are obtained using the RoBERTa model [3], [6]. The processed data is then fed into a hybrid deep learning architecture that combines Bi-LSTM and RoBERTa. The Bi-LSTM model captures sequential dependencies and patterns in the text [4], [5], while RoBERTa provides deep contextual understanding of the language [3]. The outputs

DOI: 10.30726/esij/v13.i2.2026.132006

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

from both models are integrated to generate a more accurate and reliable prediction.

Finally, the system predicts the most relevant IPC section along with a confidence score. In addition to prediction, the system retrieves the corresponding IPC title and description from the database, providing a clear explanation to the user. This end-to-end pipeline ensures efficient, accurate, and explainable legal text classification.

3. Modules Description

The proposed system follows a modular architecture, where each module is designed to perform a specific function. This approach improves system efficiency, scalability, and maintainability. The major modules of the system are described below.

The *User Module* manages all interactions between the user and the system. It provides secure authentication through login and logout features. After successful authentication, users can enter FIR case descriptions and initiate the prediction process. The module ensures a simple and user-friendly interface for input submission and result viewing.

The *Admin Module* controls system management and configuration. It includes an advanced Admin Dashboard that provides a complete overview of system performance and usage. The dashboard displays key metrics such as total predictions, number of users, and model accuracy. It also shows the model status, indicating whether the model is ready for prediction.

Additionally, the dashboard offers quick navigation options for dataset management, model retraining, user management, IPC section management, and prediction history. It also includes visual charts such as IPC distribution, confidence levels, and model usage comparison (e.g., Bi-LSTM and RoBERTa). This module helps administrators monitor and manage the system effectively.

The *Dataset Management module* allows administrators to manage the FIR dataset used for training and testing. It supports uploading datasets in CSV format and adding new case records with corresponding IPC sections. The dataset is displayed in a structured table with features like search and delete options. A key feature is the ability to retrain the model using updated or newly uploaded data. This helps improve model accuracy over time by learning from new patterns.

The *Model Training Module* is responsible for training the deep learning model using the dataset. It uses the Bi-LSTM architecture for sequential text learning. The model can be retrained whenever new data is added, ensuring continuous improvement in prediction performance.

The *FIR Upload module* allows users to upload FIR documents in PDF or TXT formats. The system extracts text from the uploaded files using text processing techniques. The extracted content is then preprocessed and passed to the prediction model. This improves usability by supporting real-world document inputs.

The *Prediction History Module* stores all previous predictions made by users. It records details such as case description, predicted IPC section, confidence score, and timestamp. This helps users review and analyze past results.

The *Auto IPC Explanation module* provides explanations for the predicted IPC sections. After prediction, the system retrieves the corresponding IPC title and description from the database. This improves transparency and helps users understand the reasoning behind the prediction.

4. Experimental Results and Performance Evaluation

The proposed system was evaluated using multiple FIR datasets to analyze its performance and reliability. The experimental results show that the hybrid deep learning model achieved significantly higher accuracy compared to traditional machine learning models such as Naive Bayes and Support Vector Machine (SVM).

The integration of Bi-LSTM and RoBERTa enabled the system to effectively process unstructured and lengthy FIR texts, resulting in more accurate IPC section predictions. The Bi-LSTM model captured sequential dependencies and patterns within the text, while the RoBERTa model improved the contextual understanding of legal language. This combination enhanced both syntactic and semantic analysis, leading to improved classification performance.

The system consistently produced predictions with high confidence scores, indicating strong reliability and robustness of the model. In addition to prediction, the system successfully generated detailed explanations for the predicted IPC sections by retrieving relevant information from the database.

Furthermore, the implementation of a PDF report generation feature allowed users to download structured prediction results, including IPC section, confidence score, case description, and explanation. This feature improves usability and makes the system suitable for real-world legal applications.

DOI: 10.30726/esij/v13.i2.2026.132006

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

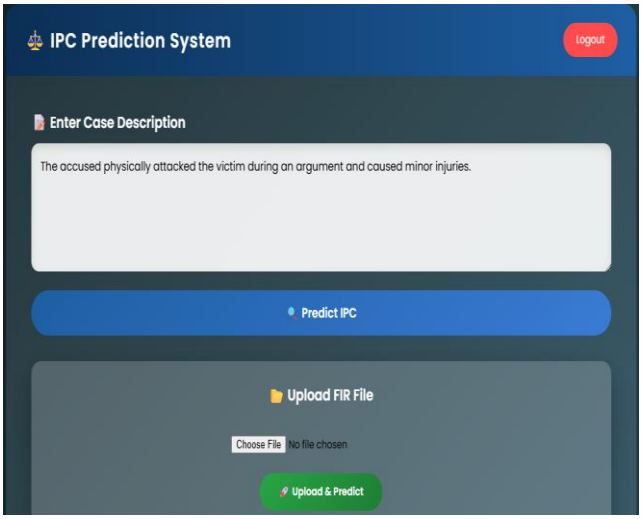

Fig.1: IPC Prediction System Dashboard

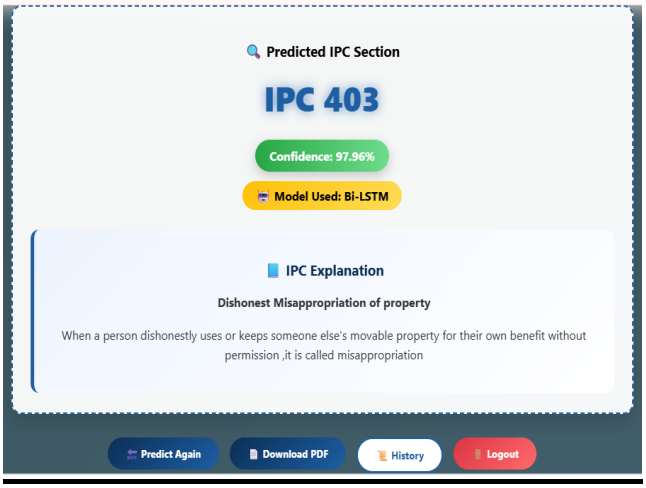

Fig.2: Result page

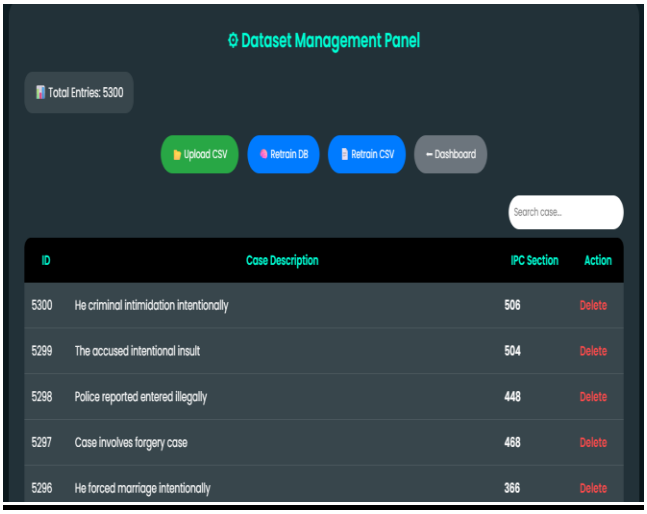

Fig.3: Data Management Panel

Fig.4: Admin Dashboard

5. Comparison of Deep Learning Module

In this research, different machine learning and deep learning models were analyzed to evaluate their effectiveness in predicting IPC sections from FIR text. However, these models rely heavily on feature engineering and fail to capture complex contextual relationships present in legal text.

Deep learning models such as Bidirectional Long Short-Term Memory (Bi-LSTM) [4], [5] provide improved performance by capturing sequential dependencies in the text. Bi-LSTM processes the input in both forward and backward directions, allowing it to understand the order and structure of words effectively. This makes it suitable for handling long FIR descriptions where the meaning depends on sequence context.

On the other hand, transformer-based models like RoBERTa [2], [3] offer superior contextual understanding by leveraging attention mechanisms [10]. RoBERTa is capable of capturing deep semantic relationships between words, even in complex and unstructured sentences.

However, each model has its limitations. Bi-LSTM struggles with capturing long-range dependencies efficiently, while RoBERTa requires high computational resources. To overcome these limitations, the proposed system uses a hybrid approach combining both models. The Bi-LSTM captures sequential patterns, and RoBERTa extracts contextual meaning. The outputs are then combined to produce the final IPC prediction [9], [11].

Experimental results show that the hybrid model outperforms individual models in terms of accuracy and reliability [12]. This demonstrates that combining sequence

DOI: 10.30726/esij/v13.i2.2026.132006

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

learning and contextual understanding leads to better performance in legal text classification tasks.

Table 1: Model Comparison

Model	Strength	Weakness	Use in System
Naive Bayes	Very fast and simple to implement	Low accuracy, assumes independence	Used as Baseline
SVM	Effective for small datasets	Cannot understand context	Used as Baseline
Bi-LSTM	Captures sequence (word order)	Limited long-range context	Used in Model
RoBERTa	Deep contextual understanding	Requires high computation power	Used in Model
Hybrid Model	Combines Bi-LSTM + RoBERTa for best accuracy	Slightly complex architecture	Final Proposed Model

6. Conclusion

This research presented an intelligent IPC prediction system that automates the classification of FIR documents using a hybrid deep learning approach. The system effectively addresses the challenges associated with manual analysis of unstructured legal text by integrating Natural Language Processing techniques [1], [8] with advanced machine learning models.

The proposed hybrid model combines the strengths of Bi-LSTM and RoBERTa to capture both sequential and contextual information from FIR data. Experimental results demonstrate that the hybrid model achieves higher accuracy and better performance compared to traditional machine learning models. The system is capable of handling long and complex textual inputs while maintaining reliable prediction results.

In addition to prediction, the system enhances interpretability by providing detailed IPC explanations and supports real-world usability through features such as FIR file upload, prediction history tracking, and PDF report generation. The inclusion of an Admin Dashboard further enables efficient dataset management and model monitoring. Overall, the proposed system improves efficiency, reduces human effort, and supports faster legal decision-making. It serves as a scalable and intelligent solution for modernizing the legal analysis process using Artificial Intelligence.

References

[1] Y. Goldberg, "A Primer on Neural Network Models for Natural Language Processing," Journal of Artificial Intelligence Research, vol. 57, pp. 345–420, 2016.
[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, 2019.
[3] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[5] Graves, "Supervised Sequence Labelling with Recurrent Neural Networks," Springer, 2012.
[6] T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.
[7] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python," O'Reilly Media, 2009.
[8] D. Jurafsky and J. H. Martin, "Speech and Language Processing," 3rd ed., Pearson, 2020.
[9] Indian Penal Code, 1860, Government of India.
[10] A.Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.
[11] F. Chollet, "Deep Learning with Python," Manning Publications, 2018.
[12] T. Brown et al., "Language Models are Few-Shot Learners," in NeurIPS, 2020.

DOI: 10.30726/esij/v13.i2.2026.132006