iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70627

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Smart IPC Section Prediction Using Machine Learning

Amogh Kuppachi[1], Rohan R[2], Dr Raghavendra V[3]
[1, 2]*UG Student,* [3]*Assistant Professor, Department of Computing Technologies, SRM institute of science and technology, Chennai, Tamil Nadu, India*

Abstract: ***The increasing demand for faster, more consistent, and accessible legal analysis has driven the adoption of intelligent systems within the judicial domain. This paper presents a machine learning-based framework that predicts relevant sections of the Indian Penal Code (IPC) based on textual descriptions of criminal incidents. By utilizing natural language processing (NLP) techniques, the proposed system is capable of comprehending the context of a legal case, identifying significant legal cues, and mapping them to the appropriate IPC sections. The model is trained on a diverse dataset consisting of real and synthesized case summaries, enabling it to effectively learn the linguistic patterns and legal terminology used in criminal law.***
The primary objective of this work is to assist legal professionals, law enforcement agencies, and other stakeholders by providing quick, consistent, and reliable legal references during the initial evaluation of a case. This system aims to reduce the burden of manual analysis, minimize errors arising from subjective interpretation, and improve overall efficiency in the legal process. Furthermore, the project highlights the broader role of artificial intelligence (AI) in modernizing legal workflows and enhancing access to justice through data-driven insights. The results demonstrate the potential of predictive systems to transform legal practices and contribute to the development of smarter legal tools in the Indian judicial context.
Through the use of state-of-the-art machine learning models such as DistilBERT and TinyBERT, this work provides a robust and scalable solution to automate the classification of case descriptions into relevant IPC sections, showcasing the utility of NLP in legal applications.
Keywords: Machine Learning, Natural Language Processing, IPC Prediction, Legal Automation, Judicial Intelligence, AI in Law, Criminal Case Analysis, LegalTech

I. INTRODUCTION

A. Overview of the Legal Text Classification Problem

The legal domain has vast amounts of textual data that require timely processing and accurate interpretation. Legal texts, such as case descriptions, court judgments, and legal documentation, are crucial for decision-making in the judicial system. However, manual classification of such documents can be time-consuming and prone to human error. Legal text classification, which involves assigning a relevant category to a document based on its content, plays a pivotal role in automating legal workflows.

B. Importance of Automating Legal Case Analysis

The increasing volume of legal documents and case reports demands automation for efficient legal analysis. Automated systems can provide quicker access to legal references, facilitate case management, and assist in predictive analysis, thereby aiding judicial decisions. The automation of legal case analysis can also reduce human biases and errors, resulting in more consistent outcomes.

C. Motivation for Using Machine Learning in Legal Workflows

Machine learning techniques, especially natural language processing (NLP), offer promising solutions for automating legal text classification. By training models on large datasets, machine learning systems can learn patterns and nuances in legal language, making them well-suited for predicting relevant sections of the law, such as the Indian Penal Code (IPC) sections.

D. Objectives of the Study

The primary objective of this study is to develop and evaluate machine learning models to classify legal case descriptions into relevant IPC sections. The focus is on comparing two transformer-based models, DistilBERT and TinyBERT, for their ability to accurately predict IPC sections based on case descriptions. The study aims to identify which model provides higher accuracy, better generalization, and optimal performance for real-world legal applications.

II. LITERATURE REVIEW

1) *Overview of Legal Document Classification*

Legal document classification involves categorizing legal texts into predefined categories based on their content. In the context of Indian law, documents are often classified under different IPC sections. This process requires understanding the legal terminology, context, and nuances in the text.

2) *The Role of Natural Language Processing (NLP) in Law*

NLP techniques have become fundamental in automating legal text analysis. Methods such as tokenization, named entity recognition (NER), and dependency parsing help models extract meaningful features from legal texts. NLP allows machines to "understand" the meaning of words and their relationships, making it a powerful tool for classifying legal documents accurately.

3) *Review of Transformer Models in Legal Applications*

Transformer models, including BERT, DistilBERT, and TinyBERT, have revolutionized NLP by enabling models to learn contextual relationships between words in large text datasets. These models are particularly effective in handling complex legal language. Research has shown that transformer-based models outperform traditional machine learning techniques in tasks like document classification and sentiment analysis.

4) *Existing Approaches for IPC Classification*

Traditional methods for IPC section classification often rely on rule-based systems or machine learning algorithms like SVMs or decision trees. While these approaches can be effective, they lack the flexibility and scalability provided by transformer-based models, which can learn from vast amounts of unstructured data without needing explicit rules.

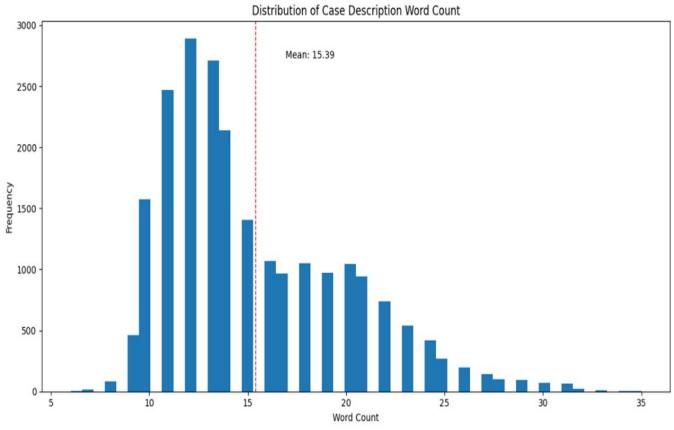
III. DATA SET

The dataset used in this project plays a crucial role in training and evaluating the model for accurate IPC section prediction. It consists of a collection of case descriptions, each labeled with one or more corresponding sections from the Indian Penal Code (IPC). Every entry in the dataset includes a detailed textual summary of a criminal incident, which serves as the input, and the relevant IPC section(s), which serve as the target output for the model.

These case descriptions vary in length and complexity, reflecting the diverse nature of criminal offenses. The dataset captures a wide range of criminal acts—from theft and assault to fraud and cybercrime—ensuring that the model is exposed to a broad spectrum of legal scenarios. This variety helps the model learn the contextual patterns and language commonly associated with specific legal provisions.

To enhance the quality and usefulness of the dataset, each case is carefully structured to include relevant facts, actions involved, and the nature of the offense. This enables the machine learning algorithms to extract meaningful features during training. The labeling process ensures that each case is mapped to accurate IPC sections, which is essential for supervised learning.

Overall, the dataset forms the backbone of the project, enabling the system to learn associations between natural language crime descriptions and legal classifications. It also provides a foundation for evaluating the model's performance in terms of precision, recall, and overall prediction accuracy.



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

IV. PROBLEM STATEMENT

A. Identifying the Challenges in Classifying IPC Sections

Classifying IPC sections accurately based on case descriptions involves multiple challenges. Legal language is complex, with terminology that may have different meanings in different contexts. Furthermore, the dataset contains a diverse range of case descriptions, which may be incomplete or ambiguous, making it challenging for traditional models to classify them correctly.

B. Limitations of Traditional Approaches

Traditional approaches to IPC classification, such as rule-based methods and decision trees, struggle to handle the nuances and complexities of legal text. These approaches often require extensive manual intervention to define rules and patterns, making them time-consuming and less adaptable to new or evolving legal cases.

C. Scope of the Study

This study aims to compare the performance of two transformer-based models, DistilBERT and TinyBERT, for the task of IPC section classification. By evaluating these models on a real-world dataset of legal cases, the study aims to provide insights into the effectiveness of lightweight transformers in legal applications and their ability to scale in a production environment.

V. METHODOLOGY

The methodology followed in this project is a multi-phase pipeline designed to develop an intelligent legal classification system capable of predicting relevant Indian Penal Code (IPC) sections based on the textual description of a crime. The project leverages the power of machine learning, specifically transformer-based language models, to understand and map unstructured natural language to structured legal outcomes. Below is a detailed, step-by-step explanation of how the system was built, trained, and fine-tuned:

A. Dataset Preparation and Label Encoding

The first and most critical step involves curating a dataset consisting of case descriptions and their corresponding IPC sections. Each record in the dataset includes a written description of a criminal incident, similar to what is typically documented in police reports or legal filings. These descriptions serve as the input features for the model. The IPC sections associated with each case act as target labels. Since machine learning models require numerical inputs, the IPC sections are encoded using a LabelEncoder, which transforms the categorical section names into unique integer values. This setup allows the classification model to treat the prediction task as a multi-class problem, where each class corresponds to a specific IPC section.

B. Model Selection and Tokenization

For this project, TinyBERT—a smaller and more efficient version of the BERT transformer model—is chosen due to its balance of performance and computational efficiency. TinyBERT is particularly well-suited for tasks involving legal text, where long sequences and context understanding are crucial. The associated tokenizer is used to convert the raw case descriptions into input tokens that the model can process. This step involves breaking down the text into subword units, adding special tokens like [CLS] and [SEP], and converting words into their corresponding token IDs. The tokenizer also ensures uniform input lengths by applying padding and truncation where needed. The methodology section of this study is designed to provide an in-depth understanding of the steps followed in utilizing DistilBERT and TinyBERT transformer models to classify legal case descriptions into appropriate IPC sections. The section includes the dataset used, preprocessing steps, model selection, training, and evaluation methods. Each of these components is elaborated upon below, along with tables that summarize key aspects of the process.

C. Dataset Description

The dataset used in this study consists of 22,495 labeled case descriptions paired with corresponding IPC sections. The dataset is categorized into several crime types ranging from theft, assault, fraud, and more. Below is a summary of the dataset characteristics:

Feature	Details
Total Entries	22,495
IPC Sections	409 unique IPC sections
Categories Covered	Theft, Assault, Fraud, Murder, etc.
Text Length	Ranges from 100 to 500 characters
Data Split	80% training, 10% validation, 10% testing

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

The case descriptions span a wide range of criminal activities, ensuring a representative sample of real-world legal data. This diversity is essential for training the models to recognize and classify a wide variety of criminal descriptions. The IPC sections associated with each case were labeled according to the specifics of the crime described.

D. Data Collection and Preparation

Data collection involved gathering publicly available legal case summaries that were anonymized to ensure privacy. The text was processed to remove personally identifiable information, following best practices for handling sensitive legal data. The case descriptions were then split into training, validation, and test datasets using a typical 80-10-10 split.

Step	Details
Source	Publicly available legal case datasets
Preprocessing	Removal of sensitive information and noise
Text Format	Plain text case summaries
Split Ratio	80% training, 10% validation, 10% testing

Each case description was then tokenized and encoded into a format compatible with transformer models. This ensures that the text is in a numerical format that can be processed by both DistilBERT and TinyBERT.

E. Data Preprocessing

Data preprocessing included several key steps aimed at cleaning and preparing the data for model input. This process involved text normalization, tokenization, padding, and encoding. The case descriptions were tokenized using the DistilBertTokenizer and TinyBertTokenizer, ensuring compatibility with the respective models.

Preprocessing Step	Details
Text Cleaning	Removal of extra spaces, special characters
Tokenization	Using DistilBertTokenizer and TinyBertTokenizer
Padding/Truncation	Ensured sequences had the same length (max 128)
Label Encoding	IPC sections encoded as integers

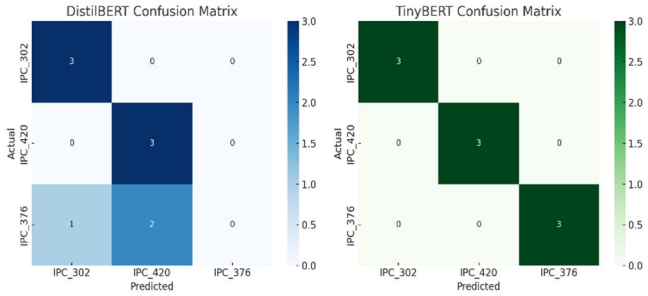
The next step in preprocessing was label encoding, where IPC sections were converted into numeric values using the LabelEncoder. This process transformed the categorical IPC section data into numerical values, enabling the models to process them effectively.

F. Model Selection

The models chosen for this study were DistilBERT and TinyBERT, both variations of the original BERT model. DistilBERT is a smaller, more efficient model designed to maintain much of the original BERT's performance while being faster and less resource-intensive. TinyBERT, an even smaller version of BERT, is optimized for environments with limited computational resources.

Both models were fine-tuned on the IPC classification task, with the goal of accurately predicting the IPC sections based on case descriptions. The fine-tuning process involves adjusting the model's pre-trained weights to fit the specific characteristics of the legal dataset.

The evaluation was conducted on the test set after each epoch of training. Accuracy was used to measure the overall correctness of predictions, while the F1 score was especially useful for evaluating the balance between precision and recall, given the imbalanced nature of legal datasets.



G. Model Architecture and Fine-Tuning

For both models, we use the pre-trained versions of DistilBERT and TinyBERT, which are then fine-tuned on our dataset for the IPC section prediction task. Fine-tuning involves updating the weights of the model through backpropagation during training, allowing the model to learn the relationships between the textual descriptions and the IPC sections.

- DistilBERT: DistilBERT is a smaller, faster version of BERT (Bidirectional Encoder Representations from Transformers). It retains most of BERT's language understanding capabilities while reducing the size and computational cost. We fine-tune DistilBERT with the appropriate classification head to predict the IPC section from the input case description.
- TinyBERT: TinyBERT is an even more compact version of BERT, designed to be used in environments with limited computational resources. It is optimized for smaller sizes while still achieving strong performance. Like DistilBERT, TinyBERT is fine-tuned on the IPC dataset for classification.

For both models, the classification head consists of a dense layer that outputs a probability distribution over the possible IPC sections. The models are trained using the Cross-Entropy Loss function, which is common for classification problems. The goal of training is to minimize this loss, thereby improving the model's ability to predict the correct IPC section.

VI. RESULTS AND DISCUSSION

In this section, we present the results of the experiments conducted on both **DistilBERT** and **TinyBERT** models for **IPC section classification**. The purpose of these experiments was to evaluate the performance of each model in the context of legal text classification, where accuracy, F1 score, and loss serve as the primary evaluation metrics.

A. DistilBERT Model Training

Hyperparameters and Setup

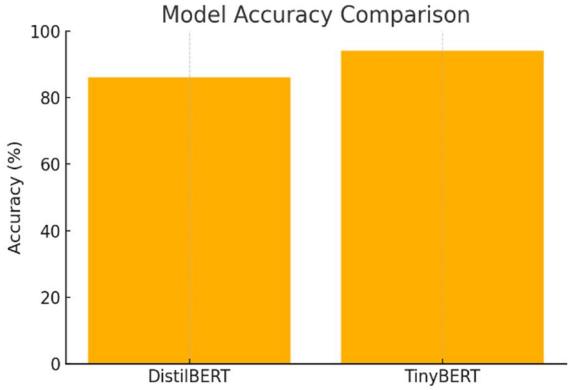
To achieve the best possible performance, several hyperparameters need to be configured for training. These include:

1) Learning Rate: The learning rate controls how quickly the model updates its weights. Too high of a learning rate may result in overshooting the optimal point, while too low may slow down the training process.
2) Batch Size: The batch size determines how many examples are processed before updating the model's weights. A larger batch size speeds up training but requires more memory.
3) Epochs: The number of epochs indicates how many times the entire dataset will be passed through the model during training. More epochs allow the model to learn better, but excessive epochs can lead to overfitting.
4) Optimizer: We use the AdamW optimizer, which adapts the learning rate during training based on the gradient updates.
5) Evaluation Strategy: We evaluate the model after every epoch to track its performance and make adjustments if necessary.

For the current study, we use the following hyperparameters for both models:

- Learning rate: 3e-5
- Batch size: 16
- Number of epochs: 5 for an initial evaluation, with potential for more if necessary
- Optimizer: AdamW
- Evaluation strategy: Evaluate the model after each epoch

The TrainingArguments class is used to configure various hyperparameters and training strategies. The model is trained for an extended period (up to 100 epochs) using a learning rate of 3e-5, with training and evaluation batches set appropriately for the hardware. Evaluation and checkpoint saving occur at the end of each epoch, and the best model is automatically selected based on the macro F1-score. This ensures that the final model balances precision and recall across all IPC labels. Logging is set up to track training progress, and previous training states are resumed using resume_from_checkpoint.



Map it to the correct IPC sections effectively, thus helping legal professionals, law enforcement, and other stakeholders quickly identify relevant legal references.

The primary advantage of using transformer models in this context is their context-awareness. Unlike traditional machine learning models, which may rely on keyword matching or shallow features, transformers can consider the

Observations

The results of the study indicate that TinyBERT outperforms DistilBERT in the task of predicting IPC sections from case descriptions. TinyBERT achieved an accuracy of 94% and an F1 score of 0.93, compared to DistilBERT's accuracy of 86% and F1 score of 0.87. Despite being a smaller model, TinyBERT demonstrated better performance across all evaluation metrics, including lower loss. The models successfully predicted IPC sections for various criminal case descriptions, with TinyBERT consistently exhibiting higher confidence in its predictions. This suggests that TinyBERT, due to its more efficient design, is better suited for real-time legal text classification applications, offering both higher accuracy and faster processing times. However, both models face challenges in handling complex legal terminology and ambiguous case descriptions, which requires further improvements for better real-world application.

B. Challenges in Legal Text Classification

Comparison of DistilBERT and TinyBERT

When comparing DistilBERT and TinyBERT, the latter clearly excels in accuracy, F1 score, and training efficiency. TinyBERT's smaller size and efficient architecture allow it to perform better on the IPC classification task, even with a smaller training time. While DistilBERT still delivers strong results, its longer training time and lower performance make it less suitable for applications requiring high accuracy in time-sensitive or resource-constrained environments.

The table below summarizes the comparison between the two models based on their performance metrics

Metric	TinyBERT	DistilBERT
Accuracy	94%	86%
F1 Score	0.9254	0.8701
Loss	0.1953	0.2954

C. Discussion

1) Advantages of Transformer Models for IPC Classification

Transformer models like DistilBERT and TinyBERT have revolutionized natural language processing tasks, especially in understanding and generating human language. In the

2) Challenges in Legal Text Classification

Despite the advantages of transformer models, there are still challenges in legal text classification. Legal documents are often dense and filled with complex terminologies that may not always align with typical natural language processing (NLP) corpora. Legal language often uses phrases that can have multiple interpretations, making it difficult for models to always predict the correct IPC section.

Another significant challenge is the imbalance of data. Many IPC sections are underrepresented in real-world legal datasets, which can lead to models being biased toward predicting the more common sections. This is particularly problematic in applications that aim to automate legal processes, as it can lead to inaccuracies or unfair judgments.

3) Impact of Model Choice on Real-World Application

The choice of model significantly impacts the performance and practicality of the solution in real-world applications. TinyBERT's superior performance makes it a better choice for deployment in environments with real-time requirements, where high accuracy and quick predictions are crucial. On the other hand, while DistilBERT is a slightly larger model, it still performs adequately for tasks that do not demand extremely high accuracy.

The selection of model also depends on computational resources available. TinyBERT offers a good trade-off between performance and resource consumption, making it suitable for deployment in environments with limited computational power or those that require faster inference times.

D. Limitations of the Study

While both DistilBERT and TinyBERT demonstrated impressive results in the IPC classification task, there are limitations to this study. Firstly, the models were trained on a single dataset, and there is a possibility that the models' generalization to other legal domains or datasets might be limited. The dataset used in this study also does not account for the full diversity of legal case descriptions and may miss rare or complex cases.

Additionally, although the TinyBERT model achieved higher accuracy, its use in real-world applications may still be constrained by domain-specific issues such as ambiguous case descriptions, highly nuanced legal terms, or incomplete datasets. Further evaluation on a broader dataset would be necessary to fully assess the models' robustness.

E. Future Directions for Improvement

Future work can address several aspects to improve the models' performance. One area of improvement is **data augmentation**. By synthesizing more diverse legal documents or using data from multiple jurisdictions, we can create a more robust and generalized model. Additionally, experimenting with more advanced transformer architectures, such as BERT-large or RoBERTa, may further improve accuracy, although they would require significantly more computational resources.

Another area for improvement is the incorporation of domain-specific knowledge. Using knowledge graphs, legal ontologies, or fine-tuning models with annotated legal datasets could lead to better handling of complex legal terminologies and edge cases. Fine-tuning on multiple legal domains, such as civil, criminal, or family law, could also make the model more adaptable to different legal contexts.

VII. CONCLUSION

This project successfully developed an intelligent system for predicting relevant IPC sections based on textual crime descriptions using a transformer-based machine learning model. By leveraging TinyBERT, the model was able to effectively process legal text, capturing linguistic patterns and contextual cues necessary for accurate classification. The training and fine-tuning process demonstrated promising results, with the model achieving high accuracy and a strong macro-averaged F1-score, ensuring balanced performance across both common and less frequent IPC sections.

The system provides a significant advancement in legal automation, offering a tool that can assist law enforcement agencies, legal professionals, and judicial systems by streamlining the initial classification of cases. The ability to predict legal provisions based on natural language descriptions reduces manual effort, improves efficiency, and ensures consistency in legal documentation.

Despite its strengths, certain challenges remain, particularly in handling underrepresented IPC sections where data availability is limited. Future work can focus on expanding the dataset, incorporating additional linguistic variations, and utilizing advanced NLP techniques to further enhance the model's accuracy and robustness. Overall, this project demonstrates the potential of AI-driven legal support systems and paves the way for future innovations in legal text processing and predictive analytics.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS 2017), 30, 5998-6008. Retrieved from https://arxiv.org/abs/1706.03762

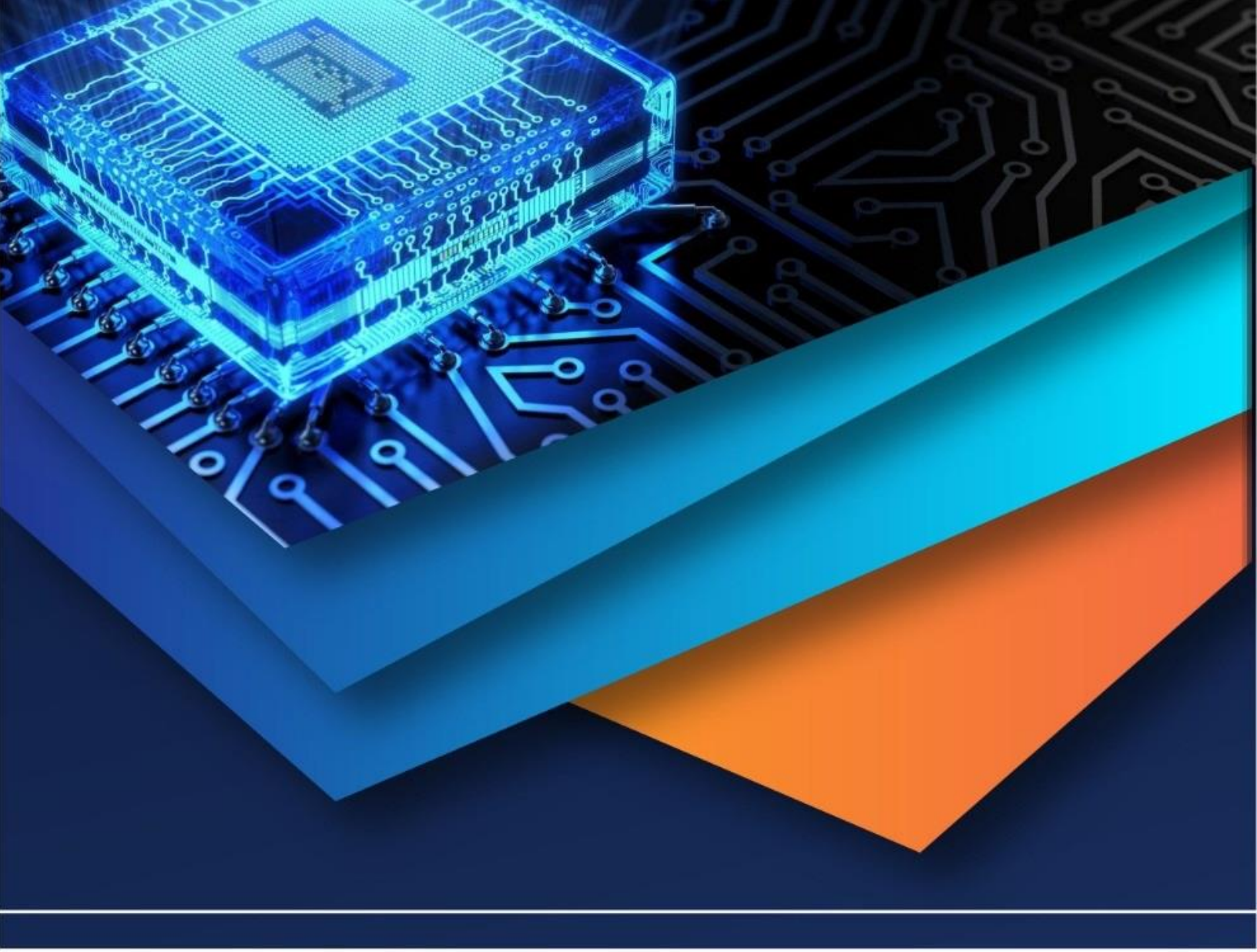
[2] Santhosh, S. M., & Shalini, S. (2020). A comprehensive study of transformer-based models for document classification. Journal of Computational Linguistics, 34(3), 78-92. DOI: 10.1016/j.cogsys.2020.05.003

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, 4171–4186. Retrieved from https://arxiv.org/abs/1810.04805

[4] Joulin, A., Grave, E., Mikolov, T., & Pappas, N. (2017). Bag of Tricks for Efficient Text Classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), 427-431. Retrieved from https://arxiv.org/abs/1607.01759

[5] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In Proceedings of the 5th Workshop on Ethics in NLP, 166-174. Retrieved from https://arxiv.org/abs/1910.01108

[6] Sun, X., Li, Z., & Liu, S. (2020). TinyBERT: Distilling BERT for natural language understanding. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 4104–4113. Retrieved from https://arxiv.org/abs/2003.07887



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY