BERT-Based Fine-Tuning for Automated Tagging of Robbery Crime Narratives

Lenin G. Falconi[1 [0000−0003−4402−6643]],
Myriam Hernandez-Alvarez[1 [0000−0003−4718−0400]], and
Ángel Leonardo Valdivieso Caraguay[1 [0000−0002−3502−020X]]

Escuela Politécnica Nacional, Quito 170525, Ecuador
https://www.epn.edu.ec
{lenin.falconi,myriam.hernandez,angel.valdivieso}@epn.edu.ec

Abstract. Accurate classification of crime narratives is vital for reliable public safety statistics. In Ecuador, Comisión Especial de Estadística de Seguridad, Justicia, Crimen y Transparencia (CEESJCT) manually categorizes robbery incident reports, which is a time-consuming process. While transformer-based models have shown success in natural language processing tasks, their application to Ecuadorian legal and security texts in the Spanish language, underexplored. This study addresses this gap by developing an automated classification system using a BERT model tailored to Spanish robbery narratives. Utilizing transfer learning and subsequent fine-tuning on an expanded, labeled dataset, the system significantly improved classification performance. Initial transfer learning achieved moderate accuracy (80.5%) but faced difficulties with semantically similar categories. Fine-tuning notably increased minority-class recall (up to 30%) with an improved accuracy (90.3%). The final implementation, which increased the number of categories to 11, achieved 95.5% accuracy with robust and consistent results on both police and judicial narratives. Collaboration with Ecuadorian institutions, including the Fiscalía General del Estado (FGE) and Instituto Nacional de Estadística y Censos (INEC), ensured model credibility.

Keywords: Legal Natural Language Processing · NLP · Natural Language Processing · transformers · fine-tuning · transfer learning.

1 Introduction

Deep learning models (DL) have yielded remarkably successful outcomes across a wide range of applications (e.g., image classification, object detection, natural language processing)[15]. In various text classification tasks, such as sentiment analysis, news categorization, question-answering, and natural language inference, DL has outperformed traditional *Machine Learning (ML)* methods[12], a testament primarily to its inherent generalization and robustness[4]. However, depending on the problem to be solved, designing a DL architecture demands a substantial volume of data to attain the desired generalization performance. For

example, in image processing and computer vision, datasets such as ImageNet (14 197 122) and Microsoft COCO (2,5 million) provide extensive data per category[15], while in *Natural Language Processing (NLP)*, resources like WebTex, composed of millions of web pages, have been pivotal in training models such as GPT-2[16].

Although the success of DL hinges on large datasets, specialized hardware [16][14], and model architecture, for niche domains with limited data, *Transfer Learning (TL)* and *Fine-Tuning (FT)* enable knowledge transfer from source to target domains by reusing early-layer features and adapting task-specific layers [26][9].The legal documentation domain is a promising field for applying NLP, where digital information can be leveraged to develop tools that optimize legal workflows. However, applications face language limitations, as LexGlue, the most widely used benchmark, focuses on English despite the growing interest in judicial outcomes prediction[13, 10, 24], legal article prediction from text and document classification [25, 7]. While transformers models like mBERT enable cross-lingual transfer, legal terminologies and system differences necessitate domain-specific pretraining (e.g., LegalBERT [6]) or hybrid approaches [11], particularly for low-resource languages [18].

In Ecuador's penal code (COIP), robbery is defined in Article 189. However, subclassifications such as home robbery[1], street robbery[2], robbery of businesses[3], vehicle part theft[4], car theft[5], and motorcycle theft[6] are used primarily for public security analytics and manually identified by the *Comisión Especial de Estadística de Seguridad, Justicia, Crimen y Transparencia* (CEESJCT) using *crime incident reports*[7]. To automate this, we propose a transformer-based classifier that uses multilingual DistilBERT [17]. We implemented a three-phase approach: **Phase 1: Classification with transfer learning** the transformer model is used as a feature extractor followed of full connecting layers and dropout achieving 80.5% accuracy (F1: 0.80). Even though the model's bidirectional attention mechanism and subword tokenization (WordPiece) proved essential for capturing semantic relationships in crime narratives, particularly when some words remain misspelled and uncorrected for legal reasons protecting the victim's statement, it struggled with minority classes. **Phase 2: Fine-Tuning for Contextual Adaptation** Using a learning rate 5×10^{-5}) improved accuracy to 90.3% (F1: 0.90) by resolving semantic ambiguities and achieving significant gains for less frequent categories. Additionally, a Flask web application was deployed for real-time predictions, which demonstrated an accuracy of 89% on unseen CEESJCT data. Stages 1 and 2 used a dataset of 431,669 robbery reports (2014–2022), tokenized into sequences of up to 300 words, and split into

[1] robo a domicilio
[2] robo a personas
[3] robo a unidades económicas
[4] robo de bienes, accesorios y autopartes de vehículos
[5] robo de automóviles
[6] robo de motocicletas
[7] Noticia del Delito

training (63%), validation (16%), and test (21%) sets. **Phase 3: Scaling to Complex Taxonomies** model is fine tuned to predict a total of 11 categories, extending the original 6. A merged validated dataset of 1.1 million narratives was developed by combining police reports with criminal complaints filed in the prosecutor's office. The model was trained using TPU to process the entire dataset. A notable improvement in accuracy performance was achieved (95.5% accuracy with an F1-score of 0.95).

The structure of this article is organized as follows. Section 2 provides a comprehensive literature review on text classification challenges, along with theoretical foundations of *Transformers*, TL and FT. Section 3 describes the methodology developed for dataset construction and model training. Experimental results are presented in Section **??**. Finally, Section **??** discusses the findings while Section **??** outlines future research directions.

2 Foundational Concepts and Related Works

In this research, we propose a transformer-based model for robbery offense classification aligned with categorical frameworks established by both the Prosecutor's Office and INEC. Legal documents present unique challenges compared to general-domain texts, characterized by three distinctive features: (1) inherent structural complexity, (2) substantial document length, and (3) specialized juridical terminology. These characteristics have led to the emergence of Legal Natural Language Processing (LNLP) as a specialized NLP subdomain [5, 27].

Within this paradigm, Legal Text Classification (LTC) refers to the task of categorizing legal documents into predefined juridical categories. While LTC shares fundamental principles with general TC, it introduces specific technical challenges including but not limited to: high class cardinality (typically ranging from dozens to hundreds of categories), multi-label classification requirements, and the need for domain-specific feature engineering. The complexity of LTC increases exponentially with the number of potential legal categories and their hierarchical relationships within juridical systems.

2.1 Legal Natural Language Processing

NLP for legal documents primarily focus on two categories of approaches: *Embedding Methods* and *Symbol-Based Methods* [27]. Embedding Methods, such as transformer-based models (e.g., BERT), derive high-level representations from large corpora and offer strong predictive performance, but their opacity raises concerns about interpretability. Conversely, Symbol-Based Methods explicitly model legal knowledge and facilitate reasoning over structured representations, thereby providing greater transparency, though often at the expense of accuracy. Recent work in explainable AI seeks to address this limitation in embedding methods.For example, attention-based techniques such as *attention rollout* and *attention flow* aggregate attention weights across layers in transformers, offering insight into model decisions and revealing a degree of inherent explainability [1].

The Transformer architecture, introduced in [22], represents a significant advancement in NLP by incorporating a self-attention mechanism that allows deep learning models to effectively capture long-range dependencies and facilitates efficient parallel training. This innovation supplanted the previous state-of-the-art recurrent neural network (RNN) approaches, including Long Short-Term Memory (LSTM) networks [21]. Transformers employ an encoder-decoder architecture augmented by self-attention and transfer learning (TL) methodologies.

BERT [8], a notable application of the Transformer encoder, generates deeply bidirectional contextual representations through exclusively encoder-based architecture. Its pre-training utilizes self-supervised objectives such as masked language modeling (MLM) and next sentence prediction (NSP), enabling the model to acquire rich contextual information from extensive corpora.

2.2 Related Works

This section reviews related work in LTC research, focusing on three critical dimensions: (1) architectural innovations, (2) dataset characteristics and domain adaptation strategies, and (3) comparative performance. Table 1 provides a comparison of these approaches. While existing research has predominantly focused on English-language legal corpora our work represents, to our knowledge, innovative in exploring transformer architectures in Spanish legal documents attached to Ecuadorian reality.

In their research work, [19] tackle large-scale multi-label legal text classification using transformer architectures (BERT, RoBERTa, DistilBERT, XLNet, M-BERT). Through systematic evaluation on the JRC-Acquis (multilingual) and EURLEX57K (English) datasets, they demonstrate that combining gradual layer unfreezing, discriminative learning rates, and domain-specific pretraining achieves state-of-the-art performance; surpassing LSTM baselines by significant margins. The authors further propose standardized dataset splits to facilitate reproducible research, establishing transformers as the new benchmark for legal document classification with complex taxonomies like EuroVoc.

[20] extends this research to zero-shot cross-lingual transfer using multilingual transformers (M-BERT, M-DistilBERT). By fine-tuning models exclusively on English EURLEX57K documents and evaluating on French/German translations, they achieve target-language performance comparable to models trained on all three languages. Key innovations include continued pretraining on legal corpora and progressive unfreezing during fine-tuning, highlighting transformers' ability to bridge linguistic gaps in low-resource legal NLP scenarios.

The cross-linguistic applicability of transformers is further validated by [2] for the classification of Turkish legal texts. Their comparative analysis reveals that transformer-based models outperform traditional ML methods (e.g., SVM, logistic regression) in accuracy, despite limited training data. While specific classification categories remain unspecified, this study crucially demonstrates that transformers' superiority extends beyond English to other languages.

Addressing the challenge of document length, [23] systematically evaluates BERT adaptations for U.S. Supreme Court decisions from the Supreme Court

Database (SCDB). They find that domain-adapted Legal-LongFormer and Legal-BERT outperform their general-domain counterparts. This underscores the significance of domain-specific pretraining and optimized strategies for handling lengthy documents in legal text classification. [19]

Table 1. Comparative analysis of legal text classification approaches

Ref	Model	Task	Dataset	Language	Key Results
[19]	BERT, RoBERTa, DistilBERT, XLNet, M-BERT	Multi-label classification	JRC-Acquis, EURLEX57K	Multilingual, English	0.661 (F1) on JRC-Acquis
[20]	M-BERT, M-DistilBERT	Cross-lingual transfer	EURLEX57K Extended	Models pre-trained in Multilingual. FT in English	34% improvement on French, 87% improvement on German
[2]	Transformer-based	Document classification	Turkish Legal Corpus	Turkish	DL models improve over traditional ML
[23]	BERT, RoBERTa, Legal-BERT, LongFormer, Legal-LongFormer	Long document classification	SCDB	English	80.1% accuracy (15 categories), 60.9% accuracy (279 categories) with Legal-BERT

2.3 Discussion, Findings and Contribution

The reviewed literature indicates that BERT models are widely utilized for legal text classification (LTC). Transformer-based architectures consistently outperform traditional machine learning methods in this domain[2]. Fine-tuning pre-trained BERT weights to the target task is a standard approach. Also, our methodology is comparable to that of [23], wherein models are trained for both broader and fine-grained legal categories. A recurrent challenge with BERT-based approaches concerns the processing of lengthy documents, due to BERT's tokenization limits.

Distinct from previous works, our study specifically addresses the classification of robbery types. Most literature focuses on broader legal categories, and while some studies may implicitly include robbery within categories such as "criminal law" or "property offenses," explicit attention to robbery classification is absent, marking our focus as potentially novel and expands the exploration in application of transformer-based models in legal domain.

Another important difference of this work is related to the nature of its data. While it is recognized that LNLP has unique characteristics compared to general NLP, the narratives in our dataset originate from three principal sources: citizen crime reports, police accounts, and prosecutor's office reports. Consequently, our dataset encompasses a blend of technical texts and non-technical victim-generated descriptions. These factors substantiate the application of BERT-based models for the current text classification task.

Building upon these foundations, this study makes three contributions to LNLP. Firstly, it focuses on the specific context of *robbery* as defined by Ecuadorian law, thus addressing a distinct and pertinent crime type within the local legal landscape. Secondly, the research tackles the complex task of narrative tagging within crime reports, which are characterized by variable lengths and the coexistence of technical and colloquial language, including frequent misspellings due to direct transcription from citizens' testimonies. Third, it introduces an original Spanish-language dataset featuring crime narratives related to robbery and official classification labels.

The implemented model has been operationally adopted by the *Dirección de Estadística y Sistemas de Información*, automating the manual classification of crime narratives in robbery and contributing to statistical analysis by the aforementioned department. This work constitutes, to the authors' knowledge, the first application of Transformer-based models for analyzing Spanish-language Ecuadorian crime narratives.

3 Methodology

This section details the development of a transformer-based robbery classification model for Ecuadorian crime reports, structured around three pillars: dataset preparation, training protocol, and performance evaluation. The model was developed in three stages: initial transfer learning (TL) on a reduced dataset of 6 categories $D_{\kappa=6}$, followed by fine-tuning (FT) on the same dataset, and concluding with FT on an expanded dataset of 11 categories $D_{\kappa=11}$.

3.1 Dataset Generation

Basic Taxonomy ($D_{\kappa=6}$) Dataset To train the model, an initial target dataset $D^T_{\kappa=6,N}$ with 6 categories is generated by updating an SQL table of robbery records from CEESJCT (as of June 8, 2022), obtaining 671 708 records. Next the crime narratives are extracted and paired with the previous table using the record's key, yielding 671 146. Next, we cleaned the text sequences by Removing non-alphanumeric characters and converting text to lowercase by default (though distilbert-base-multilingual-cased preserves case sensitivity). The statistical analysis of word counts l_w of each document X_i from the robbery narratives gathered showed a mean (μ) = 98,34, a standard deviation of (δ) = 77,38; showing high variability in narratives lengths. The histogram of the word counts presents a Bimodal distribution, with peaks at $l_w \in \{7, 100\}$. Also $\max(l_w) = 914$ words. Quartile analysis showed a median of 52 words per X_i, with $q_1 = 33$, $q_3 = 137$, and an upper threshold $l_{w_{sup}} = 293$ words. Given that the DistilBERT model has a maximum sequence length of 512 tokens and produces 768-dimensional embeddings [17], we constructed dataset $D^T_{\kappa=6,N}$ consisting of narratives where the number of words of each sample X_i satisfies $35 < l_{w_i}(X_i) \leq 300$. This selection resulted in a total of $N = 671\,708$ records.

The dataset $D^T_{\kappa=6,N}$ was partitioned into training ($D^T_{\kappa=6,train}$), validation ($D^T_{\kappa=6,valid}$), and testing ($D^T_{\kappa=6,test}$) subsets to facilitate model training and evaluation. The training set comprises 273,336 records (63.32% of the dataset), while the validation and test sets contain 68,333 (15.83%) and 90,000 (20.85%) records, respectively. The validation set was derived by splitting the training data in an 80:20 ratio.

Extended Taxonomy ($D_{\kappa=11}$) Dataset We constructed an extended target dataset comprising 11 distinct categories by integrating police reports with corresponding prosecutor's office records. This integration significantly expanded our training corpus, enhancing the model's exposure to diverse textual patterns. The final dataset contains 1 109 335 records, with the distribution across categories detailed in Table 2. The dataset creation process involved the following steps:

1. **Initial Data Extraction:**Crime reports from 2014–2022 were extracted and filtered for robbery cases from CEESJCT, yielding to 735 045 records.
2. **Narrative Retrieval:** Crime narratives are obtained for each of the previous records obtaining a total of 725 079. However, non-robbery cases are excluded if found. This produces a total of 723 435 records.
3. **Preprocessing:** lowercasing and removal of non-alphanumeric characters.
4. **Dataset Integration:**Police and prosecutor narratives were combined into a single dataset, yielding to 1 446 870 records.
5. **Quality Control:** statistical analysis over the text is carried out. Narratives with 50–400 words were retained (upper fence: 389.5 words), reducing the dataset to 1 140 728 records. Crime labels were also standardized (e.g. removing accents).
6. **Category refinement:** Non-compliant labels (48 categories) were discarded with the exception of the *OTHER ROBBERIES* category which is retained. The final dataset, with 1 109 335 records was randomly shuffled and
7. **Dataset Split:** The dataset is partition into training (807 468 samples $\approx$ 72%), validation (201 867 samples $\approx$ 18%), and test (100 000 samples $\approx$ 10%) sets.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. Proceedings of the Annual Meeting of the Association for Computational Linguistics pp. 4190–4197 (5 2020). https://doi.org/10.18653/v1/2020.acl-main.385, https://arxiv.org/abs/2005.00928v2
2. Akca, O., Bayrak, G., Issifu, A.M., Ganiz, M.C.: Traditional machine learning and deep learning-based text classification for turkish law documents using transformers and domain adaptation. 16th International Conference on INnovations in Intelligent SysTems and Applications, INISTA 2022 (2022). https://doi.org/10.1109/INISTA55318.2022.9894051

Table 2. Extended Taxonomy Dataset

Basic Taxonomy	Extended Labels	Total
HOME ROBBERY	HOME ROBBERY	172 264
STREET ROBBERY	STREET ROBBERY	421 497
BUSINESS ROBBERY	BUSINESS ROBBERY	74 088
VEHICLE PARTS AND ACCESSORIES THEFT	VEHICLE PARTS AND ACCESSORIES THEFT	154 546
CAR THEFT	CAR THEFT	90 038
MOTORCYCLE THEFT	MOTORCYCLE THEFT	119 128
NO INFORMATION	OTHER ROBBERIES	43 468
	WATER VESSEL ROBBERY	9 407
	SOCIAL ORGANIZATION ROBBERY	3 087
	EDUCATIONAL INSTITUTION ROBBERY	17 252
	PUBLIC INSTITUTION ROBBERY	4 560
Total		1 109 335

3. Allam, H., Makubvure, L., Gyamfi, B., Graham, K.N., Akinwolere, K.: Text classification: How machine learning is revolutionizing text categorization. Information 2025, Vol. 16, Page 130 **16**, 130 (2 2025). https://doi.org/10.3390/INFO16020130, https://www.mdpi.com/2078-2489/16/2/130/htm https://www.mdpi.com/2078-2489/16/2/130
4. Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Essen, B.C.V., Awwal, A.A.S., Asari, V.K.: The history began from alexnet: A comprehensive survey on deep learning approaches. CoRR **abs/1803.01164** (2018), http://arxiv.org/abs/1803.01164
5. Ariai, F., Demartini, G.: Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. ACM Computing Surveys **1** (10 2024). https://doi.org/10.1145/nnnnnnn.nnnnnnn, https://arxiv.org/abs/2410.21306v2
6. Chalkidis, I., Fergadiotis, M., Androutsopoulos, I.: Legal-bert: The muppets straight out of law school. arXiv preprint arXiv:2010.02559 (2020)
7. Clavié, B., Alphonsus, M., Clavié, C., Mundi, J.: The unreasonable effectiveness of the baseline: Discussing svms in legal text classification (2021), https://gitlab.com/jusmundi-group/public/Legal-svm-baselines
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423/
9. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification (2018), https://arxiv.org/abs/1801.06146
10. Kalia, A., Kumar, N., Namdev, N.: Classifying case facts and predicting legal decisions of the indian central information commission: a natural language processing approach. In: Advances in Deep Learning, Artificial Intelligence and Robotics, pp. 35–45. Springer (2022)
11. Liu, Z., Li, Z., Zhang, H.: Legal knowledge-enhanced language models for legal text processing. ICAIL (2022)

12. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning–based text classification. ACM Computing Surveys (CSUR) **54** (4 2021). https://doi.org/10.1145/3439726, https://dl.acm.org/doi/10.1145/3439726
13. Mumcuoğlu, E., Öztürk, C.E., Ozaktas, H.M., Koç, A.: Natural language processing in law: Prediction of outcomes in the higher courts of turkey. Information Processing & Management **58**(5), 102684 (2021)
14. Murphy, K.P.: Probabilistic machine learning: an introduction. MIT press (2022)
15. Pathak, A.R., Pandey, M., Rautaray, S.: Application of deep learning for object detection. Procedia Computer Science **132**, 1706–1717 (2018). https://doi.org/https://doi.org/10.1016/j.procs.2018.05.144, https://www.sciencedirect.com/science/article/pii/S1877050918308767, international Conference on Computational Intelligence and Data Science
16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
17. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv **abs/1910.01108** (2019)
18. Savelka, J., Ashley, K.D.: Cross-domain generalization and knowledge transfer in transformers for legal text. ICAIL (2021)
19. Shaheen, Z., Wohlgenannt, G., Filtz, E.: Large scale legal text classification using transformer models (10 2020), https://arxiv.org/abs/2010.12871v1
20. Shaheen, Z., Wohlgenannt, G., Mouromtsev, D.: Zero-shot cross-lingual transfer in legal domain using transformer models. Proceedings - 2021 International Conference on Computational Science and Computational Intelligence, CSCI 2021 pp. 450–456 (11 2021). https://doi.org/10.1109/CSCI54926.2021.00145, https://arxiv.org/abs/2111.14192v2
21. Tunstall, L., von Werra, L., Wolf, T.: Natural language processing with transformers. " O'Reilly Media, Inc." (2022)
22. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
23. Vatsal, S., Meyers, A., Ortega, J.E.: Classification of us supreme court cases using bert-based techniques (2023), https://huggingface.co/saibo/
24. Wang, Y., Gao, J., Chen, J.: Deep learning algorithm for judicial judgment prediction based on bert. In: 2020 5th International Conference on Computing, Communication and Security (ICCCS). pp. 1–6. IEEE (2020)
25. Yan, G., Li, Y., Shen, S., Zhang, S., Liu, J.: Law article prediction based on deep learning. In: 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C). pp. 281–284. IEEE (2019)
26. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? Advances in neural information processing systems **27** (2014)
27. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: How does nlp benefit legal system: A summary of legal artificial intelligence. Proceedings of the Annual Meeting of the Association for Computational Linguistics pp. 5218–5230 (2020). https://doi.org/10.18653/V1/2020.ACL-MAIN.466, https://aclanthology.org/2020.acl-main.466/