A CNN-BiLSTM Hybrid Architecture with Resampling Techniques for Arabic Legal Text Classification

Oussama Tahtah
L3IA Laboratory, Sidi Mohamed Ben Abdellah University, Fes, Morocco
oussama.tahtah@usmba.ac.ma (corresponding author)

Mohammed Bahbib
Engineering Sciences Laboratory, Sidi Mohamed Ben Abdellah University, Fes, Morocco
mohammed.bahbib@usmba.ac.ma

Ahmed Zinedine
L3IA Laboratory, Sidi Mohamed Ben Abdellah University, Fes, Morocco
ahmed.zinedine@usmba.ac.ma

Khalid Fardousse
L3IA Laboratory, Sidi Mohamed Ben Abdellah University, Fes, Morocco
khalid.fardousse@usmba.ac.ma

Received: 18 July 2025 | Revised: 8 September 2025 | Accepted: 15 September 2025

 DOI: https://doi.org/10.48084/etasr.13506

ABSTRACT

Arabic legal text classification has played a major role in improving judicial systems by automating the categorization of legal texts and facilitating access to legal information. Despite these benefits, developing a model to classify Arabic legal text faces significant challenges, including the rich morphology and the inherent complexities of the Arabic language. Additionally, the imbalanced distribution within the legal specialties adds more challenges to the development of such a model. To address these challenges, this paper proposes a hybrid Deep Learning (DL) model that combines Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks, using a pre-trained Arabic Bidirectional Encoder Representations from Transformers version 2 (AraBERTv2) model as a word embedding technique. Additionally, extensive experiments were conducted to explore the impact of resampling techniques on the legal text classification model and to achieve an equal class distribution. Furthermore, a newly collected Arabic legal dataset was used to evaluate the performance of the developed model, and several evaluation metrics were employed, including accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC). The findings demonstrate that our model yielded superior performance, with a score of more than 95% across all employed metrics. Moreover, the Random Oversampling (RO) technique showed the best results among other resampling techniques.

Keywords-Deep Learning (DL); Arabic Natural Language Processing (NLP); text classification; legal text; resampling

I. INTRODUCTION

The exponential growth of legal documents has presented a challenge for legal practitioners who must handle and evaluate massive amounts of textual information manually. Natural Language Processing (NLP) and Artificial Intelligence (AI) techniques can address these challenges by automating the processing and analysis of these texts [1]. One of the main tasks in the legal field is text classification, which allows efficient document organization, retrieval, and analysis. Several studies have developed classification models in different languages, particularly in English, Chinese, and other European languages. However, research in this field for the Arabic language remains limited due to the scarcity of Arabic legal datasets and the diversity of legal terminology among Arabic countries. These challenges add more difficulty to feature extraction and reduce the effectiveness of the classification model. Moreover, the skewed distribution among classes in legal datasets adds further challenges [2]. Although several techniques have been developed to address the problem of imbalanced datasets, including resampling techniques, the application of these techniques remains underexplored in the legal field, particularly in the classification of Arabic legal texts.

To this end, we propose a novel model for Arabic legal text classification that combines two Deep Learning (DL) architectures in conjunction with Arabic Bidirectional Encoder Representations from Transformers version 2 (AraBERTv2). In addition, three categories of resampling techniques were employed to address the issue of the imbalanced dataset. Furthermore, a newly collected dataset of Arabic legal cases was used to evaluate the efficacy of the proposed model. AraBERTv2 provides a useful contextual word representation, which can help the model handle word variations within the Arabic writing style. The combination of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks addresses the language complexity by extracting local patterns using CNNs and long dependencies using BiLSTM. Finally, a fully connected layer with a softmax activation function categorizes the input text into the corresponding class.

Earlier studies on classifying Arabic legal texts have mostly used traditional word representation methods along with Machine Learning (ML) and DL algorithms. For instance, authors in [3] classified Moroccan Arabic legislative texts across seven legal domains. They used several ML algorithms, including Support Vector Machines (SVM), Naïve Bayes (NB), Random Forest (RF), and K-Nearest Neighbor (KNN), and conducted experiments to evaluate the impact of stemming and class imbalance on model performance. The SVM classifier achieved an accuracy of 94.58% when stemming was applied, whereas KNN achieved the best results without minority classes, with 96.26% accuracy. In the same context, authors in [4] employed various ML algorithms and Term Frequency-Inverse Document Frequency (TF-IDF) to categorize Moroccan Arabic legal documents using a dataset that contained 1,452 legal texts. The SVM classifier achieved the best performance with an accuracy of 98.11%. To categorize Arabic legal questions, authors in [5] implemented Word2Vec and TF-IDF in conjunction with classifiers such as RF, SVM, and NB. They also applied Synthetic Minority Oversampling Technique (SMOTE) as an oversampling method to address the imbalanced characteristics of the dataset, which comprises 44,103 Arabic legal questions. The results indicated that the TF-IDF+RF model achieved the highest accuracy at 86%. Expanding on the same dataset, authors in [6] used a hybrid model that merges n-grams with TF-IDF and Word2Vec. The results revealed a significant improvement in classification accuracy, reaching 87.4% with the TF-IDF+SVM model. These results highlight the effectiveness of integrating n-grams with TF-IDF to improve Arabic legal text classification performance. The application of pre-trained Arabic transformer models in the legal field has not been thoroughly explored. Nevertheless, authors in [7] developed a model for the classification of Moroccan commercial court verdicts. This model integrated Convolutional Long Short-Term Memory (CLSTM) with BERT to effectively capture significant textual elements. They created a new dataset that included 2,821 Arabic documents sourced from Moroccan commercial courts. The BERT-CLSTM model attained a remarkable accuracy of 93.55%.

Several research studies have investigated the use of balancing approaches in various Arabic NLP tasks. Authors in [8] showed that the SMOTE Edited Nearest Neighbor (SMOTE-ENN) technique improved model performance for Arabic fake news detection, achieving 85% accuracy using the BiLSTM architecture. Similarly, authors in [9] used resampling techniques for Arabic sentiment analysis. They demonstrated that using oversampling methods with RF achieved a promising F1-score of 0.99, whereas undersampling techniques showed poor results compared to oversampling.

However, the impact of resampling techniques on Arabic legal text classification remains unexplored, as does the need for a hybrid model that can benefit from different DL architecture characteristics. This critical gap necessitates more effort and research. Our work addresses this gap by providing a hybrid model that leverages both CNN and BiLSTM and explores the effects of resampling techniques on Arabic legal text classification. Our main contributions are as follows:

- Construct a dataset of Arabic legal texts for classification tasks.
- Propose a CNN-BiLSTM hybrid model that effectively combines local feature extraction with sequential modeling for Arabic legal texts.
- Provide a comparative study regarding the impact of various resampling strategies on Arabic legal text classification performance.

II. METHODOLOGY

In this section, we present a detailed overview of the dataset, DL architectures, and resampling techniques employed in this work.

A. Dataset Creation and Annotation

Due to the scarcity of publicly available datasets containing Arabic legal text, this study constructed a dataset comprising 14,799 judicial cases from the Moroccan Court of Cassation. This dataset is systematically categorized into six distinct legal domains: commercial law, civil law, administrative law, criminal law, personal status law, and social law. The cases were obtained from the Court's official public website [10], which provides open-access legal archives in PDF format. A custom web-scraping procedure was implemented to systematically download the relevant case files, followed by automated PDF-to-text extraction to obtain the raw textual content. Only publicly available documents were collected, and no restricted or confidential materials were accessed.

These legal documents in the dataset follow a standardized structure characteristic of cassation court documentation, as illustrated in Figure 1. The textual organization begins with a metadata section including case reference numbers, adjudication dates, and legal category, followed by a section that contains detailed party information identifying defendants, plaintiffs, and their legal representatives. The substantive content encompasses the complete procedural history, tracing the legal journey from initial filing through successive appeals, incorporating judicial reasoning at each level and culminating in the cassation court's final decision.

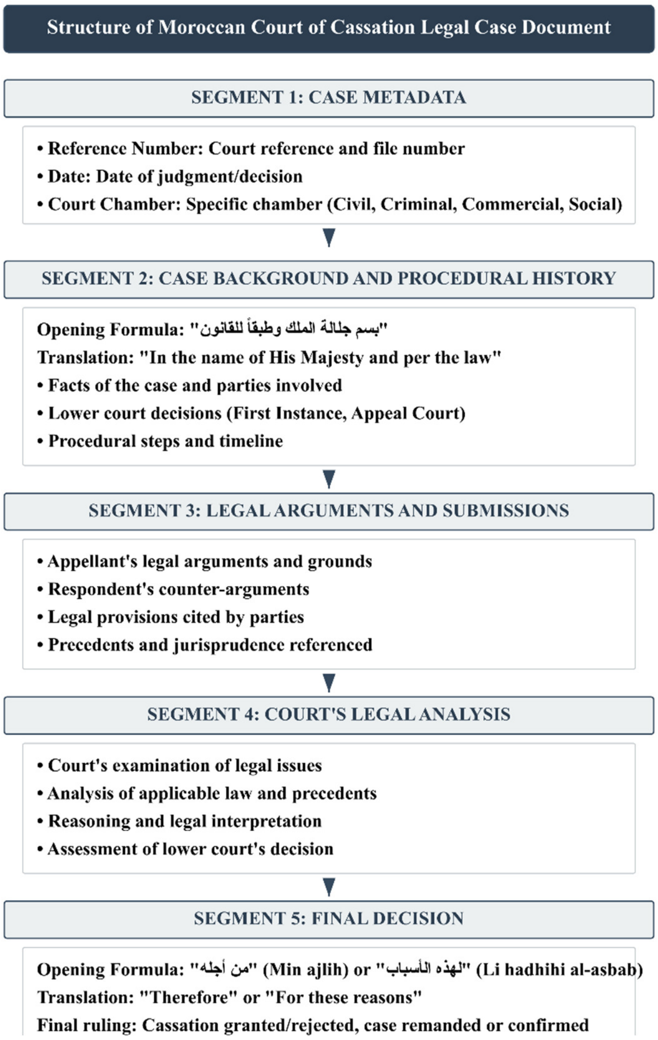

Fig. 1. Structure of a Moroccan cassation court document.

The annotation process leveraged the structure of cassation court documents through an automated extraction protocol. Each case document contains a standardized metadata header that explicitly specifies the class of the case alongside other information such as case reference and date. Our automated annotation script systematically parsed these headers to extract the ground-truth labels, subsequently removing the metadata section from the input text to prevent information leakage during model training. This approach ensured consistent and accurate labeling while eliminating the potential for human annotation errors.

The resulting corpus exhibits natural class imbalance reflective of actual cassation court caseloads, with commercial law cases comprising the majority (36.5%), followed by civil law (18.7%), administrative law (13.1%), criminal law (13%), and personal issues (9.8%), whereas social law represents the smallest category (8.8%), as shown in Figure 2.

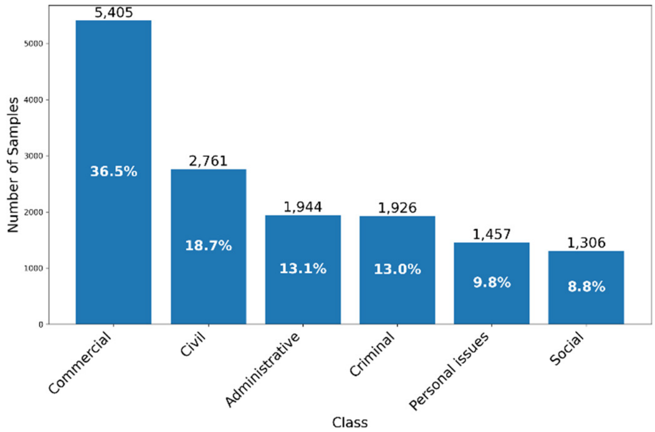

Fig. 2. Distribution of dataset classes.

B. *Text Preprocessing*

The preprocessing of Arabic legal texts presents difficulties that require a special approach to ensure data quality, preserve semantic integrity, and protect information privacy. Our preprocessing framework consists of three phases designed to address the specific characteristics of Arabic legal texts.

In the initial cleaning phase, we remove all punctuation marks, special symbols, and numerical sequences that often represent personal addresses or identity numbers of involved parties ("ر.ب.و: 123456789"). In addition, we eliminate redundant spaces to standardize text formatting. The second phase focuses on text normalization, addressing the removal of "tatweel," a typographic characteristic used for text justification in Arabic that elongates letters without affecting meaning (e.g., "محـــكمة" becomes "محكمة"). The third phase implements an enhanced stop-word removal process that extends beyond traditional linguistic stop-words. We remove common function words such as prepositions and conjunctions (e.g., "على", "من", "لكن", "و") that do not affect the semantics of legal texts. Additionally, a list of Arabic personal names (e.g., "محمد", "أحمد", "فاطمة") was added to the stop-word list. This privacy-protecting measure ensures the anonymization of case parties while maintaining the legal substance of the documents.

C. *Resampling Techniques*

The imbalanced nature of legal dataset distributions presents a significant challenge for text classification models, potentially biasing predictions towards majority classes. Resampling techniques can address this issue by achieving equal class distribution using three resampling strategies.

Oversampling techniques address class imbalance by increasing the number of minority class instances, either by duplicating existing instances or creating new synthetic ones. Several techniques have been developed to achieve a balanced distribution. For example, Random Oversampling (RO) involves duplicating samples from the minority class to achieve equal distribution within the dataset. Authors in [11] presented

SMOTE, a technique that generates new synthetic samples based on the k-nearest instances in the minority classes. Based on SMOTE, the borderline-SMOTE technique [12] generates synthetic samples only from the borderline minority instances, as they are considered more likely to be misclassified. This targeted approach avoids generating samples in areas where the minority class is already well-represented. Another notable technique is the Adaptive Synthetic (ADASYN) sampling approach [13], which adaptively generates more synthetic data for minority class samples that are harder to learn, determined by the density of majority class instances in their vicinity.

On the other hand, undersampling methods eliminate instances from the majority class to balance the dataset. The most straightforward method is random undersampling, which randomly removes majority class samples. The condensed nearest neighbor technique [14] attempts to find a consistent subset of the training data that can correctly classify all original data instances using a 1-nearest neighbor rule. ENN [15] seeks to enhance training data quality by removing instances that are misclassified by their k nearest neighbors, typically focusing on majority class instances but occasionally affecting minority class samples. Tomek Links [16] represent an aggressive undersampling technique for addressing class overlap in imbalanced datasets. A Tomek link is defined as a pair of instances from different classes that are mutual nearest neighbors, indicating ambiguous boundary regions. The method removes the majority class instance from each identified pair, thereby reducing overlapping areas between classes.

Hybrid methods have been presented as an effective approach, merging the advantages of both oversampling and undersampling. These techniques typically start with an initial oversampling of the minority group, followed by an undersampling of the majority group. Well-known examples of these hybrid techniques are the SMOTE-Tomek and SMOTE-EEN [17].

D. Proposed Model

As illustrated in Figure 3, the proposed architecture integrates four complementary neural components, each addressing a specific role in handling the aforementioned challenges of Arabic legal text:

- Embeddings layer: We employed AraBERTv2 [18], a well-known transformer-based model pre-trained on 77 GB of Arabic text, which provides contextualized embeddings that capture the nuanced meanings of words. AraBERTv2 consists of 12 transformer layers with 768-dimensional hidden states and 12 attention heads, totaling 110 million parameters.
- CNN layer: Building upon the embedding layer, we created a multi-scale CNN [19] that extracts local patterns at various granularities. The CNN architecture uses three parallel branches with filter sizes of 3, 4, and 5, each containing 100 filters. This configuration captures legal n-grams, which range from short technical terms to longer expressions common in legal texts. Each convolutional operation is followed by ReLU activation and max-pooling, which identifies the most important features in each filter's receptive field. The outputs from all three branches are concatenated, yielding a 300-dimensional feature vector containing the most discriminative local patterns in the text.
- BiLSTM layer: The sequential nature of legal documents, where conclusions often depend on arguments mentioned earlier in the text, motivates our use of a BiLSTM [20]. The BiLSTM processes the concatenated features from the CNN layer with 128 hidden units in each direction, resulting in a 256-dimensional representation that captures both forward and backward dependencies.
- Dense layers: The final classification layer consists of two fully connected layers with a dropout rate of 0.5 to prevent overfitting. The first dense layer reduces the 256-dimensional BiLSTM output to 128 dimensions with ReLU activation, whereas the second layer outputs the six legal classes with softmax activation for probability distribution over classes. Table I summarizes the key training hyperparameters and architectural specifications.

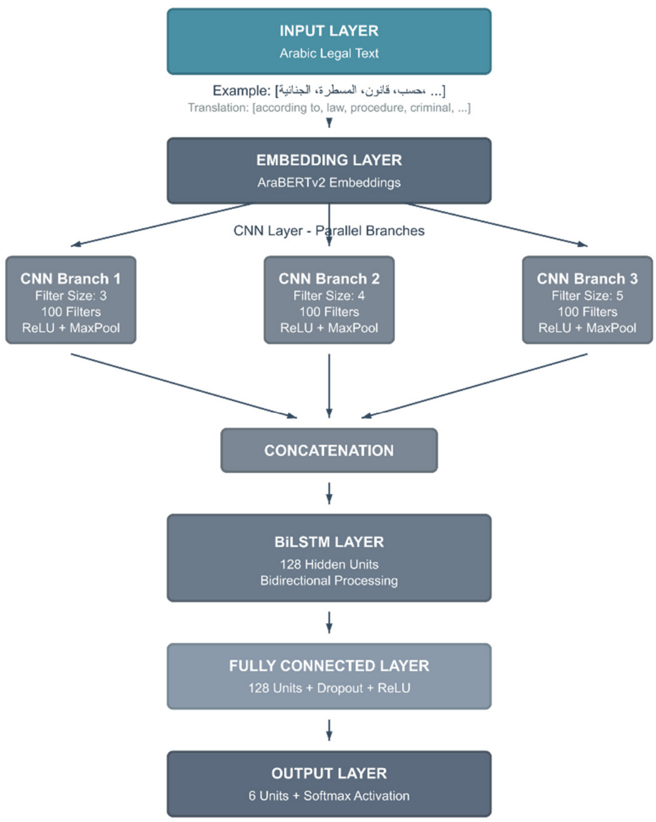

Fig. 3. Architecture of the proposed model.

E. Model Evaluation

To evaluate the performance of our model, we employed several metrics: accuracy (Acc), precision (Pre), recall (Rec), and F1-score (F1). In addition, we used the Matthews Correlation Coefficient (MCC), which offers a balanced measure that is particularly suitable for imbalanced datasets [21], as it considers all four confusion matrix categories. Moreover, the dataset was split into training (70%), validation

(15%), and test (15%) sets using stratified sampling to maintain class distribution.

TABLE I. MODEL ARCHITECTURE AND CONFIGURATION

Component	Specification	Value
Embedding layer	Pre-trained model	AraBERTv2
	Embedding dimension	768
	Maximum sequence length	512 tokens
CNN layer	Filter sizes	[3, 4, 5]
	Filters per size	100
	Activation function	ReLU
	Pooling strategy	Max-pooling
BiLSTM layer	Hidden units (per direction)	128
	Total hidden dimension	256
	Layers	1
Dense layers	Hidden layer dimension	128
	Output dimension	6 (classes)
	Dropout rate	0.5
	Activation	ReLU→softmax
Training	Optimizer	Adam
	Learning rate	2×10^{-5}
	Batch size	16
	Loss function	Cross-entropy

III. RESULTS AND DISCUSSION

This section presents the results obtained from our proposed model for Arabic legal text classification through a series of experiments. The first set of experiments compares our proposed model to traditional ML and DL models. The second set of experiments examines the impact of resampling techniques, including oversampling, undersampling, and hybrid techniques, on model performance. The third set of experiments presents an ablation study that isolates the contribution of each architectural component.

As demonstrated in Table II, the proposed CNN-BiLSTM model with AraBERTv2 embeddings achieved the highest performance across all evaluation metrics, with 95.95% accuracy, 96.07% precision, 95.04% recall, a 95.53% F1-score, and an MCC of 0.9480, significantly outperforming both traditional ML and DL baseline models. This significant improvement can be attributed to two key architectural decisions. First, the combination of CNN and BiLSTM effectively captures both local patterns and sequential dependencies present in legal terminology. Second, the use of AraBERTv2 as a word embedding technique provides contextualized representations superior to the static Word2Vec embeddings used in baseline models. Among the baseline approaches, SVM emerged as the strongest traditional classifier with an F1-score of 92.42%, whereas BiLSTM demonstrated competitive performance among DL architectures, achieving an F1-score of 93.88%, validating the importance of bidirectional context modeling for legal text classification.

Table III presents the performance of several oversampling techniques used to address class imbalance. RO demonstrated the best performance, achieving 96.22% in both accuracy and precision, with an F1-score of 95.74% and MCC of 0.9515. Borderline-SMOTE and ADASYN also yielded notable performance across multiple metrics, leveraging their sophisticated synthetic sample generation techniques. Conversely, SMOTE underperformed compared to other oversampling methods, suggesting that its interpolation-based approach may be less suitable for the complex feature space of legal texts.

TABLE II. PERFORMANCE COMPARISON WITH BASELINE MODELS

Model	Acc	Pre	Rec	F1	MCC
Word2Vec+NB	83.34	83.20	82.02	82.19	0.7873
Word2Vec+SVM	93.21	93.56	91.44	92.42	0.9128
Word2Vec+RF	91.42	91.97	89.34	90.50	0.8898
Word2vec+LR	92.26	92.65	90.13	91.28	0.9005
Word2Vec+LSTM	90.14	87.89	90.35	88.91	0.8756
Word2Vec+BiLSTM	94.41	94.75	93.12	93.88	0.9283
Proposed model	95.95	96.07	95.04	95.53	0.9480

TABLE III. PERFORMANCE USING OVERSAMPLING TECHNIQUES

Model	Acc	Pre	Rec	F1	MCC
Without resampling	95.95	96.07	95.04	95.53	0.9480
SMOTE	95.68	95.24	94.99	95.11	0.9445
Borderline-SMOTE	96.17	95.81	95.62	95.69	0.9509
RO	96.22	96.22	95.31	95.74	0.9515
ADASYN	96.04	95.43	95.59	95.50	0.9492

As detailed in Table IV, undersampling techniques negatively affected model performance, resulting in significant metric degradation. The exception was the Tomek Links technique, which achieved 95.12% accuracy and an MCC of 0.9404. This superior result is attributable to its strategy of removing borderline and overlapping majority class instances rather than random instances. However, even this best-performing Tomek Links method did not outperform the model without resampling. These results confirmed that the performance of our classification model is adversely impacted by information loss caused by undersampling techniques.

TABLE IV. PERFORMANCE USING UNDERSAMPLING TECHNIQUES

Model	Acc	Pre	Rec	F1	MCC
Without resampling	95.95	96.07	95.04	95.53	0.9480
Random undersampling	94.82	94.07	94.21	94.11	0.9336
Tomek Links	95.36	95.21	94.36	94.76	0.9404
ENN	92.84	91.46	91.92	91.55	0.9085
Condensed Nearest Neighbor	92.75	91.95	90.95	91.36	0.9069

Hybrid sampling approaches demonstrated promising results. As shown in Table V, SMOTE-Tomek and RO-Tomek achieved accuracies of 96.04% and 95.99% and recall scores of 95.37% and 95.2%, respectively.

TABLE V. PERFORMANCE USING HYBRID TECHNIQUES

Model	Acc	Pre	Rec	F1	MCC
Without resampling	95.95	96.07	95.04	95.53	0.9480
SMOTE-ENN	94.73	94.36	94.51	94.42	0.9326
SMOTE-Tomek	96.04	95.88	95.37	95.60	0.9491
RO-Tomek	95.99	95.87	95.20	95.52	0.9486

Additionally, while oversampling and hybrid methods demonstrated reliable performance, these improvements resulted in considerable computational costs, nearly doubling the training duration. This balance between effectiveness and efficiency needs to be thoughtfully examined for real-world applications. The undersampling methods, while offering computational advantages, proved unsuitable for legal text classification due to the critical information loss caused by instance removal, where majority classes often contain essential domain-specific patterns that cannot be eliminated without significant performance degradation.

Finally, to understand the contribution of each component of the proposed model, we conducted an ablation study comparing three model variants, as shown in Table VI.

TABLE VI. COMPARISON BETWEEN THE PROPOSED MODEL AND ABLATED MODELS

Model configuration	Acc	Pre	Rec	F1	MCC
Without BiLSTM	94.77	94.38	93.90	94.11	0.9330
Without CNN	91.40	89.97	90.27	90.09	0.8900
Without AraBERTv2	95.41	95.37	94.35	94.83	0.9410
Proposed model	96.22	96.22	95.31	95.74	0.9515

The key insights from the ablation study are:

- BiLSTM contribution: Removing BiLSTM (CNN only) resulted in a 1.63% decrease in F1-score, demonstrating the importance of capturing long dependencies.
- CNN contribution: Removing CNN (BiLSTM only) resulted in a 5.65% decrease in F1-score, showing that local pattern extraction is crucial.
- Language model impact: Replacing AraBERT with mBERT caused a 0.91% performance drop in F1-score, confirming the importance of an Arabic-specific pre-training embedding model.

IV. CONCLUSIONS

Classifying Arabic legal texts represents a significant challenge in the Arabic Natural Language Processing (NLP) literature due to the complexities of legal language and the morphological richness of Arabic. In this paper, we addressed this challenge by developing a Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BiLSTM) hybrid model combined with Arabic Bidirectional Encoder Representations from Transformers version 2 (AraBERTv2). Additionally, we investigated the impact of various resampling techniques to address the class imbalance inherent in the legal datasets. The findings demonstrated that the model using Random Oversampling (RO) yielded the best performance with an accuracy of 96.22%, outperforming all other evaluated resampling techniques.

Despite these results, our study identified several limitations that present opportunities for future research. The 512-token restriction of the BERT architecture poses a serious drawback since legal documents frequently exceed this threshold, potentially resulting in information loss during processing. Furthermore, data augmentation approaches using Large Language Models (LLMs) may be particularly appropriate for legal text due to their ability to generate more realistic and contextually appropriate texts.

REFERENCES

[1] T. El Moussaoui, C. Loqman, and J. Boumhidi, "Decoding legal processes: AI-driven system to streamline processing of the criminal records in Moroccan courts," *Intelligent Systems with Applications*, vol. 25, Mar. 2025, Art. no. 200487, https://doi.org/10.1016/j.iswa.2025.200487.

[2] F. Ariai, J. Mackenzie, and G. Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges." arXiv, Jul. 30, 2025, https://doi.org/10.48550/arXiv.2410.21306.

[3] A. Bouhouche, M. Esghir, and M. Errachid, "Classification of Moroccan Legal and Legislative Texts Using Machine Learning Models," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 10, pp. 1108–1114, Oct. 2024, https://doi.org/10.14569/IJACSA.2024.01510113.

[4] A. Y. Ikram and L. Chakir, "Arabic Text Classification in the Legal Domain," in *2019 Third International Conference on Intelligent Computing in Data Sciences*, Marrakech, Morocco, 2019, pp. 1–6, https://doi.org/10.1109/ICDS47004.2019.8942343.

[5] O. Tahtah, Y. Akhiat, A. Zinedine, and K. Fardousse, "Towards a Question/Answering System in Moroccan Legal Domain: data preparation and question classification phase using ML approaches," in *2023 7th IEEE Congress on Information Science and Technology*, Agadir - Essaouira, Morocco, 2023, pp. 140–144, https://doi.org/10.1109/CiSt56084.2023.10409875.

[6] O. Tahtah, Y. Akhiat, A. Zinedine, and K. Fardousse, "Towards a Question/Answering System in Moroccan Legal Domain: Investigation of the impact of tailored preprocessing techniques and alternative word representations on the question classification," in *2024 3rd International Conference on Embedded Systems and Artificial Intelligence*, Fez, Morocco, 2024, pp. 1–7, https://doi.org/10.1109/ESAI62891.2024.10913625.

[7] T. E. Moussaoui and L. Chakir, "BERT-CLSTM model for the classification of Moroccan commercial courts verdicts," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, Warsaw, Poland, 2023, pp. 281–284, https://doi.org/10.15439/2023F3561.

[8] E. Aljohani, "Enhancing Arabic Fake News Detection: Evaluating Data Balancing Techniques Across Multiple Machine Learning Models," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15947–15956, Aug. 2024, https://doi.org/10.48084/etasr.8019.

[9] M. J. Al-Khazaleh, M. Alian, and M. A. Jaradat, "Sentiment analysis of imbalanced Arabic data using sampling techniques and classification algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 1, pp. 607–618, Feb. 2024, https://doi.org/10.11591/eei.v13i1.5886.

[10] "Judicial Portal of the Kingdom of Morocco." Juriscassation. https://juriscassation.cspj.ma/.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, https://doi.org/10.1613/jair.953.

[12] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Advances in Intelligent Computing: International Conference on Intelligent Computing*, Hefei, China, 2005, pp. 878–887, https://doi.org/10.1007/11538059_91.

[13] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, 2008, pp. 1322–1328, https://doi.org/10.1109/IJCNN.2008.4633969.

[14] P. Hart, "The condensed nearest neighbor rule (Corresp.)," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 515–516, May 1968, https://doi.org/10.1109/TIT.1968.1054155.

[15] D. L. Wilson, "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-2, no. 3, pp. 408–421, Jul. 1972, https://doi.org/10.1109/TSMC.1972.4309137.

[16] "Two Modifications of CNN," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 11, pp. 769–772, Nov. 1976, https://doi.org/10.1109/TSMC.1976.4309452.

[17] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, Jun. 2004, https://doi.org/10.1145/1007730.1007735.

[18] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding." arXiv, Mar. 07, 2021, https://doi.org/10.48550/arXiv.2003.00104.

[19] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, https://doi.org/10.1109/TNNLS.2021.3084827.

[20] S. Zhang, D. Zheng, X. Hu, and M. Yang, "Bidirectional Long Short-Term Memory Networks for Relation Classification," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, 2015, pp. 73–78.

[21] C. Halimu, A. Kasem, and S. H. S. Newaz, "Empirical Comparison of Area under ROC curve (AUC) and Mathew Correlation Coefficient (MCC) for Evaluating Machine Learning Algorithms on Imbalanced Datasets for Binary Classification," in *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*, Da Lat, Vietnam, 2019, pp. 1–6, https://doi.org/10.1145/3310986.3311023.