Automated Classification of Moroccan Procedural Court Orders Using CAMeL-BERT

Saliha YASSINE, Ouafaa IBRIHICH, Mohammed ESSOUJAA, Mustapha ESGHIR
LabMIA-SI Laboratory, Faculty of Sciences, Mohammed V University in Rabat, Morocco

***Abstract*—The digitization of judicial systems creates new opportunities for automating document management tasks that have traditionally relied on manual processing. This study addresses the automated classification of Arabic procedural court orders issued by Moroccan courts of first instance into seven official procedural categories, a low-resource legal document type characterized by highly standardized procedural language and limited publicly available resources. The primary contribution is a leakage-controlled evaluation demonstrating that frozen CAMeL-BERT contextual representations, combined with lightweight conventional machine learning classifiers, achieve highly accurate classification without computationally expensive transformer fine-tuning. The proposed pipeline integrates automatic structured extraction from heterogeneous Word documents with contextual embeddings and a comparative evaluation of five classifiers under a strict train–validation–test protocol. Experimental results show that Logistic Regression achieves 99.74% test accuracy ($\kappa = 0.9966$), while even the simplest baseline, Gaussian Naive Bayes, reaches 96.66%, highlighting the strong discriminative characteristics of Moroccan procedural court orders. An ablation study demonstrates that replacing TF-IDF representations with CAMeL-BERT embeddings reduces classification errors by 71.4%, while the complete classification pipeline further improves overall performance and reduces the remaining errors. These findings show that the highly regular linguistic structure of Moroccan procedural court orders enables computationally efficient and deployment-oriented classification using frozen contextual embeddings, providing a practical solution for Arabic legal document management and supporting future e-Justice applications in Morocco.**

***Keywords*—Arabic legal NLP; CAMeL-BERT; judicial document classification; Moroccan courts; procedural court orders**

I. Introduction

Moroccan courts of first instance process a large volume of procedural documents on a daily basis. These documents, commonly referred to as ordonnances, are formal judicial orders related to procedural matters such as payment obligations, financial consignments, commercial registry operations, guardianship appointments, and other judicial procedures. Recent investments by the Moroccan Ministry of Justice in electronic case management systems have accelerated the digitization of judicial workflows, making automated document processing an increasingly important operational requirement. Nevertheless, document classification remains largely manual in many courts, where incoming files must be identified and routed to the appropriate judicial department. Misclassification may delay judicial proceedings, increase administrative workload, and negatively affect procedural efficiency. Although these documents are stored in heterogeneous Word layouts that vary across courts, clerks, and time periods, they exhibit highly regular legal formulations and category-specific terminology, making them suitable candidates for data-driven document classification.

Automating this classification task requires addressing several linguistic challenges inherent to Arabic. Written Arabic frequently omits short vowels and diacritical marks, making tokenization and morphological analysis substantially more difficult than in many European languages. Arabic morphology is also highly productive, as a single trilateral root can generate numerous surface forms through affixation and internal vowel variation [1]. The Transformer architecture [2], particularly through BERT [3], addressed many of these challenges by learning deep contextual representations from large-scale corpora. These advances have been extended to Arabic through specialized models such as AraBERT [4] and CAMeL-BERT [5], which have demonstrated strong performance across a broad range of Arabic NLP tasks [6]. More recently, these advances have stimulated the development of Legal AI applications, including legal document classification, legal information retrieval, judgment prediction, intelligent document management, and e-Justice systems.

Despite recent progress in Arabic legal NLP [7]–[10], existing studies have primarily focused on legal corpora from other Arabic-speaking countries, general-domain datasets, or legal prediction tasks. Moroccan procedural court orders remain largely unexplored despite their central role in the daily operation of Moroccan courts. Furthermore, previous studies generally assume that documents are already available as clean textual data and therefore do not address the practical challenge of extracting structured textual content from heterogeneous judicial Word documents before classification. In addition, few studies investigate whether frozen contextual embeddings combined with lightweight conventional classifiers can provide both high classification performance and computational efficiency without expensive transformer fine-tuning. These limitations reveal an important gap between recent advances in Arabic Legal AI research and the practical requirements of judicial document management in Moroccan courts.

To address these challenges, this study proposes a complete document classification pipeline that integrates automatic structured text extraction from judicial Word documents with frozen CAMeL-BERT contextual embeddings and conventional machine learning classifiers. The proposed approach is evaluated under a strict train–validation–test protocol specifically designed to prevent data leakage while assessing the impact of contextual embeddings and training-only data augmentation. Beyond achieving high classification accuracy,

the study investigates whether deployment-oriented architectures based on frozen contextual representations can provide an effective and computationally efficient solution for real-world judicial document management.

The main contributions of this work are summarized as follows:

- An automatic extraction pipeline for converting heterogeneous Moroccan judicial Word documents into structured textual data suitable for classification.
- A curated corpus of Moroccan procedural court orders covering seven official judicial procedural categories.
- A leakage-controlled evaluation of frozen CAMeL-BERT contextual embeddings combined with five conventional machine learning classifiers.
- An ablation study quantifying the contribution of contextual embeddings and the complete classification pipeline to overall classification performance.
- An analysis of classification performance on an imbalanced judicial corpus, including minority categories.
- A deployment-oriented classification framework that achieves high accuracy without transformer fine-tuning, supporting practical integration into Morocco's e-Justice initiatives.

The remainder of this study is organized as follows: Section II reviews related work on Arabic legal NLP and judicial document classification. Section III presents the proposed extraction and classification methodology. Section IV discusses the experimental results and ablation study. Finally, Section V concludes the study and outlines directions for future research.

II. Related Work

A. Pre-Trained Arabic Language Models

Transformer-based pre-trained language models have significantly advanced Arabic Natural Language Processing (NLP) by learning contextual language representations from large-scale corpora. Unlike multilingual models, Arabic-specific transformers explicitly capture the rich morphology, lexical ambiguity, and syntactic characteristics of Arabic, leading to substantial improvements across text classification, named entity recognition, sentiment analysis, and question answering tasks.

AraBERT [4] was among the first transformer models specifically developed for Arabic and demonstrated that language-specific pre-training consistently outperforms multilingual BERT on several downstream NLP tasks. Subsequently, Alammary [6] conducted a comprehensive systematic review of Arabic BERT models, highlighting the influence of corpus composition, model architecture, and pre-training strategy on downstream performance. Abdul-Mageed et al. [12] further introduced ARBERT and MARBERT, two transformer models specifically designed to capture both Modern Standard Arabic (MSA) and dialectal Arabic. Although these models achieve excellent performance on dialect-rich corpora, their emphasis on dialectal language makes them less suitable for the predominantly Modern Standard Arabic used in Moroccan judicial documents.

CAMeL-BERT [5] further extended Arabic language modeling by pre-training transformer models on a 167 GB corpus covering Modern Standard Arabic, dialectal Arabic, and Classical Arabic. This multi-register training strategy enables the CAMeL-BERT-mix variant to capture both standard morphological phenomena and domain-specific lexical variations frequently encountered in formal legal writing. Its effectiveness has been demonstrated across several Arabic NLP benchmarks. For example, Mutawa and Sruthi [11] reported that CAMeL-BERT achieved the highest performance among several transformer architectures for Arabic meter classification, illustrating its strong capability to learn complex contextual representations.

More recently, transformer-based research has expanded toward Legal AI through the development of domain-adaptive language models. LEGAL-BERT [13] demonstrated that additional pre-training on legal corpora substantially improves performance on legal-domain NLP tasks compared with general-purpose BERT models. Similarly, AraLegal-BERT [15] showed that domain-specific pre-training on Arabic legal texts further enhances semantic representations for specialized legal terminology. These studies collectively confirm that contextual language models benefit significantly from legal-domain adaptation when processing judicial documents.

The present study adopts the CAMeL-BERT-mix model because Moroccan procedural court orders are written almost exclusively in Modern Standard Arabic while containing highly regular legal terminology and institution-specific procedural expressions. Unlike approaches based on full transformer fine-tuning, this work investigates whether frozen contextual embeddings extracted from CAMeL-BERT can provide sufficiently discriminative semantic representations for highly accurate document classification when combined with lightweight conventional machine learning classifiers. This deployment-oriented strategy substantially reduces computational requirements while preserving the contextual information learned during large-scale Arabic pre-training, making it particularly suitable for practical judicial applications.

B. Methods for Legal Document Classification

Automatic legal document classification has become an increasingly important research topic because of its applications in intelligent document management, legal information retrieval, electronic case management, and e-Justice systems. Early approaches relied primarily on manually engineered lexical and syntactic features combined with conventional machine learning algorithms such as Support Vector Machines (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Naive Bayes (NB). These methods achieved competitive performance on relatively homogeneous legal corpora but required substantial feature engineering and often struggled to capture the semantic complexity of legal language, particularly in morphologically rich languages such as Arabic.

The emergence of transformer-based language models fundamentally changed this landscape by replacing manually engineered features with contextual semantic representations learned from large-scale corpora. Domain-adaptive pre-training has proved particularly beneficial for legal applications. LEGAL-BERT [13] demonstrated that additional pre-training on legal corpora substantially improves performance compared with general-purpose BERT models. MultiEURLEX [14] further showed that transformer models can successfully perform multilingual legal document classification across European legislation, while AraLegal-BERT [15] confirmed the importance of domain-specific pre-training for Arabic legal text classification.

Several recent studies have applied machine learning and transformer models to Moroccan legal documents. El Moussaoui et al. [7] fine-tuned XLM-RoBERTa for judicial chamber classification using Court of Cassation judgments and reported an accuracy of 98.48%. Tahtah et al. [17] proposed a CNN–BiLSTM architecture using AraBERTv2 embeddings together with random oversampling to address class imbalance in Arabic legal document classification. Bouhouche et al. [8] evaluated conventional machine learning algorithms for Moroccan legislative document classification and demonstrated that lexical representations remain competitive on structured legal corpora. Zahir [9] investigated judicial outcome prediction from Arabic court decisions using deep learning, whereas El Moussaoui et al. [10] developed an AI-assisted framework for criminal record processing in Moroccan courts.

Although these studies demonstrate the growing maturity of Arabic Legal AI, they address legal tasks that differ substantially from the automatic classification of procedural court orders issued by Moroccan courts of first instance. Furthermore, most previous work assumes that textual content is already available in structured digital form and therefore overlooks the practical challenge of extracting discriminative information from heterogeneous judicial Word documents before classification. Existing studies also predominantly rely on transformer fine-tuning or deep neural architectures, whereas comparatively little attention has been devoted to deployment-oriented solutions based on frozen contextual embeddings combined with lightweight conventional classifiers.

TABLE I. COMPARISON OF REPRESENTATIVE STUDIES ON ARABIC LEGAL DOCUMENT CLASSIFICATION

Study	Corpus	Doc. Type	Task	Approach	Best Score	Main Limitation
LEGAL-BERT [13]	European legal corpora	Legal documents	Document classification	Domain-adaptive BERT	Improved performance	Non-Arabic corpus
MultiEURLEX [14]	European Union	Legislative texts	Multi-label classification	Multilingual Transformer	Strong benchmark	European legislation
AraLegal-BERT [15]	Arabic legal corpus	Legal documents	Legal text classification	Domain-specific Arabic BERT	Improved representations	No Moroccan evaluation
Bouhouche et al. [8]	Morocco	Legislative texts	Document classification	Conventional ML	95.82%	Legislative corpus
Zahir [9]	Morocco	Court decisions	Judgment prediction	Deep learning	80.51%	Outcome prediction
Tahtah et al. [17]	Morocco	Court cases	Document classification	CNN–BiLSTM + AraBERTv2	96.22%	Requires fine-tuning
El Moussaoui et al. [7]	Morocco	Court of Cassation judgments	Chamber classification	Fine-tuned XLM-RoBERTa	98.48%	Different court level
This study	Moroccan Courts of First Instance	Procedural orders	7-category classification	Frozen CAMeL-BERT + LR	99.74% ($\kappa = 0.9966$)	Single national corpus

The comparison presented in Table I highlights several important research gaps. First, most existing studies focus on legal corpora that differ substantially from Moroccan procedural court orders, including legislative texts, Court of Cassation judgments, or judgment prediction tasks. Consequently, their findings cannot be directly transferred to the document-routing problem encountered in Moroccan courts of first instance. Second, previous work generally assumes that documents are already available in clean textual form and therefore overlooks the practical challenge of extracting structured information from heterogeneous judicial Word documents before classification. Third, although transformer-based models have consistently demonstrated superior performance, relatively few studies have investigated deployment-oriented architectures that combine frozen contextual embeddings with lightweight conventional machine learning classifiers, thereby reducing computational requirements while maintaining high classification performance.

To address these limitations, the present study proposes an end-to-end framework integrating automatic structured text extraction, frozen CAMeL-BERT contextual embeddings, and five conventional machine learning classifiers evaluated under a strict leakage-controlled experimental protocol. In addition, a comprehensive ablation study is conducted to quantify the contribution of contextual embeddings and the complete classification pipeline, providing practical insights into the design of efficient Arabic legal document classification systems for e-Justice applications.

III. METHODOLOGY

A. Corpus Compilation and Data Augmentation

The corpus was compiled from procedural court orders collected from several Moroccan Courts of First Instance. The documents correspond to seven official procedural categories defined by the Moroccan judicial taxonomy: Notification of Warning (تبليغ إنذار), Payment Order (أمر بالأداء), Deposit and Consignment (العرض والإيداع), Validation of Warning (المصادقة على الإنذار), Financial Withdrawal (سحب مبلغ مالي), Appointment of Guardian (تعيين قيم), and Commercial Register Deletion (التشطيب على السجل التجاري). Since these procedural categories are

officially assigned during judicial processing, no additional manual annotation was required. The labels provided by the courts were therefore used directly as the ground-truth class labels throughout the experiments.

The collected documents were stored as Microsoft Word files with heterogeneous layouts reflecting differences among courts, clerks, and document preparation practices. To prepare a structured corpus suitable for machine learning, a custom Python extraction pipeline was developed to automatically identify and extract the facts section (faits) and the decision section (dispositif) from each document using rule-based textual markers and regular-expression matching. Documents with incomplete or unsuccessful section extraction were excluded during corpus preparation, resulting in a final corpus of 3,894 original documents. Preliminary experiments showed that the extracted facts section contains the most discriminative information for procedural category identification; therefore, only this textual content was used for classification.

TABLE II. CORPUS DISTRIBUTION

Procedural Category	Original Documents
Notification of Warning — تبليغ إنذار	1480
Deposit and Consignment — العرض والإيداع	934
Validation of Warning — المصادقة على الإنذار	394
Financial Withdrawal — سحب مبلغ مالي	492
Payment Order — أمر بالأداء	432
Appointment of Guardian — تعيين قيم	120
Commercial Register Deletion — التشطيب على السجل التجاري	42
Total	3894

As shown in Table II, the original corpus exhibits a pronounced class imbalance, with an imbalance ratio of 35.24:1 between the largest and smallest categories. To reduce this imbalance while preserving the semantic meaning of each procedural category, data augmentation was applied exclusively to the training partition after train–validation–test splitting, thereby preventing any information leakage into the validation or test sets.

A per-class target of approximately 2,000 training documents was adopted. New training samples were generated by combining three label-preserving transformations specifically designed for procedural judicial documents:

- Judicial synonym substitution using a domain-specific lexicon of 40 legally equivalent term pairs validated by an experienced court clerk (e.g., التماس/طلب, الفصل/المادة);
- Date variation, replacing four-digit years with nearby values (±1 year);
- Monetary-value scaling, multiplying monetary amounts by a random factor between 0.8 and 1.2.

Date and monetary-value transformations were applied only to information that does not determine the procedural category. Since the seven target classes are defined by the underlying judicial procedure rather than by specific dates or monetary amounts, these modifications preserve the original class label. Every augmented document therefore inherits the category of its source document, and no cross-category mixing was performed.

The augmentation process increased the size of the training corpus to 13,764 documents, reducing the imbalance ratio from 35.24:1 to approximately 1.12:1 while maintaining a strict separation between original evaluation data and augmented training data.

Fig. 1 presents an anonymized excerpt of a Notification of Warning (تبليغ إنذار) procedural court order after the automatic extraction process. All personal names, case numbers, and identifying information were removed or replaced before analysis to preserve judicial confidentiality.

المحكمة الابتدائية — [×] نحن رئيس المحكمة الابتدائية بمقتضى الفصل 148
من ق.م.م. بعد الاطلاع على طلب السيد [×××] بتبليغ إنذار للمدين [×××] بأداء
مبلغ [...] درهماً أمرنا بما يلي: [تبليغ المدين بالمبلغ المذكور

Fig. 1. Anonymized excerpt of the facts section of a Notification of Warning (تبليغ إنذار) procedural court order extracted by the Python script. Personal identifiers have been replaced by ×.

The extracted facts section subsequently undergoes text normalization and contextual embedding generation before being used for document classification. The overall preprocessing pipeline is described in the following subsection.

B. Data Preprocessing and Experimental Protocol

The extracted textual data underwent several preprocessing steps before classification. Text normalization included the removal of unnecessary whitespace and formatting inconsistencies introduced during document extraction, together with the normalization of Arabic text to ensure consistent lexical representation. Documents with incomplete or unsuccessful section extraction, as well as documents containing insufficient textual content for reliable classification, were excluded during corpus preparation to improve the overall quality of the dataset. The resulting corpus contained 3,894 original procedural court orders, which served as the basis for all subsequent experiments.

The final corpus was partitioned into training (60%), validation (20%), and test (20%) sets using stratified random sampling in order to preserve the original class distribution across all partitions. The partitioning was performed before any data augmentation to prevent information leakage between the training and evaluation datasets. Consequently, every original document appears in only one partition throughout the entire experimental protocol.

Data augmentation was applied exclusively to the training partition, increasing the number of training documents from 2,336 to 13,764 while leaving the validation and test sets unchanged. This strategy preserved the semantic integrity of each procedural category while reducing the class imbalance ratio from 35.24:1 to approximately 1.12:1. Since the validation and test partitions contain only original documents, the reported performance reflects the model's ability to generalize to previously unseen judicial documents rather than to augmented samples.

To ensure a fair comparison among all classification models, the same training, validation, and test partitions were used throughout every experiment, including the ablation study. This

leakage-controlled experimental protocol guarantees that all reported performance differences result exclusively from changes in document representation or classification algorithms rather than from variations in the evaluation data.

C. Proposed Classification Framework

Fig. 2 presents the overall workflow of the proposed framework for the automatic classification of Moroccan procedural court orders. The framework consists of five sequential stages that transform heterogeneous judicial Word documents into contextual semantic representations suitable for procedural document classification.

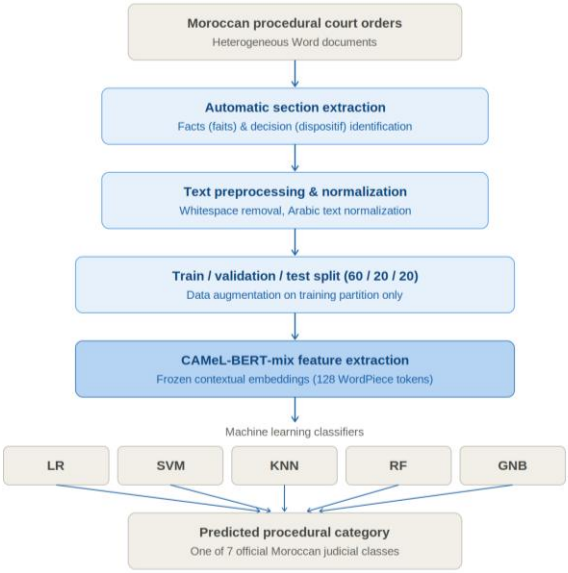

Fig. 2. Overall workflow of the proposed framework for automatic classification of Moroccan procedural court orders.

First, procedural court orders are processed by the automatic extraction module described in the previous subsection. This module identifies and extracts the facts (faits) and decision (dispositif) sections from each document. Preliminary experiments showed that the facts section contains the most discriminative procedural information; therefore, only this section is retained for the subsequent classification process.

The extracted text is then normalized to remove formatting inconsistencies while preserving its legal content. The resulting corpus is partitioned into training, validation, and test sets using a leakage-controlled experimental protocol. Data augmentation is applied exclusively to the training partition in order to mitigate class imbalance while preserving the semantic characteristics of each procedural category.

Next, each preprocessed document is converted into a contextual semantic representation using the frozen CAMeL-BERT-mix model. The resulting document embeddings are subsequently used to train five conventional machine learning classifiers: Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and Gaussian Naive Bayes (GNB).

By combining transformer-based contextual embeddings with lightweight machine learning classifiers, the proposed framework provides an effective and computationally efficient solution for the automatic routing of procedural court orders within Moroccan e-Justice systems.

D. CAMeL-BERT Feature Extraction

The proposed framework employs the CAMeL-BERT-mix model [5] as the contextual feature extractor. The main configuration parameters of the feature extraction module are summarized in Table III. CAMeL-BERT-mix is built upon the standard BERT-base architecture [3], comprising 12 transformer encoder layers, 768-dimensional hidden representations, 12 self-attention heads, and approximately 110 million parameters. The model was pre-trained on a large-scale Arabic corpus covering Modern Standard Arabic (MSA), dialectal Arabic, and Classical Arabic, making it particularly suitable for legal documents that combine formal language with institution-specific procedural terminology.

TABLE III. CAMEL-BERT CONFIGURATION USED IN THIS STUDY

Parameter	Value
Pre-trained model	CAMeL-BERT-mix
Base architecture	BERT-base
Transformer encoder layers	12
Hidden size	768
Attention heads	12
Approximate number of parameters	110 million
Tokenizer	WordPiece
Maximum sequence length	128 tokens
Document representation	Final hidden-layer [CLS] embedding
Embedding dimension	768
Transfer learning strategy	Frozen embeddings (no fine-tuning)
Feature scaling	StandardScaler (fitted on the training set only)

Classification is performed exclusively on the extracted facts section, as it contains the procedural information most relevant for distinguishing between the seven judicial categories. Each document is tokenized using the original CAMeL-BERT WordPiece tokenizer, and input sequences are truncated or padded to a maximum length of 128 tokens, ensuring a fixed-size representation compatible with the transformer architecture.

Rather than fine-tuning the transformer model, this study adopts a feature-based transfer learning strategy. The parameters of CAMeL-BERT remain frozen throughout the training process, and only the contextual document representations generated by the model are used as input features for the downstream classifiers. This strategy substantially reduces computational requirements while allowing the evaluation to focus on the discriminative capability of contextual embeddings independently of transformer optimization [18].

For each document, the contextual representation of the [CLS] token from the final hidden layer is extracted as a fixed-length 768-dimensional document embedding, which serves as the input feature vector for all classification models. Before classifier training, these embeddings are standardized using StandardScaler. The scaler is fitted exclusively on the training partition and subsequently applied, without refitting, to the

validation and test partitions, thereby preventing information leakage during preprocessing and ensuring identical feature scaling across all evaluation datasets.

The combination of frozen contextual embeddings and lightweight machine learning classifiers provides an efficient architecture for deployment-oriented judicial applications. Once document embeddings have been generated, classification can be performed without repeated transformer optimization, substantially reducing computational cost while preserving the semantic knowledge acquired during large-scale Arabic language pre-training.

E. Classification Models

To evaluate the effectiveness of the proposed contextual representations, five supervised machine learning classifiers from different algorithmic families were trained using the CAMeL-BERT document embeddings. The hyperparameter configuration of all classifiers is summarized in Table IV.

TABLE IV. HYPERPARAMETER SETTINGS OF THE CLASSIFICATION MODELS.

Classifier	Hyperparameter Configuration
Logistic Regression	L2 regularization, C = 0.5, class_weight = balanced, max_iter = 1000
Support Vector Machine	RBF kernel, C = 1.0, gamma = scale
K-Nearest Neighbors	k = 5, cosine distance
Random Forest	n_estimators = 100, max_depth = 10, min_samples_split = 10, min_samples_leaf = 5, class_weight = balanced
Gaussian Naive Bayes	Default parameters

Logistic Regression (LR) was selected as a linear baseline because of its robustness, computational efficiency, and effectiveness in high-dimensional feature spaces. The classifier employs L2 regularization with C = 0.5, class_weight = "balanced", and a maximum of 1,000 iterations.

Support Vector Machine (SVM) was configured with a radial basis function (RBF) kernel using C = 1.0 and gamma = "scale", enabling nonlinear decision boundaries to be learned in the contextual embedding space [16].

K-Nearest Neighbors (KNN) was implemented with k = 5 using the cosine distance metric, which is well suited for measuring similarity between high-dimensional contextual embeddings.

Random Forest (RF) consisted of 100 decision trees with a maximum tree depth of 10, a minimum of 10 samples required to split an internal node, a minimum of 5 samples per leaf node, and class_weight = "balanced" to reduce the impact of class imbalance [19].

Finally, Gaussian Naive Bayes (GNB) was included as a probabilistic baseline because of its simplicity and computational efficiency.

To ensure a fair comparison, all classifiers were trained using the same standardized CAMeL-BERT embeddings and the same training, validation, and test partitions. Hyperparameters were selected through grid search on the validation set and subsequently kept fixed for all experiments.

F. Evaluation Metrics

The performance of the proposed classification framework was evaluated using Accuracy, Precision, Recall, F1-score, and Cohen's Kappa coefficient (κ). Accuracy provides the overall proportion of correctly classified documents, whereas Precision and Recall evaluate the correctness and completeness of the predicted classes, respectively. The F1-score, defined as the harmonic mean of Precision and Recall, provides a balanced measure of classification performance, particularly for imbalanced datasets. Cohen's Kappa (κ) was additionally reported to quantify the agreement between predicted and true labels while accounting for agreement occurring by chance.

Let TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. The evaluation metrics are defined as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F}_1 = \frac{2\times Precision\times Recall}{Precision + Recall} \quad (4)$$

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (5)$$

where, P_o denotes the observed agreement between predicted and true labels, and P_e represents the agreement expected by chance.

To provide a comprehensive assessment of classifier performance, the evaluation was conducted on the held-out test set, while model selection and hyperparameter tuning were performed exclusively on the validation set. In addition, five-fold stratified cross-validation was employed on the training data during model development to assess the robustness and generalization capability of the proposed framework.

IV. RESULTS AND DISCUSSION

A. Overall Classifier Performance

Table V summarizes the overall performance of the five evaluated classifiers on the independent test set comprising 779 original procedural court orders. Among all evaluated models, Logistic Regression (LR) achieved the highest classification performance, reaching a test accuracy of 99.74% with a Cohen's Kappa coefficient (κ) of 0.9966. Support Vector Machine (SVM) achieved the second-best performance with 99.49% accuracy (κ = 0.9932), followed by K-Nearest Neighbors (KNN) at 99.61%, Random Forest (RF) at 98.84%, and Gaussian Naive Bayes (GNB) at 96.66%.

The consistently high cross-validation accuracies, together with the very low standard deviations (±0.06%–±0.37%), demonstrate that the proposed framework generalizes well across different training folds and that the reported performance is not dependent on a particular train–test partition. Similarly, the high Cohen's Kappa values (LR: 0.9966, SVM: 0.9932,

KNN: 0.9949, RF: 0.9848, and GNB: 0.9562) indicate an almost perfect agreement between predicted and true labels, confirming that the excellent classification performance is not an artifact of class imbalance.

TABLE V. OVERALL PERFORMANCE OF THE EVALUATED CLASSIFICATION MODELS

Classifier	Validation Accuracy	Test Accuracy	5-Fold Cross-Validation Accuracy
Naive Bayes	97.30%	96.66%	98.34% ±0.37%
K-Nearest Neighbors	98.97%	99.61%	99.89% ±0.07%
Random Forest	98.84%	98.84%	99.91% ±0.07%
Support Vector Machine	99.49%	99.49%	99.96% ±0.07%
Logistic Regression	99.49%	99.74%	99.97% ±0.06%

Overall, the proposed framework demonstrates that combining frozen CAMeL-BERT contextual embeddings with lightweight machine learning classifiers provides highly accurate and robust classification of Moroccan procedural court orders. The strong performance can be attributed to three complementary factors: 1) the rich contextual semantic representations learned by CAMeL-BERT; 2) the highly discriminative legal vocabulary contained in the extracted facts section; and 3) the balanced training corpus obtained through targeted data augmentation. In addition, the low mean pairwise Jaccard similarity (12.3%) among the seven procedural categories indicates limited lexical overlap between categories, facilitating their discrimination in the embedding space. Finally, the consistent ranking of the evaluated classifiers (LR ≈ SVM > KNN > RF > GNB) across validation, test, and cross-validation experiments indicates that this is primarily attributable to the discriminative power of the contextual embeddings rather than classifier complexity.

B. Effect of Data Augmentation

Data augmentation increased the size of the training set from 2,336 to 13,764 documents while reducing the class imbalance ratio from 35.24:1 to 1.12:1. This strategy was specifically designed to improve the representation of minority procedural categories without modifying the validation and test sets, thereby preserving the integrity of the evaluation protocol.

The effectiveness of data augmentation is reflected in the per-category results reported in Table VII, where the two minority classes, تعيين قيم (24 test documents) and التشطيب على السجل التجاري (8 test documents), both achieved an F1-score of 1.000 despite their limited number of original training examples. These results indicate that the proposed augmentation strategy successfully compensated for severe class imbalance while preserving the discriminative characteristics of each procedural category.

The contribution of data augmentation is further confirmed by the ablation study presented in Table VI. Although the overall classification accuracy remains nearly unchanged, augmentation improves class-balanced performance by eliminating the remaining misclassifications affecting the minority class تعيين قيم. Consequently, the only residual errors are observed between المصادقة على الإنذار and العرض والإيداع, the two procedural categories exhibiting the highest semantic similarity.

Overall, these findings demonstrate that the principal contribution of data augmentation is not to increase the already high overall accuracy, but rather to enhance the robustness and reliability of the classifier across all procedural categories. This improvement is particularly important in judicial applications, where reliable recognition of minority case types is essential for practical deployment.

TABLE VI. ABLATION STUDY OF THE PROPOSED FRAMEWORK

Configuration	Accuracy	Macro-F1	F1 (تعيين قيم)	F1 (التشطيب)	Errors
TF-IDF + SVM	98.20%	0.9562	1.000	0.769	14/779
CAMeL-BERT + SVM (no augmentation)	99.49%	0.9933	0.979	1.000	4/779
CAMeL-BERT + LR (full pipeline)	99.74%	0.9970	1.000	1.000	2/779

C. Performance by Category

Table VII presents the per-category classification performance of the best-performing model, Logistic Regression, on the independent test set comprising 779 original procedural court orders. Overall, the proposed framework achieved perfect or near-perfect performance across all seven procedural categories. Five categories obtained an F1-score of 1.000, while المصادقة على الإنذار and العرض والإيداع achieved 99.46% and 98.09%, respectively.

TABLE VII. PER-CLASS PERFORMANCE OF THE LOGISTIC REGRESSION MODEL ON THE INDEPENDENT TEST SET

Procedural Category	Precision	Recall	F1-score	Support
Notification of Warning (تبليغ إنذار)	99.00%	100.00%	99.50%	296
Deposit and Consignment (العرض والإيداع)	100.00%	98.93%	99.46%	187
Financial Withdrawal (سحب مبلغ مالي)	100.00%	100.00%	100.00%	99
Payment Order (أمر بالأداء)	100.00%	100.00%	100.00%	86
Validation of Warning (المصادقة على الإنذار)	98.72%	97.47%	98.09%	79
Appointment of Guardian (تعيين قيم)	100.00%	100.00%	100.00%	24
Commercial Register Deletion (التشطيب على السجل التجاري)	100.00%	100.00%	100.00%	8

Note: Support denotes the number of original test documents in each procedural category.

The only two remaining misclassifications occurred between المصادقة على الإنذار and العرض والإيداع, which share similar procedural terminology and legal expressions. This observation suggests that the residual errors are primarily attributable to semantic similarity between these categories rather than to limitations of the contextual representations learned by CAMeL-BERT.

The two minority categories, تعيين قيم (24 test documents) and التشطيب على السجل التجاري (8 test documents), both achieved an F1-score of 1.000, demonstrating that the proposed framework generalizes effectively even for underrepresented procedural categories. Although these results should be interpreted with

appropriate caution because of the relatively small number of test instances, their consistent recognition provides additional evidence of the robustness of the proposed classification framework.

D. Error Analysis

The best-performing classifier (Logistic Regression) made only two errors on the independent test set of 779 original procedural court orders, corresponding to an error rate of 0.26%, while the SVM classifier produced four errors. Since both classifiers exhibit a very similar residual confusion pattern, the confusion matrix of the CAMeL-BERT + SVM configuration is presented in Fig. 3 as a representative visualization of the remaining ambiguity in the embedding space.

The four SVM misclassifications are exclusively concentrated between العرض والإيداع and المصادقة على الإنذار. Specifically, two documents belonging to العرض والإيداع were predicted as المصادقة على الإنذار, while two documents from المصادقة على الإنذار were classified as العرض والإيداع. No errors were observed for the remaining five procedural categories, including the minority classes تعيين قيم and التشطيب على السجل التجاري.

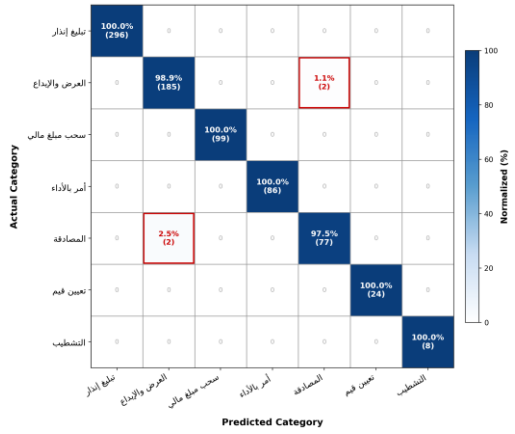

Fig. 3. Normalized confusion matrix of the CAMeL-BERT + SVM model on the independent test set (779 original documents).

The confusion matrix confirms that all residual errors are restricted to two procedurally related categories sharing similar legal terminology and document structure. The final Logistic Regression model further reduces these misclassifications from four to two while preserving the same overall confusion pattern, indicating that the remaining errors originate from semantic similarity between the categories rather than from limitations of the contextual representations.

From a practical perspective, these rare ambiguous cases could be addressed by incorporating a human verification step for predictions with low confidence scores. Such a strategy would enable reliable deployment in judicial document management systems while maintaining the efficiency of the proposed automated classification framework.

E. Corpus Structure and Accuracy

The 99.74% test accuracy achieved by the proposed Logistic Regression model is substantially higher than the performance reported in many previous studies on Arabic legal document classification and judicial prediction tasks. For example, judicial outcome prediction for European Court of Human Rights (ECHR) cases has reported F1-scores of approximately 79% [20], [21], while Zahir et al. [9] achieved an F1-score of 80.51% for Arabic court decision classification. Although these studies address different legal tasks and datasets, the comparison provides useful context for interpreting the present results.

The exceptionally high performance observed in this study can largely be attributed to the structural characteristics of Moroccan procedural court orders. First, the mean pairwise Jaccard similarity between category vocabularies is only 12.3%, indicating limited lexical overlap across procedural categories. Second, the contextual representations produced by CAMeL-BERT capture semantic and positional information beyond surface lexical features, resulting in highly separable document embeddings. Third, each procedural category follows standardized drafting conventions and contains recurring legal expressions that are consistently used within that category, thereby reducing intra-class variability. Finally, the documents are drafted by trained court clerks according to institutional writing practices, which further limits stylistic variation and linguistic ambiguity.

Taken together, these characteristics create a classification task that is particularly well-suited to contextual transformer embeddings combined with a linear classifier. This interpretation is consistent with previous studies on Moroccan legal documents, which also reported strong classification performance on structurally similar corpora, including 98.48% by El Moussaoui et al. [7] and 95.82% by Bouhouche et al. [8].

F. Deployment Prospects and Limitations

The proposed framework achieves an error rate of only 0.26% (2 misclassifications out of 779 test documents), suggesting that it is well suited for deployment as a decision-support tool rather than as a fully autonomous system. In practical settings, predictions associated with low confidence scores can be forwarded to court clerks for verification before final assignment, thereby combining the efficiency of automated document classification with human oversight.

Two deployment scenarios are envisioned within Morocco's ongoing e-Justice initiatives. The first is real-time document routing, where newly issued procedural court orders are automatically assigned to the appropriate departmental workflow while uncertain predictions are flagged for manual review. The second is retrospective digital archiving, where historical procedural court orders are automatically indexed according to their legal category, facilitating efficient search, retrieval, and document management.

An additional practical advantage of the proposed framework is its computational efficiency. Since CAMeL-BERT is used only for offline feature extraction and the final classifier is a lightweight Logistic Regression model, inference requires only standard CPU resources without GPU acceleration. This makes the system compatible with the computing infrastructure commonly available in Moroccan courts of first instance.

The proposed framework nevertheless has several limitations. First, the extraction pipeline was developed and validated using procedural court orders collected from

Moroccan courts of first instance between 2020 and 2025. Its applicability to documents from other court levels or jurisdictions may therefore require adaptation of the extraction component. Second, the classifier was trained to recognize only seven procedural categories. Extending the framework to additional legal document types would require new annotated data and retraining of the classification pipeline. Finally, the study evaluates the system on a single national corpus; future work should investigate its generalization to broader collections of Arabic legal documents and to judicial systems beyond Morocco.

V. CONCLUSION

This study presented an end-to-end framework for the automated classification of Moroccan procedural court orders. The proposed pipeline addresses a practical challenge that has received limited attention in previous studies: the lack of standardized document structure in Arabic legal Word documents. A custom Python extraction script automatically converts heterogeneous documents into structured representations by extracting the facts, decision, and confidence score, enabling the classifier to focus on the most informative textual content.

The framework was evaluated on a manually collected corpus of 3,894 procedural court orders from Moroccan courts of first instance (2020–2025), covering seven official procedural categories. After expanding the training set to 13,764 documents through domain-specific data augmentation, five machine learning classifiers were evaluated using frozen CAMeL-BERT embeddings. Logistic Regression achieved the best performance, reaching 99.74% test accuracy with a Cohen's κ of 0.9966. The ablation study further demonstrated that the strong performance results from the combination of highly discriminative contextual embeddings and the structural characteristics of Moroccan procedural court orders, while data augmentation effectively improves robustness for minority categories.

Although the proposed framework demonstrates excellent performance, several directions remain for future research. Extending the corpus to additional court levels, legal document types, and Arabic-speaking jurisdictions would improve the generalizability of the approach. A fine-tuned CAMeL-BERT baseline was not included in this study. Consequently, the effectiveness of frozen contextual embeddings is demonstrated for this corpus but was not directly compared with end-to-end transformer fine-tuning. Future work will investigate this comparison.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education, Scientific Research, and Innovation; the Digital Development Agency (DDA); and the National Centre for Scientific and Technical Research (CNRST) of Morocco under grant Alkhawarizmi/2020/25.

REFERENCES

[1] N. Y. Habash, Introduction to Arabic Natural Language Processing, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2010.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. NeurIPS, Long Beach, CA, Dec. 2017, pp. 5998–6008.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, Jun. 2019, pp. 4171–4186.

[4] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in Proc. 4th Workshop on Open-Source Arabic Corpora and Processing Tools, Marseille, May 2020, pp. 9–15.

[5] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," in Proc. 6th Arabic NLP Workshop, Kyiv, Apr. 2021, pp. 92–104.

[6] A. S. Alammary, "BERT models for Arabic text classification: A systematic review," Applied Sciences, vol. 12, no. 11, Art. no. 5720, Jun. 2022.

[7] T. El Moussaoui, A. Es-salmi, J. Boumhidi, and C. Loqman, "Automating legal document classification with XLM-RoBERTa: A case study on the Moroccan Court of Cassation," in Digital Technologies and Applications, ICDTA 2025, Lecture Notes in Networks and Systems, vol. 1640, Cham: Springer, 2026, ch. 30. doi: 10.1007/978-3-032-07785-1_30.

[8] A. Bouhouche, M. Esghir, and M. Errachid, "Classification of Moroccan legal and legislative texts using machine learning models," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 10, pp. 1108–1114, 2024.

[9] J. Zahir, "Prediction of court decision from Arabic documents using deep learning," Expert Systems, vol. 40, no. 6, Art. no. e13236, Jun. 2023.

[10] T. El Moussaoui, C. Loqman, and J. Boumhidi, "Decoding legal processes: An AI-driven system for criminal records in Moroccan courts," Intell. Syst. Appl., vol. 25, Art. no. 200487, Mar. 2025.

[11] A. M. Mutawa and S. Sruthi, "A comparative evaluation of transformers and deep learning models for Arabic meter classification," Applied Sciences, vol. 15, no. 9, Art. no. 4941, Apr. 2025.

[12] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT and MARBERT: Deep bidirectional transformers for Arabic," in Proc. 59th Annual Meeting of the ACL, Aug. 2021, pp. 7088–7105.

[13] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," in Findings of ACL: EMNLP 2020, Nov. 2020, pp. 2898–2904.

[14] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "MultiEURLEX—A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer," in Proc. EMNLP 2021, Online and Punta Cana, Nov. 2021, pp. 6974–6996.

[15] M. Al-Qurishi, S. AlQaseemi, and R. Souissi, "AraLegal-BERT: A pretrained language model for Arabic legal text," in Proc. Workshop on Natural Legal Language Processing, Abu Dhabi, Dec. 2022, pp. 338–344.

[16] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[17] O. Tahtah, M. Bahbib, A. Zinedine, and K. Fardousse, "A CNN-BiLSTM hybrid architecture with resampling techniques for Arabic legal text classification," Eng. Technol. Appl. Sci. Res., vol. 15, no. 6, pp. 29062–29068, Dec. 2025.

[18] Y. Sun, C. Qiu, Y. Xu, Z. Wang, and H. Chen, "How to fine-tune BERT for text classification?" in Chinese Computational Linguistics, Cham: Springer, Oct. 2019, pp. 194–206.

[19] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[20] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in English," in Proc. 57th Annual Meeting of the ACL, Florence, Italy, Jul. 2019, pp. 4317–4323.

[21] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective," PeerJ Comput. Sci., vol. 2, Art. no. e93, Oct. 2016.