Fine-Tuning Arabic and Multilingual BERT Models for Crime Classification to Support Law Enforcement and Crime Prevention

Njood K. Al-harbi, Manal Alghieth
Department of Information Technology-College of Computer, Qassim University, Buraydah, Saudi Arabia

***Abstract*—Safety and security are essential to social stability since their absence disrupts economic, social, and political structures and weakens basic human needs. A secure environment promotes development, social cohesion, and well-being, making national resilience and advancement crucial. Law enforcement struggles with rising crime, population density, and technology. Time and effort are required to analyze and utilize data. This study employs AI to classify Arabic text to detect criminal activity. Recent transformer methods, such as Bidirectional Encoder Representation Form Transformer (BERT) models, have shown promise in NLP applications, including text classification. Applying these models to crime prevention motivates significant insights. They are effective because of their unique architecture, especially their capacity to handle text in both left and right contexts after pre-training on massive data. The limited number of crime field studies that employ the BERT transformer and the limited availability of Arabic crime datasets are the primary concerns with the previous studies. This study creates its own X (previously Twitter) dataset. Next, the tweets will be pre-processed, data imbalance addressed, and BERT-based models fine-tuned using six Arabic BERT models and three multilingual models to classify criminal tweets and assess optimal variation. Findings demonstrate that Arabic models are more effective than multilingual models. MARBERT, the best Arabic model, surpasses the outcomes of previous studies by achieving an accuracy and F1-score of 93%. However, mBERT is the best multilingual model with an F1-score and accuracy of 89%. This emphasizes the efficacy of MARBERT in the classification of Arabic criminal text and illustrates its potential to assist in the prevention of crime and the defense of national security.**

***Keywords*—Artificial intelligence; deep learning; natural language processing; bidirectional encoder representation from transformer; crime classification; crime prevention; tweets; text classification; transformer; Arabic; X**

I. Introduction

Safety and security are essential pillars of a stable and functioning society, serving as fundamental prerequisites for meeting basic human needs. In their absence, individuals and communities face significant challenges in achieving personal and collective goals. A secure environment fosters social cohesion, economic development, and overall well-being, enabling societies to progress and thrive. Presently, as mobile data technology advances and social media users gain easier access to the internet, the volume of crime-related data that requires analysis increases proportionally. Much of this information is unorganized and presented as free text. Consequently, approaches are developed to manage unstructured data [1]. In the rapidly growing field of mobile technology, social media platforms have gained huge popularity as a preferred means of communication for exchanging private messages and sharing thoughts, videos, and images. Furthermore, it has the potential to serve as a reliable and comprehensive platform for global news coverage including both political and social aspects. Numerous social media networks, such as X, Facebook, Instagram, and Snapchat, are currently in use. X is a popular application as a means of sharing informal messages, and thoughts, as well as facilitating the transmission of political news as approved by [2], X is the most common platform for sharing political activities by 43% [2]. This platform is widely utilized by a significant number of individuals globally, including Arab nations. This renders it a suitable platform for use in the present research, which aims to examine Arabic tweets on the X platform to analyze criminal activities and develop an appropriate solution for the detection of crime. Consequently, this contributes to the prevention of crime and the enhancement of law enforcement within the nation.

Law enforcement faces significant challenges mostly associated with the rise of crime-related data, due to increased crime rates, population density, and technological advancements. The analysis and utilization of the data require an extensive amount of effort and time. Text classification is a crucial task in NLP across several applications, including topic classification, question answering, and sentiment analysis, which is of most popular use [3], to achieve the state-of-the-art result in our text classification of crimes, we will use BERT-based models, a Transformer model that was widely used recently. The BERT framework was introduced through two steps: pre-training and fine-tuning. Initially, during pre-training, the model was trained on unlabeled data across various tasks. Subsequently, in the fine-tuning step, the BERT model was initialized with the pre-trained parameters, and it was then refined using labeled data for downstream tasks. BERT large and base are introduced in the original BERT paper. Each version supports cased and uncased text. The training uses only raw English text for labeling without human intervention [4]. BERT has been developed to accommodate several languages, including multilingual BERT (mBERT) [4], XLM-RoBERTa (XLM-R) [5], and DistilBERT [6]. Some of these versions are specifically designed for distinct languages, such as AraBERT [7], MARBERT, ARBERT [8], and ArabicBERT for Arabic. This research will collect a new dataset from the X platforms.

The dataset will be labeled and prepared for the model, and it will be used to fine-tune various Arabic BERT-based models and multiple support BERT-based models to identify the optimal model.

AI has transformed a variety of industries, including law enforcement. NLP has been increasingly employed in the context of predictive policing and criminal detection. Several studies have investigated the NLP in ML and DL methods for the analysis of criminal reports and social media texts.

Despite advancements, current research still faces challenges in collecting large amounts of sensitive data from law enforcement agencies, identifying patterns in various languages, and ensuring ethical considerations, such as bias in predictions due to unbalanced data. Additionally, numerous law enforcement agencies continue to depend on traditional criminal detection methods, which restrict the potential of AI-driven approaches. Considering these problems, it is essential to enhance crime detection and NLP tasks via BERT-based models to improve accuracy, efficiency, and fairness in crime prediction. The current research aims to fine-tune different Arabic BERT-based models for crime classification by collecting new Arabic data from the X platform, balancing data, and discussing ethical AI considerations to help law enforcement make data-driven decisions.

The purpose of the proposed solution is to enhance crime prevention and law enforcement in Arabic. The aims and objectives encompass:

- Aims:
  - Develop an effective technique for crime classification by comprehensive analysis of Arabic language texts from social media platforms.
  - Enhance crime prevention and law enforcement in Arabic by the application of NLP leveraging transformer techniques.
- Objectives:
  - Analyze previous Arabic studies to identify the most appropriate methodology and determine the research gap.
  - Based on the analysis, gather the Arabic data from the X platform and prepare it to be applied to the chosen technique.
  - Design the proposed solution utilizing the selected technique, which is the BERT transformer.
  - Implement and evaluate the reliability of results to enable seamless integration with legal standards.

According to the aims and objectives, the main research questions are as follows: Q1: Can AI discover and assist in classifying crimes in Arabic textual data? Q2: Can transformer BERT improve the effectiveness of Arabic crime classification based on pre-existing models? Q3: Can the Arabic language be identified despite its difficulties?

The research focuses on crime and law enforcement in Arabic tweets collected from the X platform, aiming to categorize Arabic criminal text, utilizing Arabic and multilingual BERT-based models to assess optimal performance. Hence, this research can be leveraged by the Ministry of Interior or the public prosecution.

The subsequent sections of the study are organized as follows: Section II provide background information about the model. Section III provides an overview of the related work. Section IV outlines the methodologies and materials employed for experiments. Section V presents the result and Section VI presents the discussion of the research. And finally, Section VII presents the conclusion and outlines recommendations for future research.

II. BACKGROUND

Large language models (LLMs), which are models designed to comprehend and produce text at the level of the human language, are constructed using a massive amount of data for training. The term "large" denotes an LLM with a huge number of parameters. LLMs are utilized in a variety of contexts and possess immersive capabilities within NLP tasks: 1) Natural language understanding (NLU), including sentiment analysis and text classification. 2) Generation of texts, including chatbots and question-and-answer systems. LLMs are built based on DL architectures like transformers, allowing them to learn from and process vast amounts of data. BERT is incorporated into LLM models, and both play a significant role in NLP tasks involving sequential text understanding and improvement [9].

BERT is a pre-trained language model (PLM) based on transformer architecture. The transformer architecture incorporates an attention mechanism that acquires learning of the contextual relationships among words and sub-words inside the given text. It consists of two separate mechanisms: the encoder and the decoder. The encoder is responsible for processing the textual input, while the decoder generates the output for the task.

BERT is an abbreviation for bidirectional encoder representation built around the transformer architecture. In 2018, researchers affiliated with Google AI authored twelve publications, which showcases promising outcomes in NLP tasks including text classification. It is a framework that was introduced through two steps: pre-training and fine-tuning. Initially, during pre-training, the model was trained on unlabeled data across various tasks. Subsequently, in the fine-tuning step, the BERT model was initialized with the pre-trained parameters, and it was then refined using labeled data for downstream tasks. The BERT learning process is illustrated in Fig. 1, with the initial step training on a large amount of text and the subsequent step training on a specific task with a labeled dataset.

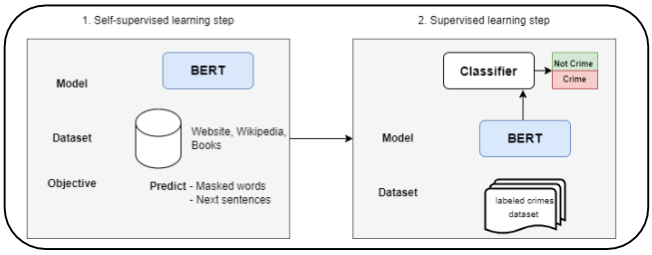

Fig. 1. BERT Learning steps [10].

The BERT model architecture is a multi-layer bidirectional transformer encoder based on the original implementation by the authors [11]. It was introduced in two variations: BERT base (L=12, H=768, A=12, Total Parameters=110M) and BERT large. (L=24, H=1024, A=16, Total Parameters=340 million). L represents the number of layers, H represents the hidden size, and A represents the number of thirteen self-attentions. Each variation supports both cases and uncases input text. The training process exclusively utilizes raw English text without any human involvement in the labeling process [4]. However, there exist several versions of BERT that have been developed to support different languages such as multilingual BERT (mBERT) [4], XLM-RoBERTa (XLM-R) [5], and DistilBERT [6]. Some of these versions are specifically designed for distinct languages, such as AraBERT [7], MARBERT, ARBERT [8], and ArabicBERT for Arabic, bert-base-chinese for Chinese [4], AM-BERT, AM-RoBERTa for Amharic [12], FlauBERT for Frensh [13], and BanglaBERT for Bangla [14].

Furthermore, there are specialized versions of BERT for certain domains, such as ClinicalBERT for the clinical domain [15], LEGAL-BERT for the legal domain [16], and FinBERT for the financial domain [17].

Self-Attention Mechanism is one of the most critical concepts in BERT. It is regarded as a unique form of attention that was initially introduced with the transformer model, which is an attention mechanism that calculates the contextual representation from each sequence by linking distinct elements in a single sequence. Single self-attention is illustrated in Fig. 2, which displays the word "it" in each sentence. Another mechanism is multi-head attention. The link will be linearly projected multiple times with various learned linear projections rather than a single link. This will allow for the joint attention of information from multiple representation subspaces at different positions.

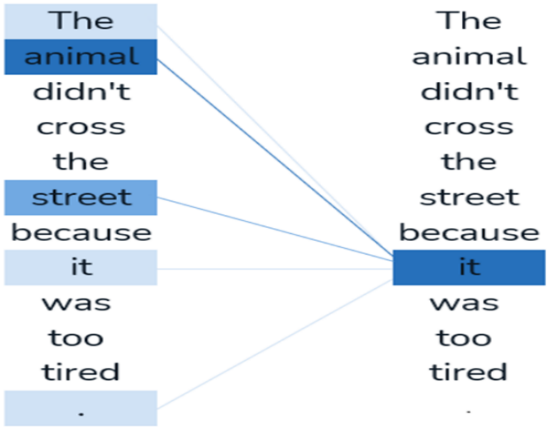

Fig. 2. Single self-attention mechanism [11].

As indicated in the definition of BERT, it is a model consisting of multiple stacked encoder layers. Each layer processes the input using multi-head self-attention and feed-forward network (FFN) layers to capture contextual information. Multi-head self-attention is dependent on three critical components: Query (Q), Key (K), and Value (V), which are linear transformations of the data input, as illustrated in Eq.(1):

$$MultiHead(Q,K,V) = Concat(head_1,\dots,head_h)W^O$$

$$where\ head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

$$Where\ the\ projections\ are\ parameter\ matrices\ W_i^Q \in \mathbb{R}^{d_{model}\times d_k}, W_i^K \in \mathbb{R}^{d_{model}\times d_k}, W_i^V\ \mathbb{R}^{d_{model}\times d_v}\ and\ W^O \in \mathbb{R}^{hd_v\times d_{model}}. \quad (1)$$

The output of the attention block is fed to the Feed-Forward network (FNN) following the multi-head self-attention mechanism. The position-wise fully connected FFN is applied to each location individually and identically, including two linear transformations. The subsequent Eq. (2) demonstrates that:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (2)$$

In BERT, the FFN employs the GELU activation function rather than ReLU, and it is specified, as indicated in Eq. (3).

$$GELU(x) = 0.5x\left(1 + tanh\left(\sqrt{2/\Pi}(x + 0.044715x^3)\right)\right) \quad (3)$$

A residual connection is implemented around each of the two sublayers by the encoder layer in BERT. The normalization layer follows, with the output of each sublayer defined as in Eq. (4) and Sublayer(x) representing the function implemented by the sublayer.

$$LayerNorm(x + Sublayer(x)) \quad (4)$$

The sublayers will generate outputs that are equivalent in size to *dmodel* to facilitate residual connections. Fig. 3 illustrates the architecture of the BERT base and BERT large alongside the configuration of a single encoder.

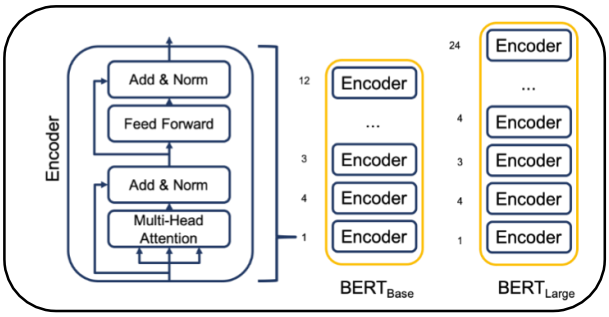

Fig. 3. Single BERT architecture with single encoder configuration [10].

The process and encoding of raw text in a format that is comprehensible and conducive to learning is referred to as input representation in BERT. It is a structured combination of three primary components: token, segment, and position embeddings [4].

- Token embedding is the first step in BERT when text is transformed into tokens through a tokenizer. The tokenizer divides the text into smaller chunks known as tokens (words or subwords). Each token is associated with a vector (embedding) using a pre-trained embedding matrix. BERT employs WordPiece embedding, which comprises a vocabulary of 30,000 and sixteen tokens. A special classification token [CLS] is included as the initial token in each sentence. The second special token is [SEP], which has two purposes: first, to

separate two sentences, and second, to denote the end of a sentence. The third special token is [PAD], which denotes empty tokens. When sentences are shorter than the predetermined fixed max length, the remaining tokens will be filled with padding tokens.

- Segment embedding is employed to manage pairs of sentences, namely sentences A and B. Each token will be assigned a segment ID of either 0 or 1 to denote its corresponding sentence. Assist in distinguishing between the two segments of the input sequences, as all tokens in sentence A are assigned an ID of 0, whereas sentence B is assigned an ID of 1. It is useful in question-answering tasks and sentence classification.
- Position embeddings represent token positions in the sequence of the input. These embeddings happen through pre-training. The final input representation combines each token's position embedding with the token and segment embeddings. The input representation in Fig. 4 is the sum of the embeddings for tokens, segments, and positions.

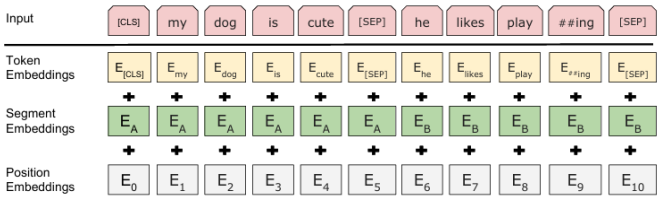

Fig. 4. Input representation for BERT [4].

BERT is pre-trained using two objectives: the Masked Language Model (MLM) and the Next Sentence Prediction (NSP) in conjunction. MLM is the process of randomly masking some words in input data. The goal is to estimate the original vocabulary of the masked words by considering the context provided by the unmasked words. MLM reads the sentence bidirectionally, both from left to right and from right to left, allowing the model to gain a comprehensive understanding of the language context. That's contrasted to other pre-trained language models that only read unidirectionally from left to right or from right to left [10].The NSP involves feeding the model two sentences and asking it to determine whether the second sentence in the pair is the subsequent sentence to the first. This method facilitates the comprehension of the relationships between sentences, and it is a crucial capability in NLP tasks such as summarization, question answering, and text classification. Fig. 5 shows the Comprehensive BERT pre-training and fine-tuning [10].

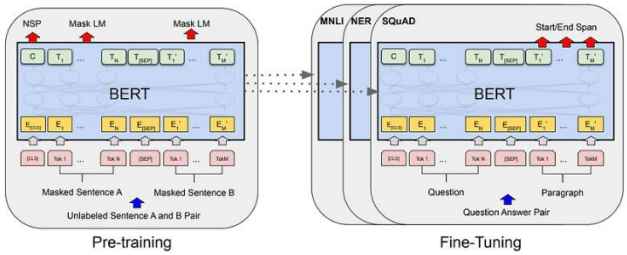

Fig. 5. Input representation for BERT [4].

III. Related Work

A systematic review was conducted for BERT to collect studies that focused on low-resource languages, particularly Arabic, for criminal classification. The studies that were compiled are detailed in the subsequent sections.

A. Description of the Model and Classification

Low-resource languages are those that possess a limited number of resources and require further exploration, such as Urdu, Arabic, and Estonian. High-resource languages are those that possess substantial support and resources, such as English, Chinese, Japanese, and Spanish.

This section provides comprehensive information on the model used in each study, including the year of publication (in chronological order), the model's name, the languages supported, the classification task, and the data type, as demonstrated in Table I for the low-resource languages.

The models employed in low-resource languages primarily include mBERT, with eleven studies supporting various languages. AraBERT is utilized in four studies, exclusively supporting language. Distlm-BERT is utilized in three studies, encompassing 104 languages. XLM-RoBERTa is utilized in two studies, encompassing 100 languages. Additionally, there is a single study for each of the following models.

MarBERT, which is specifically designed for the Arabic language. Bangla-BERT which was particularly designed for the Bengali language. The XLM-100 is designed to provide support for 100 languages. The classification task varies depending on the content of the data, although all fall under the umbrella of illegal actions in broad terms. The data mostly consists of social media posts and text, with additional kinds including court or tribunal documents and newsgroup data.

B. Description of the Dataset

This section offers a thorough overview of the entire dataset process, encompassing details such as the language of the data, data source (whether it is available online or newly collected), number of categories (binary or multi), types of data categories (including various crime types), dataset size, labeling method (manual or pre-annotated), whether the dataset is balanced or not, the technique employed if the dataset is not balanced, the size of the dataset after balancing, and finally, the specifics of data splitting. All the details are outlined in Table II. The majority of the dataset languages belong to Arabic, accounting for 38.46% (five studies). Bengali follows with a rate of 23.08% (three studies), and then Hindi with a rate of 15.38% (two studies) while the remaining languages, including Estonian, Italian, Greek, Polish, and Urdu for Pakistan, each have a rate of 7.69% (one study) for each. Certain datasets are obtained from Kaggle, while others are not accessible online. The number of categories varies, with some being binary and others being multi categorical (ranging from three to eleven categories). Additionally, the size of the data varies. Among these, the smallest dataset consists of 1,670 entries, while the largest datasets contain millions of entries. However, the small datasets are labeled manually, whereas the largest dataset is pre-annotated. Out of the total of thirteen datasets, 53.84% of them are unbalanced. Among these unbalanced datasets, 46% (six datasets) do not employ any technique to balance the data, while

just one study 7.69% (one dataset) makes use of the over and under sampling technique. Out of the total number of datasets, which is two, are balanced, and the remaining four datasets are balanced and not balanced using a different dataset. When splitting the dataset 38.46% of the studies employed the method of train, test, and validation splitting. 30.76% of the studies employ train and test splitting, while the remaining studies do not disclose the splitting details. When employing the three splitting techniques, the splitting percentages are as follows: for training, 80%, and 70%; for testing 20%, and 10%; and for validation 10%. If the two splitting techniques are applied (train 90%, 80%, and test 20%, 10%).

TABLE I. DESCRIPTION OF THE MODEL AND CLASSIFICATION FOR LOW-RESOURCE LANGUAGES

Article	Model Name	Supported Language	Classification Task	Type of Data
[18]	mBERT (uncased)	104	Domain classification	Documents
[19]	AraBERT base	Arabic	Detecting offensive language	Social media Tweets
[20]	mBERT(cased), XLM-100 (cased), Distilm-BERT (cased), XLM-R base	104 100 104 100	Text classification	Paragraphs
[21]	AraBERT base mBERT	Arabic 104	Detecting offensive language	Social media Tweets
[22]	mBERT	104	hate speech classification	Social media Tweets
[23]	mBERT	104	Racist and xenophobic hate speech classification	Social media texts
[24]	mBERT,	104	Aggressive text detection	Social media Posts
[25]	mBERT (uncased), Distilm-BERT (cased), Bangla-BERT base, XLM-R base	104 104 Bengali 100	Aggressive content detection	Social media texts
[26]	mBERT	104	Aggressive, hate, and abuse detection	Social media texts
[27]	AraBERT base MarBERT	Arabic	Detecting offensive language	Social media Tweets
[28]	mBERT AraBERT base	104 Arabic	Multilingual Offensive Language Detection task	Social media Tweets
[29]	mBERT	104	Hateful content detection	Social media Tweets
[30]	mBERT	104	Crime Text Classification and Drug Modeling	Bengali News Articles

TABLE II. DESCRIPTION OF THE DATASET FOR LOW-RESOURCE LANGUAGES

Article	Language	Source	Categories	Categories of Data	Size	Way of labeling	Balance or no	Use A Technique to Balance the Data or no	After Balance	Splitting the Data
[18]	Bengali	Open source BARD, OSBC ProthomAlo	5 11 6	Include crime	50,560 78,796 128,761	Annotated	No	No	-	Train:80% Test:20%
[19]	Arabic	X Platform	2	Offensive Not offensive	10,000	Annotated	No	Yes, Over/under-sampling	-	Train:70% Test:20% Validation:10 %
[20]	Estonian	Postimees Estonian newspaper	4	Include crime Topic: Negative Ambiguous Positive Neutral	4,088	Annotated with sentiment & with rubric labels.	No	No	-	Train:70% Test:20% Validation:10 %
[21]	Arabic	X Platform	4	- Offensive - Vulgar - Hate speech - Clean	10,000	Experience annotator	No	No	-	N/A
[22]	Indonesian Polish Arabic	Public Dataset from X platform	2	-hate speech -normal	13,169&713 9,788 4,120&1,670	Annotated	Yes No No & YES	No	-	Train:70% Test:20% Validation:10 %

[23]	Greek Italian	PHARM datasets	2	- racist/xenophobic hate tweets -non-racist/xenophobic hate tweets	10,399 10,752	Manual	Yes	-	-	N/A
[24]	Hindi	TRAC	3	Non-aggressive (NAG), Overtly Aggressive (OAG), & Covertly Aggressive (CAG)	15,001	Annotated	Yes	-	-	N/A
[25]	Bengali	BAD Facebook & YouTube	2 4	AG & NoAG ReAG & PoAG & VeAG & GeAG.	14,443	Manual	Yes No	No	-	Train:80% Test:10% Validation:10 %
[26]	Hindi	8 datasets from social media	2 3	Hate, normal Hostile, non-hostile CAG, NAG, OAG Abusive, hate, natural	Different sizes	Annotated	Some are balanced and some not	No	-	Train:80% Test:20%
[27]	Arabic	X Platform	2	Offensive Not offensive	12,700	Annotated	No	No	-	Train:70% Test:20% Validation:10 %
[28]	Arabic	SemEval'2020 competition Arabic dataset	2	1=Offensive 0=Not offensive	7,800	Annotated	No	No	-	Train:80% Test:20%
[29]	Urdu (Pakistan)	X Platform	2	Hatful Neutral	21,759	Manual	No	No	-	Train:90% Test:10%
[30]	Bengali	Kaggle & Bangla Newspaper	2 4	Crime & Others Murder, Drug, Rape & Others	approximately 5.3 million entries	Annotated & Manual	N/A	N/A	-	N/A

C. Model Evaluation

In this section, the evaluation of a model is initially determined by the type of evaluation metrics employed in each study. Secondly, in their publication, do they include a comparison with other ML, DL, or Transformer-based models? Thirdly, comparison with prior studies, and finally, the most robust result. These are detailed in Table III. As the table below demonstrates, the primary evaluation metric is the F1 score, which is utilized in 77% of the studies. The precision and recall metrics are the second most used, with a rate of 54%. The accuracy metric is the third most used, appearing in 46% of studies. Other metrics include micro F1 scores, weighted F1 score (WF), and macro F1 scores. When it comes to evaluating comparisons in studies, 84.61% (eleven studies) do not provide comparisons with previous studies, whereas 15.38% (two studies) do include comparisons. When it comes to comparing with their study, 76.92% of studies compare, whereas 23.07% do not.

TABLE III. MODEL EVALUATION FOR LOW-RESOURCE LANGUAGES

Article	Evaluation Metrics	Compare the Result with Other Models in Their Paper or not	Compare the Result with Previous Studies or not	Best Result
[18]	Precision, Recall, F1 score, Accuracy	Yes, ELECTRA	No	ELECTRA gets the best accuracy in all datasets
[19]	Macro-F1 score	No	No	AraBERT achieved 90%
[20]	Accuracy	Yes, fastText	No	XLM-RoBERTa achieved the highest & DistilmBERT the lowest
[21]	Precision, Recall, F1 score	Yes, fastText, SVM, Decision Tree, Random Forest, GaussianNB, Perceptron, AdaBoost, Gradient Boosting, Logistic Regression	No	AraBERT achieved the highest F1 score of 83%, while mBERT achieved 76%.
[22]	F1 score	Yes, MUSE + CNN-GRU Translation + BERT LASER + LR mBert	No	mBERT is superior in Arabic with a f1 score of 83% and in Indonesian with a f1 score of 81%. In Polish, the translation with bert is superior with a score of 71%.
[23]	Accuracy and F1-score	No	No	mBERT achieves an accuracy rate of 91% in Italian and 81% in Greek.

[24]	Precision, Recall, F1 score, WF	Yes, 16 traditional & deep neural classifiers	No	CNN is better with a WF of 64%
[25]	Precision, Recall, F1 score, Error, WF	Yes, LR, RF, NB, SVM, CNN, BiLSTM & CNN + BiLSTM	Yes	In WF metrics The 2-class XLM-RoBERTa is the Best In 4 classes Bangla-BERT is the best Outperforms previous studies
[26]	weighted-F1	Yes, MuRIL, M-BERT-Bilstm, MuRIL-Bilstm, and cross-lingual information.	Yes	The best model is cross-lingual information with a rate of 95%
[27]	Precision, Recall, F1 score, Accuracy	Yes, baseline models	No	MarBERTv2 outperforms AraBERT and other baseline models with 84% F1-score and 86% accuracy
[28]	Accuracy and F1-score	Yes, CNN, RNN, bidirectional RNN, ULMFiT, ELMo, SVM & combined models	No	AraBERT has the highest F1 score 93% in Arabic, surpassing models they compared it against, and 91% accuracy.
[29]	Precision, Recall, F1 score, Accuracy	Yes, NB, SVM, LR, RF, CNN, LSTM and BiLSTM	No	mBERT is the most effective model, obtaining an F1 score of 0.83%.
[30]	Precision, Recall, F1 score	No	No	mBERT in Crime classification had 96% Precision and crime type had 98% recall.

D. Model Hyperparameters

This section presents the hyperparameters derived from the author's specifications in each study, comprising the hidden size, learning rate, batch size, epoch number, and model name. Details are outlined in Table IV. The epoch number is a value that falls inside a range, which can be a tiny number (3 to 8), or a medium number (16 to 20). The batch size options are twelve, sixteen, and thirty-two. The learning rates most utilized are 2e-5 and range between 2e-5 and 5e-6. The batch size does not surpass thirty-two, and the number of epochs is moderate.

TABLE IV. MODEL HYPERPARAMETERS FOR LOW-RESOURCE LANGUAGES

Article	Model Name	Epoch Number	Batch Size	Learning Rate	Hidden Size
[18]	mBERT (uncased)	20	16	N/A	768
[19]	AraBERT base	5	32	N/A	
[20]	mBERT(cased), XLM-100 (cased), Distilm-BERT (cased), XLM-R base	(8-16)	N/A	(5e-5, 3e-5, 1e-5, 5e-6, 3e-6)	768 1024 768 768
[21]	mBERT AraBERT base	30	N/A	5e-1	768
[22]	mBERT	(1- 5)	16	(2e-5, 3e-5, 5e-5)	
[23]	mBERT,	3	N/A	3e-5	
[24]	mBERT	3	32	2e-5	
[25]	mBERT (uncased), Distilm-BERT (cased), Bangla-BERT base, XLM-R base	20	12	2e-5	
[26]	mBERT	2	30	2e-5	
[27]	AraBERT base MarBERT	100 with an early stopping patience of 10	N/A	2e-5	
[28]	AraBERT base	5	32	N/A	
[29]	mBERT	3	32	768	
[30]	mBERT	N/A	N/A	N/A	

E. Critical Analysis

In four studies, it was determined that BERT-based Arabic models outperformed other ML and DL models and improved accuracy across various datasets. The initial study [21] in the Arabic language, aimed to identify offensive language through social media tweets. The study's primary contribution is the construction of a dataset comprising 10,000 tweets annotated with experiential annotators. The AraBERT and mBERT models are utilized to identify offensive tweets, and their performance is compared to several ML and DL models. The results indicate that AraBERT outperforms all the ML, DL models, and mBERT, thirty-two achieving an F1-score of 83%. The obstacle is that the sample is unbalanced, and the F1-score improves by employing balancing techniques. Using the same dataset from the authors of study [21], the second study [19] for detecting offensive social media tweets employs the AraBERT. They employ balance techniques that involve over- and undersampling. The results indicated that AraBERT has a 90% F1 score, which is superior to the result of [21]. However, they are not comparable to any of the preceding results or ML and DL models. Their findings suggest that balanced data is beneficial for the precise detection of offensive tweets. A further study [27], employs the same models as the prior study,

AraBERT, and incorporates other Arabic models, namely MarBERT, to identify offensive tweets. The dataset comprises 12,700 imbalanced tweets. In comparison to baseline models, the results indicate that MarBERT outperforms AraBERT and baseline models, achieving an F1-score of 84%. The outcome surpasses that of [21], despite the dataset's imbalance, showing that MarBERT outperforms AraBERT by 1%. However, when addressing the imbalance, the findings of [19], provide a superior performance with a 6% improvement. It is necessary to balance tweets to assess the efficacy of MarBERT in comparison to AraBERT. The last study in the Arabic language [28], was conducted using AraBERT and mBERT to offensive Arabic tweets. The class is binary, and the dataset is pre-annotated with a range of 5,000 to 7,000 entries. The issue at hand is the presence of unbalanced data, which has not been effectively addressed. They compare the ML and DL models but do not compare them to previous studies. The study indicates that AraBERT outperforms the other models, including mBERT with an F1-score of 93%, which is better than all previous Arabic studies. From one perspective, the researchers utilized an English dataset and then translated it into Arabic. This implies that there is currently no existing dataset available in Arabic and therefore, there is a need to generate one. Additionally, evaluate the balance of the data to determine if it affects the findings. On top of that, it is important to note that there is a lack of comparisons with previous studies, which may be due to the limited availability of studies in this thirty-three specific field. This factor should be taken into consideration. An additional consideration is that when balancing the dataset, one must assess the accuracy differences between MarBERT, AraBERT, and other Arabic models.

IV. METHODOLOGY

The Arabic language is considered a morphologically rich language, but it is low in resources compared to high-resource languages such as English. Consequently, the use of Arabic in NLP tasks is challenging. However, the emergence of transformers, such as BERT-based models, has contributed to effective language understanding. The present research will utilize six variants of BERT-based Arabic models and three models that support multiple languages. Arabic models' architecture is derived from the original BERT-based model, which is elaborated upon in detail in the Background section. BERT is a language model that is pre-trained and relies on transformer architecture. It is pre-trained with MLM and NSP objectives concurrently. MLM randomly mask words in incoming data. 15% of N input tokens are substituted. Those tokens are replaced 80% with [MASK], 10% with a random token, and 10% with the original token. The objective is to infer the original vocabulary of the obscured words by analyzing the context supplied by the unmasked ones. The MLM analyzes the sentence from left to right and right to left to understand the linguistic context. This differs from pre-trained language models that read left-to-right or right-to-left [10]. In the NSP, the model is given two sentences and asked to determine if they follow each other. This method improves inter-sentential connection understanding for NLP tasks, including summarization, question answering, and text classification. Fig. 5 shows BERT's extensive pre-training and fine-tuning [10]. The methodology employed in the proposed solution is structured into distinct steps, as shown in the following Fig. 6.

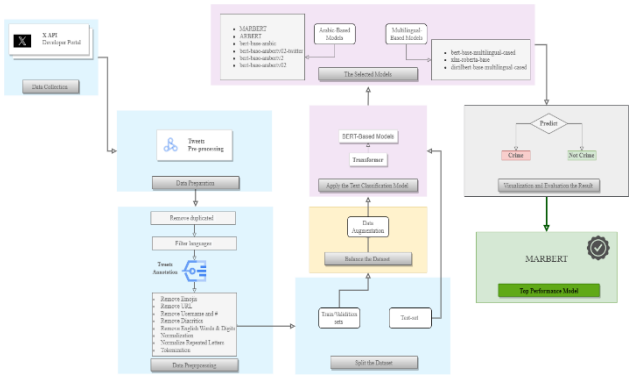

Fig. 6. The Proposed solution.

A. Data Collection Tool

The process of retrieving data using X API (Application Programming Interface), allows programmatic access to X in unique and advanced procedures. Utilizing it to analyze, learn from, and deal with tweets, direct messages, users, and other essential X resources involves the consideration of crime incidents and locations of interest [31]. Upon registration with the X developers and gaining access to the X API by using the Python library (tweetpy) to get access through API access, which enables to develop applications or scripts capable of executing diverse tasks, including: a) Retrieve tweets, users, and additional information from X, b) Post new tweets, retweets, and replies, c) Follow or unfollow users, d) Like or dislike tweets, e) Search for tweets containing specific keywords or hashtags, f) Stream real-time tweets according to specific criteria, and g) Examine tweet data, user profiles, and much more.

B. The Procedure of Searching Keywords and Hashtags

Next, extract tweets that include specific words or hashtags, as indicated in Table V, which presents the selected words and hashtags. Several Arab nations, including Saudi Arabia, Kuwait, Iraq, Jordan, Algeria, Tunisia, and Egypt, identified the most frequently committed offenses, according to the study [32]. The authors categorize crimes into major and minor offenses, which include assault, murder, smuggling, attempted murder, rape, fraud, theft, disturbing the peace, driving offenses, drunkenness, draft dodging, sexual assault, and other categories.

TABLE V. THE PROCEDURE OF SEARCHING KEYWORDS AND HASHTAGS

Keywords in Arabic	تلصص, ابتزاز, ترحيل, طرد, اغتيال, احتجاز, عنف, العنف المنزلي, اعتداء جسدي, تزوير, ,أدوات حادة, تعاطي المخدرات, اغتيال, تحرش هروب, حيازة المخدرات, اختلاس,غسيل الأموال, التخريب,اختطاف
Keywords Translation in English	Domestic Violence, Voyeurism, Blackmail, Deportation, Expulsion, Assassination, Detention, Violence, Sharp Objects, Drug Use, Harassment, Physical Assault, Forgery, Escape, Drug Possession, Embezzlement, Money Laundering, Vandalism, Kidnapping
Hashtags	#فساد, #قتل, #سطو, #مداهمة, #تهريب, #سرقة, #خيانة, #قتلة, #جريمة, #هروب, #سارق, #جرائم, #خونة, #لصوص, #جاسوس, #اغتصاب #اختناق, #طعن, #اضطهاد, #مذبح
Hashtags Translation in English	#Crime, #Murder, #Robbery, #Raid, #Smuggling, #Theft, #Treason, #Killers, #Corruption, #Thieves, #Spy, #Rape, #Traitors, #Crimes, #Escape, #Thief, #Massacre, #Suffocation, #Stabbing, #Persecution,

The retrieved tweets span the period from February 20, 2024, to March 12, 2024, with a total of 3,405 samples encompassing the ID, creation date, text, source, language, name, username, location, verification status, description, and URL. Subsequently, eliminate insignificant, unfilled columns, retain the text and language, and discard the source and location due to their emptiness. The language is retained because, even when specifying Arabic (ar), additional languages such as Urdu (ur), Sindhi (si), Pashto (pas), and Farsi (fa), which share a similar structure, are automatically retrieved.

C. Data Preparation

This part focuses on the pre-processing required to ensure the dataset is clean and suitable for the proposed model and the annotation procedure for categorizing incidents as criminal or non-criminal.

1) Data Pre-processing. Data preprocessing is an important component in NLP tasks, encompassing data cleaning, formatting, and transformation tasks. Moreover, it features engineering and selection. High-quality data is a crucial step in ML and DL, directly impacting the model's performance ability. Preprocessing data is a crucial step that must be undertaken before feeding it into a model or tool [33]. Data cleaning refers to the procedure of eliminating erroneous, corrupted, duplicate, or missing values from a dataset [33]. Pre-processing of the tweets was conducted utilizing NLP tools, employing multiple methods on tweets as shown in Fig. 7.

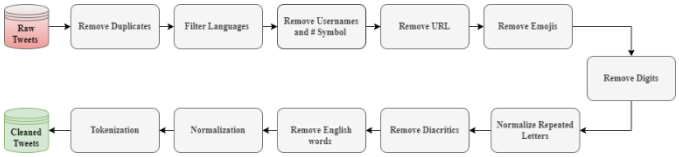

Fig. 7. Data Pre-processing.

The additional preprocessing employed regular expressions (RE), a built-in Python library, to manage string matching and replacements via the re.sub method. Eliminate punctuation including underscores, quotation marks, periods, and commas. Normalize successive identical characters to a single character and standardize letters such as "[إأآا]" to "ا" ,"ؤ" to "ء" ,"ئ" to "ء" ,"ة" to "ه" ,"ي" to "ى" ,"گ" to "ك". Furthermore, eliminate any emojis from tweets, as well as numerals and English letters or sentences. Furthermore, eliminate the URL, hashtag symbol, and mention. Finally, eliminate Arabic diacritics: Tashdid (ّ), Fatha (َ), Tanwin Fath (ً), Damma(ُ), Tanwin Damm(ٌ) , Kasra (ِ), Tanwin Kasr (ٍ), Sukun. Tokenization breaks up the text into discrete words, called tokens, to simplify processing and extract relevant information from tweets.

2) Data annotation. After eliminating irrelevant tweets, each tweet will be assigned a label indicating its association with a crime (0 = not crime, 1 = crime). The labeling procedure is done manually following the authors' guidelines [34] according to the presence of a description of a real crime. In this context, crime is generally defined to encompass reports of experienced or personally observed torture, interrogation, death, assault, psychological violence, military attacks, village damage, looting, and forced displacement. We limit our focus to binary classification, meaning that various acts of crimes were not further subdivided into subcategories.

D. Data Splitting

Splitting the dataset is an essential procedure for the model; during the training process, 80% of the data will be allocated for the training set and 20% for the testing set, while 10% of the training set will be reserved for the validation set.

E. Data Balancing

In this step, data balance is an essential part of developing an accurate model. Imbalanced data, where one class has more samples than the other, can result in biased predictions and suboptimal performance. This section will experiment with the oversampling technique to address the variations in sample sizes between the two classes. The dataset lacks balance. Fig. 8 illustrates the distribution of the training set. Therefore, this issue can be addressed by employing data augmentation approaches that involve solely oversampling the training set while preserving the test data.

Fig. 8. Train set before the balance.

V. Results

Fig. 9 displays the dataset following a balanced oversampling technique that improves the minority label by adding 817 samples, resulting in a total of 1,061 samples for both labels, crime and not crime. This technique utilizes an AraBERT-based augmentation, which employs substitution methods to replace words.

A. The Implementation

The environment employed is the Google colab Pro version. The hyperparameters are chosen based on the experiment and utilize many modules, including sklearn.model_selection. This module employs the ParameterSampler function, which generates random combinations of parameters sampled from specified distributions. The remaining hyperparameters were selected based on various experimental combinations. Table VI displays the Final Hyperparameter selection obtained from the different combinations of codes used during the experiment.

Fig. 9. Train set after the balance.

TABLE VI. FINAL HYPERPARAMETER CONFIGURATION FOR MODEL EVALUATION

Parameter Name	Final Parameter
Drop out	0.5
Learning Rate	3e-5
Number of Epochs	10
Batch size	32
Optimizer	Adam
Validation split	0.1
Weighted decay (L2 regularization)	0.1
Early stop patient parameter	2
Shuffle	True
Padding max length	70
Random seed	34

The model's performance will be evaluated using standard criteria for comparison with previous studies, including accuracy, recall, precision, and F1-score. The criteria are delineated and elucidated in Eq. (5), (6), (7), and (8).

Accuracy is a crucial evaluation statistic that measures the ratio of instances properly identified by the model. It is determined by the calculation outlined in Eq. (5):

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TN+FN+TP+FP}} \quad (5)$$

The recall metric measures the ability of a model to accurately identify and retrieve every instance belonging to a specific class within a given dataset, and it is calculated as follows in Eq. (6):

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

Precision is defined as the ratio of relevant instances to the total number of retrieved instances, as expressed by Eq. (7):

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

The F1-score is determined as the harmonic mean of precision and recall, represented by the following Eq. (8):

$$\text{F1 Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall+ Precision}} \quad (8)$$

B. Fine-Tuning Results

The most proficient Arabic model is MARBERT, attaining 93% accuracy, alongside equivalent F1-score, recall, and precision measures. Subsequently, bert-base-arabertv02 attained a score of 92%, while bert-base-arabic scored 91%, and the other models demonstrated comparable competence at 90%. The model with the lowest performance was bert-base-arabertv2, achieving a score of 88% as illustrated in Table VII.

TABLE VII. THE RESULTS OF ARABIC MODELS

Model	*Accuracy*	*F1-score*	*Recall*	*Precision*	*Loss*
bert-base-arabertv02	92%	92%	92%	92%	20%
MARBERT	**93%**	**93%**	**93%**	**93%**	25%
ARBERT	90%	90%	90%	90%	**19%**
bert-base-arabic	91%	91%	91%	91%	21%
bert-base-arabertv02-twitter	90%	90%	90%	90%	21%
bert-base-arabertv2	88%	88%	88%	89%	29%

Fig. 10 presents a comparison of accuracy and loss during training and validation against the test set. Train accuracy: 97% and Validation accuracy: 97%. Stopping training at epoch 4 by implementing early stopping to preserve optimal loss.

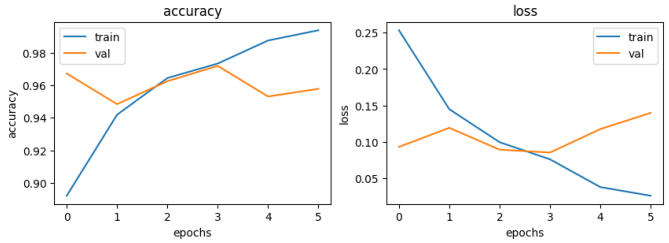

Fig. 10. Accuracy and loss results of MARBERT model.

Among models built for multilingual support, mBERT achieves the highest performance at 89%, followed by XLM-Roberta, and DistilBERT with scores of 87% and 85%, respectively as shown in Table VIII.

TABLE VIII. THE RESULTS OF MULTILINGUAL SUPPORT MODELS

Model	*Accuracy*	*F1-score*	*Recall*	*Precision*	*Loss*
mBERT	**0.8868**	**0.8887**	**0.8868**	**0.8912**	**0.2542**
XLM-R	0.8746	0.8722	0.8746	0.8703	0.2713
DistilmBERT	0.8501	0.8527	0.8501	0.8557	0.3644

Fig. 11 displays the accuracy and loss comparison of mBERT, indicating a model test loss of 0.27, a training loss of 0.15, and a validation loss of 0.11. Training accuracy: 0.9377 and Validation accuracy: 0.9437. Restoring the optimal loss and stopping at epoch 6.

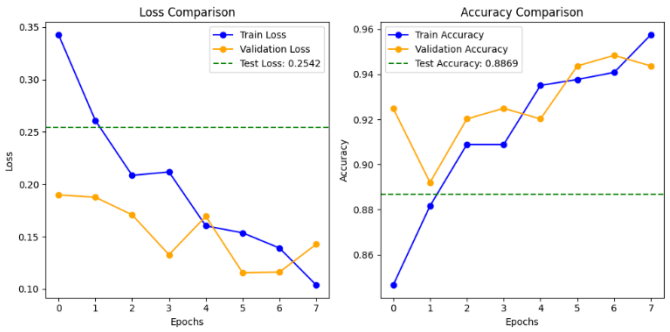

Fig. 11. Accuracy and loss results of the mBERT model.

VI. Discussion

Based on various experiments, MARBERT outperformed all selected Arabic BERT-based models as well as multiple support BERT-based models, achieving an F1-score and accuracy of 93%. The oversampling method and hyperparameter optimization improve performance and mitigate overfitting, which is the primary advantage of our proposed model. Following MARBERT is AraBERT version 02, and then ArabicBERT, although mBERT was the most effective among many multilingual support language models. The main finding is that almost all Arabic BERT-based models outperform mBERT.

A. Comparison with Previous Studies

Table I in the related work indicated that five studies utilized Arabic. Four studies employed a specialized Arabic model incorporating multiple support languages, whereas one study utilized solely a multiple support languages model. Table IX summarizes these models and their accuracy metrics compared to our best model.

TABLE IX. Comparison of Arabic Model Classification Results with Previous Studies

Model or Study	*Accuracy*	*F1-score*	*Recall*	*Precision*
AraBERT/ [21]	-	83%	82%	85%
MARBERT/ [27]	86%	84%	84%	84%
AraBERT/[28]	91%	93%	-	-
MARBERT/ ours	**93%**	**93%**	**93%**	**93%**
AraBERT/ ours	92%	92%	92%	92%

The MARBERT model outperformed all our Arabic models and previous studies in terms of accuracy, as it was trained differently from the other Arabic models. We employ identical model versions of previous studies utilizing diverse datasets and maintain consistent batch sizes and epochs, while optimizing some hyperparameters to achieve improved accuracy; nevertheless, they do not specify the loss, preventing comparison in this metric.

The majority of models were trained with numerous parameters and vocabulary sizes. ARBERT and MARBERT, as detailed in the authors' study [8], possesses 163 million parameters and a vocabulary size of 100,000, surpassing AraBERT, which has 136 million parameters and a vocabulary size of 60,000. The remaining models are outlined in the table. Another significance pertains to dialects, as the Arabic language encompasses numerous dialects, including Gulf Arabic, Egyptian Arabic, Levantine Arabic, Maghrebi Arabic, Iraqi Arabic, Sudanese Arabic, and Yemeni Arabic; thus, this model was trained on AD and MSA. The distinction between ARBERT and MARBERT lies in the fact that ARBERT focused exclusively on MSA, which accounts for its weakness compared to MARBERT. Ultimately, the most effective model classification for social media messages is the model that is trained in dialects and MSA.

B. Address the Research Questions

The research questions that are based on aims and objectives will be addressed in the subsequent paragraph. The primary aim of the present research is to improve crime prevention and law enforcement in Arabic by developing an effective technique, Arabic BERT-based models for crime classification, through multiple objectives: comprehensive analysis of previous Arabic studies, data construction, solution design, and evaluation of the results. Q1: Can AI discover and assist in the classification of crimes in Arabic textual data? As seen in Table VII, all the results were 90% and above, except for one study that got an 88% F1-score, which is a good result in the Arabic language, and all the Arabic studies do not exceed 93%, as seen in the comparison Table IX. So, AI helps in the classification of crimes. Q2: Can transformer BERT improve the effectiveness of Arabic crime classification based on pre-existing models? BERT-based models have demonstrated notable outcomes despite the constrained data size, whereas numerous prior models necessitate extensive datasets to achieve satisfactory accuracy. As the study by [35] on crime classifications across various ML and DL models attained the highest accuracy of 79% with the SVM model, even after preprocessing and balancing the dataset. Moreover, there is a substantial volume of tweets, around 37,000. Furthermore, the study [36] encompasses 1,555 Arabic tweets, achieving an average F1-score of 87% among several ML and DL models. The last study [1] received 92% of the 8K tweets. MARBERT achieved a 93% accuracy rate with a minimal number of tweets, indicating that pre-existing models enhance the classification of Arabic crime, even when the number of datasets is minimal. Q3: Can the Arabic language be identified despite its difficulties? The pre-trained models have been developed using broad Arabic datasets and dialects. In contrast to the challenges faced by traditional ML and DL, they have been trained to address various AD and MSA. Furthermore, several models were trained, especially on both MSA and DA, whereas others were trained exclusively on one. For instance, ArabicBERT was exclusively trained in MSA, but the others were trained in both types. Furthermore, certain models, particularly AraBERT version 02 Twitter, were trained on social media content from the X platform. The distinction is in the volume of training data, the parameters, and the vocabulary sizes that evolve with each model version, with MARBERT being the extensively trainable model.

C. The Constraint of Research

One of the most significant limitations of this research is the difficulty of data collection. As a result of the sensitivity of the data and the difficulty of acquiring it from various sources, we utilized social media platforms, specifically the X platform, to gather it. Additionally, one of the constraints we encountered

was the restricted quantity of tweets that were collected from the X platform. Furthermore, duplicated and unrelated languages were gathered due to the Arabic hashtag being used in other languages. This process resulted in the deletion and filtering of a significant number of non-Arabic languages. Ultimately, the experimentation of numerous models, which required a significant amount of time to accumulate all the results, is the final and most significant limitation.

VII. Conclusion and Future Work

Safety and security are essential to human needs and the stability of economic, social, and political systems. Technology, density, and increased criminality make law enforcement challenging. Safety can be achieved with AI. This research aimed to classify the Arabic language for the detection of criminal activities employing different Arabic BERT-based models. BERT has recently attracted considerable attention from researchers and practitioners, demonstrating notable effectiveness in various NLP tasks, including text classification. This efficacy can be attributed to its unique architectural features, particularly its ability to process text using both left and right context, having been pre-trained on extensive datasets. In the context of the criminal domain, the classification of data is a crucial activity, and transformers are increasingly recognized for their potential to support law enforcement efforts.

There was a limited number of crime field studies that employed the Arabic BERT transformer and a restricted number of Arabic crime datasets, considering the areas for improvement in previous studies. Hence, it is imperative to analyze the availability and efficacy of BERT in Arabic.

This was done by building newly posted tweets from X and classifying them into criminal and non-criminal categories. Subsequently, these tweets were processed using NLP tools, and their imbalance was resolved through the oversampling technique. Afterward, fine-tuning of BERT-based models, six variations of Arabic BERT models, and three multilingual models were utilized to classify tweets of criminal behavior, and the best-performing model was evaluated based on its accuracy and other performance metrics. Consequently, this contributed to the nation's enforcement of the law and the prevention of illicit activity. The results indicated that most Arabic models surpassed the multilingual models in efficacy. MARBERT, the leading Arabic model, attained an accuracy and F1-score of 93%, followed by AraBERTv02 by 92%, while ArabicBERT at 91%, and both ARBERT and AraBERTv02-twitter with an identical accuracy of 90%. The final Arabic model, AraBERTv2, had the lowest accuracy of 88%.

Nonetheless, mBERT is the most proficient model accommodating multiple languages, with an F1-score and accuracy of 89%, surpassing both XLM-R and DistilBERT, as well as just one Arabic variant, AraBERTv2. Furthermore, the processes of balancing and pre-processing facilitated the achievement of optimal findings while reducing overfitting, as the MARBERT model surpassed previous research in accuracy without exhibiting overfitting.

A. Future Works and Recommendations

For future research direction, the dataset will be expanded to include a third category, "lead to crime", to evaluate the model in multiclassification and optimize its performance. Assess the models using real reports from law enforcement, if accessible. Furthermore, assess additional developing Arabic language models and conduct a thorough comparison with other transformers, including ELECTRA, GPT, T5, and several other transformer models. Data availability remains a significant challenge for low-resource languages, particularly Arabic. Although the scarcity of datasets poses a barrier, the widespread use of social media platforms offers a potential solution. Platforms such as X and Facebook provide APIs for collecting posts, which could facilitate dataset creation. Furthermore, models such as BERT can be leveraged with a limited dataset; however, the data must be authentic and contain multiple categories to evaluate the model.

One critical aspect that requires investigation is the creation of a BERT-based crime model that is specifically tailored to Arabic crime and is trained on criminal activities to optimize its performance, such as the Legal-BERT for the English language. This research identifies critical areas that require further exploration and improvement in the application of BERT-based models for crime classification.

References

[1] A.-S. Hissah and H. Al-Dossari, “Detecting and classifying crimes from arabic twitter posts using text mining techniques,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 10, 2018.

[2] A. Waston, “Leading sources audiences pay the most attention to when consuming news on social networks worldwide as of February 2023,” statista. Accessed: Oct. 22, 2023. [Online]. Available: https://www.statista.com/statistics/1352912/top-news-audiences-pay-most-attention-to-social-media/

[3] R. Kora and A. Mohammed, “A Comprehensive Review on Transformers Models For Text Classification,” in 2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), 2023, pp. 1–7. doi: 10.1109/MIUCC58832.2023.10278387.

[4] J. Devlin, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[5] A. Conneau et al., “Unsupervised cross-lingual representation learning at scale,” arXiv preprint arXiv:1911.02116, 2019.

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arXiv preprint arXiv:1910.01108, 2019.

[7] W. Antoun, F. Baly, and H. Hajj, “Arabert: Transformer-based model for arabic language understanding,” arXiv preprint arXiv:2003.00104, 2020.

[8] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, “ARBERT & MARBERT: Deep bidirectional transformers for Arabic,” arXiv preprint arXiv:2101.01785, 2020.

[9] M. Zhou, J. Tan, S. Yang, H. Wang, L. Wang, and Z. Xiao, “Ensemble transfer learning on augmented domain resources for oncological named entity recognition in Chinese clinical records,” IEEE Access, 2023.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” CoRR, vol. abs/1810.04805, 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[11] A. Vaswani et al., “Attention is all you need,” Adv Neural Inf Process Syst, vol. 30, 2017.

[12] S. M. Yimam, A. A. Ayele, G. Venkatesh, I. Gashaw, and C. Biemann, “Introducing various semantic models for amharic: Experimentation and evaluation with multiple tasks and datasets,” Future Internet, vol. 13, no. 11, p. 275, 2021.

[13] H. Le et al., “FlauBERT: Unsupervised Language Model Pre-training for French,” in PROCEEDINGS OF THE 12TH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2020), N. Calzolari, F. Bechet, P. Blache, K. Choukri, C. Cieri, T.

Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., 55-57, RUE BRILLAT-SAVARIN, PARIS, 75013, FRANCE: EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA, 2020, pp. 2479–2490.

[14] A. Bhattacharjee et al., "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," arXiv preprint arXiv:2101.00204, 2021.

[15] E. Alsentzer et al., "Publicly available clinical BERT embeddings," arXiv preprint arXiv:1904.03323, 2019.

[16] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," arXiv preprint arXiv:2010.02559, 2020.

[17] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models. arXiv 2019," arXiv preprint arXiv:1908.10063, 2019.

[18] M. M. Rahman, M. A. Pramanik, R. Sadik, M. Roy, and P. Chakraborty, "Bangla Documents Classification using Transformer Based Deep Learning Models," in 2020 2ND INTERNATIONAL CONFERENCE ON SUSTAINABLE TECHNOLOGIES FOR INDUSTRY 4.0 (STI), 345 E 47TH ST, NEW YORK, NY 10017 USA: IEEE, 2020. doi: 10.1109/STI50764.2020.9350394.

[19] M. Djandji, F. Baly, W. Antoun, and H. Hajj, "Multi-task learning using AraBert for offensive language detection," in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 97–101.

[20] C. Kittask, K. Milintsevich, and K. Sirts, "Evaluating Multilingual BERT for Estonian," in HUMAN LANGUAGE TECHNOLOGIES - THE BALTIC PERSPECTIVE (HLT 2020), A. Utka, J. Vaicenoniene, J. Kovalevskaite, and D. Kalinauskaite, Eds., in Frontiers in Artificial Intelligence and Applications, vol. 328. NIEUWE HEMWEG 6B, 1013 BG AMSTERDAM, NETHERLANDS: IOS PRESS, 2020, pp. 19–26. doi: 10.3233/FAIA200597.

[21] H. Mubarak, A. Rashed, K. Darwish, Y. Samih, and A. Abdelali, "Arabic offensive language on twitter: Analysis and experiments," arXiv preprint arXiv:2004.02192, 2020.

[22] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "A deep dive into multilingual hate speech classification," in Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V, Springer, 2021, pp. 423–439.

[23] C. Arcila-Calderón, J. J. Amores, P. Sánchez-Holgado, L. Vrysis, N. Vryzas, and M. Oller Alonso, "How to detect online hate towards migrants and refugees? Developing and evaluating a classifier of racist and xenophobic hate speech using shallow and deep learning," Sustainability, vol. 14, no. 20, p. 13094, 2022.

[24] S. Modha, P. Majumder, and T. Mandl, "An empirical evaluation of text representation schemes to filter the social media stream," JOURNAL OF EXPERIMENTAL & THEORETICAL ARTIFICIAL INTELLIGENCE, vol. 34, no. 3, pp. 499–525, May 2022, doi: 10.1080/0952813X.2021.1907792.

[25] O. Sharif and M. M. Hoque, "Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers," Neurocomputing, vol. 490, pp. 462–481, Jun. 2022, doi: 10.1016/j.neucom.2021.12.022.

[26] P. Kapil and A. Ekbal, "A transformer based multi-task learning approach leveraging translated and transliterated data to hate speech detection in Hindi," Data Science and Machine Learning, pp. 191–207, 2022.

[27] A. Shapiro, A. Khalafallah, and M. Torki, "Alexu-aic at arabic hate speech 2022: Contrast to classify," arXiv preprint arXiv:2207.08557, 2022.

[28] F. El-Alami, S. O. El Alaoui, and N. E. Nahnahi, "A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model," JOURNAL OF KING SAUD UNIVERSITY-COMPUTER AND INFORMATION SCIENCES, vol. 34, no. 8, B, pp. 6048–6056, Sep. 2022, doi: 10.1016/j.jksuci.2021.07.013.

[29] M. H. Akram, K. Shahzad, and M. Bashir, "ISE-Hate: a benchmark corpus for inter-faith, sectarian, and ethnic hatred detection on social media in Urdu," Inf Process Manag, vol. 60, no. 3, p. 103270, 2023.

[30] Md. M. Hossain, Z. R. Chowdhury, S. M. Rezwanul Haque Akib, Md. Sabbir Ahmed, Md. Moazzem. Hossain, and A. S. M. Miah, "Crime Text Classification and Drug Modeling from Bengali News Articles : A Transformer Network-Based Deep Learning Approach," in 2023 26th International Conference on Computer and Information Technology (ICCIT), Dec. 2023, pp. 1–6. doi: 10.1109/ICCIT60459.2023.10441195.

[31] "About X's APIs," X developer platform. Accessed: Oct. 28, 2023. [Online]. Available: https://help.twitter.com/en/rules-and-policies/x-api

[32] H. S. Albarbari, H. M. Al - Awami, A. A. Bazroon, H. H. Aldibil, S. M. Alkhalifah, and R. G. Menezes, "Criminal behavior and mental illness in the Arab world," J Forensic Sci, vol. 66, no. 6, pp. 2092‑2103, 2021.

[33] D. ' 'Kumar, "Introduction to Data Preprocessing in Machine Learning," Towards Data Science. Accessed: May 19, 2023. [Online]. Available: https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d

[34] M. Schirmer, U. Kruschwitz, and G. Donabauer, "A New Dataset for Topic-Based Paragraph Classification in Genocide-Related Court Transcripts," in LREC 2022: THIRTEEN INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, N. Calzolari, F. Bechet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, H. Mazo, H. Odijk, and S. Piperidis, Eds., 55-57, RUE BRILLAT-SAVARIN, PARIS, 75013, FRANCE: EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA, 2022, pp. 4504–4512.

[35] A. Algefes, N. Aldossari, F. Masmoudi, and E. Kariri, "A text-mining approach for crime tweets in Saudi Arabia: from analysis to prediction," in 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA), IEEE, 2022, pp. 109–114.

[36] M. A. AlGhamdi and M. A. Khan, "Intelligent analysis of Arabic tweets for detection of suspicious messages," Arab J Sci Eng, vol. 45, pp. 6021–6032, 2020.