Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061674

ARTICLE

Leveraging Transformers for Detection of Arabic Cyberbullying on Social Media: Hybrid Arabic Transformers

Amjad A. Alsuwaylimi[1,*] **and Zaid S. Alenezi**[2]

[1]Department of Computer Science, College of Science, Northern Border University, Arar, 91431, Saudi Arabia
[2]Information Technology Management, Northern Border University, Arar, 91431, Saudi Arabia

*Corresponding Author: Amjad A. Alsuwaylimi. Email: amjad.alsuwaylimi@nbu.edu.sa

Received: 30 November 2024; Accepted: 20 February 2025; Published: 16 April 2025

ABSTRACT: Cyberbullying is a remarkable issue in the Arabic-speaking world, affecting children, organizations, and businesses. Various efforts have been made to combat this problem through proposed models using machine learning (ML) and deep learning (DL) approaches utilizing natural language processing (NLP) methods and by proposing relevant datasets. However, most of these endeavors focused predominantly on the English language, leaving a substantial gap in addressing Arabic cyberbullying. Given the complexities of the Arabic language, transfer learning techniques and transformers present a promising approach to enhance the detection and classification of abusive content by leveraging large and pretrained models that use a large dataset. Therefore, this study proposes a hybrid model using transformers trained on extensive Arabic datasets. It then fine-tunes the hybrid model on a newly curated Arabic cyberbullying dataset collected from social media platforms, in particular Twitter. Additionally, the following two hybrid transformer models are introduced: the first combines CAmelid Morphologically-aware pre-trained Bidirectional Encoder Representations from Transformers (CAMeLBERT) with Arabic Generative Pre-trained Transformer 2 (AraGPT2) and the second combines Arabic BERT (AraBERT) with Cross-lingual Language Model - RoBERTa (XLM-R). Two strategies, namely, feature fusion and ensemble voting, are employed to improve the model performance accuracy. Experimental results, measured through precision, recall, F1-score, accuracy, and Area Under the Curve-Receiver Operating Characteristic (AUC-ROC), demonstrate that the combined CAMeLBERT and AraGPT2 models using feature fusion outperformed traditional DL models, such as Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM), as well as other independent Arabic-based transformer models.

KEYWORDS: Cyberbullying; transformers; pre-trained models; arabic cyberbullying detection; deep learning

1 Introduction

In recent years, social media platforms have become integral to our daily lives. These platforms offer numerous benefits such as fostering connections, enabling instant communication, and providing access to vast amounts of information. Furthermore, social media empowers individuals to express themselves freely and participate in global discussions. Despite these advantages, the rapid growth of social media has introduced profound challenges, particularly those related to user safety and privacy. One of the most pressing issues arising from social media use is cyberbullying [1–3]. Cyberbullying affects both children and adults and leads to severe emotional and psychological consequences. For children, this can result in low self-esteem, academic difficulties, and, in extreme cases, suicidal thoughts. Adults are also not immune, as cyberbullying can impact mental health, work performance, and personal relationships. The pervasive nature of online platforms amplifies these effects, making it difficult for victims to escape harassment. The term cyberbullying is common among teenagers, particularly those who stay for an extended time at home using the Internet and social media networks [4,5].

The cyberbullying issue has captured the attention of researchers who have proposed many approaches to tackling this issue. Scholarly Machine Learning (ML) classifiers [6–8], Deep Learning (DL) models [9–12], and Natural Language Processing (NLP) [13–15] have emerged as powerful approaches in the fight against cyberbullying. ML algorithms are traditional methods that can analyze large datasets to identify patterns and predict instances of cyberbullying with remarkable accuracy. In ML, classifiers such as Support Vector Machines (SVM), Naive Bayes (NB), and Logistic Regression (LR) are commonly utilized to categorize content as bullying or non-bullying based on word frequencies and other features. DL, with its advanced neural networks, enhances this capability by automatically learning complex features from data, making it particularly effective in detecting subtle forms of online harassment compared to ML classifiers. In DL models, Convolutional Neural Networks (CNNs) [16–18] and Long Short-Term Memory Networks (LSTMs) [19] enhance detection by capturing local features and sequential patterns in text data, compared with ML classifiers.

In both ML and DL approaches, NLP focuses on text analysis, enabling the identification of harmful language, hate speech, and abusive content in real time. Together, these approaches can monitor social media platforms, flag potential incidents of cyberbullying, and even provide automated responses to prevent the escalation of harmful behaviors. In NLP methods, word embeddings such as Word2Vec [20], fast text [21], and GloVe [22] further contribute by analyzing the semantic similarities within text that can improve the classification task. However, both ML and DL approaches suffer from limited High-Quality Annotated Datasets, which are time-consuming and labor-intensive owing to Arabic language complexity and dialectal variations.

Recently, the transfer learning approach has achieved better results in terms of accuracy in detecting bullying texts. In this approach, transformers save scholars' time and effort because they are pre-trained models that train a large amount of data. The most popular architecture is Bidirectional Encoder Representations from Transformers (BERT) [23–26], and its variations offer state-of-the-art performance in understanding the context and nuances of language. Additionally, hybrid approaches such as ensemble learning and multimodal methods combine different models and data types to provide robust and comprehensive detection systems. These methods are integral to automated content moderation systems on social media platforms, where they help to monitor and flag potentially harmful content, thereby contributing to safer online environments. Many studies have focused on detecting cyberbullying in English, whereas few have focused on Arabic. This scenario remains a challenge, as Arabic is a morphologically rich and complex language with many dialects and variations in vocabulary, rendering traditional machine learning models less effective. Therefore, this study aims to address the Arabic cyberbullying issue by proposing a hybrid transformer to detect and classify Arabic cyberbullying texts. The following two architectures yield the proposed model: the first combines CAMelBERT and AraGPT2, whereas the second combines AraBERT and XML-R. In both models, the same two methods are used: feature fusion and voting ensemble. In addition, A novel Arabic dataset has been collected from Twitter (now referred to as 'X'). This dataset focuses on Arabic cyberbullying in a variety of Arabic dialects. The initial size of the dataset consists of 95,512 rows, then after the filtering process has been performed, the dataset has been reduced to 43,122 rows.

Additionally, several NLP methods have been applied to the dataset; data cleaning, data annotation and data preprocessing, which further reduces the dataset to 17,670 rows. Moreover, the dataset consists of two classes: bullying and non-bullying. The dataset is considered an imbalanced dataset. Furthermore, three Arabic native speakers have helped in the annotation process to ensure the manual annotation process is conducted accordingly. In order to evaluate the results of the proposed models, several experiments were conducted using the proposed Arabic dataset collected. The experimental results show that the proposed models outperform the DL models (LSTM and BiLSTM), as well as the transformers as separate models. The first model using CAMBelBERT and AraGPT2 achieved a better performance in terms of accuracy, reaching 97% using the feature fusion method. Our research advances Arabic cyberbullying detection through three key contributions: (1) proposing two Arabic hybrid transformer models based on the BERT architecture, incorporating CAMeLBERT, AraGPT2, AraBERT, and XML-R; (2) introducing an Arabic dataset consisting of 17,670 entries categorized as bullying or non-bullying; and (3) evaluating the effectiveness of feature fusion and voting ensemble methods in detecting Arabic cyberbullying using transformer-based models.

The remainder of this paper is organized as follows. Section 2 provides an overview of related studies, and Section 3 outlines the phases of this study. Section 4 describes the experimental setup and presents the results, along with a discussion. Finally, the conclusion is presented in Section 5.

2 Related Studies

This section provides an overview of the related studies and methods focused on cyberbullying detection. In [1], authors used a hybrid deep learning approach called DEA-RNN for cyberbullying detection on Twitter and achieved 86.61% accuracy, 85.94% precision, 84.14% recall, and 85.54% F1-score. The dataset used in this study consists of tweets that go through preprocessing and data-cleaning phases to enhance feature extraction and classification. The dataset of this study comprised 10,000 labeled tweets, 6508 (65%) non-cyberbullying tweets, and 3492 (35%) cyberbullying tweets. The language processing techniques used included noise removal, out-of-vocabulary cleaning, and tweet transformation. Similarly, using a deep learning approach, authors of [9] developed a Multichannel Deep Learning Framework for Cyberbullying Detection on Twitter, which achieved 88% accuracy. This study highlights the need for an automatic method for detecting cyberbullying using NLP and advanced machine learning. The dataset used for evaluation had 55,788 tweets, of which 56.2% were offensive and 43.8% were non-offensive. The dataset used posts from social media, specifically Twitter.

Similarly, the work, in [11] developed a Cyberbullying Detection System (CDS) using DL to detect and classify different types of online cyberbullying. Two datasets were used. The first was the Binary Aggressive Cyberbullying dataset (115,661 post samples) and the second was the Multiclass Cyberbullying dataset (39,869 tweet samples). The datasets had various classes such as religion, age, gender, ethnicity, and non-bullying content. The experiments showed that the models performed well, and the BiLSTM model achieved up to 99% accuracy in some classes. Correspondingly, authors in [12] presented a study on cyberbullying detection using a deep neural network in the Bengali language. The dataset consisted of 44,001 comments from public Facebook posts targeting male and female victims in different professions. Comments were classified as non-bully or sexual harassment. The model achieved 87.91% accuracy for binary classification and 85% accuracy for multiclass classification.

Likewise, the study, suggested in [6], presented a method for detecting cyberbullying using machine learning, specifically classifiers such as SVM and neural networks. The authors used a cyberbullying dataset from Kaggle, 12773 conversation messages from Formspring. The Neural Network performed 92.8% with 3-grams and outperformed the SVM, which was 90.3% with 4-grams. The Neural Network also achieved an average F-score of 91.9%. In the same vein, the study of [27] presented a Machine Learning-based cyberbullying detection model to detect bullying messages on Facebook and Twitter. The model used NLP for data preprocessing and feature extraction using Bag-of-Words (BoW) and TF-IDF. Various machine learning algorithms have been used. The accuracy rates were 66.7% for probabilistic modeling and 78.5% for language-based cyberbullying detection.

The work in [28] focused on detecting Cyber-Bullying and Cyber-Harassment in Arabic social media using supervised machine learning. The dataset contained 6138 tweets and Facebook posts: 2138 from Facebook and 4000 from Twitter. The language is Arabic. the study got high accuracy with F1-measure ranging from 0.67 to 0.99 for different machine learning algorithms on the dataset.

The study of [29] focused on detecting cyberbullying in Arabic social media streams using machine learning, specifically the Naive Bayes classifier algorithm. The dataset used was comments from Arabic social media platforms such as Twitter and YouTube, which are known for discussions about sensitive topics. The language was Arabic, and the model achieved a 95.9% accuracy in detecting cyberbullying comments.

Author of [30] presented a study on the detection of cyberbullying in Arabic social media using sentiment analysis and lexicon-based approaches. The authors used a dataset of 100,327 tweets and YouTube comments that were classified into bullying and non-bullying categories. The authors achieved an accuracy of 81% in detecting cyberbullying, while the Chi-square and Entropy were 62.11% and 39.14%, respectively.

The research presented in [31] investigated the classification of cyberbullying texts in Arabic language. The authors used the AJComments dataset, which contains both cyberbullying and non-cyberbullying content. The primary language of the dataset is Arabic, which poses challenges owing to dialectal diversity and informal language use. They used various machine learning and deep learning models and achieved notable accuracy; some models got an F1-measure of 0.93 on an oversampled dataset and 85% on a balanced dataset.

The study of [32] investigated the detection of offensive languages in Arabic social media content, specifically utilizing a dataset from the fourth workshop on Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) at the LREC 2020, which includes annotated tweets. It employs various machine learning models, including a Support Vector Machine (SVM), Logistic Regression, and decision trees, along with ensemble methods, such as bagging, random forest, and AdaBoost. The bagging classifier achieved the highest F1-score of 88%, whereas the SVM classifier achieved an F1-score of 82%.

The work in [33] introduced a CNN-BiLSTM deep learning model designed to detect cyberbullying content in tweets across three languages: English (Latin script), Hindi (Devanagari script), and Hinglish (Latin script). It utilizes a multilingual dataset comprising tweets in these languages and employs a combination of GloVe and FastText word embeddings to enhance the model performance.

All the previously mentioned research on detecting cyberbullying, especially in Arabic social media situations. Although the reviewed studies show noteworthy advancements in cyberbullying detection using a variety of machine learning and deep learning techniques, the majority concentrate on datasets or languages, ignoring the difficulties like dialectal variances and informal usage presented by the diversity of the Arabic language. Although research like [28–32] shows that machine learning can be used to detect cyberbullying in Arabic. Those studies mostly depend on single deep learning architectures or conventional machine learning models, which might not adequately account for the linguistic complexity of Arabic. The suggested hybrid models seek to improve feature representation and make use of context-rich embeddings by fusing the advantages of pre-trained transformer models like CAMELBERT and ARAGPT or ARABERT and XLM-R. This will provide better generalization across a range of Arabic dialects and social media platforms. By combining contextual adaptability and strong language knowledge, these models close a significant research gap and advance the identification of cyberbullying in underrepresented languages.

3 Methods and Materials

This section outlines the methods employed in this study, as illustrated in Fig. 1. It covers all phases, including data acquisition, annotation, preprocessing, dataset splitting, model building, and performance evaluation.

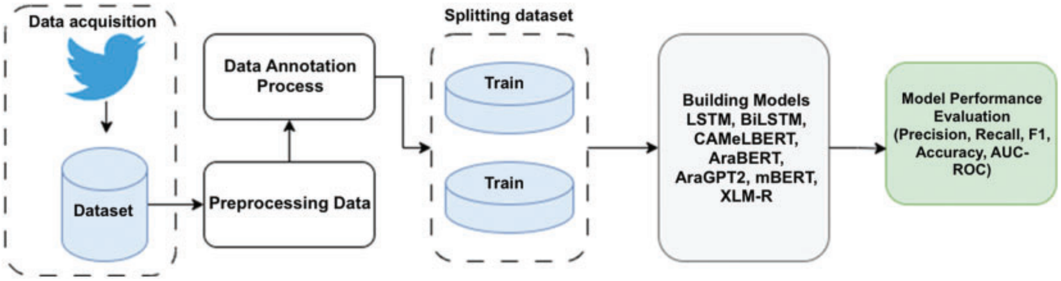
Data acquisition
Dataset
Data Annotation Process
Preprocessing Data
Splitting dataset
Train
Train
Building Models LSTM, BiLSTM, CAMeLBERT, AraBERT, AraGPT2, mBERT, XLM-R
Model Performance Evaluation (Precision, Recall, F1, Accuracy, AUC-ROC)

Figure 1: Methods and phases of this study

3.1 Data Acquisition Phase

In the initial phase of this study, data were collected from the Twitter platform (i.e., X.com) using Python. The dataset originally comprised 43,122 user comments, compiled into a single CSV file. After data cleaning and an annotation process conducted by three native Arabic speakers, the dataset was reduced to 17,670 comments, leading to an imbalanced dataset. Table 1 presents the final composition of the dataset and examples of the dataset are shown in Table 2.

Table 1: Dataset description

Item(s)	Description	Max. length	Min. length
Bullying	10,807	168 words	15 words
Not bullying	6863		
Total	**17,670**		

Table 2: Examples of the dataset

No.	Comments in Arabic	Translated to English	Class
1	اول شيء، طرح للاسف مقرف ثاني شيء ليه ينشر شيء مقرف في صفحته تقريبا تتكلم عنكم انتم..	First of all, the post is unfortunately disgusting. Secondly, why would he publish something disgusting on his page that is almost talking about you?	Bullying
2	هوا انت اهبل انت مش من المفروض انك تفتح الموضوع دة اصلا انا لغيت الاشتراك بسبب الهبل الى انت بتقولة دة	Are you an idiot? You shouldn't have talked about this topic in the first place. I canceled my subscription because of the stupidity you're saying.	Bullying

(Continued)

Table 2 (continued)

No.	Comments in Arabic	Translated to English	Class
3	يعطيك العافيه خليتني حب البرمجه والبايثون اكتر واكتر علقد ماشرحك بسيط وسهل الفهم شكرا كتير الك كل الدعم استاذنا.	May God bless you. You made me love programming and Python more and more. Your explanation is simple and easy to understand. Thank you very much for all the support, our professor.	Not bullying
4	يا اخي شرحك كتير حلو وسلس هلق انا مبرمج جافا يمكن علي سهله بس جد البايثون اسهل بكتير من اللغات التانيه مشكور اخي	My brother, your explanation is very nice and smooth. I am a Java programmer now, so it may be easy for me, but seriously, Python is much easier than other languages. Thank you, my brother.	Not bullying

3.2 *Pre-Processing Phase*

The preprocessing phase involves several essential steps to ensure the dataset is properly prepared for feature selection and that the model is effectively trained during this process. First, the text is tokenized to enable further analysis. Next, non-essential punctuation marks that do not contribute to the meaning of the text are removed. Arabic stop words [34], which do not carry significant information, are then excluded. Irrelevant numerical values and special characters are also removed.

3.3 *Building Models*

In this study, several deep learning transformer models were employed to evaluate the proposed approach. These models include LSTM, BiLSTM, CAMeLBERT, AraGPT2, AraBERT, XLM-R, and mBERT. A detailed description of each model is provided in this section.

LSTM is a type of recurrent neural network (RNN) designed to handle sequential data by retaining information from previous time steps. It addresses the vanishing gradient problem commonly encountered in standard RNNs, making it well-suited for learning long-term dependencies in text. LSTM effectively identifies sequential patterns in Arabic text [35]. BiLSTM extends LSTM by processing sequences in both forward and backward directions, allowing the model to incorporate context from both preceding and succeeding words. This bidirectional approach enables BiLSTM to outperform traditional LSTMs, particularly in capturing Arabic's complex syntactical structure, particularly in sentiment analysis and classification applications.

CAMeLBERT is a transformer-based model that was pre-trained for the Arabic language using a masked language modeling (MLM) objective. It is based on the BERT architecture. By using this method, the model may learn rich, contextualized word representations by predicting masked tokens within a phrase [36]. Pre-training on Arabic literature allows CAMeLBERT to grasp the distinctive features of the language, such as its unique syntax, morphology, and derivational patterns. Additionally, it considers the notable differences between regional dialects and Modern Standard Arabic (MSA), guaranteeing excellent performance across a variety of Arabic text genres and styles.

CAMeLBERT's language-specific design allows it to perform very well on Arabic Natural Language Processing (NLP) tasks like sentiment analysis, named entity recognition (NER), and text categorization. The model is especially useful for addressing the complexities of Arabic since it can manage ambiguous situations, morphologically rich structures, and text orientation from right to left. Additionally, it outperforms many multilingual models on Arabic datasets due to its flexibility and good contextual comprehension, solidifying its status as a cutting-edge solution for Arabic language processing problems [37].

AraGPT2 is based on GPT-2 architecture serves as the foundation for the generative model, which has been specially trained for Arabic text. In contrast to bidirectional models, it employs a unidirectional transformer architecture that sequentially processes text and predicts a sentence's next word by analyzing the context that comes before it [38]. Because of its design, AraGPT2 is especially useful for text generating activities including composing stories, creating conversation, and modeling Arabic language usage. It can produce content that is coherent and pertinent to its context because of its extensive and varied pre-training on Arabic corpora, which gives it a solid grasp of the language's syntax, grammar, and semantics.

Despite being primarily made for text creation, AraGPT2 can also be used for other NLP tasks. The model can be adjusted to meet classification and categorization problems, including question answering, topic classification, and sentiment analysis, by fine-tuning it on datasets. AraGPT2 can catch the intricacies of Arabic morphology, dialectal variances, and stylistic subtleties because of its versatility and language-specific pre-training. Consequently, it bridges the gap between generating capabilities and task-specific performance, making it a flexible tool for researchers and practitioners working on Arabic NLP applications.

AraBERT is based on the BERT architecture that has been pre-trained for Arabic utilizing a variety of large and varied Arabic corpora. AraBERT learns deep contextual embeddings that capture the syntactic and semantic nuances of Arabic text by using a masked language modeling (MLM) aim [39]. AraBERT can easily handle the intricacies of Arabic morphology, syntax, and extensive derivational patterns thanks to this pre-training technique. It is an extremely powerful model for a variety of natural language processing (NLP) applications due to its capacity to analyze and comprehend Modern Standard Arabic (MSA) and dialectal variations.

Across a range of NLP applications, including Named Entity Recognition (NER), sentiment analysis, and text classification, optimized versions of AraBERT consistently attain excellent accuracy. The model is excellent at representing Arabic's finer points, such as word forms with complicated morphology, confusing structures, and context-dependent interpretations. It outperforms several general-purpose multilingual models and achieves some of the top results on Arabic text categorization tasks. AraBERT is a top option for researchers and practitioners working on Arabic NLP problems because of its efficiency, versatility, and language-specific architecture, which establishes a high standard for Arabic language comprehension.

Based on the RoBERTa architecture, XLM-R (XLM-RoBERTa) is a transformer-based model that was trained on multilingual data in more than 100 languages, including Arabic [40]. It builds strong contextual representations in a variety of languages by using self-supervised learning to predict missing words in phrases. Because of its multilingual pre-training, XLM-R is quite adaptable and can function well in both cross-lingual and monolingual activities. Although it might not adequately account for the distinct morphological and syntactic intricacies of the language, XLM-R exhibits competitive performance for Arabic text categorization, capturing the essential linguistic patterns of Arabic text.

XLM-R is still a useful tool for Arabic NLP tasks, even though models like AraBERT or CAMeLBERT are better suited for Arabic because of their language-specific pre-training, particularly in situations that call for cross-lingual transfer or multilingual applications. When optimized, XLM-R can perform well on monolingual Arabic tasks including named entity recognition (NER), sentiment analysis, and text categorization. It is especially helpful for multilingual research and applications because of its cross-linguistic generalization, which fills in gaps in languages with limited resources and supports a variety of linguistic settings. While mBERT is a pre-trained version of BERT that supports 104 languages, including Arabic. It uses a masked language modeling objective to predict missing words in input text across multiple languages. While mBERT performs reasonably well on Arabic text, it is less specialized than models trained exclusively on Arabic data, such as AraBERT. It can be useful in multilingual contexts, though it may not perform as well as Arabic-specific models like AraBERT or CAMeLBERT.

The proposed models use transfer learning to enhance performance and accuracy, leveraging large pre-trained models trained on extensive datasets. These models employ hybrid transformer architectures with two variations, one that combines CAMeLBERT and AraGPT2, and another that combines AraBERT and XLM-R. Both models utilize two techniques: feature fusion, as shown in Fig. 2, and voting ensemble, as shown in Fig. 3. To improve classification accuracy, both soft and hard voting ensemble methods were applied.

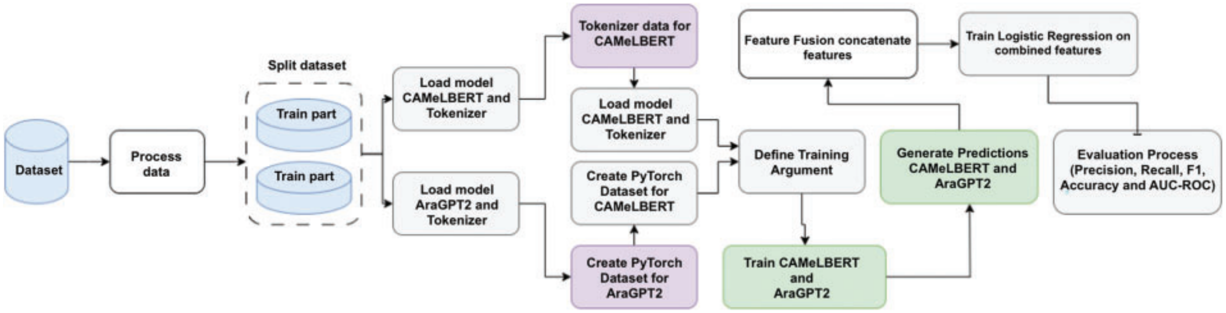
Dataset
Process data
Split dataset
Train part
Train part
Load model CAMeLBERT and Tokenizer
Load model AraGPT2 and Tokenizer
Tokenizer data for CAMeLBERT
Load model CAMeLBERT and Tokenizer
Create PyTorch Dataset for CAMeLBERT
Create PyTorch Dataset for AraGPT2
Feature Fusion concatenate features
Train Logistic Regression on combined features
Define Training Argument
Generate Predictions CAMeLBERT and AraGPT2
Evaluation Process (Precision, Recall, F1, Accuracy and AUC-ROC)
Train CAMeLBERT and AraGPT2

Figure 2: First Model Architecture

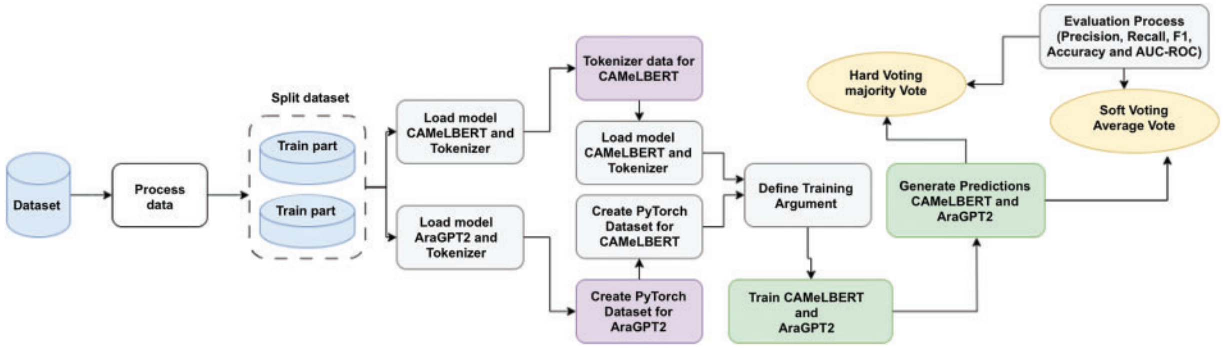
Dataset
Process data
Split dataset
Train part
Train part
Load model CAMeLBERT and Tokenizer
Load model AraGPT2 and Tokenizer
Tokenizer data for CAMeLBERT
Load model CAMeLBERT and Tokenizer
Create PyTorch Dataset for CAMeLBERT
Create PyTorch Dataset for AraGPT2
Define Training Argument
Train CAMeLBERT and AraGPT2
Generate Predictions CAMeLBERT and AraGPT2
Hard Voting majority Vote
Evaluation Process (Precision, Recall, F1, Accuracy and AUC-ROC)
Soft Voting Average Vote

Figure 3: Second model architecture

Fig. 2 illustrates the feature fusion technique using a Logistic Regression classifier for final predictions. Two proposed hybrid models are used: CAMeLBERT-da and AraGPT2, or AraBERT and XLM-R. In both models, the dataset, which contains Arabic comments labeled for cyberbullying, is preprocessed by encoding the target labels and splitting the data into training and testing sets. In the first model, CAMeLBERT-da, a sentiment analysis model, and AraGPT2, a generative language model with a classification head, are fine-tuned independently on the training data. In the second model, AraBERT, a transformer model pre-trained on Arabic language data, and XLM-R, a multilingual transformer model capable of handling Arabic text, are used.

Then, the text data is tokenized separately for each model, respecting their unique tokenization schemes. After fine-tuning, the logits (predicted probabilities before applying softmax) from both models are concatenated to create combined feature vectors for each input. These fused features are normalized using a standard scaler and passed to a Logistic Regression classifier, which is trained to make the final classification.

While Fig. 3 shows the voting mechanisms using both hard voting and soft voting mechanisms. The models initialized with their respective tokenizers, and the text data is tokenized according to their specific requirements. The models are fine-tuned on the training data using PyTorch's Trainer class with predefined training arguments, including batch size, learning rate, and the number of epochs. After fine-tuning, predictions are obtained from both models in the form of logits, which are converted into probabilities using the softmax function. These probabilities are used for ensemble decision-making through two mechanisms: hard voting, which assigns the class label based on the majority vote, and soft voting, which averages the probabilities from both models to determine the final predictions.

Fig. 4 illustrates the attention maps and the ways in which the models concentrate on various aspects of the input text while classifying it. Because they highlight the tokens or phrases that have the most influence on decision-making, attention maps are especially helpful for comprehending the inner workings of transformer models. We can show how the Arabic-specific design of CAMeLBERT and the contextual knowledge of AraGPT2 work together to create greater performance by showing attention weights.

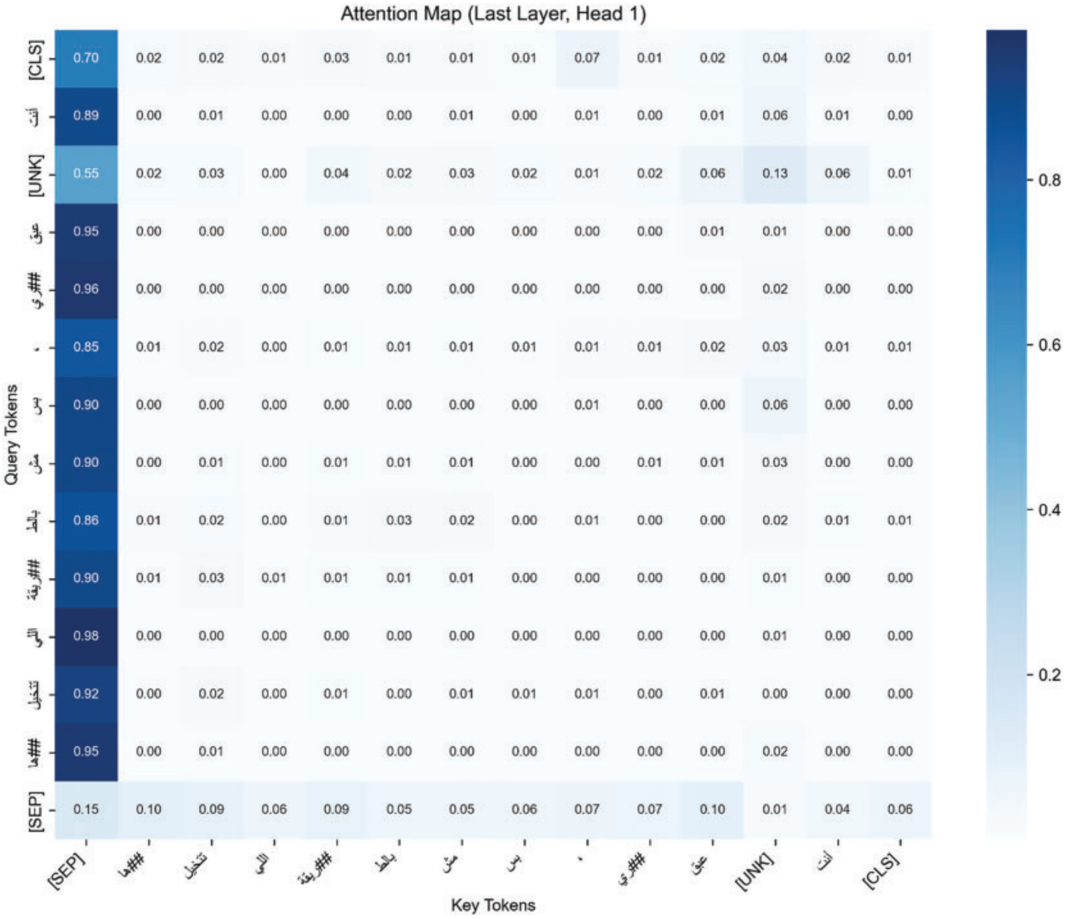

Figure 4: The attention map shows the Classification Token (CLS), a special token added at the beginning of the input sequence; the Separator Token (SEP), which separates the sentences; and the Unknown Token (UNK), which represents an Out-of-Vocabulary word

For instance, consider a sentence with subtle sarcasm, such as: "أنت فعلاً عبقري، بس مش بالطريقة اللي تتخيلها" ("You're really a genius, but not in the way you think"). An attention map visualization of CAMeLBERT + AraGPT2 as seen in Fig. 4 may reveal that the model places a lot of emphasis on crucial terms that are essential for identifying the sarcastic tone, such as "عبقري" (genius) and "بس مش" (but not). This illustrates how CAMeLBERT can recognize linguistically important Arabic tokens and how AraGPT2 can place these tokens in the context of the sentence's overall meaning. An attention map for AraBERT + XLM-R on the same input, on the other hand, would show a more dispersed concentration or an excessive dependence on less important words, suggesting difficulties in capturing subtle or implicit meanings, particularly in informal or dialectal Arabic.

The bottom layers of CAMeLBERT are probably going to concentrate on the morphological structure of Arabic words, successfully capturing root patterns, prefixes, and suffixes. Building semantic and contextual linkages becomes more important as the layers advance, and AraGPT2's generative capabilities help to further improve this. In complicated tasks like identifying cyberbullying, where informal language, code-switching, and implicit cues are common, the CAMeLBERT + AraGPT2 model performs exceptionally well thanks to this gradual refining. However, because XLM-R is designed for multilingual contexts rather than Arabic subtleties, AraBERT + XLM-R may not pay as much attention to Arabic-specific linguistic elements as it does to broad text classifications.

3.4 Model Evaluation

In this section, the most commonly used metrics for evaluating machine learning and deep learning models are presented. The model's performance was assessed using the F1-score, precision, accuracy, confusion matrix, and Area Under the Curve-Receiver Operating Characteristic (AUC-ROC) [41]. Precision measures the ratio of true positives (TP) to the sum of true positives and false positives (FP). It indicates the percentage of user comments classified as cyberbullying that are actually cyberbullying, as shown in Eq. (1).

$$Precison = \frac{Number\ of\ the\ correct\ cyberbullying\ user\ comments\ classified}{Total\ number\ of\ relevant\ cyberbullying\ user\ comments} \tag{1}$$

Recall is the ratio of true positives (TP) to the sum of true positives and false negatives (FN). It indicates the percentage of cyberbullying comments that were correctly identified, as shown in Eq. (2).

$$Recall = \frac{Number\ of\ the\ correct\ cyberbullying\ user\ comments\ classified}{Total\ number\ of\ cyberbullying\ user\ comments\ classified} \tag{2}$$

The F1-score, defined in Eq. (3), represents the harmonic mean of precision and recall, providing a balanced measure of model performance when both metrics are important. The AUC-ROC metric is determined by plotting the true positive rate (TPR) against the false positive rate (FPR), offering a comprehensive assessment of the model's ability to discriminate between bullying and non-bullying user comments.

$$F1 - score = 2 \times \frac{Precision + Recall}{Precision \times Recall} \tag{3}$$

4 Experimental Settings and Results

This section outlines the experimental setup and presents the results obtained from experiments conducted with transformer models and the proposed hybrid transfer learning approach.

4.1 Experimental Setup

All experiments were conducted to evaluate the performance of the proposed models, divided into two categories: deep learning models and transformer models. In the deep learning category, the primary models used were LSTM and BiLSTM, both recognized for their effectiveness in text classification tasks and demonstrating high performance, particularly in terms of accuracy. In all experiments, synthetic samples for the minority class (SMOT) were used to balance the dataset and address the issue of imbalance. Table 3 outlines the hyperparameters used in both experimental categories. All experiments were executed in the Google Colab environment with GPU support, using Python as the programming language. Table 4 lists the libraries utilized, and Table 5 specifies the hyperparameters for the transformer models. The libraries employed for the transformers include `AutoTokenizer`, `AutoModelForSequenceClassification`, `Trainer`, and `TrainingArguments`.

Table 3: Hyperparameter LSTM and BILSTM

Hyperparameter	Value
Max words	5000
Max length (Padding)	100
LSTM units (Layer 1)	128
LSTM units (Layer 2)	64
Dropout rate	0.2
Batch size	32
Epochs	30
Optimizer	Adam
Loss function	Binary crossentropy
Validation split	0.2

Table 4: Libraries used

Library	Purpose
pandas	Data manipulation and analysis
numpy	Numerical operations
matplotlib.pyplot	Data visualization
seaborn	Enhanced data visualization
re	Regular expressions for text processing
string	String manipulation
arabic_reshaper	Arabic text reshaping
sklearn	Machine learning utilities
tensorflow.keras	Deep learning model building

Table 5: Hyperparameters of transformers

Hyperparameter	Value
per_device_train_batch_size	8
per_device_eval_batch_size	8
num_train_epochs	5
evaluation_strategy	Epoch
logging_steps	200
save_steps	10,000
save_total_limit	2
max_length	512
num_labels	len(set(labels))
ignore_mismatched_sizes	True
test_size	0.2 (train-test split)
random_state	42 (train-test split)

4.2 Experimental Results

As mentioned earlier, the effectiveness of LSTM and BiLSTM models in identifying and categorizing text data through word dependencies contributed to their use. Transformers were also employed because they are contemporary, pretrained models trained on large corpora, which enhances their performance on language tasks.

The proposed models consist of two hybrid transformer architectures: the first combines CAMelBERT and XLM-R, while the second merges AraBERT and AraGPT2. Both models utilize two methods—Feature Fusion and Voting Ensemble. To further improve classification accuracy, both soft and hard voting techniques were employed in the Voting Ensemble.

In the initial testing of deep learning (DL) models, Table 6 below illustrates the effectiveness of these models in identifying and categorizing Arabic text related to cyberbullying. Key metrics, including accuracy, precision, recall, and F1-score, were used to assess model performance. The LSTM model achieved a precision of 92.86%, recall of 95.84%, F1-score of 94.33%, and overall accuracy of 93.10%. In contrast, the BiLSTM model yielded slightly better results, with 95.43% accuracy, 94.80% recall, 95.12% F1-score, and 94.17% precision. These results highlight the models' effectiveness in identifying and categorizing Arabic text related to cyberbullying. Figs. 5 and 6 display the confusion matrix and the validation and training accuracy, respectively.

Table 6: Comparison between precision, recall, F1-score, and accuracy of the DL models

Model(s)	Precision	Recall	F1-score	Accuracy
LSTM	95.43%	94.80%	95.12%	94.17%
BiLSTM	92.86%	95.84%	94.33%	93.10%

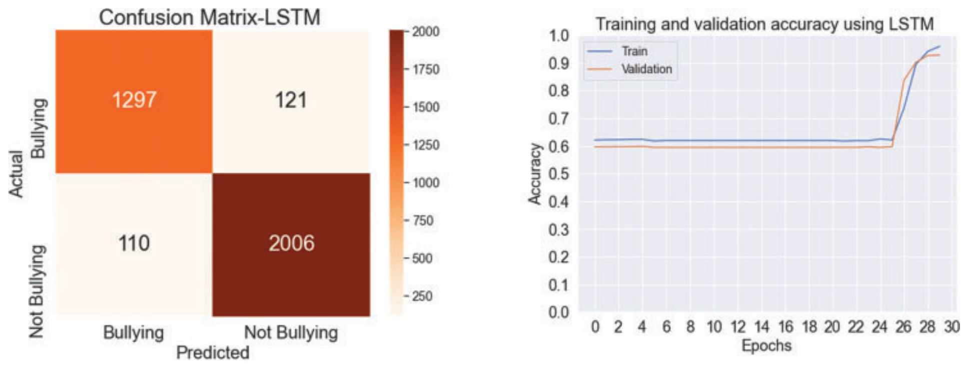

Figure 5: Confusion matrix and accuracy using LSTM

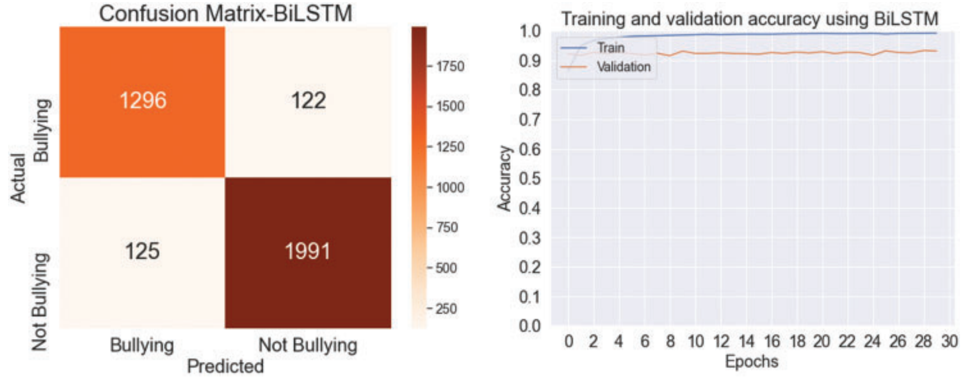

Figure 6: Confusion matrix and accuracy using BiLSTM

The second experiment focused on transformer models, which significantly outperform deep learning models such as LSTM and BiLSTM. Transformers are also referred to as pretrained models, trained on large volumes of data. Table 7 displays the effectiveness of various transformer models in identifying and categorizing Arabic text related to cyberbullying. CAMeLBERT emerged as the most effective model, achieving an accuracy of 96.97%, F1-score of 96.84%, recall of 96.70%, and precision of 96.99%. AraBERT also demonstrated impressive performance, with 96.71% accuracy, 96.57% recall, 96.64% F1-score, and 96.71% precision. XLM-R, MBERT, and AraGPT2 performed well too, although with slightly lower metrics. Specifically, XLM-R and MBERT achieved accuracies of 94.07% and 94.65%, respectively, while AraGPT2 recorded a precision of 94.65% and accuracy of 94.74%. The confusion matrices for the transformer models are shown in Fig. 7.

Table 7: Comparison between precision, recall, F1-score, and accuracy of the transformer models

Model(s)	Precision	Recall	F1-score	Accuracy
CAMeLBERT	96.99%	96.70%	96.84%	96.97%
AraGPT2	94.65%	94.37%	94.51%	94.74%
XLM-R	94.83%	94.01%	94.92%	94.07%
MBERT	94.39%	94.50%	94.44%	94.65%
AraBERT	96.71%	96.57%	96.64%	96.77%

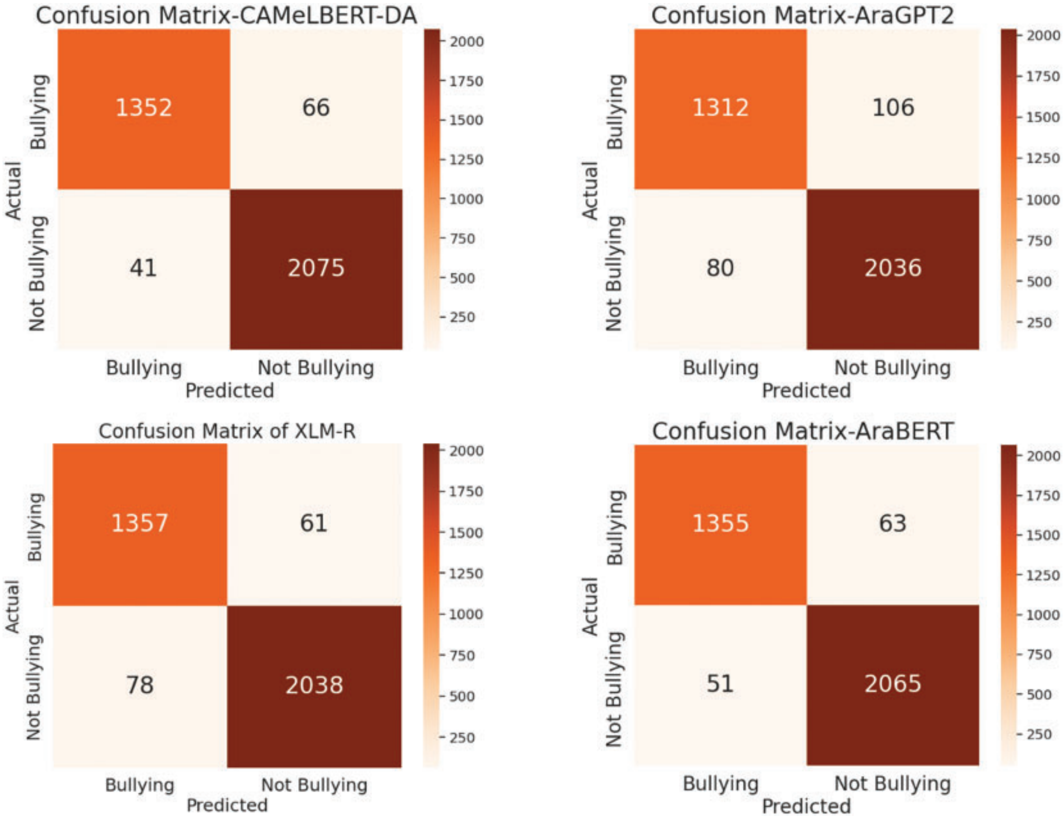

Figure 7: Confusion matrix of transformers

Two hybrid transfer learning models were proposed in the third experiment: CAMeLBERT + AraGPT2 and AraBERT + XLM-R. The results of this experiment evaluated the effectiveness of these hybrid models for identifying and categorizing Arabic cyberbullying, as shown in Table 8. The models assessed were AraBERT combined with XLM-R and CAMeLBERT paired with AraGPT2. The two distinct approaches used to test the models were the voting ensemble and feature fusion. The confusion matrix and AUC-ROC curves are presented in Figs. 8 and 9, respectively.

Table 8: Comparison between precision, recall, F1-score, and accuracy of the proposed hybrid transformers models

Model(s)	Precision	Recall	F1-score	Accuracy
CAMelBERT + AraGPT2 (Feature Fusion)	97.43%	97.51%	97.47%	97.57%
AraBERT + XLM-R (Feature Fusion)	95.14%	94.83%	94.98%	95.19%
CAMelBERT + AraGPT2 (Voting)-Hard	96.26%	96.83%	96.52%	96.63%
AraBERT + XLM-R (Voting)-Hard	93.77%	94.77%	94.16%	94.31%
CAMelBERT + AraGPT2 (Voting)-Soft	97.32%	97.08%	97.20%	97.31%
AraBERT + XLM-R (Voting)-Soft	95.92%	95.70%	95.81%	95.98%

To improve classification accuracy, feature fusion entails combining the feature representations from the two models. By applying both hard and soft voting procedures, the voting ensemble strategy, by contrast, applies a combination of predictions from many models to determine the final outcome.

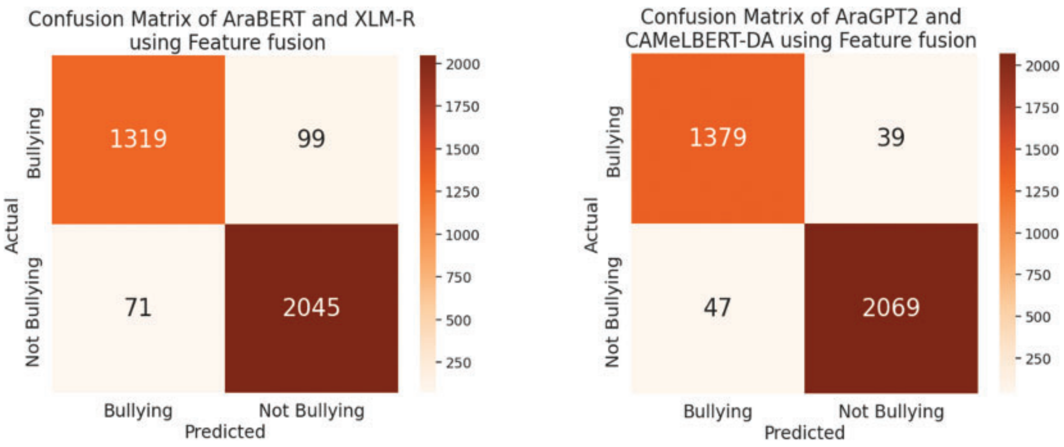

Figure 8: Confusion matrix of both proposed using feature fusion

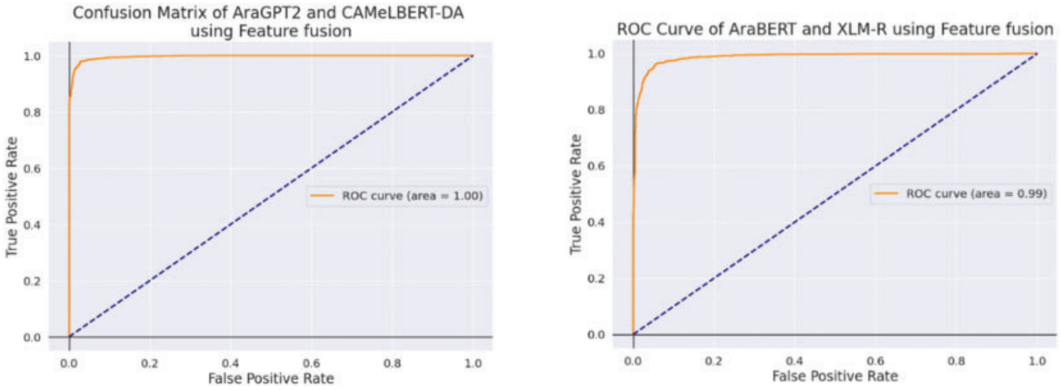

Figure 9: AUC-ROC for both proposed using feature fusion

Table 8 shows that, across all methodologies and metrics, the CAMeLBERT + AraGPT2 model consistently outperforms the AraBERT + XLM-R model. In particular, CAMeLBERT + AraGPT2 using feature fusion surpasses the voting ensemble techniques in terms of accuracy, recall, F1-score, and precision. For example, CAMeLBERT + AraGPT2 achieved an accuracy of 97.57% with the feature fusion method, whereas AraBERT + XLM-R obtained 95.19%. Similarly, in the voting ensemble method, the AraBERT + XLM-R model achieved an accuracy of 95.98%, while the CAMeLBERT + AraGPT2 model reached 97.31% with soft voting.

Additionally, the soft-voting ensemble method consistently outperformed the hard-voting ensemble method. This is evident when comparing the accuracy rates for both the CAMeLBERT + AraGPT2 and AraBERT + XLM-R models using hard and soft voting. For instance, CAMeLBERT + AraGPT2 achieved an accuracy of 97.31% with soft voting, compared to 96.63% with hard voting. A similar trend is observed with the AraBERT + XLM-R model: it reached an accuracy of 95.98% with soft voting, but its accuracy dropped to 94.31% with hard voting. All these results are summarized in Table 7 and confusion matrix presented Fig. 10. The AUC-ROC for the proposed models using the voting ensemble methods is presented in Figs. 11 and 12.

These results show that the CAMeLBERT + AraGPT2 hybrid model is more effective at detecting and classifying Arabic cyberbullying than the AraBERT + XLM-R model, exhibiting superior performance across all evaluation metrics.

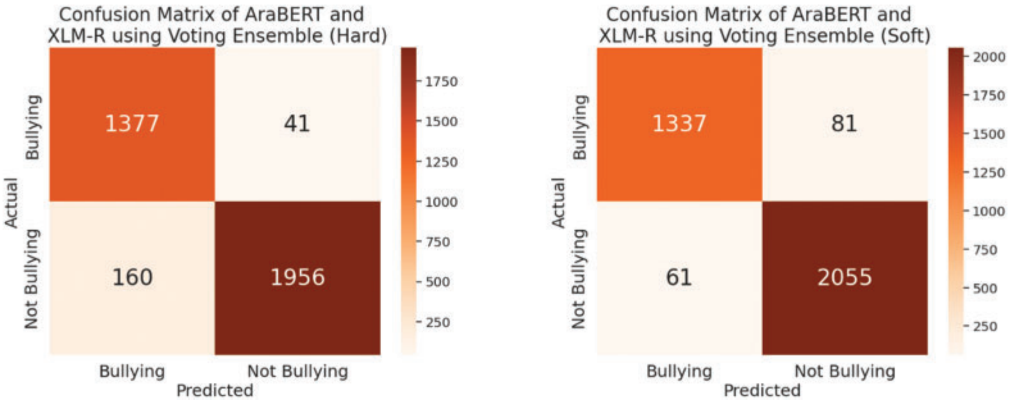

Figure 10: Confusion matrix using the AraBERT+XLM-R model using the voting ensemble (Hard and Soft methods)

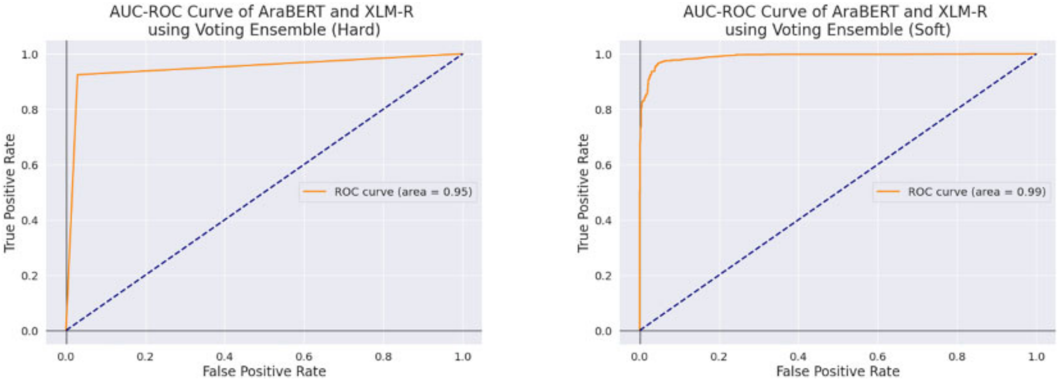

Figure 11: AUC-ROC using AraBERT and XLM-R using Voting Ensemble method

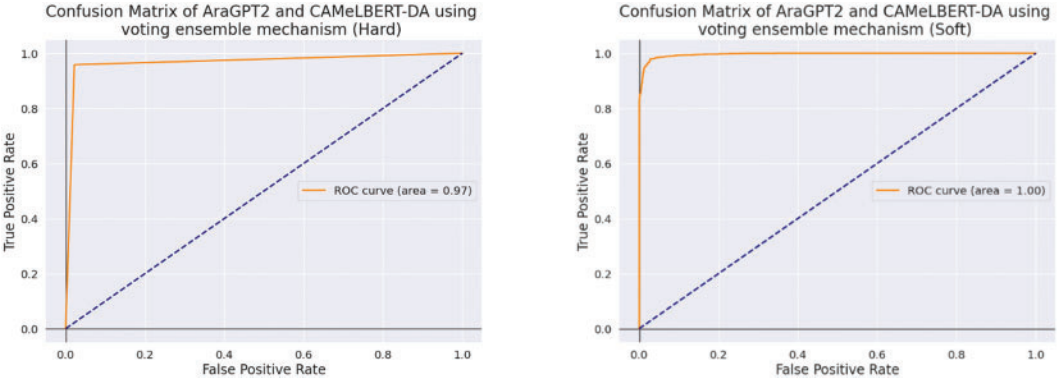

Figure 12: AUC-ROC using AraGPT2 and CAMeLBERT using Voting Ensemble method

4.3 Results Discussion

In the first experiment, the performance of deep learning models, particularly LSTM and BiLSTM, was evaluated in identifying cyberbullying in Arabic. In terms of accuracy, F1-score, precision, and recall, the BiLSTM model outperformed the LSTM model. LSTM achieved an accuracy of 93.10%, while BiLSTM reached 94.17%. The main advantage of BiLSTM is its bidirectional architecture, which allows it to capture contextual information from both past and future words, thereby enhancing its understanding of the complexities of Arabic text. However, when compared to transformer-based models, both LSTM and BiLSTM show limitations. While effective, these models do not fully leverage the contextual depth that transformers can provide.

In the second experiment, Transformer models such as CAMeLBERT, AraGPT2, XLM-R, MBERT, and AraBERT were evaluated for their performance. These Transformer models significantly outperformed the deep learning models. Among them, CAMeLBERT and AraBERT achieved the highest accuracy and F1-scores, with AraBERT slightly surpassing CAMeLBERT, indicating its superior capacity to handle complex linguistic patterns. Although XLM-R and MBERT did not perform as well as CAMeLBERT and AraBERT, they still demonstrated strong results. Due to their extensive pre-training on large corpora, Transformer models excel in capturing the nuanced features of Arabic cyberbullying text, highlighting their advantage over other models.

In the third experiment, hybrid transfer learning models were investigated using voting ensembles and feature fusion. Compared to individual models, the hybrid models—CAMeLBERT + AraGPT2 and AraBERT + XLM-R—demonstrated improved performance. The Voting-Soft method achieved an impressive 97.31% accuracy, establishing the CAMeLBERT + AraGPT2 combination as the most accurate hybrid strategy. This result indicates that feature fusion and voting ensembles, when combining multiple models, can significantly enhance detection and classification performance. The effectiveness of the Voting-Soft technique lies in its use of probabilistic predictions from multiple models, resulting in more balanced and accurate classifications. However, training and integrating multiple models can be computationally intensive and complex, which may limit their practical application.

In summary, Transformer-based methods outperform deep learning algorithms in identifying Arabic cyberbullying, due to their ability to capture intricate linguistic nuances. Models such as CAMeLBERT and AraBERT, in particular, exhibit strong performance. Additionally, hybrid models employing feature fusion and voting ensembles further enhance results, with the Voting-Soft technique proving especially effective. These findings highlight the advanced capabilities of Transformer and hybrid models, offering promising directions for further research and practical applications in cyberbullying detection.

The study found that CAMeLBERT + AraGPT2 outperformed AraBERT + XLM-R in specific metrics due to their complementary strengths. CAMeLBERT's tokenizer which is optimized for Arabic-specific tokenization had an impact on the results, alongside its understanding of Arabic language structures. AraGPT2's generative and precise contextual capabilities complement this by bringing generative capabilities and broader contextual understanding, this enables effective handling of implicit and informal text in cyberbullying detection. In addition, Table 9 presents a comparison of the accuracy of existing models in related studies for Arabic cyberbullying detection.

Table 9: Comparison between precision, recall, F1, and accuracy of the transformer models

Author(s)	Dataset size	Model(s)	Accuracy
[42]	17,748 tweets	Support Vector Machine (SVM)	81%
[43]	Balanced dataset	Ensemble Deep Learning (CNN + Bi-LSTM + GRU)	92.75%
[44]	35,000 tweets	Logistic Regression (LR)	90.57%
[45]	Public dataset	Deep learning with LSTM model	87.57%
[46]	Public dataset	Ensemble machine learning model	94%

In the implementation of the proposed models, the high computational requirements must be met while preserving performance. We used Google Colab's GPU resources in our studies, which allowed for effective model training and inference. However, in practical situations, such access might not always be possible. Deployment becomes more feasible when optimizations like quantization, pruning, and knowledge distillation lower memory consumption and computing demands. Efficiency is further increased by substituting lightweight models like MobileBERT or TinyBERT for one of the models in the hybrid setup and optimizing the input sequence (e.g., dynamic truncation). Performance and resource consumption can also be balanced by deploying AI accelerators or by employing hybrid techniques that combine edge devices and cloud-based fallback systems. Using GPU resources, such as Colab, during testing offers insights on efficiently deploying CAMeLBERT + AraGPT2 in real-world activities like cyberbullying detection, even though these optimizations may include accuracy trade-offs.

5 Conclusion

This study proposed hybrid Transformer models to enhance the identification and categorization of Arabic cyberbullying content. The experimental results indicate that the proposed models, particularly CAMeLBERT and AraBERT, outperform conventional deep learning models such as LSTM and BiLSTM in terms of precision, recall, F1-score, and accuracy. Additionally, they surpass individual Transformer models. The feature fusion method proved especially effective, outperforming the soft-voting ensemble method in terms of model accuracy. These findings underscore the potential of using feature fusion with hybrid Transformers to address the challenges of cyberbullying identification. Future research should focus on several key areas to build on these results. First, incorporating additional contextual and semantic variables may further enhance model accuracy and robustness. Valuable insights could also be gained by evaluating hybrid approaches with emerging Transformer models. Expanding the models to encompass a wider range of Arabic dialects could improve generalizability. Addressing the computational complexities associated with developing and deploying sophisticated models will be essential for real-world applications. Ensuring high performance while optimizing model scalability and efficiency will be crucial when implementing these advanced methods in practical settings.

Acknowledgement: The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, Saudi Arabia for funding this research work.

Funding Statement: This study was funded by the Deanship of Scientific Research at Northern Border University, Arar, Saudi Arabia, through the project number "NBU-FFR-2025-1197-01".

Author Contributions: The authors confirm contribution to the paper: Amjad A. Alsuwaylimi: Conceptualization, Methodology, Resources, Writing—original draft preparation, Writing—review and editing, Project administration, Funding acquisition; Zaid S. Alenezi: Validation, Formal analysis, Investigation, Data curation, Writing—original draft preparation, Visualization. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/amjadalsuwaylimi/arabic-cyberbullying-dataset-ver01 (accessed on 28 January 2024).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Nomenclature

ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
SVM	Support Vector Machine
NB	Naive Bayes
LR	Logistic Regression
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
CAMeLBERT	A BERT-based model trained specifically for Arabic language tasks
AraGPT2	Arabic Generative Pretrained Transformer 2
AraBERT	Arabic-specific BERT model
XLM-R	Cross-lingual Model-RoBERTa, a multilingual transformer model
mBERT	Multilingual BERT, a pre-trained model for multiple languages including Arabic
AUC-ROC	Area Under the Curve-Receiver Operating Characteristic
GloVe	Global Vectors for Word Representation, a word embedding method
RNN	Recurrent Neural Network
DEA-RNN	Deep Learning-based hybrid model for social media text analysis
BoW	Bag-of-Words, a technique for text representation
TF-IDF	Term Frequency-Inverse Document Frequency, a statistical measure for text relevance
MLM	Masked Language Modeling
MSA	Modern Standard Arabic
NER	Named Entity Recognition
CLS	Classification Token
SEP	Separator Token
UNK	Unknown Token

References

1. Murshed BAH, Abawajy J, Mallappa S, Saif MAN, Al-Ariki HDE. DEA-RNN: a hybrid deep learning approach for cyberbullying detection in Twitter social media platform. IEEE Access. 2022;10(2):25857–71. doi:10.1109/ACCESS.2022.3153675.
2. Rosa H, Pereira N, Ribeiro R, Ferreira PC, Carvalho JP, Oliveira S, et al. Automatic cyberbullying detection: a systematic review. Comput Hum Behav. 2019;93(2):333–45. doi:10.1016/j.chb.2018.12.021.
3. Iwendi C, Srivastava G, Khan S, Maddikunta PKR. Cyberbullying detection solutions based on deep learning architectures. Multimed Syst. 2023;29(3):1839–52. doi:10.1007/s00530-020-00701-5.
4. Alqahtani SI, Yafooz WM, Alsaeedi A, Syed L, Alluhaibi R. Children's safety on YouTube: a systematic review. Appl Sci. 2023;13(6):4044. doi:10.3390/app13064044.
5. Hasan MT, Hossain MAE, Mukta MSH, Akter A, Ahmed M, Islam S. A review on deep-learning-based cyberbullying detection. Fut Internet. 2023;15(5):179. doi:10.3390/fi15050179.
6. Hani J, Nashaat M, Ahmed M, Emad Z, Amer E, Mohammed A. Social media cyberbullying detection using machine learning. Int J Adv Comput Sci Appl. 2019;10(5):703–7. doi:10.14569/IJACSA.2019.0100589.
7. Ali A, Syed AM. Cyberbullying detection using machine learning. Pakistan J Eng Technol. 2020;3(2):45–50. doi:10.51846/vol3iss2pp45-50.
8. Alhejaili R, Alsaeedi A, Yafooz WM. Detecting hate speech in Arabic tweets during COVID-19 using machine learning approaches. In: Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022; 2022 Nov; Singapore: Springer Nature Singapore. p. 467–75. doi:10.1007/978-981-16-7324-1_39.

9. Alotaibi M, Alotaibi B, Razaque A. A multichannel deep learning framework for cyberbullying detection on social media. Electronics. 2021;10(21):2664. doi:10.3390/electronics10212664.
10. Agrawal S, Awekar A. Deep learning for detecting cyberbullying across multiple social media platforms. In: European Conference on Information Retrieval; 2018; Cham, Switzerland: Springer International Publishing. p. 141–53. doi:10.1007/978-3-319-76941-7_12.
11. Aldhyani TH, Al-Adhaileh MH, Alsubari SN. Cyberbullying identification system based on deep learning algorithms. Electronics. 2022;11(20):3273. doi:10.3390/electronics11203273.
12. Ahmed MF, Mahmud Z, Biash ZT, Ryen AAN, Hossain A, Ashraf FB. Cyberbullying detection using deep neural network from social media comments in Bangla language. 2021. doi:10.48550/arXiv.2106.04506.
13. Raj C, Agarwal A, Bharathy G, Narayan B, Prasad M. Cyberbullying detection: hybrid models based on machine learning and natural language processing techniques. Electronics. 2021;10(22):2810. doi:10.3390/electronics10222810.
14. Afrifa S, Varadarajan V. Cyberbullying detection on Twitter using natural language processing and machine learning techniques. Int J Innov Technol Interdiscip Sci. 2022;5(4):1069–80. doi:10.17148/IJITIS.2022.5414.
15. Bhatia B, Verma A, Anjum A, Katarya R. Analysing cyberbullying using natural language processing by understanding jargon in social media. In: Sustainable Advanced Computing: Select Proceedings of ICSAC 2021; 2022; Singapore: Springer Singapore. p. 397–406. doi:10.1007/978-981-16-7565-8_37.
16. Gada M, Damania K, Sankhe S. Cyberbullying Detection using LSTM-CNN architecture and its applications. In: 2021 International Conference on Computer Communication and Informatics (ICCCI); 2021 Jan; Coimbatore, India: IEEE. p. 1–6. doi:10.1109/ICCCI51007.2021.9406730.
17. Ghosh S, Chaki A, Kudeshia A. Cyberbully detection using 1D-CNN and LSTM. In: Proceedings of International Conference on Communication, Circuits, and Systems: IC3S 2020; 2021; Singapore: Springer. p. 295–301. doi:10.1007/978-981-33-4963-6_36.
18. Daraghmi EY, Qadan S, Daraghmi Y, Yussuf R, Cheikhrouhou O, Baz M. From text to insight: an integrated CNN-BiLSTM-GRU model for Arabic cyberbullying detection. IEEE Access. 2024. doi:10.1109/ACCESS.2024.3166951.
19. Dass A, Daniel DK. Cyberbullying detection on social networks using LSTM model. In: 2022 International Conference on Innovations in Science and Technology for Sustainable Development (ICISTSD); 2022; Kollam, India: IEEE. p. 293–6. doi:10.1109/ICISTSD55179.2022.9720716.
20. Asqolani IA, Setiawan EB. Hybrid deep learning approach and Word2Vec feature expansion for cyberbullying detection on Indonesian Twitter. Ingénierie des Systèmes d'Information. 2023;28(4):887–95. doi:10.18280/isi.280410.
21. Wang K, Cui Y, Hu J, Zhang Y, Zhao W, Feng L. Cyberbullying detection, based on the fasttext and word similarity schemes. ACM Trans Asian Low-Resour Lang Info Process. 2020;20(1):1–15. doi:10.1145/3345955.
22. Umer M, Alabdulqader EA, Alarfaj AA, Cascone L, Nappi M. Cyberbullying detection using PCA extracted GLOVE features and RoBERTaNet transformer learning model. IEEE Trans Comput Soc Syst. 2024;1–10. doi:10.1109/TCSS.2024.3422185.
23. Paul S, Saha S. CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification. Multimed Syst. 2022;28(6):1897–904. doi:10.1007/s00530-020-00710-4.
24. Yadav J, Kumar D, Chauhan D. Cyberbullying detection using pre-trained BERT model. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC); 2020; Coimbatore, India: IEEE. p. 1096–100. doi:10.1109/ICESC48915.2020.9156092.
25. Ahmed T, Ivan S, Kabir M, Mahmud H, Hasan K. Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying. Soc Netw Anal Min. 2022;12(1):99. doi:10.1007/s13278-022-00934-4.
26. Yafooz WM, Al-Dhaqm A, Alsaeedi A. Detecting kids' cyberbullying using transfer learning approach: transformer fine-tuning models. In: Kids cybersecurity using computational intelligence techniques; Cham: Springer International Publishing; 2023. p. 255–67. doi:10.1007/978-3-031-31507-4_17.

27. Islam MM, Uddin MA, Islam L, Akter A, Sharmin S, Acharjee UK. Cyberbullying detection on social networks using machine learning approaches. In: 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE); 2020; Gold Coast, Australia: IEEE. p. 1–6. doi:10.1109/CSDE50250.2020.9346846.
28. Kanan T, Aldaaja A, Hawashin B. Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in Arabic social media contents. J Internet Technol. 2020;21(5):1409–21. doi:10.3966/160792642020102105020.
29. Mouheb D, Albarghash R, Mowakeh MF, Al Aghbari Z, Kamel I. Detection of Arabic cyberbullying on social networks using machine learning. In: 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA); 2019; Abu Dhabi, United Arab Emirates: IEEE. p. 1–5. doi:10.1109/AICCSA47709.2019.8962582.
30. AlHarbi BY, AlHarbi MS, AlZahrani NJ, Alsheail MM, Alshobaili JF, Ibrahim DM. Automatic cyberbullying detection in Arabic social media. Int J Eng Res Technol. 2019;12(12):2330–5. doi:10.17577/IJERTV12IS120763.
31. Rachid BA, Azza H, Ghezala HHB. Classification of cyberbullying text in Arabic. In: 2020 International Joint Conference on Neural Networks (IJCNN); 2020; Glasgow, UK: IEEE. p. 1–7. doi:10.1109/IJCNN48605.2020.9207354.
32. Husain F. Arabic offensive language detection using machine learning and ensemble machine learning approaches. 2020. doi:10.48550/arXiv.2005.08946.
33. Raj M, Singh S, Solanki K, Selvanambi R. An application to detect cyberbullying using machine learning and deep learning techniques. SN Comput Sci. 2022;3(5):401. doi:10.1007/s42979-022-01308-5.
34. Alsuwaylimi AA. Enhancing Arabic phishing email detection: a hybrid machine learning based on genetic algorithm feature selection. Int J Adv Comput Sci Appl. 2024;15(8). doi:10.14569/IJACSA.2024.0150832.
35. Yafooz W, Alsaeedi A. Leveraging user-generated comments and fused BiLSTM models to detect and predict issues with mobile apps. Comput Mater Contin. 2024;79(1):735–59. doi:10.32604/cmc.2024.048270.
36. Inoue G, Alhafni B, Baimukan N, Bouamor H, Habash N. The interplay of variant, size, and task type in Arabic pre-trained language models. 2021. doi:10.48550/arXiv.2103.06678.
37. Alammary AS. BERT models for Arabic text classification: a systematic review. Appl Sci. 2022;12(11):5720. doi:10.3390/app12115720.
38. Antoun W, Baly F, Hajj H. AraGPT2: pre-trained transformer for Arabic language generation. 2020. doi:10.48550/arXiv.2012.15520.
39. Yafooz WM. Enhancing Arabic dialect detection on social media: a hybrid model with an attention mechanism. Information. 2024;15(6):316. doi:10.3390/info15060316.
40. Conneau A. Unsupervised cross-lingual representation learning at scale. 2019. doi:10.48550/arXiv.1911.02116.
41. Alhejaili R, Alhazmi ES, Alsaeedi A, Yafooz WM. Sentiment analysis of the COVID-19 vaccine for Arabic tweets using machine learning. In: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO); 2021 Sep; Noida, India: IEEE. p. 1–5. doi:10.1109/ICRITO51393.2021.9596517.
42. Almutiry S, Abdel Fattah M. Arabic cyberbullying detection using arabic sentiment analysis. Egypt J Lang Eng. 2021;8(1):39–50. doi:10.21608/ejle.2021.50240.1017.
43. Ahmed MS, Maher SM, Khudhur ME. Arabic cyberbullying detecting using ensemble deep learning. Indones J Electr Eng Comput Sci. 2023;32(2):1031–41. doi:10.11591/ijeecs.v32.i2.pp1031-1041.
44. Alduailaj AM, Belghith A. Detecting arabic cyberbullying tweets using machine learning. Mach Learn Knowl Extr. 2023;5(1):29–42. doi:10.3390/make5010003.
45. Worlali Azumah S, Elsayed N, ElSayed Z, Ozer M, La Guardia A. Deep learning approaches for detecting adversarial cyberbullying and hate speech in social networks. 2024. doi:10.48550/arXiv.2406.17793.
46. Alqahtani AF, Ilyas M. A machine learning ensemble model for the detection of cyberbullying. 2024. doi:10.48550/arXiv.2402.12538.