Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2024.052291

ARTICLE

Enhancing Arabic Cyberbullying Detection with End-to-End Transformer Model

Mohamed A. Mahdi[1], Suliman Mohamed Fati[2,*], Mohamed A.G. Hazber[1], Shahanawaj Ahamad[3] and Sawsan A. Saad[4]

[1]Information and Computer Science Department, College of Computer Science and Engineering, University of Ha'il, Ha'il, 55476, Saudi Arabia

[2]Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, 11586, Saudi Arabia

[3]Software Engineering Department, College of Computer Science and Engineering, University of Ha'il, Ha'il, 55476, Saudi Arabia

[4]Computer Engineering Department, College of Computer Science and Engineering, University of Ha'il, Ha'il, 55476, Saudi Arabia

*Corresponding Author: Suliman Mohamed Fati. Email: sgaber@psu.edu.sa; smfati@yahoo.com

Received: 29 March 2024 Accepted: 04 July 2024 Published: 27 September 2024

ABSTRACT

Cyberbullying, a critical concern for digital safety, necessitates effective linguistic analysis tools that can navigate the complexities of language use in online spaces. To tackle this challenge, our study introduces a new approach employing Bidirectional Encoder Representations from the Transformers (BERT) base model (cased), originally pretrained in English. This model is uniquely adapted to recognize the intricate nuances of Arabic online communication, a key aspect often overlooked in conventional cyberbullying detection methods. Our model is an end-to-end solution that has been fine-tuned on a diverse dataset of Arabic social media (SM) tweets showing a notable increase in detection accuracy and sensitivity compared to existing methods. Experimental results on a diverse Arabic dataset collected from the 'X platform' demonstrate a notable increase in detection accuracy and sensitivity compared to existing methods. E-BERT shows a substantial improvement in performance, evidenced by an accuracy of 98.45%, precision of 99.17%, recall of 99.10%, and an F1 score of 99.14%. The proposed E-BERT not only addresses a critical gap in cyberbullying detection in Arabic online forums but also sets a precedent for applying cross-lingual pretrained models in regional language applications, offering a scalable and effective framework for enhancing online safety across Arabic-speaking communities.

KEYWORDS

Cyberbullying; offensive detection; Bidirectional Encoder Representations from the Transformers (BERT); continuous bag of words; Social Media; natural language processing

1 Introduction

In recent years, the proliferation of social media and online communication platforms has been accompanied by a surge in cyberbullying, a form of harassment that occurs in the digital space [1–3]. This phenomenon poses significant psychological and social risks to individuals, particularly in

Arabic-speaking online communities, where the intricacies of the language add a layer of complexity to the identification and understanding of cyberbullying instances [4,5]. Cyberbullying in these communities has not been as extensively studied or addressed as in English-speaking environments, highlighting a critical gap in both research and practical applications for online safety [4]. The detection of cyberbullying in Arabic poses unique challenges, primarily due to the semantic and contextual richness of the language, which is often laden with dialectal variations and colloquialisms [6]. Traditional cyberbullying detection methods, which largely rely on keyword-based approaches or simple linguistic models, fall short of capturing these nuances, leading to high rates of false positives and negatives [7]. This limitation underscores the need for more sophisticated, context-aware methods that can adapt to the linguistic characteristics of Arabic [8].

Conventional methods for detecting cyberbullying, typically reliant on keyword-based approaches or simplistic linguistic models, prove inadequate in the face of Arabic's semantic richness and contextual depth, characterized by dialectal variations and colloquialisms [7]. These methods often result in inaccurate detections, either missing instances of cyberbullying or falsely identifying benign interactions as harmful [9,10]. This deficiency underscores the urgent need for more sophisticated, contextually aware detection methods capable of comprehending the complexities of Arabic [4].

BERT (Bidirectional Encoder Representations from Transformers) has been chosen as the baseline model for this study due to its robust performance in a wide range of natural language processing (NLP) tasks. BERT's architecture, which is based on the Transformer model, allows it to understand the context of words in a sentence by looking at both the preceding and succeeding words simultaneously. This bidirectional approach provides a deeper understanding of language context, making BERT particularly effective in tasks that require contextual comprehension, such as sentiment analysis, question answering, and text classification.

The Transformer model, introduced in "Attention is All You Need" by [11], revolutionized NLP by utilizing self-attention mechanisms to handle the dependencies between words in a sentence efficiently. The self-attention mechanism enables the model to weigh the importance of different words in a sentence when making predictions, thus improving the model's ability to capture long-range dependencies and contextual information. This innovation is fundamental to BERT's success, as it allows the model to encode rich contextual representations of text, which are crucial for accurately detecting cyberbullying that often involves subtle cues and varying contexts. Therefore, the primary goal of this study is to develop a robust, effective model for identifying cyberbullying in Arabic social media (X platform) content using enhanced Bidirectional Encoder Representations from Transformers (E-BERT), contributing significantly to online safety and mental well-being. This approach not only addresses a crucial research gap but also pioneers the cross-lingual application of advanced NLP technologies in diverse linguistic contexts.

The remainder of this paper is organized as follows: Section 2 reviews the related work in cyberbullying detection and NLP models. Section 3 describes the methodology, including the adaptation of the BERT model and the integration of various methods. Section 4 presents the experimental setup and results, followed by a discussion. Finally, Section 5 concludes the paper with insights and potential directions for future research.

2 Related Works

Cyberbullying, an escalating issue on social media platforms, has prompted numerous initiatives to employ machine learning techniques for its detection and prevention, especially in the context of Arabic online interactions [12,13]. Traditional machine learning methods, including Support Vector

Machines (SVMs), decision trees, and random forests, have been previously applied to identify cyberbullying incidents [7,14]. These techniques, initially effective, face challenges when grappling with the intricacies and subtleties of natural language, particularly in Arabic, which has unique linguistic features and dialectal variations. Consequently, there is a limited capability in these methods to recognize new or evolving forms of cyberbullying, an issue that is increasingly relevant in the dynamic landscape of Arabic social media interactions [13–15].

For example, Alduailaj et al. [16] employed machine learning to automatically detect cyberbullying in Arabic social media contexts. Their method uses an SVM classifier, trained, and tested on real datasets from YouTube and Twitter. To address the linguistic complexities of Arabic, the Farasa tool is integrated, enhancing the detection capabilities of their system in identifying cyberbullying incidents. The study used traditional machine learning (ML) and the model has limited performance.

Khairy et al. [17] evaluated the efficacy of various single and ensemble machine learning algorithms in detecting foul language and cyberbullying in Arabic text. They tested three single classifiers and three ensemble models on three Arabic datasets, including two publicly available offensive datasets and one specifically created for this research. The findings indicate that ensemble methods, particularly the voting technique, outperform single classifiers. The voting ensemble achieved higher accuracy (71.1%, 76.7%, and 98.5%) compared to the best single classifiers (65.1%, 76.2%, and 98%) across these datasets. Additionally, they enhanced the voting technique's effectiveness through hyperparameter tuning specifically for the Arabic cyberbullying dataset. Mubarak et al. [18] developed a dynamic training dataset aimed at detecting offensive tweets, harnessing a seed list of offensive terms. They utilized character n-grams to train a deep learning model that achieved a 90% F1 score in classifying tweets. Additionally, they released a novel dataset of dialectal Arabic news comments sourced from various social media platforms, including Twitter, Facebook, and YouTube. Their research delves into the distinct lexical features of abusive comments, particularly focusing on the use of emojis. The study revealed that this multi-platform news commentary dataset could effectively capture the diversity present in different dialects and domains. Beyond assessing the model's ability to generalize, Mubarak et al. [18] provided a detailed analysis of how emojis, especially those from the animal category, are used in offensive commentary, echoing similar patterns found in lexical studies conducted by Chowdhury et al. [19].

Alakrot et al. [20] established an Arabic dataset composed of YouTube comments designed to identify offensive language within a machine learning framework. The dataset was curated adhering to criteria of availability, diversity, representativeness, and balance to ensure it could effectively train predictive models for detecting abusive online communication in Arabic. In a subsequent study, Haddad et al. [21] tackled the detection of abusive language and hate speech. They proposed a methodology for data preprocessing and rebalancing, utilizing models such as the bidirectional Gated Re-current Unit (GRU) and Convolutional Neural Network (CNN). Their results demonstrated that the bidirectional GRU model, enhanced with an attention layer, outperformed other models. On a labeled dataset of Arabic tweets, this approach achieved an F1 score of 85.9% for offensive language detection and 75% for hate speech detection.

Abozinadah et al. [22] conducted an evaluation of different machine learning algorithms to identify abusive accounts using Arabic tweets. For this analysis, the dataset was compiled using the top five Arabic profanity terms, resulting in a collection of 255 unique user accounts. The Naïve Bayes classifier, configured with 10 tweets and 100 features, delivered the most effective results, boasting a 90% accuracy rate in identifying abusive accounts. Building on this work, Abozinadah et al. [22] introduced an Arabic word correction technique aimed at circumventing internet censorship and content-filtering system vulnerabilities. This technique demonstrated a high level of accuracy, reaching 96.5% in their tests. Further advancing their research, Abozinadah et al. [23] employed a statistical learning approach to detect adult content disseminators among Arabic-speaking social media accounts. By analyzing the use of obscenities, vulgar language, slang, and swear words in Arabic tweets, they were able to pinpoint abusive accounts. The statistical method applied not only yielded a predictive precision of 96% but also successfully addressed the limitations associated with the bag-of-words (BOW) model.

While deep learning models have improved the detection of offensive content on social media platforms [24,25], their application to offensive tweets in Arabic presents specific challenges. These challenges include the linguistic diversity and complexity of Arabic dialects, the contextual understanding necessary to distinguish offensive from non-offensive content accurately, and the scarcity of large, annotated datasets for model training. Additionally, the generalization of models across various social media platforms and types of user-generated content remains a challenging task. These issues underscore the need for continued research to enhance the robustness and adaptability of models tailored for offensive tweet detection. This study aims to bridge these research gaps, with a focus on developing sophisticated models capable of effectively identifying offensive tweets in Arabic, thereby contributing to safer online environments.

3 Materials and Methods

In this study, we introduced a new BERT model called E-BERT to detect cyberbullying in Arabic texts (Fig. 1). The E-BERT model underwent pretraining with BookCorpus, which is composed of 11,038 books that have not been published, alongside the text of English Wikipedia, with the exclusion of lists, tables, and headers. The training procedure begins with preprocessing, where texts are tokenized utilizing the WordPiece model with a 30,000-word vocabulary. The model's input format comprises sequences: [CLS] Sentence A [SEP] Sentence B [SEP]. There is a 50% chance that Sentence A and Sentence B are sequential sentences from the original corpus; otherwise, Sentence B is a random sentence from elsewhere in the corpus. Here, the term 'sentence' refers to a span of text longer than a single traditional sentence, bound by the condition that the two combined 'sentences' total less than 512 tokens.

During the masking phase of the training, 15% of the tokens undergo masking. In this process, 80% of the time, the masked tokens are substituted with a [MASK] token. In 10% of instances, a random token replaces the masked token, ensuring it is different from the original. For the remaining 10%, the masked tokens are kept unchanged. This technique is designed to help the model better understand context and predict masked words within the text. The E-BERT model undergoes two primary phases: pre-training and fine-tuning. In both stages, the model utilizes similar architectures, except for the output layers. The pre-training phase involves using a set of initial model parameters, which are then applied to initialize the models for various downstream tasks. When it comes to fine-tuning, all these parameters are meticulously adjusted to optimize performance. Special symbols play a crucial role in this process: '[CLS]' is added at the beginning of each input example as a unique identifier, while '[SEP]' serves as a separator token, which is particularly useful in distinguishing different elements in the data, such as questions and answers. Fig. 1 shows the architecture of the E-BERT model for detecting cyberbullying in Arabic text. The process includes input preprocessing, BERT-based contextual embedding generation, feature extraction, dimensionality reduction, and final cyberbullying detection.

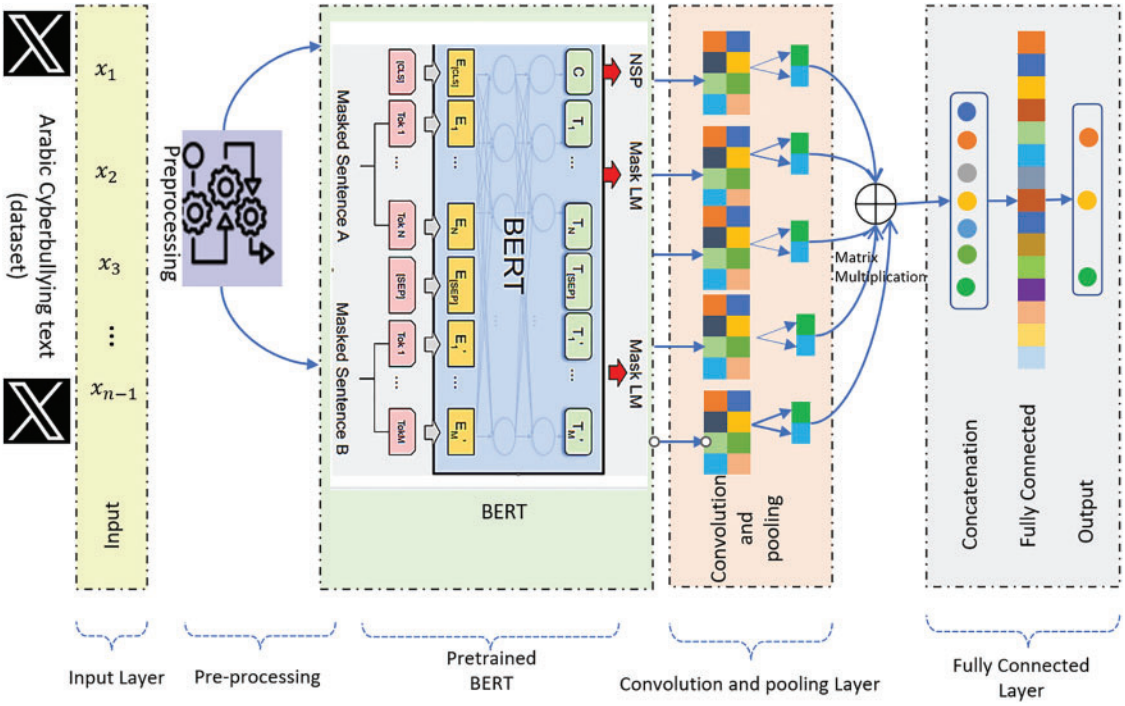

Figure 1: Architecture of the E-BERT model for detecting cyberbullying in Arabic text. The process includes input preprocessing, BERT-based contextual embedding generation, feature extraction, dimensionality reduction, and final cyberbullying detection

3.1 Datasets and Input Layer

This study uses an Arabic dataset collected from the “X” SM platform [26] and can be found on Kaggle. It is a collection designed to study and identify instances of cyberbullying in Arabic content. Comprising various forms of potentially abusive content, this dataset serves as a valuable resource for researchers and technologists aiming to develop machine learning models to detect cyberbullying behavior. The data is mainly text-based and contains examples labeled for bullying traits, making it possible to train supervised learning algorithms on pattern recognition in Arabic textual data. Given that the Arabic language poses unique challenges due to its rich morphology and various dialects, this dataset is particularly significant for efforts to improve online social interactions among Arabic-speaking communities. Table 1 presents a sample of tweets from the “X” platform dataset that exemplify instances of cyberbullying.

Table 1: Example of cyberbullying samples in “X” platform dataset

ID	Cyberbullying tweet samples	Pred	Label
1	أنت حمار ولا تعرف شيئًا	Offensive (Cyberbullying)	1
2	وزنك زاد وصرتي قبيحه جدا	Offensive (Cyberbullying)	1
3	الحرية حق و ليست منحة يمنحها من لا يملك لمن يستحق	Non-offensive (non-bullying)	0
4	أنتي عاهره ومتبرجه	Offensive (Cyberbullying)	1

(Continued)

Table 1 (continued)

ID	Cyberbullying tweet samples	Pred	Label
5	انت قمر	Non-offensive (non-bullying)	0
6	أتمنى لك يومًا سعيدًا ومليئًا بالنجاح	Non-offensive (non-bullying)	0

3.2 Preprocessing

In preparing data for E-BERT, especially for a task such as detecting cyberbullying in Arabic text, the first crucial step is tokenization. Tokenization transforms raw text into a format that E-BERT can understand, namely, a series of tokens. The tokenizer function, T, breaks down the input string S into a sequence $T(S) = [t_1, t_2, \ldots, t_n]$, where each t_i is an individual token [27]. For Arabic text, E-BERT employs the WordPiece tokenizer, which is adept at managing the language's morphological complexity by breaking down words into smaller, meaningful subwords that exist within a predefined vocabulary [28]. Following tokenization, the sequence is augmented with special tokens that BERT recognizes. The '[CLS]' token is prepended to the start of every sequence, serving as a unique identifier that the model uses to understand that the output of this token should be used for classification purposes. Additionally, '[SEP]' tokens are inserted to separate different sentences or textual segments, allowing E-BERT to differentiate between them, which is particularly vital for tasks involving multiple distinct text inputs [29].

The uniformity of input is essential for the E-BERT model. Hence, sequences are adjusted to a fixed length, L, typically 512 tokens for E-BERT. If a sequence exceeds this length, it is truncated, and if it is shorter, it is extended with '[PAD]' tokens. The padded sequence, denoted S'', is thus standardized across the dataset, ensuring consistent structure in the input data [30]. Attention masks, represented as M, are binary sequences that enable the model to focus on the meaningful content of the input and disregard the '[PAD]' tokens. For each token in S'', the corresponding element in M is set to 1 if the token is not a '[PAD]' token and O if it is a '[PAD]' token [31]. This mechanism is a critical component of E-BERT 's attention system, which discerns which parts of the input should contribute to the understanding of the context. Segment IDs are another aspect of the preprocessing routine. They provide a means for the model to distinguish between multiple sentences within a single input. For a single-sentence input, the Segment ID vector, C, consists entirely of zeros. In contrast, for paired sentences, the Segment IDs switch from 0 to 1 after the first '[SEP]' token, effectively demarcating the beginning of the second sentence [32].

Lastly, handling out-of-vocabulary (OOV) words is crucial. E-BERT's tokenizer deconstructs OOV words into known sub-tokens that are present in the vocabulary [33]. This is particularly important for Arabic text, where a single word can have numerous variants due to its rich inflectional system. By representing words as combinations of sub-tokens, BERT can capture the meaning and nuances of words it has not encountered before.

In the WordPiece tokenization algorithm [34], the input word 'w' represents the word to be tokenized, serving as the starting point for the tokenization process. The vocabulary 'V' is a predefined set of known tokens, including subwords and special characters, that the algorithm uses to decompose 'w' into recognizable pieces. The token list 'T' is the output of the process, initially empty and progressively filled with tokens derived from 'w'. As 'w' is processed, the algorithm identifies the longest substring that matches an entry in 'V', appending it to 'T' and updating 'w' by removing the matched portion. If a segment of ‘w’ isn’t in ‘V’, the algorithm appends a special ‘[UNK]’ token to ‘T’, indicating an unknown subword. The final output is the token list ‘T’, a sequence of tokens that collectively represent the original word ‘w’ in a form suitable for subsequent NLP tasks within BERT models. Each component–‘w’, ‘V’, and ‘T’–plays a critical role in translating the raw text into a structured format that preserves semantic meaning and accommodates the model’s vocabulary constraints.

Algorithm 1: WordPiece tokenization algorithm

```
Input:  A word w to be tokenized, a vocabulary V of token to index mappings
Output:  A list of tokens T representing the word w
procedure TOKENIZE (w, V)
    T $\leftarrow$ empty list
    while w is not empty do
        subword $\leftarrow$ the longest substring in w that is also in V (starting from the first  character)
        if subword does not exist then
            T.append(‘[UNK]’)
            break
        else
            T.append(subword)
            w $\leftarrow$ suffix of w after removing subword
        end if
    end while
    return T
end procedure
```

3.3 Pre-Training of Deep Bidirectional Transformers Model

BERT [35] is a transformer-based model pretrained on a vast corpus of English data using a self-supervised approach. This pretraining involves no human labeling, leveraging raw texts to automatically generate inputs and labels. For the context of cyberbullying detection, the BERT model’s pretraining can be mathematically described through its two main objectives.

3.3.1 Masked Language Modeling (MLM)

In MLM, a sentence S is taken, and approximately 15% of the words are randomly masked. Let $S = \{w_1, w_2, \ldots, w_N\}$ be a sentence with N words. The model creates a masked version S', where each word w_i has a probability p (15% in this case) of being replaced by a mask token. The objective function of MLM can be represented as:

$$L_{MLM}(S, S') = -\sum_{i=1}^{N} m_i \log P\left(w_i \mid S'\right) \tag{1}$$

Here, m_i is an indicator function that is 1 if the word w_i is masked and 0 otherwise. The model predicts the masked words by maximizing the likelihood of the original words given the masked sentence. This approach differs from traditional recurrent neural networks (RNNs), which process words sequentially, and autoregressive models like GPT, which mask future tokens. BERT’s bidirectional nature enables it to learn a more comprehensive representation of the sentence.

3.3.2 Next Sentence Prediction (NSP)

In NSP, the model concatenates two sentences, A and B, during pretraining. The model has to predict whether Sentence B follows A in the original text. Let Concat(A, B) represent the concatenation of Sentences A and B, and y be a binary label indicating whether B follows A (1 if true, 0 if false). The objective function of NSP is:

$$L_{NSP}(A, B) = -\log P(y \mid \mathrm{Concat}(A, B)) \tag{2}$$

This ability is crucial in understanding the flow of conversation in online interactions, which is a key element in identifying instances of cyberbullying. By combining these two tasks, BERT effectively learns complex language patterns, which is essential for cyberbullying detection. For the downstream task of cyberbullying classification, the feature vector $F(S)$ extracted from BERT for a sentence S can be utilized. If C is a classifier and “Loss” is an appropriate loss function (like cross-entropy), the training objective for cyberbullying detection can be expressed as:

$$L_{\text{cyberbullying}}(S, \text{label}) = \mathrm{Loss}(C(F(S)), \text{label}) \tag{3}$$

Here, “label” indicates whether the sentence S contains cyberbullying content. By fine-tuning BERT on cyberbullying-specific data, the model learns to discern the subtle linguistic cues of online harassment, making it an effective tool for identifying and mitigating cyberbullying in digital communication platforms. The BERT model undergoes two primary phases: pre-training and fine-tuning (Fig. 2). In both stages, the model utilizes similar architectures, except for the output layers. The pre-training phase involves using a set of initial model parameters, which are then applied to initialize the models for various downstream tasks. When it comes to fine-tuning, all these parameters are meticulously adjusted to optimize performance. Special symbols play a crucial role in this process: ‘[CLS]’ is added at the beginning of each input example as a unique identifier, while ‘[SEP]’ serves as a separator token, which is particularly useful in distinguishing different elements in the data, such as questions and answers [35].

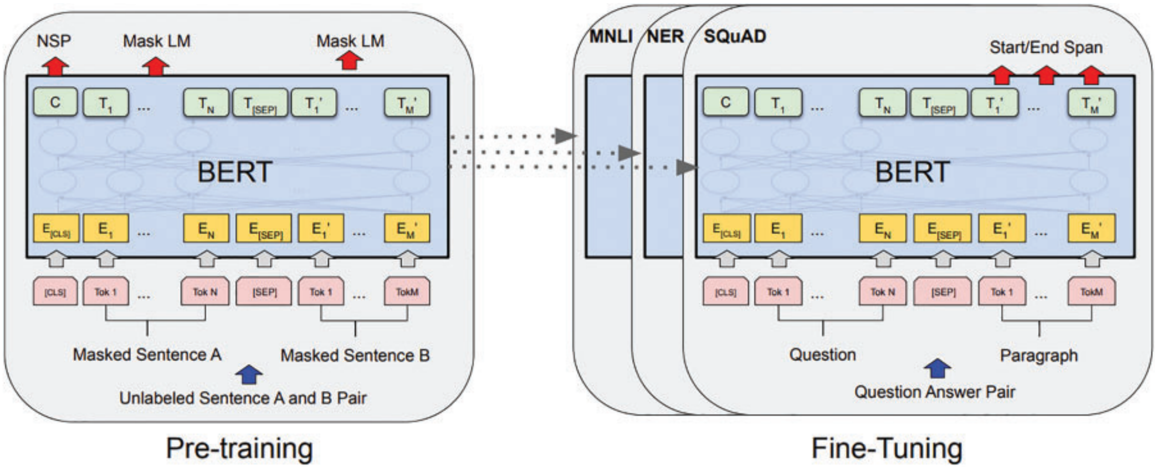

Figure 2: Overview of BERT’s pre-training and fine-tuning stages. Pre-training involves learning from unlabeled sentences using Masked LM and NSP. Fine-tuning adapts the model to specific tasks with labeled data

3.4 Fully Connected Layer (FC)

In an FC, the feature vector representation derived from the weight vector of the concatenated pooling layers is mapped to the input vector through a weights matrix to learn the bullying for building the cyberbullying model. The FC includes multiple dense layers, non-linear activation, softmax, and prediction function to obtain the correct bully classification as follows:

$$\mathbb{H}_t = \text{SoftMax}\left(w_t h_{t-1} + b_t\right) \tag{4}$$

where w_t and b_t are parameters learned in training, $\mathbb{H}_t$ is the obtained from the pooled concatenated feature vector and h_{t-1} is the feature map received from the E-BERT layers. The output layer performs the correct classification using the SoftMax function. The cross-entropy loss was minimized to learn the model parameters as the training objective using the Adam optimization algorithm [36]. It is provided by:

$$\text{CrossEntropy (p, q)} = -\sum\nolimits_{p}(x)\log(q(x)) \tag{5}$$

Given a true distribution p, which represents a one-hot vector representing characters in messages posted on social media, and a SoftMax output q, the negative log probability of the true bullies can be computed.

Pretrained models like BERT use tokenizers that are optimized for the language they were trained on. English tokenizers may not effectively capture the nuances of Arabic, leading to suboptimal token representations. This mismatch can result in poor handling of Arabic words, especially those with complex morphological patterns.

4 Results and Analysis

4.1 Experimental Setting

In our experimental setup, we leverage the flexibility of BERT's self-attention mechanism within its Transformer architecture, enabling it to adapt seamlessly to a variety of downstream tasks. This adaptability is particularly crucial for tasks involving either single texts or pairs of texts, as it allows for straightforward modifications to the inputs and outputs as required. Unlike approaches that encode text pairs separately before implementing bidirectional cross-attention—such as those proposed by Parikh et al. [37] and Seo et al. [38]—BERT integrates these processes by encoding concatenated text pairs using self-attention. This method inherently encompasses bidirectional cross-attention between the pair of sentences [39].

For each specific task in our study, we adapt BERT by inputting task-relevant data and fine-tuning all parameters in an end-to-end manner. This involves handling diverse types of text pairs, similar to various scenarios observed in pre-training, such as paraphrasing (sentence pairs), entailment (hypothesis-premise pairs), question answering (question-passage pairs), and text classification or sequence tagging (text-Ø pairs). On the output side, token representations are utilized for tasks requiring token-level analysis, like sequence tagging or question answering. For classification tasks such as entailment or sentiment analysis, the [CLS] token representation is employed and fed into a dedicated output layer for effective classification. The model was fine-tuned using a learning rate schedule, with a batch size of 32, and trained for 40 epochs. The Adam optimizer, incorporating a weight decay of 0.01, was utilized for optimization.

4.2 Accuracy, Precision, Recall and F1 Score

In this study, we concentrated on the assessment of our proposed model's capacity to differentiate between offensive and non-offensive content, utilizing a range of evaluation metrics. We developed and compared deep learning-based models for cyberbullying detection, including a baseline BERT model and our enhanced version, E-BERT. Evaluation metrics are vital for comprehensively understanding how different models perform and are recognized within the scientific community. We employed the following commonly accepted criteria to evaluate the effectiveness of cyberbullying detection classifiers on Twitter (currently, X platform):

- Accuracy is the metric that indicates the ratio of correctly identified tweets to the overall count of tweets analyzed by cyberbullying prediction models. The subsequent formula is employed for its calculation.

$$\text{Accuracy} = \frac{(\text{tp} + \text{tn})}{(\text{tp} + \text{fp} + \text{tn} + \text{fn})} \tag{6}$$

where 'fp' stands for false positives, which are instances where the model incorrectly predicts cyberbullying; 'fn' stands for false negatives, where the model fails to identify actual cyberbullying occurrences. Conversely, 'tp' represents true positives, correctly identified instances of cyberbullying, and 'tn' refers to true negatives, which are non-cyberbullying instances that the model correctly does not flag as cyberbullying.

- Precision measures the accuracy of the positive predictions made by the model; it is the fraction of true positive results among all tweets labeled as cyberbullying.
- Recall, also known as sensitivity, evaluates the model's ability to correctly detect all relevant instances; it is the fraction of true positive results relative to the number of actual cyberbullying cases.
- The F1 score is a composite metric that combines precision and recall into a single value by calculating their harmonic mean, thus ensuring a balance between the two.
- These metrics—accuracy, precision, recall, and F1 score—are extensively utilized in literature to evaluate the effectiveness of cyberbullying prediction models. They are computed as follows:

$$\text{Precision} = \frac{\text{tp}}{(\text{tp} + \text{fp})} \tag{7}$$

$$\text{Recall} = \frac{\text{tp}}{(\text{tp} + \text{fn})} \tag{8}$$

$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{recision} + \text{recall}} \tag{9}$$

Our model for Arabic cyberbullying detection demonstrates promising trends in both the training and validation phases (see Fig. 3). Initially, the model's training accuracy increases sharply, indicating efficient learning from the training dataset. It then plateaus, suggesting that the model has reached a state of convergence where further learning on the training set yields minimal gains. The validation accuracy, while slightly lower, mirrors the overall upward trend, plateauing similarly. This parallel progression between training and validation accuracy is indicative of good model generalization with minimal overfitting, though the presence of fluctuations in the validation accuracy suggests that the model's response to the validation set may benefit from further stabilization measures.

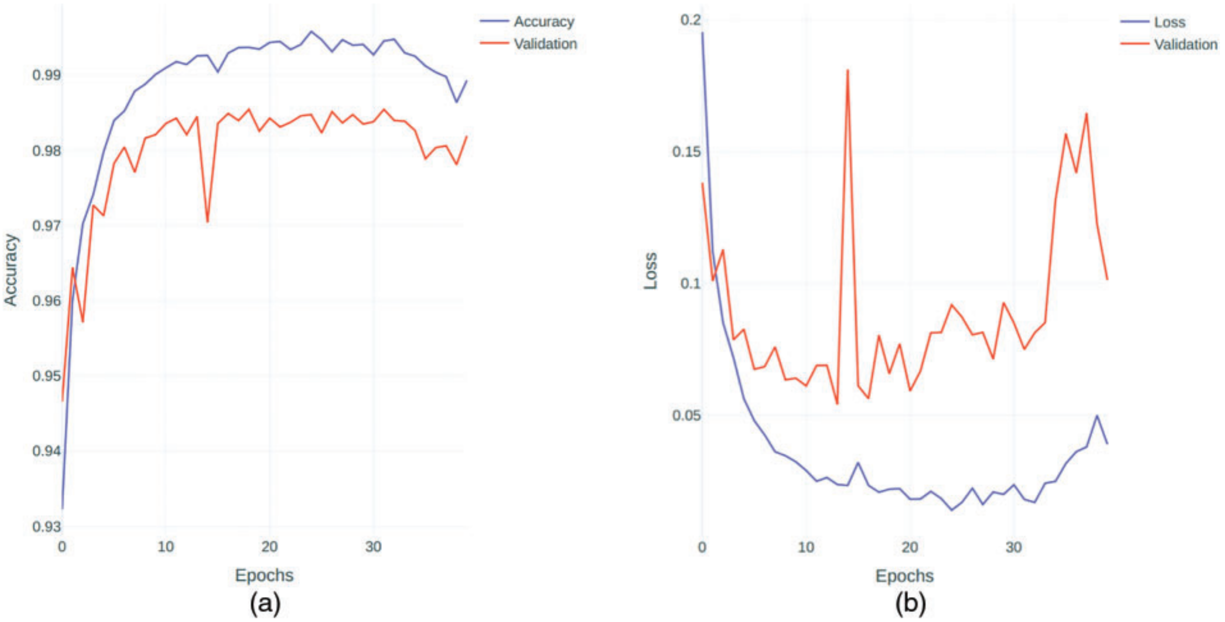

Figure 3: Testing accuracy and loss of the proposed enhanced BERT model on the "X" platform dataset. (a) Proposed enhanced BERT model testing accuracy. (b) Proposed enhanced BERT model loss

The loss plots present a more nuanced picture. The training loss decreases consistently, aligning with the expected optimization trajectory of a well-fitting model. However, the validation loss exhibits notable volatility, with spikes that suggest sensitivity to the composition of the validation set. This could imply that the model while performing well on average, may be prone to performance dips depending on the specific data it encounters. Such behavior points to the potential need for further model tuning, incorporating regularization techniques or adjusting the learning rate, to achieve a more consistent loss reduction in the validation phase and improve the model's robustness and reliability in detecting cyberbullying within Arabic text.

The matrix reveals a predominant diagonal concentration of values, which is indicative of a model with a high predictive accuracy. Specifically, the model exhibits a true positive rate of 0.92, meaning that 92% of the offensive instances were correctly identified. This high detection rate is crucial for the application of cyberbullying, where the correct identification of offensive content is necessary to protect individuals from harm. Conversely, the true negative rate is 0.99, suggesting that the model correctly identified 99% of non-offensive content. This is equally important in minimizing the number of false alarms, where benign content is misclassified as cyberbullying, which could lead to unnecessary censorship or other unintended consequences. However, a false negative rate of 0.08 indicates that 8% of offensive content is not being correctly identified by the model. This can be attributed to the challenges posed by subtle or coded language often used in cyberbullying, which may appear benign out of context. Also, the diversity in Arabic dialects and slang can cause the model to miss certain expressions not adequately represented in the training data. In the context of cyberbullying detection, this could mean that a small proportion of harmful content might go undetected, potentially allowing negative interactions to persist. This aspect of model performance might require attention, potentially by improving the representativeness of offensive content in the training data or by fine-tuning the model's threshold for classifying text as offensive. The low false positive rate of 0.01 demonstrates the model's precision in not misclassifying non-offensive content as offensive, which is desirable to prevent over-policing of the content and to maintain freedom of expression. Fig. 4 is the confusion matrix of the Arabic cyberbullying detection model.

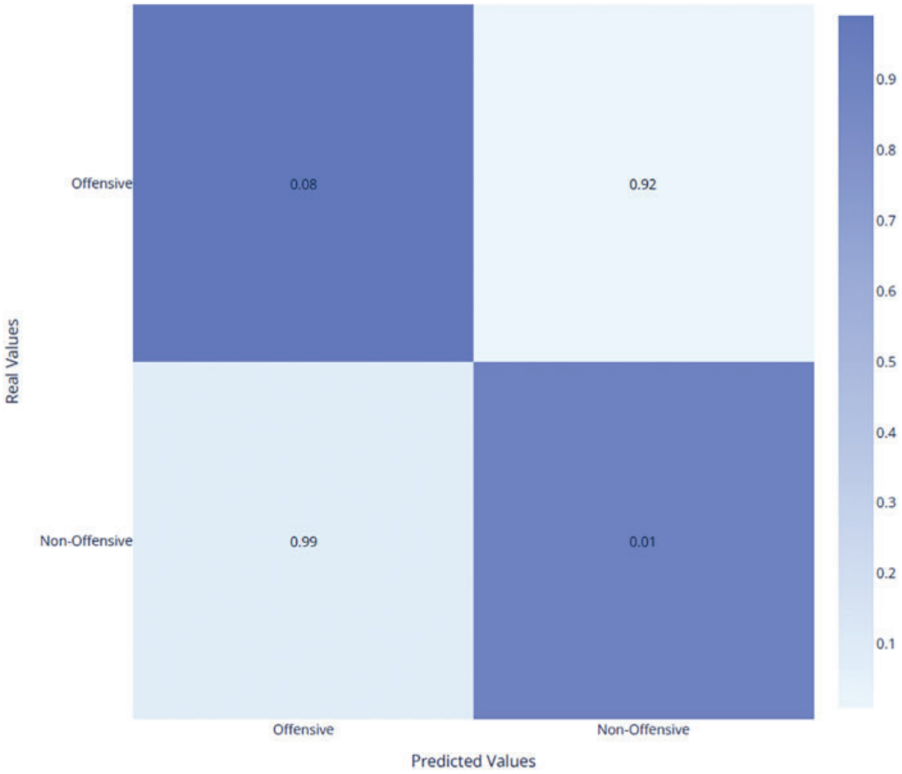

Figure 4: Confusion matrix of the Arabic cyberbullying detection model

4.3 Precision, Recall and F1 Score Threshold Curves

Fig. 5 shows that the precision for non-offensive content (indicated by the red line) remains consistently high across all thresholds. This suggests that when the model predicts content as non-offensive, it is correct most of the time. For offensive content (blue line), precision starts high when the threshold is low but begins to decrease slightly as the threshold increases. This behavior typically indicates that at lower thresholds, the model is more conservative in predicting content as offensive, but as the threshold is raised, it becomes more selective, leading to more false negatives (offensive content not identified). Fig. 6 presents the recall curve for offensive content (blue line) starts high and begins to decline as the threshold increases, which is expected since a higher threshold requires stronger evidence to classify content as offensive, potentially leading to more false negatives. The recall for non-offensive content (red line) is almost perfect across all thresholds, suggesting the model is highly capable of identifying all non-offensive instances. The F1 score is the harmonic mean of precision and recall, providing a single measure that balances both. In Fig. 7, the F1 score for offensive content (blue line) remains relatively high across the thresholds but shows a slight decline as the threshold increases, indicating a balanced trade-off between precision and recall at lower thresholds. The F1 score for non-offensive content (red line) is high and stable, which, combined with the high precision and recall, suggests that the model is particularly adept at identifying non-offensive content.

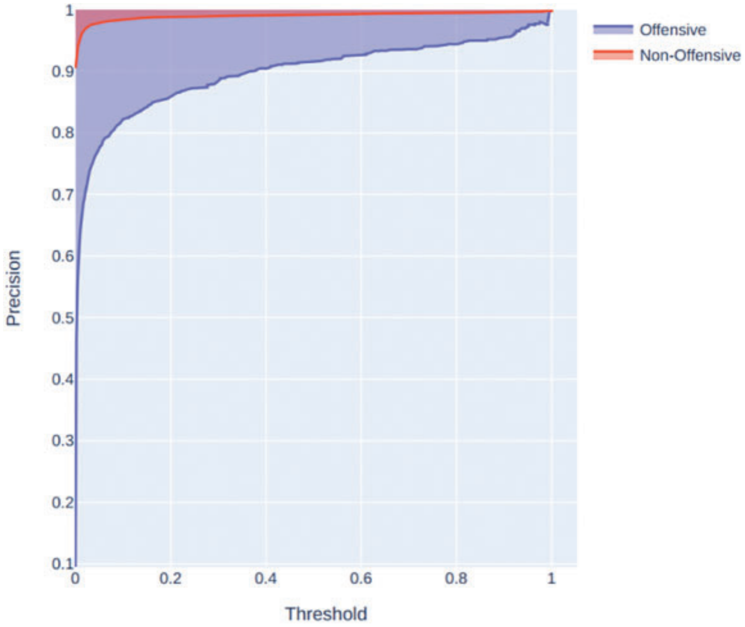

Figure 5: Precision curve for Arabic cyberbullying classification model

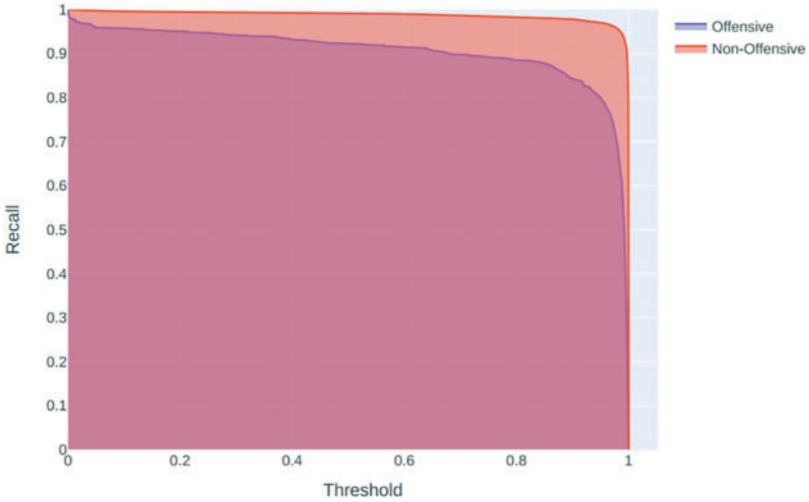

Figure 6: Recall curve for Arabic cyberbullying classification model

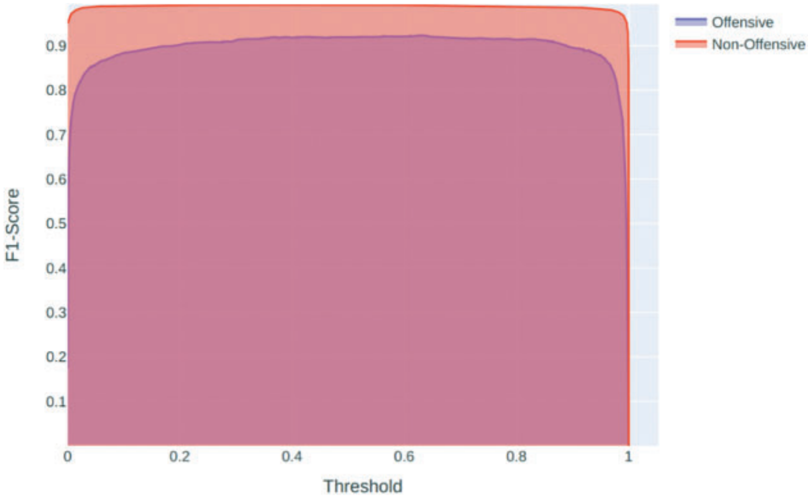

Figure 7: F1 score curve for Arabic cyberbullying classification model

Additionally, these curves suggest that the classifier performs with a high degree of reliability, particularly in identifying non-offensive content. The model demonstrates a high precision for non-offensive predictions across all thresholds, which is paramount in applications where the cost of false positives, and wrongfully censoring content is high. However, the model shows a decline in precision for offensive content as the threshold increases, which may require careful consideration to balance the need for high precision without overly sacrificing recall. The high recall for non-offensive content indicates very few false positives, which is desirable. Yet, the model must also maintain sufficient recall for offensive content to ensure that cyberbullying instances are not overlooked. The F1 score suggests the model achieves a good balance between precision and recall for offensive content at lower thresholds. For practical application, a threshold that maximizes the F1 score may be chosen to achieve an optimal balance, ensuring that cyberbullying is detected accurately without an excessive number of false alarms.

4.4 Precision-Recall Curve

The precision-recall curve is a pivotal tool for evaluating binary classifiers in tasks such as cyberbullying detection, where the balance between precision; the proportion of true positives among all positive predictions and recall, and the proportion of true positives among all actual positives is critical. This curve plots precision against recall at various threshold settings, illustrating the trade-off between capturing all relevant instances and ensuring that the predictions made are relevant.

For cyberbullying detection, a high area under the curve (AUPRC) is indicative of a model's effective performance, with a value closer to 1.0 signaling superior capability. The provided figure showcases two curves, one for each class of the binary problem: offensive and non-offensive. The non-offensive class achieves an AUPRC of 1.00, denoting a perfect precision-recall balance, while the offensive class shows a slightly lower (Fig. 8), yet still strong, AUPRC of 0.96. This disparity suggests the model is exceptionally adept at confirming instances of non-offensive content while still maintaining commendable accuracy in identifying offensive content. such precision-recall characteristics would be interpreted as a robust indicator of the model's utility, especially in scenarios where missing an instance of cyberbullying (false negatives) is as significant as incorrectly flagging benign content (false positives). As shown in Fig. 8, both classes approach the ideal point (1,1) in precision-recall space, with the non-offensive class nearly overlapping it, which implies an excellent performance. The offensive class, while not as close, still demonstrates a high degree of reliability. This suggests that the model is well-calibrated for the sensitivity required in cyberbullying detection, able to discern most true cases of offensive content with a high certainty of correctness, which is essential for practical applications.

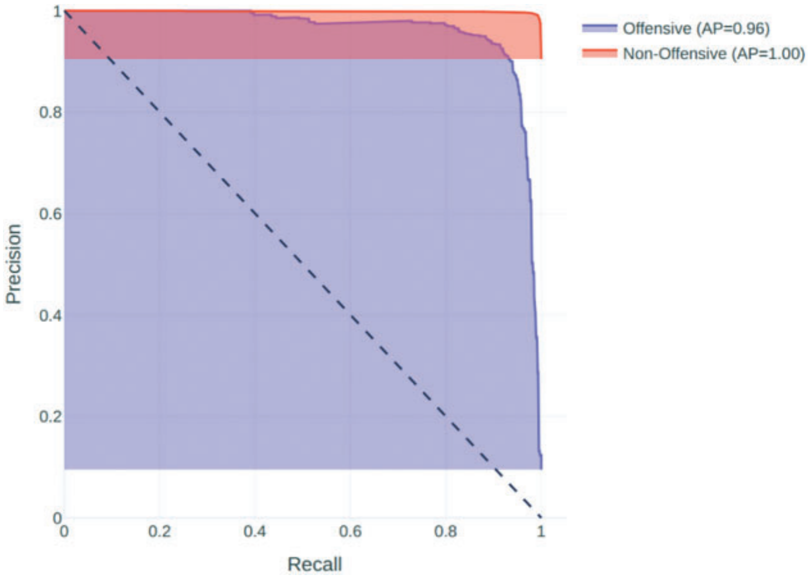

Figure 8: Precision-recall curves for offensive and non-offensive classifications in cyberbullying detection with near-optimal model performance

4.5 Area under the Curve (AUC)

The Area Under the Curve (AUC) is a widely used metric for evaluating the efficacy of binary classifiers, particularly relevant in cyberbullying detection on social media platforms. It quantifies a classifier's aggregate capability by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Within the cyberbullying context, the TPR quantifies the proportion of verified offensive instances that are accurately detected, whereas the FPR indicates the proportion of non-offensive instances that are incorrectly flagged as such. The AUC metric spans from 0 to 1, with 1 denoting an ideal classifier that impeccably distinguishes between bullying and non-bullying content without error, and 0.5 signifying a classifier whose predictive performance is equivalent to random guessing.

Additionally, the greater the AUC, the better the model's efficiency in differentiating the positive and negative samples [40]. Fig. 9 shows AUC curves for the proposed enhanced BERT model in our study, where (a) demonstrates the model performance in our own cyberbully X dataset and (b) demonstrates the model performance in terms of AUC for the "X" platform dataset with 0.99 and 0.99, respectively.

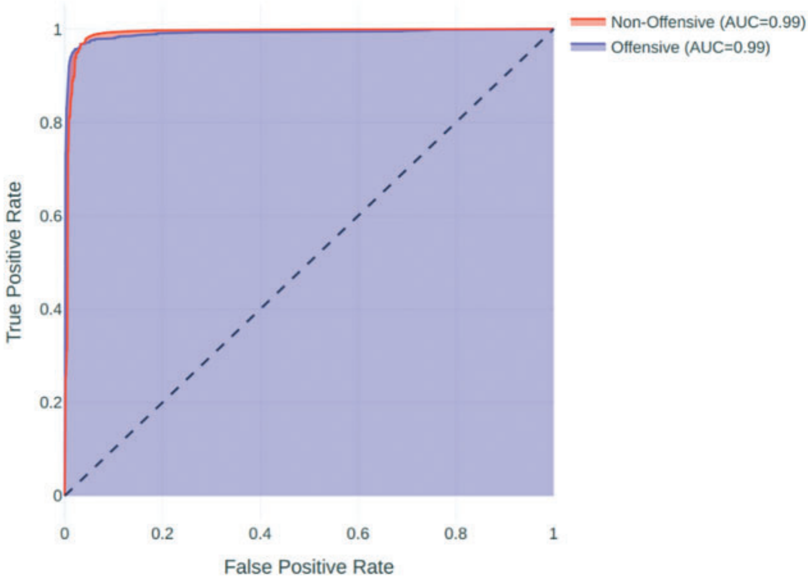

Figure 9: Receiver operating characteristic (ROC) curve for an Arabic cyberbullying classification model

4.6 *Performance Result of Baseline Model vs. Proposed Model*

Fig. 10 presents a comparative analysis between the baseline BERT model and the enhanced BERT model developed in this study, designated as E-BERT, on the X cyberbullying dataset. The baseline BERT demonstrates robust performance with an accuracy of 93.4%, a precision of 91.5%, a recall of 91.5%, and an F1 score of 92.9%. In contrast, our proposed E-BERT model shows a marked improvement across all metrics: it achieves an accuracy of 98.45%, indicating a higher overall rate of correct predictions. Precision, which reflects the model's ability to identify only relevant instances of cyberbullying, is notably higher at 99.17%. The recall rate of 99.10% suggests that our model is exceptionally proficient at detecting the most positive instances of cyberbullying in the dataset. Finally, the F1 score, a harmonic means of precision and recall, stands at 99.14%, underscoring the superior balance between precision and recall in our E-BERT model. These results highlight the effectiveness of the modifications implemented in our E-BERT model for identifying cyberbullying content in Arabic text.

4.7 *Comparison between the Proposed Model and Related Literature Contributions*

Table 2 delineates the performance metrics of various models, facilitating a comprehensive understanding of each model's effectiveness. The Baseline BERT model sets a standard with an accuracy of 93.4%, a precision and recall of 91.5%, and an F1 score of 92.9%. Our enhanced model, E-BERT, surpasses this benchmark, achieving an accuracy of 98.45%, precision of 99.17%, recall of 99.10%, and an F1 score of 99.14%. This indicates a substantial improvement, particularly in precision, which suggests that E-BERT is better at minimizing false positives in the detection process. When compared to other literature contributions, such as CNN and Bidirectional Gated Recurrent Unit augmented with attention layer (Bi-GRU_ATT) models with accuracies of 92% and 93%, respectively, E-BERT demonstrates superior performance, highlighting the effectiveness of our modifications. Notably, the precision of CNN is significantly lower at 63%, which may point to a higher rate of false positive classifications. The long short-term memory (LSTM) and Bi-LSTM models show a marked decrease in accuracy compared to BERT-based models, reflecting the potential limitations of these architectures in capturing the contextual nuances necessary for Arabic cyberbullying detection.

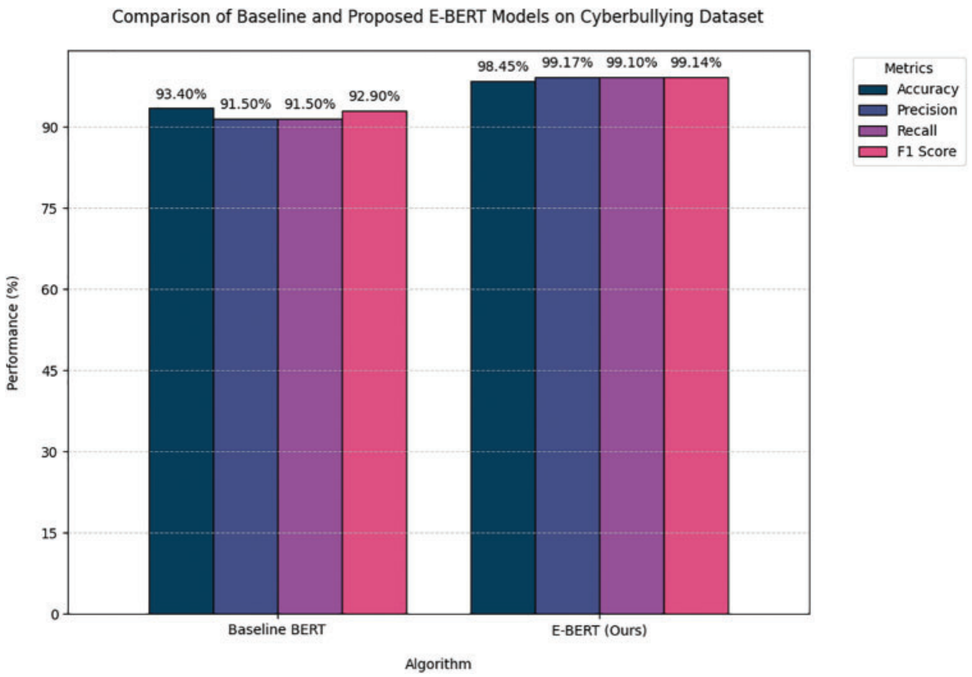

Figure 10: Comparison analysis between baseline and proposed E-BERT model on X cyberbullying dataset

Table 2: Comparison analysis between the proposed model and related literature contributions

No.	Algorithm	Accuracy (%)	Precision	Recall	F1 score
1	Baseline BERT	0.934	0.915	0.915	0.929
2	E-BERT (Ours)	**0.9845**	**0.9917**	**0.9910**	**0.9914**
3	CNN [21]	0.92	0.63	0.84	0.85
4	Bi-GRU_ATT [21]	0.93	0.91	0.83	0.86
5	LSTM [4]	83.18	83.00	–	83.00
	Bi-LSTM [4]	82.12	81.00	–	82.00
6	SVM [16]	95.742	–	–	–
7	J48 [22]	84.00	85.00	85.00	85.00

Furthermore, applying English-pretrained models to Arabic presents challenges, such as linguistic differences, vocabulary mismatch, and dialectal variations. These issues can be addressed through custom tokenization, dialectal adaptation, cultural embedding, and data augmentation techniques. Additionally, the research presented herein offers significant theoretical implications for the field of offensive content detection in SM. It enhances the current body of knowledge by demonstrating the adaptability of contextualized word embeddings across different languages and cultural milieus. This is especially relevant for the Arabic language, which poses distinct challenges due to its dialectical diversity and script complexity. By successfully adapting a model pre-trained in English to comprehend Arabic text, this research contributes to a deeper understanding of NLP's capabilities and limitations in dealing with linguistically rich and structurally complex languages. Furthermore, the study underscores the theoretical potential of cross-lingual transfer learning. The successful fine-tuning of a model initially pre-trained in English to perform tasks in Arabic bolsters the case for the transferability of language models across linguistic boundaries. This serves to broaden the horizon of transfer learning efficacy and lays the groundwork for subsequent explorations into developing language-agnostic models. Such advancements could lead to more universal NLP systems, capable of understanding and processing multiple languages with minimal need for language-specific adjustments.

Deploying the BERT-based model for cyberbullying detection in Arabic requires significant computational resources due to the model's complexity and the volume of data it processes. For real-time applications, the model should be able to process input data swiftly, which typically requires GPUs (Graphics Processing Units) or TPUs (Tensor Processing Units) for efficient parallel processing. Depending on the deployment scale, we recommend at least one NVIDIA V100 GPU or equivalent for moderate traffic, while larger deployments might benefit from multiple GPUs or TPU pods. For non-real-time applications, such as batch processing of large datasets, high-performance computing clusters equipped with sufficient CPU and GPU resources are necessary to handle parallel processing of data batches, reducing overall processing time. The BERT model, especially when fine-tuned for specific tasks, requires substantial memory, with a minimum of 16 GB of RAM recommended to handle the model and the input data without memory bottlenecks. Larger models or more intensive tasks may require up to 32 GB or more.

5 Challenges and Potential Pitfalls of Cross-Lingual Pretraining

One of the primary challenges in applying English-pretrained models to other languages is the significant linguistic differences between languages. English and Arabic, for instance, differ in syntax, morphology, and script. Arabic is a Semitic language with a rich morphology and root-based word formation, which can pose challenges for models initially trained in English. Pretrained models like BERT use tokenizers optimized for the language they were trained on. English tokenizers may not effectively capture the nuances of Arabic, leading to suboptimal token representations. This mismatch can result in poor handling of Arabic words, especially those with complex morphological patterns. Arabic has numerous dialects, each with distinct vocabulary and grammatical rules. An English-pretrained model fine-tuned on Modern Standard Arabic may struggle with dialectal variations common in social media and online communications. This adds complexity to the model's adaptation process. Language models also capture cultural nuances embedded in the text they are trained on. English-pretrained models may not fully grasp the cultural and contextual references unique to Arabic-speaking communities, impacting the model's ability to accurately classify context-specific content. High-quality, annotated datasets in languages other than English are often scarce. Fine-tuning an English-pretrained model on limited Arabic data can result in overfitting and reduced generalizability. Ensuring diverse and representative datasets for fine-tuning is crucial to overcoming this challenge.

6 Conclusions

To conclude, the present study has successfully demonstrated the efficacy of leveraging pretrained BERT models in the task of detecting cyberbullying in Arabic text. The comparative analysis, as presented in the results, highlights a significant improvement in the detection performance when using our enhanced BERT model. The proposed approach outperforms the baseline BERT model across all the evaluated metrics, including accuracy, precision, recall, and F1 score. Specifically, our model achieved a remarkable accuracy of 98.45%, precision of 99.17%, recall of 99.10%, and an F1 score of 99.14%, which indicates a notable advancement over the baseline metrics. These results not only underscore the effectiveness of integrating contextualized word embeddings with deep learning architectures but also pave the way for more robust and nuanced models in the realm of automated content moderation. One of the primary challenges in applying English-pretrained models to other languages is the significant linguistic differences between languages. English and Arabic, for instance, differ in syntax, morphology, and script. Arabic is a Semitic language with a rich morphology and root-based word formation, which can pose challenges for models initially trained in English, a Germanic language with relatively simpler morphological structures. Future research may build upon these findings to explore the generalizability of the model across different dialects and forms of Arabic, potentially leading to a more inclusive and comprehensive solution for combating cyberbullying in the Arab-speaking online community. For future research direction, we aim to explore the use of meta-heuristic algorithms for feature selection to enhance the identification of nuanced cyberbullying indicators within large, unstructured datasets. Also, we aim to integrate user feedback loops into the model's deployment environment. Moreover, we plan to enhance transparency by developing an explainable AI component that can provide users with insights into the model's decision-making process.

Acknowledgement: This research has been funded by Scientific Research Deanship at University of Ha'il-Saudi Arabia through Project Number RG-23 092.

Funding Statement: This research has been funded by Scientific Research Deanship at University of Ha'il-Saudi Arabia through Project Number RG-23 092.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Mohamed A. Mahdi, Sawsan A. Saad and Suliman Mohamed Fati; Data curation, Shahanawaj Ahamad and Sawsan A. Saad; Methodology, Mohamed A. Mahdi and Suliman Mohamed Fati; Project administration, Mohamed A. Mahdi and Suliman Mohamed Fati; Resources, Sawsan A. Saad; Software, Mohamed A. Mahdi; Validation, Mohamed A. Mahdi; Visualization, Mohamed A.G. Hazber; Writing—original draft, Mohamed A. Mahdi, Shahanawaj Ahamad and Sawsan A. Saad; Writing—review & editing, Mohamed A.G. Hazber and Suliman Mohamed Fati. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data can be accessed at: https://www.kaggle.com/datasets/yasserhessein/arabic-cyberbullying (accessed on 11 May 2023).

Ethics Approval: Not applicable.

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

References

1. Ahmed MT, Urmi AS, Rahman M, Islam AZ, Das D, Rashed MG. Cyberbullying detection based on hybrid ensemble method using deep learning technique in bangla dataset. Int J Adv Comput Sci Appl. 2023;14(9):545–51. doi:10.14569/issn.2156-5570.

2. Lange K. Improving the fairness of cyberbullying detection for sexism on social media while keeping predictive power. Tilburg University: Netherlands, 2020.
3. Muneer A, Alwadain A, Ragab MG, Alqushaibi A. Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. Information. 2023 Aug 18;14(8):467. doi:10.3390/info14080467.
4. Alzaqebah M, Jaradat GM, Nassan D, Alnasser R, Alsmadi MK, Almarashdeh I, et al. Cyberbullying detection framework for short and imbalanced Arabic datasets. J King Saud Univ-Comput Inf Sci. 2023 Sep 1;35(8):101652. doi:10.1016/j.jksuci.2023.101652.
5. Alsunaidi N, Aljbali S, Yasin Y, Aljamaan H. Arabic cyberbullying detection using machine learning: state of the art survey. In: Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering, 2023 Jun 14; p. 499–504.
6. Husain F, Uzuner O. A survey of offensive language detection for the Arabic language. ACM Transact Asian and Low-Res Lang Inform Process (TALLIP). 2021 Mar 9;20(1):1–44. doi:10.1145/3421504.
7. Muneer A, Fati SM. A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet. 2020 Oct 29;12(11):187. doi:10.3390/fi12110187.
8. Al-Saif HF, Al-Dossari HZ. Exploring the role of emotions in Arabic rumor detection in social media. Appl Sci. 2023 Jul 30;13(15):8815.
9. Barlett CP, Scott JE. Racism behind the screen: examining the mediating and moderating relationships between anonymity, online disinhibition, and cyber-racism. J Pers Soc Psychol. 2023 Sep 21;6:1332–50.
10. Machova K, Srba I, Sarnovský M, Paralič J, Kresnakova VM, Hrckova A, et al. Addressing false information and abusive language in digital space using intelligent approaches. In: Towards digital intelligence society: a knowledge-based approach. Switzerland: Springer International Publishing, 2021. p. 3–32.
11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:3–32.
12. AL Nuaimi A. Effectiveness of cyberbullying prevention strategies in the UAE. In: ICT Analysis and Applications: Proceedings of ICT4SD 2020, Springer Singapore, 2021; vol. 2, p. 731–9.
13. Al-Ibrahim RM, Ali MZ, Najadat HM. Detection of hateful social media content for Arabic language. ACM Trans Asian Low Resour Lang Inf Process. 2023 Sep 22;22(9):1–26.
14. Talpur BA, O'Sullivan D. Cyberbullying severity detection: a machine learning approach. PLoS One. 2020 Oct 27;15(10):e0240924.
15. Kim S, Razi A, Stringhini G, Wisniewski PJ, De Choudhury M. A human-centered systematic literature review of cyberbullying detection algorithms. In: Proceedings of the ACM on Human-Computer Interaction, 2021 Oct 18; p. 1–34.
16. Alduailaj AM, Belghith A. Detecting Arabic cyberbullying tweets using machine learning. Mach Learn Knowl Extract. 2023 Jan 5;5(1):29–42.
17. Khairy M, Mahmoud TM, Omar A, Abd El-Hafeez T. Comparative performance of ensemble machine learning for Arabic cyberbullying and offensive language detection. Lang Resour Eval. 2023 Aug;13:1–8.
18. Mubarak H, Rashed A, Darwish K, Samih Y, Abdelali A. Arabic offensive language on twitter: analysis and experiments. arXiv preprint arXiv:2004.02192. 2020 Apr 5.
19. Chowdhury SA, Mubarak H, Abdelali A, Jung SG, Jansen BJ, Salminen J. A multi-platform Arabic news comment dataset for offensive language detection. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020 May, Marseille, France, European Language Resource Association; p. 6203–12.
20. Alakrot A, Murray L, Nikolov NS. Dataset construction for the detection of anti-social behaviour in online communication in Arabic. Procedia Comput Sci. 2018 Jan 1;142:174–81.
21. Haddad B, Orabe Z, Al-Abood A, Ghneim N. Arabic offensive language detection with attention-based deep neural networks. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing

Tools, with a Shared Task on Offensive Language Detection. Language Resources and Evaluation Conference (LREC 2020), 2020 May; p. 76–81.

22. Abozinadah EA, Mbaziira AV, Jones J. Detection of abusive accounts with Arabic tweets. Int J Knowl Eng-IACSIT. 2015 Sep;1(2):113–9. doi:10.7763/IJKE.2015.V1.19.
23. Abozinadah EA. Improved micro-blog classification for detecting abusive Arabic Twitter accounts. Int J Data Min & Know Manag Process. 2016;6(6):17–28. doi:10.5121/ijdkp.2016.6602.
24. Abozinadah EA, Jones Jr JH. A statistical learning approach to detect abusive twitter accounts. In: Proceedings of the International Conference on Compute and Data Analysis, 2017 May 19; p. 6–13.
25. Fati SM, Muneer A, Alwadain A, Balogun AO. Cyberbullying detection on twitter using deep learning-based attention mechanisms and continuous bag of words feature extraction. Mathematics. 2023 Aug 17;11(16):3567. doi:10.3390/math11163567.
26. Almutiry S, Abdel Fattah M. Arabic cyberbullying detection using arabic sentiment analysis. Egyptian J Lang Eng. 2021 Apr 1;8(1):39–50. doi:10.21608/ejle.2021.50240.1017.
27. Abdul-Mageed M, Elmadany A, Nagoudi EM. ARBERT & MARBERT: deep bidirectional transformers for Arabic. arXiv preprint arXiv:2101.01785. 2020 Dec 27.
28. Song X, Salcianu A, Song Y, Dopson D, Zhou D. Fast wordpiece tokenization. arXiv preprint arXiv:2012.15524. 2020 Dec 31.
29. Qiang Y, Pan D, Li C, Li X, Jang R, Zhu D. AttCAT: explaining transformers via attentive class activation tokens. Adv Neural Inform Process Syst. 2022 Dec 6;35:5052–64.
30. Yang CF, Chen YC, Yang J, Dai X, Yuan L, Wang YC, et al. LACMA: language-aligning contrastive learning with meta-actions for embodied instruction following. arXiv preprint arXiv:2310.12344. 2023 Oct 18.
31. Sarracén GL, Rosso P. Offensive keyword extraction based on the attention mechanism of BERT and the eigenvector centrality using a graph representation. Pers Ubiquitous Comput. 2023 Feb;27(1):45–57. doi:10.1007/s00779-021-01605-5.
32. Myilvahanan K, Shashank B, Raj T, Attanti C, Sahay S. A study on deep learning based classification and identification of offensive memes. In: 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2023 Feb 2; Coimbatore, India, IEEE; p. 1552–6.
33. Zhuang Z, Liang Z, Rao Y, Xie H, Wang FL. Out-of-vocabulary word embedding learning based on reading comprehension mechanism. Natural Lang Process J. 2023 Dec 1;5(2):100038. doi:10.1016/j.nlp.2023.100038.
34. Alyafeai Z, Al-shaibani MS, Ghaleb M, Ahmad I. Evaluating various tokenizers for Arabic text classification. Neural Process Lett. 2023 Jun;55(3):2911–33. doi:10.1007/s11063-022-10990-8.
35. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
36. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014 Dec 22.
37. Parikh AP, Täckström O, Das D, Uszkoreit J. A decomposable attention model for natural language inference. arXiv preprint arXiv:1606.01933. 2014 Dec 22.
38. Seo M, Kembhavi A, Farhadi A, Hajishirzi H. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603. 2016 Nov 5.
39. Hadi MU, Qureshi R, Shah A, Irfan M, Zafar A, Shaikh MB, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. 2023 Dec 7. doi:10.36227/techrxiv.23589741.v4.
40. Ling CX, Huang J, Zhang H. AUC: a statistically consistent and more discriminating measure than accuracy. In: International Joint Conference on Artificial Intelligence (IJCAI), 2003 Aug; vol. 3, p. 519–24.