JOURNAL OF CYBERSECURITY

Journal of Cybersecurity, 2025, tyaf030
https://doi.org/10.1093/cybsec/tyaf030

Research paper

AraBERT-based approach for detecting Arabic cyberbullying in Facebook comments

Rania Ibrahim Hithnawi[1], Mohammad M.N. Hamarsheh[2,*], Mohammed Maree[3]

[1]Department of Engineering and Technology Sciences, Arab American University, Jenin, P.O. Box 240, Palestine
[2]Department of Computer Science and Networks Security, Arab American University, Jenin, P.O. Box 240, Palestine
[3]Department of Information Technology, Arab American University, Jenin, P.O. Box 240, Palestine

*Corresponding author. Department of Computer Science and Networks Security, Arab American University, Jenin, P.O. Box 240, Palestine. E-mail: mohammad.hamarsheh@aaup.edu

Received 5 March 2025; revised 28 August 2025; accepted 1 October 2025

Abstract

Cyberbullying is a significant issue on social media platforms. It poses serious emotional consequences and harassment to victims. Conventional pre-trained language models, such as Bidirectional Encoder Representations from Transformers (BERT), have achieved significant success in detecting cyberbullying through the analysis of natural language texts, especially with resource-rich languages such as English. However, for low-resource languages, such as Arabic, there has been limited attention given to the detection of cyberbullying. This research investigates the effectiveness of Arabic BERT (AraBERT), a pre-trained language model, for detecting Arabic cyberbullying comments. It also explores the trade-off between computational resources and model performance through various fine-tuning and freezing strategies. From an initial pool of >40 000 collected comments, we constructed a high-quality, balanced dataset of 20 000 Facebook comments written in Arabic. This subset was then manually labeled as either bullying or non-bullying to ensure data reliability and to facilitate robust model training. We employed fine-tuning techniques to adapt AraBERTv2 to the cyberbullying detection task. Through experimentation with layer freezing technique, we explored the trade-off between leveraging pre-trained knowledge and adapting the model to the specific task. Our findings demonstrate that fine-tuning all layers of AraBERTv2, which involves adjusting the weights and biases of each layer during training, achieved the highest performance. This approach offers a flexible method for applying a pre-trained model to new problems, resulting in an accuracy of 91.9% and an F1 score of 92.8%.

Keywords: pre-trained language models; AraBERT; cyberbullying detection; social media platforms; deep learning; neural network layers freezing; natural language processing

Introduction

The pervasive issue of bullying, traditionally confined to physical environments, has evolved significantly with the advent of digital technologies, giving rise to the complex phenomenon of cyberbullying. Bullying could happen in public spaces such as playgrounds, bus stops, or during school hours. However, advances in modern technology enable perpetrators to harass and intimidate their targets beyond the limitations of physical space by utilizing electronic devices such as computers and cell phones. This allows aggressors to relentlessly target their victims persistently, in a new form of abuse, known as cyberbullying [1,2]. Cyberbullying victimization is an escalating concern that has been steadily increasing over the years and affects both individuals and societies as depicted in Fig. 1. Accordingly, concerted efforts are required from all stakeholders, including parents, teenagers, school administrators, and other responsible individuals in positions of influence and responsibility. Often, the best approaches to reduce and even eliminate the widespread problem of cyberbullying in our culture are public education and intervention. Parents, children, and teenagers need to be aware of cyberbullying, its consequences, and the individuals prone to engaging in such behavior, so they can take steps to avoid becoming victims [3,4]. Numerous efforts have been undertaken for the prevention, detection, or mitigation of cyberbullying.

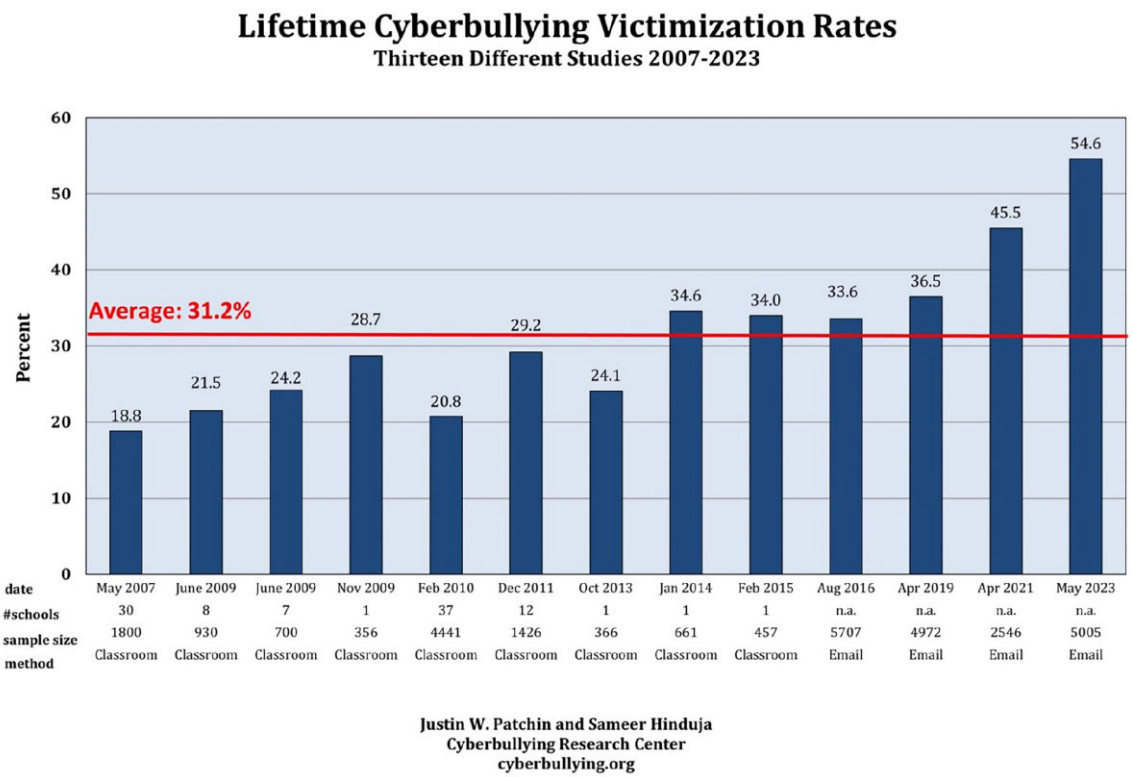

Figure 1. Lifetime cyberbullying victimization rates [55].

Cyberbullying is influenced by online disinhibition, a psychological concept where the lack of face-to-face interaction lowers the social inhibitions. Anonymity allows cyberbullies to act without fear of consequences, particularly in the Arab collectivist societies where shame is culturally significant [5]. Online toxicity is a result of linguistic, cultural norms, and the interaction between them. Arabic cyberbullying involves complex sociolinguistic codes, including the use of dialectal slurs that are mainly based on the country, but also varies based on tribal, regional, or sectarian identity. Gender abusive terms are used more frequently in Arabic cyberbullying, and they are more severe than the English ones based on the cultural norms [6].

The multifaceted nature of cyberbullying on social media, encompassing images, harmful comments, and videos, presents significant challenges for its management. This research focuses on detecting cyberbullying expressed in Arabic text on social media platforms. Many efforts for developing and implementing automated cyberbullying detection systems using Machine Learning (ML) and Deep Learning (DL) algorithms, which enable the identification or accurate classification of new relevant instances after training them with enough data. The increasing availability of large datasets has facilitated the development of more effective detection systems [7–9]. ML relies on identifying patterns from predefined features and thus requires careful feature selection. This dependency limits its ability to capture complex linguistic nuances and contextual understanding [10]. On the other hand, DL can automatically learn complex features from data, especially by using models such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). These models are highly effective in detecting subtle forms of cyberbullying. However, DL models often require large amounts of data to achieve high performance and may be computationally expensive to train [11].

Models have been trained for rich-resources languages such as English. However, for low-resource languages such as Arabic, Bengali, Hindi, or Urdu, efforts to address cyberbullying detection are limited [12,13], as shown in Fig. 2. Although >500 million in 21 countries speak Arabic in the world [14], there are unique challenges for the researcher in this field due to complexities in the morphological structure of the Arabic language, as well as the presence of different dialects [15]. In this context, Arabic Natural Language Processing (NLP) tasks such as Sentiment Analysis and Named Entity Recognition have proven particularly challenging [16,17]. With the advancement of transformer-based models in NLP, the pre-training of language-specific models, such as Bidirectional Encoder Representations from Transformers (BERT), has shown to be highly effective in achieving high performance on various NLP tasks [18,19]. By pre-training BERT specifically for the Arabic language, researchers aim to overcome the challenges faced by Arabic NLP and achieve similar success to what BERT has achieved for the English language. The newly developed model, Arabic BERT (AraBERT), has shown state-of-the-art performance on various Arabic NLP tasks compared to multilingual BERT models and other state-of-the-art approaches [20,21].

AraBERT, a model based on the BERT architecture, has become a widely adopted model for a variety of NLP tasks. It is an Arabic pre-trained model derived from the Google BERT architecture. AraBERT exists in various versions (e.g. v0.1, v1, v0.2, v2, Base, and Large), differing in training data size and vocabulary, thereby offering options to balance performance and efficiency [21].

This research contributes to the field of cyberbullying detection through two primary ways: (1) the creation of a large, balanced dataset of Arabic Facebook comments, labeled as either bullying or non-bullying content; and (2) the implementation and evaluation of AraBERT-based models for Arabic cyberbullying detection, specifically investigating the effectiveness of fine-tuning strategies on the newly constructed dataset to improve the performance of the model.

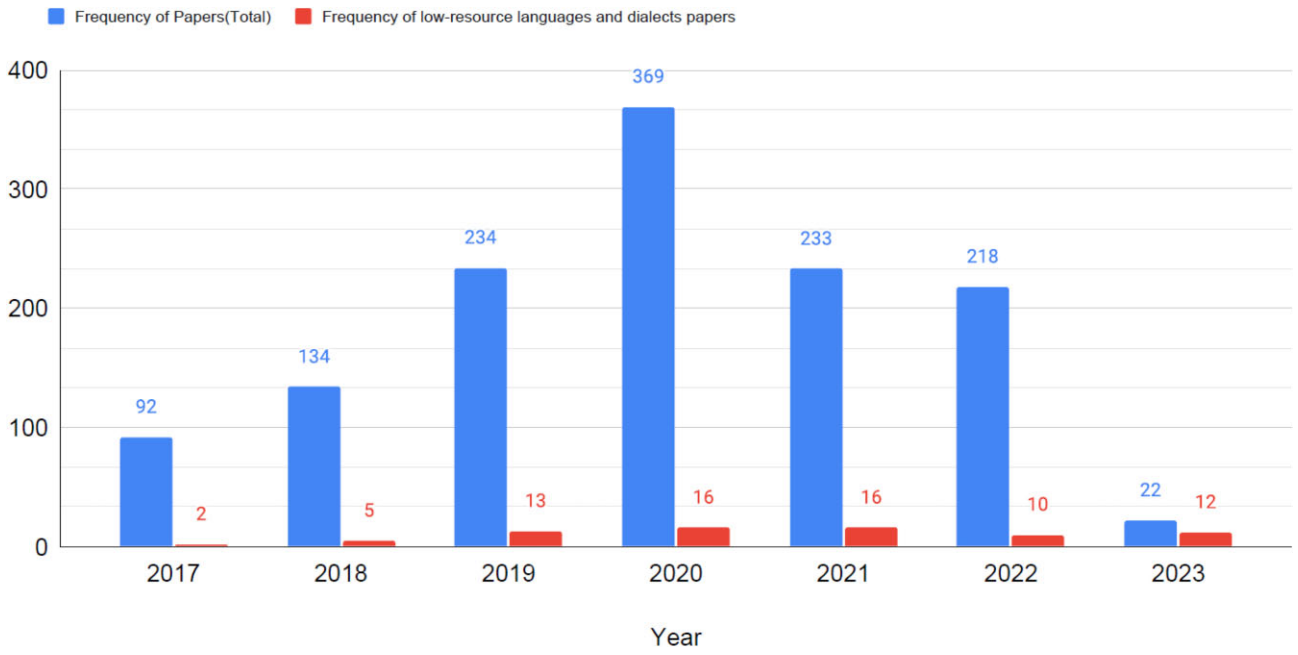

Figure 2. Cyberbullying detection for low resources languages datasets in years 2017–2023 [13].

Literature review

Theoretical underpinnings of cyberbullying in the arabic context

Detecting cyberbullying is not merely a technical challenge of text classification; it requires an understanding of the complex psychological and sociolinguistic factors that define online aggression. A key concept is the online disinhibition effect, where anonymity and asynchronicity lower users' inhibitions, leading to behaviors they would not exhibit face-to-face [5]. This effect is particularly pronounced on platforms such as Facebook, where the perceived distance can embolden aggressors. In the context of Arab collectivist societies, where social reputation and honor are paramount, this disinhibition can manifest in unique ways. Cyberbullying often employs culturally specific forms of aggression, such as attacks on family honor, the use of potent dialectal slurs, and gender-based insults that carry significant weight [6]. Therefore, a detection model must go beyond literal word meaning to grasp these nuances. Furthermore, cyberbullying is not monolithic. Aggression typologies distinguish between *direct aggression* (e.g. explicit insults, threats) and *indirect or relational aggression* (e.g. sarcasm, malicious rumors, social exclusion). The latter is often harder to detect automatically as it is highly context dependent. Our research, therefore, considers these different manifestations, especially the subtle forms prevalent in the informal, dialect-rich environment of Facebook comments.

Methodologies and paradigms for cyberbullying detection

The automated detection of cyberbullying has progressed through several distinct methodological paradigms, each building upon the shortcomings of its predecessor. Understanding this evolution provides a clear rationale for our choice of a fine-tuned, domain-specific transformer model. Early efforts predominantly relied on keyword-based approaches, which use predefined dictionaries of offensive, hateful, or abusive terms to flag potentially harmful content [22]. The primary advantage of this method is its simplicity and computational efficiency. However, its effectiveness is severely limited by its inability to comprehend context. For instance, a keyword-based system cannot distinguish between a genuine insult and sarcasm, irony, or the use of reclaimed slurs within a community. This contextual blindness leads to a high rate of both false positives (benign content flagged as bullying) and false negatives (FNs; subtle or novel forms of bullying being missed). Furthermore, the dynamic and ever-evolving nature of online slang requires constant manual curation of these dictionaries, a task that is both labor-intensive and perpetually lagging linguistic trends [23,24].

To address the limitations of keyword-based approaches regarding contextual understanding, researchers turned to classical ML models, such as Naive Bayes, Logistic Regression, and Support Vector Machines [25–27]. Unlike static keyword lists, these supervised models learn to identify patterns from labeled data. Their primary limitation, however, stemmed from their dependence on manual feature engineering. Researchers must design and extract features such as Term Frequency-Inverse Document Frequency, N-grams (sequences of words), and sentiment scores [27]. While these features provide more information than simple keywords, they offer only a limited representation of language. For example, N-grams capture local word order but fail to model long-range dependencies or semantic relationships between non-adjacent words, ultimately limiting the model's ability to grasp complex sentence structures.

The advent of DL marked a significant advancement, enabling models to automatically learn hierarchical feature representations directly from the textual data [10,28]. Architectures such as Long Short-Term Memory (LSTM) networks and their bidirectional variants (BLSTM) were particularly influential [12,29]. By processing text sequentially, these models maintain a "memory" of prior context, thereby capturing a degree of word order and dependency. BLSTMs further improved on this by processing the text in both forward and backward directions, providing a more complete contextual picture for each word. However, the sequential nature of these models remains a constraint. For long or complex sentences, information from the beginning of a sequence can become "diluted" by the time the model reaches the end, a challenge known as handling

long-range dependencies. More recently, transformer-based models such as BERT have transformed the field of NLP by overcoming this limitation [18,30]. Instead of sequential processing, transformers employ a self-attention mechanism, which allows the model to weigh the importance of all words in a sentence simultaneously when processing any single word. This enables the direct modeling of relationships between any two words, regardless of their distance, providing a deep and robust contextual understanding [31]. For a morphologically rich and dialectally diverse language such as Arabic, a specialized model is crucial. The AraBERT model was pre-trained specifically on a massive Arabic corpus, giving it an intrinsic understanding of Arabic syntax, semantics, and morphology [21]. Unlike multilingual models, AraBERT and its dedicated tokenizer are designed to handle the complexities of Arabic, such as prefixes and suffixes, which is critical for accurate interpretation social media text. Studies have consistently shown AraBERT's superior performance on various Arabic NLP tasks, including sentiment analysis and hate speech detection [32–34]. This proven performance, combined with its inherent ability to process the dialectal and informal language prevalent in Facebook comments, logically motivates our choice to employ and fine-tune AraBERT. Our work builds on this foundation by creating a new, large-scale dataset for Arabic cyberbullying and systematically investigating fine-tuning strategies to adapt AraBERT's powerful pre-trained knowledge to this specific, challenging domain.

The AraBERT model

AraBERT, a pre-trained language model specifically designed for Arabic, demonstrates a deep understanding of the language, enabling it to capture subtle linguistic nuances often missed by generic models. Trained on an extensive Arabic corpus, AraBERT exhibits a comprehensive understanding of the language facilitating high-accuracy performance in language translation, classification of text, and sentiment analysis [21]. AraBERT can be fine-tuned for specific applications with minimal additional training data, thus reducing the time and resources required to develop and deploy NLP models. AraBERT can be readily integrated into existing NLP pipelines, making it suitable for a wide range of domains. The capability of fine-tuning pre-trained models such as AraBERT facilitates transfer learning, where knowledge from one task can be applied to another related task. This approach reduces the required training data and simultaneously enhances the performance [20,21,35].

The authors in [32] used AraBERT to examine Arabic tweet analysis for forecasting customer sentiment and feedback for Saudi Arabian telecommunication companies. The results indicated that AraBERT accurately predicts customer sentiment and outperforms CNN and RNN, particularly achieving the highest accuracy on the Mobily Saudi Telecom Company datasets. The methodology presented by the authors in [33] also leverages the AraBERT language model. A pre-processed text from the ARev dataset, which contains >40 000 comments and reviews, is segmented using Farasa segmentation, then the AraBERT model is implemented with well-tuned parameters. They achieved an accuracy of 92.5%, which represents a competitive outcome. Future efforts focus on resolving the Arabic text segmentation issue and improving the Farasa segmentation version. On the other hand, the authors in [36] Emphasize the need for automatic detection of toxic contents, an Arabic Tunisian dataset is developed and a model based on AraBERT is proposed. The experimental results show that the AraBERT model performed well and achieved an F1 score of 0.99.

Other research focuses on the challenging task of multilingual offensive language detection by leveraging the power of transfer learning from transformer fine-tuning model [37], or the challenges of detecting sarcasm and sentiment in Arabic tweets [38,39], or the automated detection of hate speech and abusive content in Arabic tweets [40]. The systematic review article [41] identifies BERT models employed for Arabic text classification and compare their performance with each other and then assessed their effectiveness relative to the original English BERT models. To our knowledge, most of the research on cyberbullying detection using AraBERT has predominantly focused on Twitter data. Facebook comments pose distinct challenges due to their informal nature and diverse contexts. Fine-tuning AraBERT on Facebook comment data enhances its comprehension of the specific linguistic nuances and cyberbullying expressions prevalent in this context. This research investigates the effectiveness of AraBERT for detecting cyberbullying in Facebook comments.

Datasets and proposed methodology

The initial phase of this framework is the collection of an Arabic dataset to facilitate the development and implementation of an AraBERT—based model for detecting Arabic cyberbullying comments. The majority of the publicly available cyberbullying datasets are in English or comprise data from Twitter or YouTube platforms. The inherent complexities of Arabic dialects in the context of user-generated online content pose significant challenges. The diversity of Arabic dialects, coupled with their widespread use in online platforms, can lead to data sparsity and hinder the performance of NLP models [42,43]. Therefore, we constructed a dataset comprising Arabic Facebook comments, a subset of which was manually annotated. Then the dataset goes under a few steps of preprocessing before the AraBERT is applied and fine-tuned for classification as summarized in Fig. 3.

Dataset collection

The dataset was collected using the Apify tool, which supports Python and JavaScript libraries. Apify facilitates data extraction from Facebook pages via URLs and enable the download of extracted data in various formats such as JSON, CSV, and Excel files [44]. In addition, a custom Facebook Pages Scraper was developed to collect comments from specific Facebook Page URLs. Figure 4 shows the basic steps in using the Facebook Pages Scraper to extract the data. The maximum size allowed of each extraction is 5000 rows. The key features extracted from Facebook comments on a post are date, Facebook ID, Facebook URL, feedback ID, comment ID, likes count, post title, profile ID, profile name, and profile picture.

We collected the Facebook comments from addresses of Arabic pages that have millions of followers such as Aljazeera channel, Roya kitchen, MTV Lebanon, Ramallah News, and other public pages in different subjects and times. We also chose posts that have >1000 bullying and non-bullying comments to build the dataset with 40 000 comments.

Dataset annotation process

While our initial data collection from various Facebook pages yielded a substantial corpus of 40 000 comments, we selected a subset of 20 000 comments for the rigorous manual annotation process. This subset selection was primarily driven by two critical factors: (1) The need for a balanced dataset is essential for training an unbiased

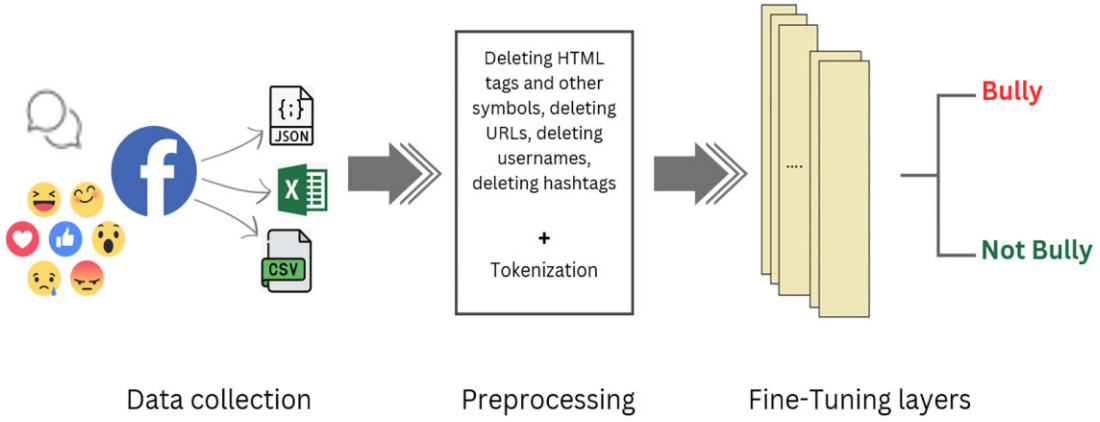

Figure 3. Proposed methodology.

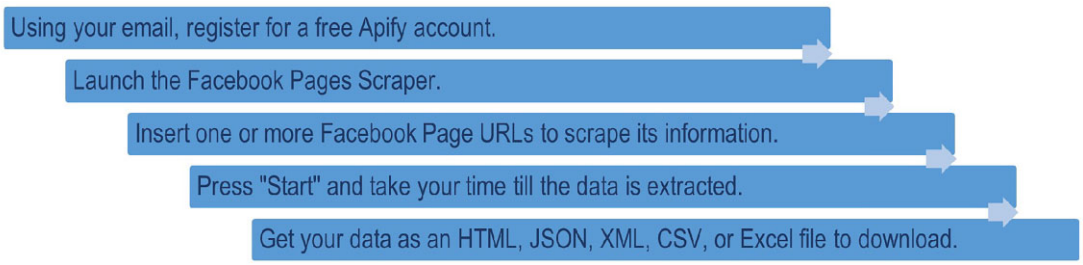

Figure 4. Using APIFY tool to scrape Facebook data.

classifier: Raw social media data often exhibits skewness, and curating the dataset enabled the achievement of an approximate 50/50 split between bullying and non-bullying records. (2) The resource-intensive nature of high-quality annotation: the 20 000-comment dataset was manually labeled in collaboration with psychological specialists to ensure accuracy. This rigorous process was vital for quality but infeasible to apply to the entire initial collection. Annotators were provided with a clear definition of cyberbullying, adapted from existing literature, encompassing direct insults, harassment, threats, and targeted derogatory remarks. To ensure our annotations captured the multifaceted nature of online aggression, our guidelines, developed in collaboration with psychological specialists, were informed by established aggression typologies. We instructed annotators to label not only direct aggression such as insults and threats but also forms of indirect aggression. For instance, sarcasm used to attack or belittle an individual was labeled "Bullying," as it represents a common form of relational aggression that is particularly challenging for automated systems. In contrast, general profanity used for emphasis without a clear target was labeled "Non-Bullying" to distinguish targeted harassment from expressive language. To ensure high-quality labels, the final label for each comment was determined by a majority vote. 52% of the dataset is bullying and as shown in Fig. 5 and a sample of the dataset is shown in Table 1. While this rigorous, manual annotation process was critical for creating a high-fidelity dataset, we acknowledge that the 20 000-comment sample size may not fully encompass the vast linguistic diversity across all Arabic dialects or the highly informal nature of Facebook comments. This is a recognized limitation of the study, further discussed in the conclusion.

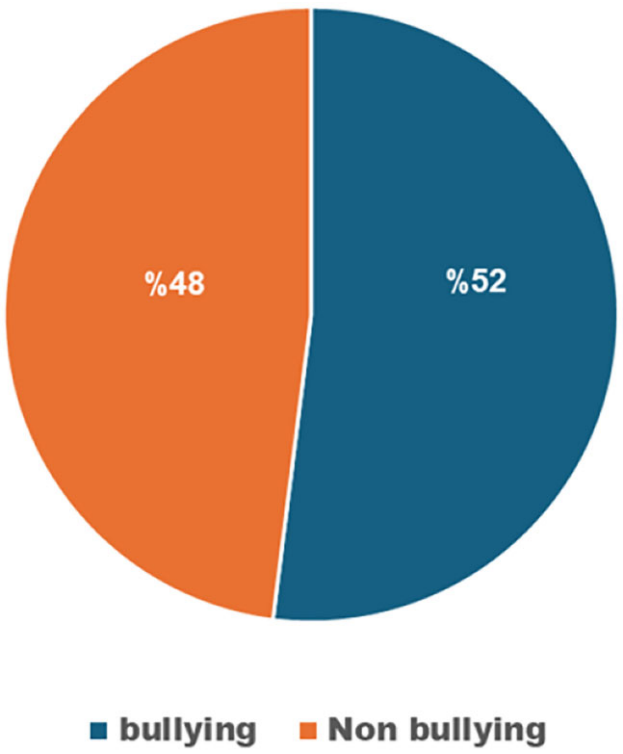

Figure 5. The percentage of bully and not bully comments in the dataset.

Table 1. A sample of our dataset.

Index	Text	Label
3143	الله لا يبارك لا فيك ولا فيها	1
19546	يلعن حرمكم شو انكم قرون	1
6556	تسلم الأيادي أبدعتي يا شيف	0
2448	الا لعنه الله عليك وعلى من هو على شاكلتك	1
7087	جمييييل	0

The initial step in the preprocessing stage involves eliminating noise and minimizing data sparsity through the removal of HTML tags, URLs, usernames, hashtags, mentions, and other extraneous symbols from the dataset. Usually, datasets are represented in a computer-understandable format, with tokens, vectors, and embeddings playing crucial roles. Sentence tokenization is performed, generating tokens that can be single words, bi-grams, or tri-grams. These tokens are subsequently converted into numerical vector representations, which are then used to train a DL model to identify contextual relationships and semantic meanings among them [45]. Pre-trained on an extensive corpus of Arabic text, AraBERT captures important semantic details. Compared to training a model from scratch, this approach significantly reduces the training time. Furthermore, its exclusive training on Arabic data is crucial for addressing dialectal variations and comprehending the language's nuances.

Our implementation leveraged pre-trained AraBERT model via the *HuggingFace* transformers library. The text preprocessing pipeline involved several key steps. First, we cleaned the raw comments by removing noise such as HTML tags, URLs, and usernames. Subsequently, sentence tokenization was performed using the *BertTokenizerFast* class, optimized for BERT-based models. This tokenizer converts Arabic text into numerical vector representations that capture the contextual relationships between tokens. This pre-trained embedding is crucial for our task, as it enables the model to leverage its extensive knowledge of the Arabic language and its dialects, reducing training time and enhancing performance compared to training a model from scratch [45]. The dataset was split into an 80/20 ratio for training and testing, respectively, and the *bert-base-arabertv2* model was initialized for training the AraBERT model. To evaluate our proposed model, we have considered the following four metrics:

- Precision: measures the proportion of correctly identified cyberbullying comments among all the comments the model predicted as cyberbullying.
- Recall: measures the proportion of actual cyberbullying comments that the model correctly identified.
- F1: the average of the harmonics of recall and precision gives the F1 score, which offers an accurate evaluation of the performance of the model.
- Accuracy: The total proportion of correctly classified comments, cyberbullying, and non-cyberbullying.

The *HuggingFace* trainer object encapsulates the default transformer fine-tuning approach, allowing customization by passing training arguments such as learning rate, number of epochs, and batch size. Logging steps were set to 20 to enable frequent evaluation of model performance on the validation set throughout training. Two key components, *TrainingArguments* and *Trainer*, were imported From the *HuggingFace* Transformers library. These components are essential for training and evaluating ML models, especially those associated with NLP tasks. The *TrainingArguments* class is needed to define and configure training hyperparameters for the model including learning rate, number of epochs, batch size and optimizer. The *Trainer* class serves as a high-level wrapper for executing various DL operations, encompassing both model training and evaluation.

The hyperparameters are set for training AraBERT model. The learning rate, which governs the update of model weights during the training process, was set to 2e-5, a common starting point for fine-tuning. The number of epochs, which defines the number of times the training dataset passes through the model in the training process, was set to 3. The batch size is set to 16, which indicates the number of examples processed together in the training step. FP16 was disabled to prevent mixed precision training, and the GPU was utilized to accelerate training without compromising accuracy. Finally, the evaluation strategy was set to "steps," meaning model evaluation occur after a fixed number of training steps without relying on epochs.

Experiments and results

In this section, we explain the experimental process and present the results of manual annotation for detecting Arabic cyberbullying using AraBERTv2 model. The investigation focuses on the effectiveness of applying freezing and fine-tuning to different number of layers in the model. The performance of the proposed model is evaluated using several metrics, including F-measure, accuracy, recall, and precision. A 20 000-comment balanced dataset of Arabic cyberbullying Facebook comments was utilized, 10 246 of them labeled as (1) for annotating bullying comments.

BERT layers progressively improve the understanding of the relationships between words, thereby extracting the primary context of the text. Thus, BERT and its variations are effective tools for a range of NLP applications [46]. According to [47], there is a decrease in the understanding of linear word order in layer 4 of BERT-base. Several research with different tasks agreed that syntactic information is predominantly captured in the middle layers of BERT [48], while the final layers are tasks specific [49]. Figure 6 indicates how the different layers of BERT transfer the learning knowledge for various tasks of NLP, where the columns representing distinct probing tasks.

Despite their well-defined and complex structure, BERT models have been a subject for several research studies examining how BERT models can be used to improve model performance and whether fine-tuning all pre-trained layers is required for efficient NLP tasks [46]. Fine-tuning more layers in the model facilitates the learning of a combined representation of information from both deep and output layers of the BERT model, thereby potentially capturing aspects missed by focusing only on a single layer. Furthermore, using information from all BERT layers with assigned weights enables the model to prioritize specific layers based on their task [50]. To evaluate our model, we assessed different techniques involving the fine-tuning of different numbers of layers in AraBERTv2 model and freezing the remaining layers. Layer freezing, a common technique in pre-trained models such as AraBERT, reduces the number of training parameters and consequently decreases training time [51]. In our experiments,

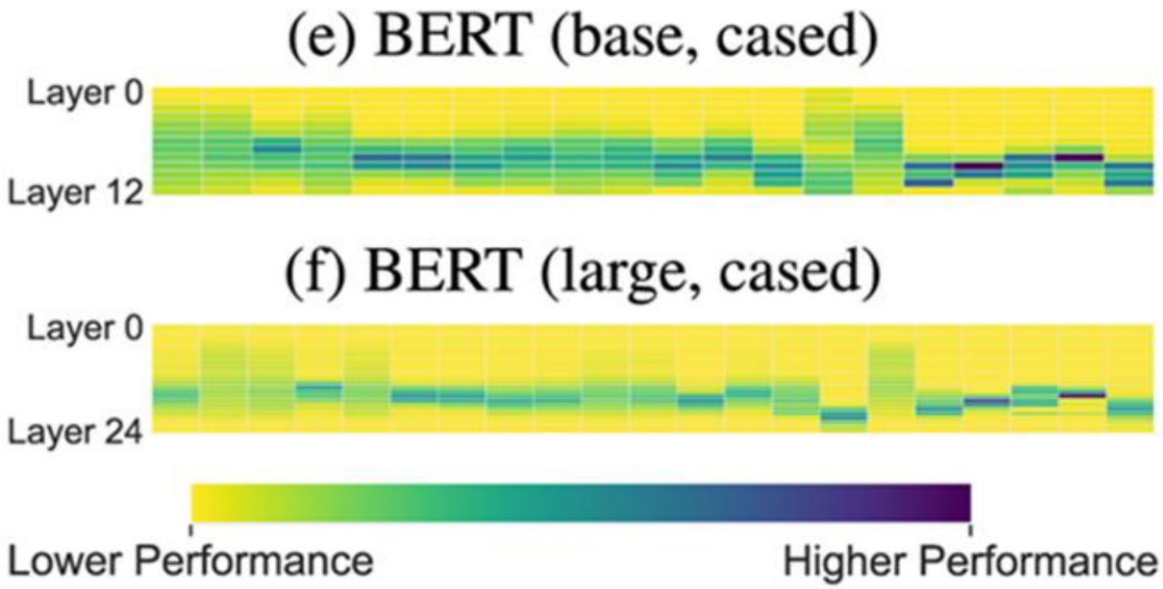

Figure 6. BERT layers transferability by task [50].

Table 2. Computational load and training time used in the experiments.

Strategy	Training time (minute)	System RAM (12.7 GB)	Gpu ram (15 gb)	System hard disk (78.2 GB)
Freeze all layer but final layer	24.5	3.6	7.5	35.2
Freeze 9 layers	25.54	3.6	7.9	35.7
Freeze 7 layers	26.42	3.3	8.2	36.3
Freeze 5 layers	27.40	3.6	8.5	36.7
Unfreeze layers	30.22	3.6	9.3	38.0

Table 3. Performance of freezing strategies in the evaluation process.

Strategy	Accuracy	F1	Precision	Recall
Freeze all layer but final layer	81.7	83.5	86.7	80.5
Freeze 9 layers	83.0	84.8	87.0	82.8
Freeze 7 layers	85.3	86.6	90.8	82.8
Freeze 5 layers	87.7	88.7	93.8	84.5
Unfreeze layers	91.9	92.8	94.7	90.9

the same training parameter settings were applied in all runs. In Table 2, we compare the freezing strategies used in the model according to computational load and training time. It is observed that tuning more layers requires using additional system resources, specifically GPU RAM, and Hard Disk space. Unfreezing layers increases computational demand, therefore enlarging the training dataset may lead to resource overload. System RAM usage remained nearly constant across all runs due to the same batch size utilized.

During the fine-tuning of AraBERT for Arabic cyberbullying detection, the model requires freezing or unfreezing of layers. Performance was evaluated using accuracy as a key metric. As shown in Table 3, unfreezing all layers resulted in the best testing accuracy of 91.9%. This significant enhancement implies that unfreezing layers enables the model to more accurately capture the complexity of Arabic cyberbullying. Figure 7 shows the evaluation matrix of our model results.

It is important to point out that our experiments were designed not merely to identify the optimal configuration but also to quantify the practical trade-offs between model performance and computational cost. The results of our layer-freezing strategies are presented in Table 2 (Computational Load) and Table 3 (Performance Metrics). Consistent with observations in large pre-trained models, unfreezing all 12 layers yielded the highest predictive performance, achieving an accuracy of 91.9% and an F1-score of 92.8%. This configuration allows the model to fully adapt its parameters to the specific nuances of our Arabic cyberbullying dataset. However, this peak performance incurs the highest computational cost, requiring 30.22 min of training on our hardware and consuming 9.3 GB of GPU RAM. More insightful findings emerge from the analysis of the intermediate strategies. For instance:

- High-performance, lower-cost alternative: Freezing the first 5 layers (fine-tuning the subsequent 7) yielded an an accuracy of

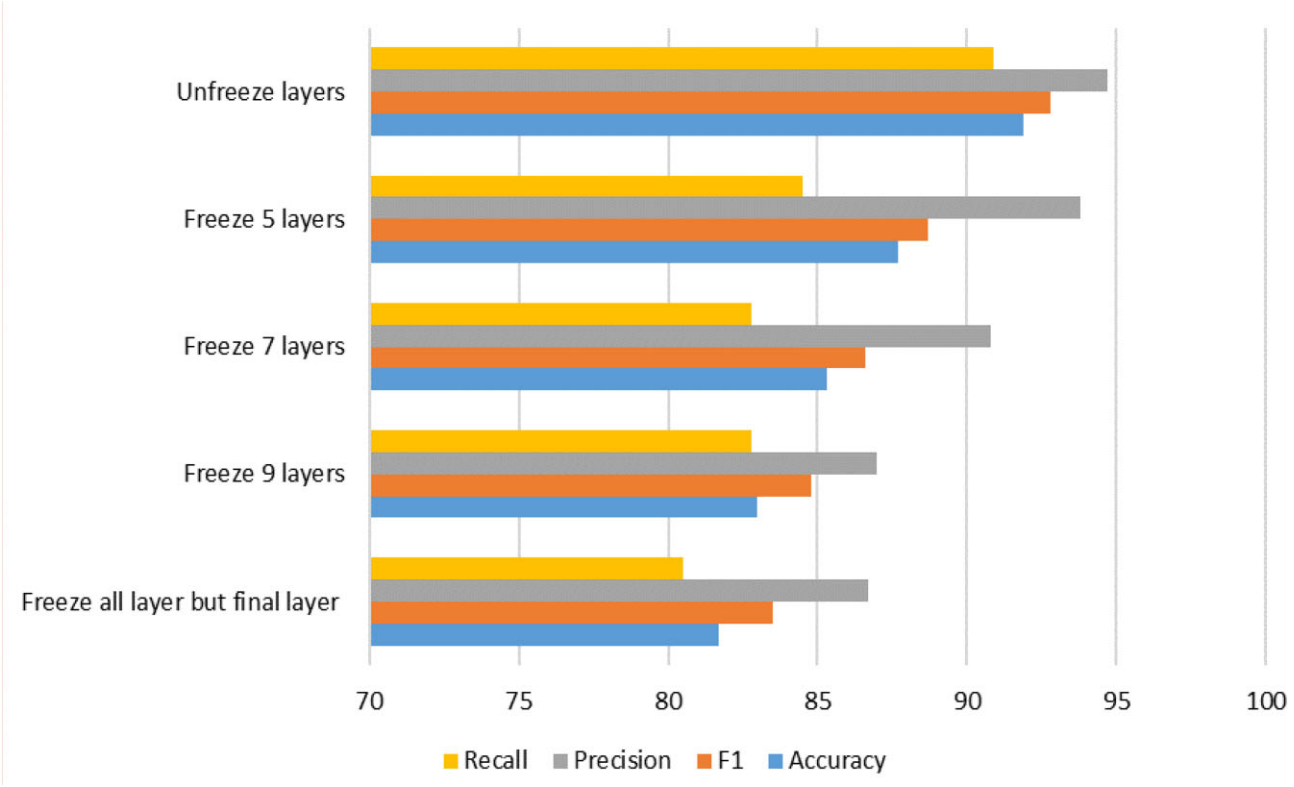

Figure 7. The evaluation matrix of our model results.

87.7%. This represents a performance reduction of only 4.2 percentage points from the maximum, yet it decreased training time by nearly 10% to 27.40 min and reduced GPU RAM usage. This strategy offers a compelling balance for scenarios prioritizing high accuracy under moderate resource constraints.

- Resource-efficient baseline: Freezing all layers except the final one represent the most computationally efficient approach, requiring only 24.5 min for training. Although the accuracy of 81.7% is the lowest among the tested configurations, it establishes a robust baseline and may be the only feasible option for environments with very limited GPU memory or a requirement for extremely fast training iterations.
- The strategy of freezing the first 7 layers (fine-tuning the subsequent 5) provides a notable performance improvement over freezing 9 layers (an increase from 83.0% to 85.3% accuracy) with only a marginal increase in training time.

Our results thus confirm that a greater number of tunable parameters generally enhances the performance. However, our quantitative analysis provides a more valuable, practical insight: the relationship between tunable parameters and performance is not linear. Practitioners can achieve >95% of the peak accuracy (87.7% vs. 91.9%) while conserving significant computational resources. This data-driven guidance empowers developers to make informed decisions based on their specific project constraints, balancing considerations such as maximum accuracy, training speed, and hardware costs.

Error analysis was conducted using confusion matrices and sample-level analysis to gain deeper insight into the model's behavior. A confusion matrix serves as an effective visualization tool for errors during the evaluation of a model's performance in classification tasks. It clearly demonstrates the number of correctly classified samples for each class [true positives (TP) and true negatives (TN)]. A confusion matrix was generated with true and predicted classes represented by rows and columns, respectively. Each cell in the matrix indicates the number of samples from the testing set corresponding to their true and predicted cyberbullying labels. For instance, a high number of FNs suggests that the model failed to identify a considerable number of actual cyberbullying comments [52,53]. Examination of the distribution of values (TP, TN, FP, FN) from the model experiments, as shown in Fig. 8, provides deeper insight into the model's performance.

While both FP and FN increased with freezing layers, FP shows a little bit more increase rate, reaching a difference of 48 between unfreezing and all freezing layers, while FN reaches only 40. This indicates that the model with unfreezing layers becomes highly effective at correctly identifying positive cases (cyberbullying) and minimizing false positives. Recall also improves, but at a slightly slower rate, suggesting that while the model gets better at finding actual cyberbullying instances, it is moving slower in reducing FNs (missing actual cyberbullying comments).

FP steadily increases from 21 (Unfreeze all layers) to 69 (Freezing 11 layers). This absolute improvement of 48 highlights the benefit of fine-tuning more layers. The largest jumps in FP occur when moving from freezing 9 layers to 7 layers (10) and from freezing 7 layers to 5 layers (15), suggesting that the unfrozen layers in these transitions contribute significantly to the model's discriminative power.

Error analysis revealed that unfreezing layers enhanced the model's ability to identify specific types of cyberbullying comments, such as sarcasm or indirect language as shown in Table 4. Thus, unfreezing layers is an appropriate approach for significant improvement in accuracy and focusing on efficiency. Notably, the experiments demonstrate that unfreezing layers enhances the model's ca-

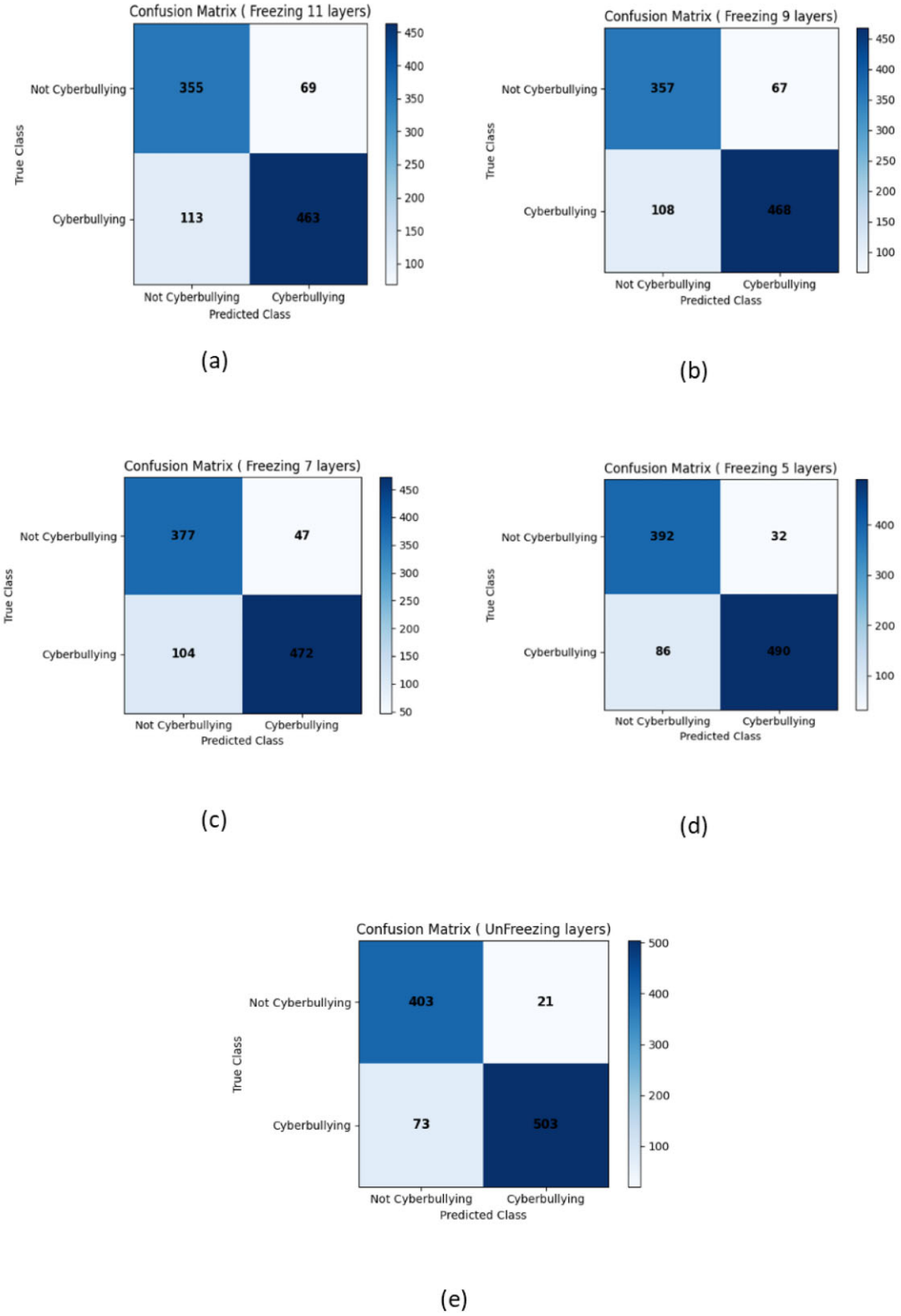

Figure 8. Confusion matrix for the model experiments.

Table 4. A sample of sarcastic and indirect cyberbullying comments.

Text	Comment
!!وصمت الاحمق نعمه	This is a sarcastic proverb implying it's better when a foolish person stays silent.
الاطفال يمكنهم تطوير تفاعلات ايجابية، استفيد من خبرتهم؟ !!	Sarcastic tone implies that the speaker believes the addressed person is more immature than a child.
انا بحاجة لإيجابية في حياتي، وانتِ عاتق ومعلق هاهاهاها	Using laughter to add a mocking sarcastic edge
انت مثل العث في حياتنا، نحن نحتاج الى التخلص منك ونرشك	A metaphor comparing someone to moths, clearly sarcastic and insulting
انا فعلا تعبت من تفاؤلك السلبي	Juxtaposition of "optimism" and "negativity" for sarcastic effect
الحيوانات ما بتحب تتعامل معاك	Indirect insult comparing someone to animals
احيانا تحسين إنك متخصصة في اشعال النقاشات السلبية	Indirect blame for starting conflict.

Table 5. Performance of benchmark models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN	80.6	81.1	81.6	81.4
LSTM	52.0	52.0	100.0	68.4
BLSTM	79.8	78.2	84.8	81.3
Our Model	91.9	94.7	90.9	92.8

pability to detect cyberbullying comments involving sarcasm or indirect language, whereas the frozen model often misclassified these more complex types of cyberbullying. This indicates that unfreezing layers enables the model to accurately capture the nuances of Arabic cyberbullying speech.

This finding aligns with cyberpsychology theories, as the model's enhanced adaptability reflects its improved ability to learn the subtle, context-dependent cues characteristic of indirect aggression. Such forms of aggression, amplified by the online disinhibition effect, are often missed by models trained only on general language patterns, but become detectable when fine-tuned on domain-specific data that reflects these complex social dynamics. This is consistent with the observed increase in testing accuracy, as the model learns domain-specific features of Arabic cyberbullying text that are not fully captured by the pre-trained AraBERT model.

Our investigation revealed that fine-tuning the full AraBERTv2 model yielded the highest accuracy for Arabic cyberbullying detection. Since AraBERTv2 is a pre-trained model, its first layers extract general linguistic features from a large Arabic corpus. These layers serve as a basis for higher layers by acting as a feature extractor in the model. However, AraBERTv2 is not initially trained for a specific task, consequently its higher layers are not optimized for cyberbullying detection. Fine-tuning each layer enables the model to learn features that differentiate cyberbullying text from non-cyberbullying text. This aligns with the research by [54], which indicates that tasks requiring the learning of complex patterns, such as text classification or sentiment analysis, can benefit by fine-tuning all layers of a pre-trained model. In our context, fine-tuning enables the model to learn complex subtle linguistic clues that extend beyond basic language features and are necessary for cyberbullying detection.

Regarding feature extraction and adaptation, the initial layers of the model capture fundamental language features, including word embeddings and syntactic structures, which are essential for comprehending the overall meaning of the text. Subsequent layers focus on higher-level features and inter-word relationships. Fine-tuning these layers enables the model to adapt more effectively to the specific task of cyberbullying detection. This facilitates the acquisition of key cyberbullying characteristics, such as specific word choices, sentiment, and sentence structure patterns.

To contextualize the performance of our proposed approach, a benchmark was established by evaluating our fine-tuned AraBERTv2 model against three widely used DL architectures, specifically, CNN, LSTM, and a BLSTM. All models were trained and evaluated on the same 80/20 training/testing split of our manually annotated dataset. The comparative performance is summarized in Table 5.

The results clearly demonstrate the superior performance of our fine-tuned AraBERTv2 model. Among the baseline models, the CNN and BLSTM achieved robust and comparable results, with the CNN exhibiting the highest performance with an F1-score of 81.4%. In contrast, the standard LSTM model failed to learn effectively, its 100% recall and low 52.0% precision indicates a trivial classification strategy of labeling nearly all comments as "bullying," thereby underscoring the inherent difficulty of the task for simpler sequential models.

Our proposed model achieved an F1-score of 92.8%, significantly outperforming the best baseline (CNN) by >11 percentage points. This substantial improvement underscores the advanced capabilities of a fine-tuned, domain-specific transformer model. The failure of the standard LSTM model, coupled with the comparatively weak performance of the CNN and BLSTM, highlights the limitations of sequential or localized feature-based models in capturing the long-range contextual dependencies characteristic of sociolinguistic phenomena such as sarcasm and culturally-specific insults. The transformer architecture's self-attention mechanism is better suited to model these complex relationships, which are critical for distinguishing sophisticated cyberbullying from benign text. This confirms that leveraging pre-trained knowledge of the Arabic language is crucial for understanding the contextual and linguistic nuances required for accurate cyberbullying detection, significantly exceeding the performance of traditional DL architectures trained from scratch.

Conclusion and future work

This research investigates the effectiveness of the AraBERTv2 pre-trained model in detecting Arabic cyberbullying detection. Various fine-tuning approaches were explored, involving the freezing of specific model layers while keeping others trainable. Our results indicate that the highest accuracy for Arabic cyberbullying detection is achieved when all layers of AraBERTv2 are fine-tuned as expected. However, while freezing certain layers appeared to limit the model's ability to learn the complex patterns required for optimal cyberbullying detection, this strategy provides the developers with a trad-off to balance the performance with resources available. This suggests that leveraging the model's full capacity helps in capturing the nuances of Arabic cyberbullying language but requires more resources. These findings contribute to the expanding body of research on DL applications in Arabic NLP, particularly within the domain of cyberbullying detection. Additionally, as part of this study, an Arabic dataset containing cyberbullying Facebook comments was constructed to enhance the availability of Arabic-language resources for future NLP applications.

Despite its contribution, this research is subject to certain limitations that present clear challenges for future work. A primary limitation is the size and scope of the dataset. While our 20 000-comment dataset is manually annotated, high-quality, and balanced, it may not be large enough to capture the full spectrum of linguistic variations inherent in the numerous Arabic dialects and the evolving, informal slang used in Facebook comments. Consequently, the generalizability of our model to dialects or contexts not well represented in our dataset may be limited. This trade-off between rigorous manual validation and comprehensive data coverage is a key constraint of this study. Future work should explore the impact of larger and more dialectally diverse datasets on the efficacy of AraBERTv2 and other models in Arabic cyberbullying detection. A potential next step could also be integrating the fine-tuned AraBERTv2 model into practical applications for online content monitoring or educational platforms. Practical implementation would need evaluating the model's performance in such settings and addressing potential challenges such as scalability and simplicity. The potential impact of this research is its contribution to developing applications using DL and AraBERT for text analysis in Arabic language. In addition, the results can produce automated tools that can effectively help in addressing and detecting Arabic cyberbullying comments on Facebook posts and other online platforms, that will constitute a safer online environment for Arabic speaking communities

Author contributions

Rania Ibrahim Hithnawi (Conceptualization [equal], Data curation [equal], Formal Analysis [equal], Investigation [equal], Methodology [equal], Resources [equal], Software [equal], Visualization [equal], Writing – original draft [lead]), Mohammad M. N. Hamarsheh (Conceptualization [equal], Data curation [equal], Formal Analysis [lead], Investigation [equal], Methodology [equal], Project administration [lead], Resources [equal], Software [equal], Supervision [lead], Validation [lead], Visualization [equal], Writing – original draft [equal], Writing – review & editing [lead]), Mohammed Maree (Data curation [supporting], Formal Analysis [equal], Investigation [supporting], Methodology [equal], Software [equal], Validation [equal], Writing – review & editing [equal])

Conflict of interest: None declared.

Funding

None declared.

References

1. Ray G, McDermott CD, Nicho M. Cyberbullying on social media: definitions, prevalence, and impact challenges. *J Cybersecur* 2024;**10**:26. https://doi.org/10.1093/cybsec/tyae026
2. Cowie H. Cyberbullying and its impact on young people's emotional health and well-being. *Psychiatrist* 2013;**37**:167–70. https://doi.org/10.1192/pb.bp.112.040840
3. Muneer A, Fati SM. A comparative analysis of machine learning techniques for cyberbullying detection on Twitter. *Future Internet* 2020;**12**:187. https://doi.org/10.3390/fi12110187
4. Zhu C, Huang S, Evans R. *et al.* Cyberbullying among adolescents and children: a comprehensive review of the global situation, risk factors, and preventive measures. *Front Public Health* 2021;**9**:634909. https://doi.org/10.3389/fpubh.2021.634909
5. Aljasir S. Effect of online civic intervention and online disinhibition on online hate speech among digital media users. *Online J Commun Media Technol* 2023;**13**:e202344. https://doi.org/10.30935/ojcmt/13478
6. Alshalabi N, Lahiani H, Yasin A. The role of culture in abusive language on social media: examining the use of english and arabic derogatory terms. *tpls* 2024;**14**:3057–66. https://doi.org/10.17507/tpls.1410.06
7. Raj M, Singh S, Solanki K. *et al.* An application to detect cyberbullying using machine learning and deep learning techniques. *Sn Comput Sci.* 2022;**3**:1–13. https://doi.org/10.1007/s42979-022-01308-5
8. Muneer AM, Alwadain A, Balogun AO. *et al.* Cyberbullying detection on Twitter using deep learning-based attention mechanisms and continuous bag of words feature extraction. *Mathematics* 2023;**11**:3567.
9. Balakrisnan V, Kaity M. Cyberbullying detection and machine learning: a systematic literature review. *Artif Intell Rev* 2023;**56**:1375–416. https://doi.org/10.1007/s10462-023-10553-w
10. Hasan MT, Hossain MAE, Mukta MSH. *et al.* A review on deep-learning-based cyberbullying detection. *Future Internet* 2023;**15**:179. https://doi.org/10.3390/fi15050179
11. Khafajeh H. Cyberbullying detection in social networks using deep learning. *IAJIT* 2024;**21**. https://doi.org/10.34028/iajit/21/6/9
12. Iwendi C, Srivastava G, Khan S. *et al.* Cyberbullying detection solutions based on deep learning architectures. *Multimedia Syst* 2023;**29**:1839–52. https://doi.org/10.1007/s00530-020-00701-5
13. Mahmud T, Ptaszynski M, Eronen J. *et al.* Cyberbullying detection for low-resource languages and dialects: review of the state of the art. *Inform Process Manage* 2023;**60**:103454. https://doi.org/10.1016/j.ipm.2023.103454
14. SIL International. *Ethnologue: Languages of the World*. 25th ed. SIL International, 2022.
15. Shaalan K, Siddiqui S, Alkhatib M. *et al.* Challenges in arabic natural language processing. *Comput Linguist Speech Image Process Arabic Lang* 2018:59–83. https://doi.org/10.1142/10693
16. El Gougi B, Ridouani M, Hassouni L. Arabic named Entity recognition: approaches, datasets, and comparative study. *Lecture Notes Netw Syst* 2024;**1048**:418–27.
17. Lakhfif A, Laskri MT. A frame-based approach for capturing semantics from Arabic text for text-to-sign language MT. *Int J Speech Technol* 2016;**19**:203–28. https://doi.org/10.1007/s10772-015-9290-8
18. Devlin J, Chang M-W, Lee K. *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North* 2019:4171–86. https://doi.org/10.18653/v1/N19-1423
19. Hao Y, Dong L, Wei F. *et al.* Visualizing and understanding the effectiveness of BERT. *EMNLP-IJCNLP 2019–2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* 2019:4143–52.
20. Faraj D, Faraj D, Abdullah M. SarcasmDet at Sarcasm Detection Task 2021 in Arabic using AraBERT pretrained model. 2021:345–50.
21. Antoun W, Baly F, Hajj H. AraBERT: transformer-based model for arabic language understanding. 2020.

22. Bashir E, Bouguessa M. Data mining for cyberbullying and harassment detection in arabic texts. *IJITCS* 2021;**13**:41–50. https://doi.org/10.5815/ijitcs.2021.05.04
23. Mouheb D, Albarghash R, Mowakeh MF. *et al.* Detection of Arabic cyberbullying on Social networks using machine Learning. *ACS/IEEE International Conference on Computer Systems and Applications* 2019;**2019**, https://doi.org/10.1109/AICCSA47632.2019.9035276
24. Kaarthika R, Hemamalini R, Sujithra Kanmani R. Enhancing cyberbullying detection through keyword filtering: a comparative study of ML and DL approaches. *2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication, IConSCEPT 2024–Proceedings* 2024, https://doi.org/10.1109/ICONSCEPT61884.2024.10627823
25. AlHarbi BY, AlHarbi MS, AlZahrani NJ. *et al.* Automatic cyber bullying detection in arabic social media. *Int J Engineer Res Technol* 2019;**12**:2330–5.
26. Alakrot A, Murray L, Nikolov NS. Dataset construction for the detection of anti-social behaviour in online communication in Arabic. *Procedia Comput Sci* 2018;**142**:174–81. https://doi.org/10.1016/j.procs.2018.10.473
27. Alakrot A, Murray L, Nikolov NS. Towards accurate detection of offensive language in online communication in Arabic. *Procedia Comput Sci* 2018;**142**:315–20. https://doi.org/10.1016/j.procs.2018.10.491
28. Aldhyani THH, Al-Adhaileh MH, Alsubari SN. Cyberbullying identification system based deep learning algorithms. *Electronics (Basel)* 2022;**11**:3273.
29. Alzaqebah M, Jaradat GM, Nassan D. *et al.* Cyberbullying detection framework for short and imbalanced arabic datasets. *J King Saud Univ Comput Inform Sci* 2023;**35**:101652. https://doi.org/10.1016/j.jksuci.2023.101652
30. Zhang A, Li B, Wang W. *et al.* MII: a novel text classification model combining deep active learning with BERT. *Comput Mater Continua* 2020;**63**:1499–514. https://doi.org/10.32604/cmc.2020.09962
31. Saini H, Mehra H, Rani R. *et al.* Enhancing cyberbullying detection: a comparative study of ensemble CNN–SVM and BERT models. *Soc Netw Anal Min* 2024;**14**:1–18. https://doi.org/10.1007/s13278-023-01158-w
32. Aftan S, Shah H. Using the AraBERT model for customer satisfaction classification of telecom sectors in Saudi Arabia. *Brain Sci* 2023;**13**:147. https://doi.org/10.3390/brainsci13010147
33. EL Moubtahij H, Abdelali H, Tazi EB. AraBERT transformer model for Arabic comments and reviews analysis. *IJ-AI* 2022;**11**:379. https://doi.org/10.11591/ijai.v11.i1.pp379-387
34. Salomon PO, Kechaou Z, Wali A. Arabic hate speech detection system based on AraBERT. *2022 IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*. IEEE, 2022,208–13. https://doi.org/10.1109/ICCICC57084.2022.10101577
35. Nada AMA, Alajrami E, Alsaqqa A. *et al.* Arabic text summarization using AraBERT model using extractive text summarization approach. *International Journal of Academic Information Systems Research (IJAISR)* 2020, https://doi.org/10.5281/zenodo.3978186
36. Salomon P, Kechaou Z, Wali A. Arabic hate speech detection system based on AraBERT. *IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, 2022, 208–13.
37. El-Alami F, Ouatik El Alaoui S, En Nahnahi N. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *J King Saud Univ Comput Inform Sci* 2022;**34**:6048–56. https://doi.org/10.1016/j.jksuci.2021.07.013
38. Wadhawan A. AraBERT and Farasa segmentation based approach for sarcasm and sentiment detection in arabic tweets. *WANLP 2021–6th Arabic Natural Language Processing Workshop, Proceedings of the Workshop* 2021:395–400.
39. Bashmal L, Alzeer DH. ArSarcasm shared task: an ensemble BERT model for SarcasmDetection in arabic tweets. 2021:323–8.
40. Aldjanabi W, Dahou A, Al-Qaness MAA. *et al.* Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. *Informatics* 2021;**8**:69. https://doi.org/10.3390/informatics8040069
41. Alammary AS. BERT models for Arabic text classification: a systematic review. *Appl Sci* 2022;**12**:5720. https://doi.org/10.3390/app12115720
42. Althobaiti MJ. Creation of annotated country-level dialectal Arabic resources: an unsupervised approach. *Nat Lang Eng* 2022;**28**:607–48. https://doi.org/10.1017/S135132492100019X
43. Althobaiti MJ. Automatic Arabic dialect identification Systems for written texts: a survey. 2020.
44. APIFY. Apify. 2024.
45. Metzger S. A beginner's guide to tokens, vectors, and embeddings in NLP. 2020.
46. Rogers A, Kovaleva O, Rumshisky A. A primer in BERTology: what we know about how BERT works. *Trans Assoc Comput Linguist* 2020;**8**:842–66. https://doi.org/10.1162/tacl_a_00349
47. Lin Y, Tan YC, Frank R. Open Sesame: getting inside BERT's linguistic Knowledge. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019,241–53. https://doi.org/10.18653/v1/W19-4825
48. Hewitt J, Manning CD. A structural probe for finding syntax in word representations. *Proceedings of the 2019 Conference of the North*. Vol **1**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, 4129–38. https://doi.org/10.18653/v1/N19-1419
49. Liu NF, Gardner M, Belinkov Y. *et al.* Linguistic knowledge and transferability of contextual representations. *Proceedings of the 2019 Conference of the North*. Vol **1**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, 1073–94. https://doi.org/10.18653/v1/N19-1112
50. You Y, Li J, Reddi S. *et al.* Large batch optimization for deep learning: training BERT in 76 minutes. *8th International Conference on Learning Representations, ICLR 2020* 2019.
51. Al-Ghamdi S, Al-Khalifa H, Al-Salman A. Fine-tuning BERT-based pretrained models for Arabic dependency parsing. *Appl Sci* 2023;**13**:4225. https://doi.org/10.3390/app13074225
52. Düntsch I, Gediga G. Confusion matrices and rough set data analysis. *J Phys Conf Ser* 2019;**1229**:012055, https://doi.org/10.1088/1742-6596/1229/1/012055
53. Bhandari A. Understanding & interpreting confusion matrix in machine learning. 2020.
54. Howard J, Ruder S. Universal language model fine-tuning for text classification. *ACL 2018–56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 2018;**1**:328–39.
55. Summary of our cyberbullying research (2004-2022).