RESEARCH ARTICLE OPEN ACCESS

Understanding the Impact of Dialect and Dataset Shift in Arabic Cyberbullying Detection Using Transformer Models

Eman Bashir

Department of Computer Science, Sudan University of Science and Technology (SUST)
Khartoum, Sudan

ABSTRACT

The widespread use of social media platforms has intensified concerns regarding cyberbullying, particularly among Arabic-speaking communities, where linguistic diversity, dialectal variation, and informal writing styles make automated cyberbullying detection highly challenging. This study offers an empirical comparison of deep learning and transformer-based models for Arabic cyberbullying detection. A prior study developed an LSTM model with Word2Vec embeddings, while this research fine-tunes and evaluates two pre-trained transformer architectures, AraBERT and MARBERT, on Arabic Twitter-based datasets. The most notable finding of this study is the substantial performance difference between the two datasets, from over 91% on ArbCyD to approximately 75–76% on the Arabic Twitter corpus, when each model is fine-tuned and evaluated separately on its own dataset. AraBERT achieves 91.9% accuracy on ArbCyD and 75% accuracy on the Twitter corpus; MARBERT achieves 93.2% and 76%, respectively, compared to 72% for the LSTM baseline on the Twitter corpus. Because each transformer model was fine-tuned on the dataset it was tested on, this gap reflects differences in dataset difficulty rather than a demonstrated failure to generalize and should not be read as a cross-dataset transfer result. The LSTM baseline was not re-implemented under identical conditions, so this comparison should be treated as indicative rather than fully controlled. Among all experiments, MARBERT shows marginally more stable performance, though this difference is small and should be treated as tentative.

***Keywords*:** cyberbullying detection; Arabic NLP; AraBERT; MARBERT; LSTM; transformer models; Arabic Twitter; dialect variation; multi-dataset comparison

I. INTRODUCTION

The internet has grown rapidly in recent years and provides many important benefits. However, its misuse, especially on social media platforms, has become a serious societal concern. These platforms are increasingly exploited for cyberbullying, which occurs across many languages and is used to threaten, intimidate, and humiliate individuals. Given its significant social and psychological consequences, automated detection of cyberbullying has become an important area of research within natural language processing (NLP).

Research on English cyberbullying detection has made considerable progress, while Arabic remains comparatively understudied. This is largely due to the linguistic complexity of Arabic, its rich morphological structure, and the coexistence of Modern Standard Arabic (MSA) alongside numerous regional dialects.

Most previous studies relied on traditional or deep learning approaches, such as Long Short-Term Memory (LSTM) networks, for cyberbullying detection. Although these models can perform well, they depend on fixed word representations and have limited capacity to model contextual meaning. More recently, transformer-based models such as AraBERT and MARBERT have substantially improved NLP performance by learning dynamic contextual representations. MARBERT is particularly well-suited for social media text, as it was pre-trained on large-scale Arabic dialectal Twitter data. However, there remains a limited understanding of how pre-training on different data domains affects model robustness across datasets.

Most studies that compare transformer models for Arabic cyberbullying detection evaluate them on a single dataset, which limits the ability to draw reliable conclusions about real-world performance [1–3] and may produce misleading results when models are applied to data from different sources or dialects. While cross-dataset evaluation has been explored in related Arabic NLP tasks such as sentiment analysis and dialect identification [4–5], this approach has rarely been applied to cyberbullying detection under strictly controlled conditions where all models share identical training settings [1]. This study contributes toward that gap by fine-tuning and evaluating AraBERT, MARBERT, and an LSTM baseline separately on two independent Arabic Twitter datasets under a unified experimental setup, providing a controlled comparison of performance across datasets rather than a direct cross-dataset transfer test. However, since the LSTM baseline was not re-implemented under identical conditions, comparisons involving the LSTM should be interpreted as indicative rather than fully controlled.

This study makes the following modest empirical contributions.

- A comparative evaluation of LSTM and transformer-based models, specifically AraBERT and MARBERT, for Arabic cyberbullying detection under consistent evaluation settings.
- A comparison of model performance across two independently collected Arabic cyberbullying datasets, showing how strongly reported accuracy depends on the dataset used and motivating genuine cross-dataset transfer evaluation as a direction for future work.
- An empirical investigation suggesting the benefit of dialect-aware pre-training in improving model robustness on Arabic social media text.

- A reproducible experimental pipeline using publicly available HuggingFace transformer models and openly accessible datasets.
- A qualitative error analysis identifies recurring failure patterns in model predictions, including sensitivity to sarcasm, code-switching, and short-text ambiguity.

This paper is organized into the following sections: Section II reviews related work, Section III describes the datasets and preprocessing steps, Section IV presents the methodology, Section V reports on the experimental results and evaluation, Section VI discusses the limitations, Section VII concludes the study, and Section VIII outlines potential future work.

II. RELATED WORK

Early approaches to cyberbullying detection relied primarily on classical machine learning methods using manually constructed feature representations, such as TF-IDF weighting and bag-of-words histograms [6]. Deep learning methods, particularly Long Short-Term Memory (LSTM) networks [7], improved upon these baselines by capturing sequential dependencies within text. When combined with distributed word representations such as Word2Vec embeddings [8], these models demonstrated stronger semantic understanding and better overall classification performance. Agrawal and Awekar [9] showed that bidirectional CNN-LSTM architectures consistently outperform unidirectional recurrent models across multiple English cyberbullying benchmarks. Despite these advances, most early neural approaches were developed exclusively for English data and relied on static word embeddings, leaving Arabic cyberbullying detection largely unaddressed. A systematic review by Rosa et al. [10] confirmed that feature-based and early neural methods offer only marginal real-world improvements, identifying key weaknesses including limited contextual understanding, shallow lexical representations, and poor generalizability across datasets.

Research on Arabic offensive language detection has evolved through several distinct phases. Among the earliest contributions, Mubarak et al. [5] developed a lexicon of offensive terms using Log-Odds Ratio scoring to identify abusive users on Arabic Twitter, producing one of the first large-scale annotated Arabic resources for this task. However, vocabulary-driven approaches are inherently constrained by their fixed word lists and cannot detect implicit, sarcastic, or contextually offensive language. Abozinadah et al. [11] took a complementary approach by combining content-level features, account metadata, and social graph signals, achieving 90% accuracy using a probabilistic classifier on a small, balanced dataset. The reliance on an artificially balanced and limited sample, however, reduces the credibility of these results in realistic deployment conditions where abusive content represents a minority class within a predominantly benign data stream.

These earlier approaches struggle to generalize beyond predefined lexical resources. Later work shifted toward deep learning methods to address these limitations. [12] introduced LSTM networks for Arabic cyberbullying detection, demonstrating clear improvements over traditional classifiers and establishing a strong baseline against which the present study is evaluated. However, LSTM-based models still face limitations in capturing long-range dependencies and handling the linguistic complexity of Arabic text, particularly due to dialectal variation and informal language usage on social media. To address some of these limitations, hybrid architectures combining convolutional and recurrent layers were proposed. Elzayady et al. [13] applied a CNN-LSTM model to Arabic opinion mining and demonstrated improvements in local feature extraction. Nevertheless, these models continued to rely on static word embeddings, which do not adequately capture contextual meaning or dialectal variation in Arabic social media text.

The diversity of Arabic dialects has emerged as a key factor in cyberbullying detection. Al-Hassan and Al-Dossari [14] demonstrated that datasets tailored to Egyptian and Levantine dialects significantly enhance detection effectiveness compared to models trained on Modern Standard Arabic. Their findings show that a mismatch between the dialects used in training and evaluation data leads to a clear decline in performance, which highlights the need for dialect-aware model selection.

In a related context, Aljohani and Yafooz [15] conducted a comparative study involving LSTM, Bi-LSTM, and BERT models using a dataset of 10,662 posts from X. They found that transformer-based models performed notably better than recurrent architectures. The SVM model also achieved high accuracy of 95.7% on lexically distinct data. However, their study relied on data from a single platform, which limits how well the findings can be applied to other domains or datasets.

A recent systematic review by Aljalaoud et al. [1], which covered 35 peer-reviewed Arabic cyberbullying studies, reported accuracy ranges of 73–96% across different platforms. The review also emphasized that macro-averaged F1 and per-class metrics are essential when evaluating imbalanced datasets, which directly supports the evaluation choices made in this study.

Several studies have confirmed the strong performance of transformer models in this domain. Mozafari et al. [16] showed that BERT-based models outperform LSTM models in hate speech detection tasks. Mahdi et al. [2] proposed an E-BERT model achieving 98.45% accuracy for Arabic cyberbullying detection on a single platform dataset without cross-dataset validation, making generalization assessment difficult. More recently, Daouadi et al. [4] systematically investigated pre-trained language models for hate speech detection across multiple Arabic Twitter datasets, finding that dialectal models consistently outperform MSA models in cross-domain settings. Hithnawi et al. [3] applied AraBERT to Arabic cyberbullying detection in Facebook comments, achieving 91.9% accuracy, though without cross-dataset evaluation.

What distinguishes this study from most previous work is that it evaluates all three models on two separate datasets using the same training conditions, which allows for a more controlled comparison. However, it remains an empirical

benchmarking study rather than a methodological advance. Table I shows how this study compares to the most closely related recent studies.

Inoue et al. [17] introduced CAMeLBERT, a family of Arabic pre-trained language models trained on a large corpus of 167GB covering Modern Standard Arabic, classical Arabic, and various dialectal varieties. Although this model was designed to provide broad linguistic coverage, it has not been extensively fine-tuned for cyberbullying detection on social media platforms, which limits its direct applicability in this domain. This gap motivates the present study, which builds upon and fine-tunes AraBERT [18] and MARBERT [19] using the HuggingFace Transformers library [20], specifically targeting Arabic cyberbullying detection across multiple datasets and providing a controlled multi-dataset comparison that remains largely absent from most prior work in this field.

TABLE I Comparison of the Present Study with Closely Related Recent Work

Study	Model	Dataset	Cross	Key Focus
Hithnawi et al. [3]	AraBERT	Facebook	No	Single-platform fine-tuning
Daouadi et al. [4]	Multiple PLMs	Multiple Twitter	Partial	Systematic PLM comparison
Aljohani & Yafooz [15]	LSTM, BERT	X (Twitter)	No	Model architecture comparison
Mahdi et al. [2]	E-BERT	Single dataset	No	Custom transformer
This study	LSTM, AraBERT, MARBERT	ArbCyD + Twitter	Partial	Multi-dataset performance comparison

III. DATASETS

This study uses two Arabic Twitter-based datasets for cyberbullying detection:

A. ArbCyD Dataset (Primary)

This study uses the ArbCyD (Arabic Cyberbullying Dataset), obtained from Kaggle, as the primary evaluation dataset. The dataset consists of Arabic social media posts collected from Twitter and labelled as either bullying or non-bullying. It covers a range of Arabic varieties, including Modern Standard Arabic as well as dialectal texts from the Gulf, Levantine, and Egyptian regions [21]. The presence of multiple dialects makes this dataset a challenging and representative benchmark for evaluating transformer-based models on real-world Arabic social media content. Table II summarizes the dataset statistics. The dataset exhibits a moderate class imbalance, with 38% bullying instances and 62% non-bullying instances across all splits. This imbalance reflects realistic social media distributions and directly motivates the use of macro F1-score as the primary evaluation metric, as it weighs both classes equally regardless of their frequency [10].

TABLE II ArbCyD Dataset Statistics

Split	Total Sample	Bullying	Not Bullying
Training (70%)	7,000	2,657 (38%)	4,343 (62%)
Validation (15%)	1,500	570 (38%)	930 (62%)
Test (15%)	1,500	569 (37.9%)	931 (62.1%)
Total	10,000	3,796 (38%)	6,204 (62%)

B. Arabic Twitter Corpus (Secondary Dataset)

The Arabic Twitter corpus used as a secondary dataset was originally introduced by [12]. This dataset comprises approximately 32,428 tweets collected between October 2018 and February 2019 using a real-time keyword streaming technique, with labels assigned through polarity scoring and manual verification. For compatibility with our binary classification pipeline, we converted the original multi-class polarity scores into a two-class scheme. Specifically, negative values were mapped to the cyberbullying class (1), while zero and positive values were mapped to the non-cyberbullying class (0).

This corpus serves as a secondary benchmark, enabling direct comparison with the LSTM baseline reported by [12], which achieved 72% accuracy using Word2Vec embeddings [8]. The corpus was divided into 22,700 training samples (70%), 4,864 validation samples (15%), and 4,864 test samples (15%), with an approximate class distribution of 45% cyberbullying and 55% non-cyberbullying after binarization.

C. Data Splitting and Preprocessing

Both datasets were divided using stratified sampling to preserve class distribution across splits, as described in Sections III-A and III-B.

The following preprocessing steps were applied before tokenization. First, Arabic script normalization was performed, including unifying alef variants (ا → آ ,إ ,أ), converting alef maqsura (ي → ى), and replacing ta marbuta (ه → ة). Second, URLs, user mentions, and hashtag symbols were removed while preserving the hashtag content itself. Third, emojis, non-Arabic characters, and elongated or repeated letters were removed. Finally, punctuation marks, diacritics, and extra whitespace were removed to reduce noise.

After preprocessing, all texts were tokenized using the Wordpieces tokenizer specific to each model. All sequences were then padded and truncated to a maximum length of 64 tokens. The preprocessing pipeline was developed using Python 3.10, along with the PyArabic library and HuggingFace tokenizers. The choice of 64 tokens as the maximum length was based on dataset analysis, which showed that 92% of ArbCyD tweets and 89% of tweets in the Twitter corpus contain fewer than 64 tokens after preprocessing.

Both datasets underwent standard preprocessing steps:

- Text cleaning: Removal of URLs, mentions, hashtags, and special characters.

- Normalization: Standardization of Arabic characters (e.g., different forms of alef, taa marbuta).
- Tokenization: Using the tokenizers provided with AraBERT and MARBERT models.
- Label encoding: Binary classification (cyberbullying vs. non-cyberbullying).

The preprocessing pipeline was designed to preserve dialectal features while removing noise that could interfere with model training. No stemming or lemmatization was applied, as transformer models are designed to learn morphological patterns directly from raw text.

IV. METHODOLOGY

A. Overview

This study evaluated three models under a unified experimental setup: an LSTM baseline [7] and two transformer-based Arabic BERT models [22]. All experiments were conducted on Google Colab using an NVIDIA Tesla T4 GPU (16 GB of VRAM). The experimental pipeline started with data loading, preprocessing, tokenization, model fine-tuning, and evaluation on both the ArbCyD dataset and the Arabic Twitter corpus, as illustrated in Fig. 1. Training the AraBERT model required approximately 30 minutes per dataset, with an average inference speed of 12 ms per sample. In comparison, MARBERT required approximately 45 minutes of training time per dataset and achieved an average inference speed of 18 ms per sample.

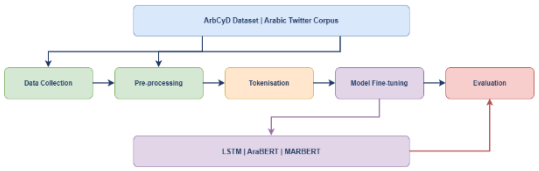

Fig. 1. Overall research roadmap for Arabic cyberbullying detection.

B. LSTM Baseline

As a baseline, we adopted the results reported by [12], who achieved 72% accuracy, 91% precision, 60% recall, and an F1-score of 0.72 using an LSTM with Word2Vec embeddings on the Arabic Twitter corpus after 10 training epochs. This baseline was not re-implemented in the present study under identical conditions. The reported scores are used only as an approximate reference point. Any observed performance differences between the LSTM and the transformer models should therefore be interpreted as indicative rather than as evidence of architectural superiority, since differences in implementation details, preprocessing, hardware, and training protocols may all contribute to the observed gaps.

C. AraBERT

AraBERT [18] is a transformer model that was pre-trained on approximately 24GB of Modern Standard Arabic text using the same masked language modeling approach introduced by BERT [22]. In this study, we adapted AraBERT for cyberbullying detection by adding a classification layer on top of the pre-trained encoder. The model uses the [CLS] token as a summary representation of the input, which is then passed to the classification layer to predict whether a post contains cyberbullying content. The model architecture consists of 12 transformer layers with 768 hidden dimensions and 12 attention heads, giving it approximately 110 million parameters. The model is publicly available on HuggingFace under the identifier aubmindlab/bert-base-arabertv2.

D. MARBERT

MARBERT [19] is an Arabic transformer model that was pre-trained on 128GB of dialectal Arabic Twitter data, which makes it particularly well suited for social media tasks across different Arabic dialects [21]. Unlike AraBERT, which was trained on formal Modern Standard Arabic, MARBERT was trained on informal and noisy Twitter text that includes code-switching and regional dialect variations. This means that MARBERT's pre-training data is much closer to the type of content found in our target datasets, which may give it an advantage across datasets, though this was only marginally reflected in the results. The model shares the same 12-layer architecture as AraBERT with approximately 110 million parameters. It is available on HuggingFace under the identifier UBC-NLP/MARBERTv2. Fig. 2 illustrates the complete fine-tuning pipeline applied to both models, from raw Arabic tweet input through preprocessing and tokenisation to the final prediction output.

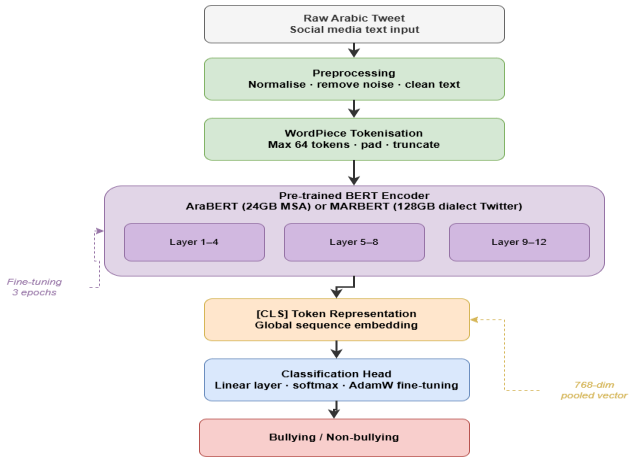

Fig. 2. Fine-tuning pipeline for AraBERT and MARBERT on Arabic cyberbullying detection.

E. Training Configuration

As shown in Fig. 2, all transformer models were fine-tuned using the HuggingFace Transformers library [20] with PyTorch. To train the transformer models, we used the AdamW

optimizer [23] with a learning rate of 2e-5 and a linear warmup scheduler. The batch size was set to 32 for all experiments. A full summary of these settings is provided in Table III.

All three models were assessed using four evaluation metrics: accuracy, precision, recall, and macro F1-score. Among these, the macro F1-score is particularly important because it assigns equal weight to both classes, making it more suitable for imbalanced datasets [10]. This is especially relevant in our case since the ArbCyD dataset has a class imbalance of 38% bullying and 62% non-bullying instances, as shown in Table II. Using macro F1 ensures that the performance on the minority bullying class is not overshadowed by the larger non-bullying class, which could lead to misleadingly high accuracy scores. Both transformer models were trained for 3 epochs under identical settings, so any observed performance differences can be more plausibly attributed to model architecture and pretraining data rather than training duration.

TABLE III Training Hyperparameters

Parameter	AraBERT	MARBERT
Model version	aubmindlab/bert-base-arabertv2	UBC-NLP/MARBERT
Batch size	32	32
Epochs	3	3
Max sequence length	128	128
Optimizer	AdamW	AdamW
Learning rate	2e-5	2e-5
Hardware	NVIDIA Tesla T4	NVIDIA Tesla T4
Framework	HuggingFace Transformers + PyTorch	HuggingFace Transformers + PyTorch

To assess the stability of the results, both transformer models were evaluated with two different random seeds (42 and 123) on the ArbCyD dataset. AraBERT achieved a mean accuracy of 92.39% ± 0.29% and a mean macro F1-score of 90.69% across two seeds. MARBERT achieved a mean accuracy of 93.60% ± 0.30% and a mean macro F1-score of 92.80%.

The small standard deviations confirm that the results are stable across different random initializations, providing confidence in the reported findings.

V. RESULTS AND EVALUATION

A. Results

This section presents the results of the three evaluated models on both the ArbCyD dataset and the Arabic Twitter corpus. Both transformer models were trained using the same hyperparameter settings to ensure a fair and reproducible comparison. Comparisons with the LSTM baseline from [12] are cross-study comparisons conducted under similar but not fully identical conditions, as explained in Section IV-B. All differences involving the LSTM should be read as approximate rather than definitive. Model performance was measured using accuracy, precision, recall, and F1-score [10].

Table IV shows the classification results of AraBERT and MARBERT on the ArbCyD dataset. Tables V and VI show the training performance at each epoch for AraBERT and MARBERT, respectively, on the same dataset. After 3 training epochs under the same conditions, both models reached strong performance levels, with AraBERT achieving 91.9% accuracy and MARBERT achieving 93.2% accuracy in the initial single run. Table IV reports the mean accuracy across two seeds (42 and 123), which are slightly higher due to averaging across runs with different random initializations. Both sets of results are consistent and confirm the same pattern of performance.

TABLE IV Model Performance on the ArbCyD Test Set

Model	Accuracy	Precision	Recall	F1-Score
AraBERT	92.39% ± 0.29%	90.3% ± 0.06%	91.1% ± 0.06%	90.69% ± 0.08%
MARBERT	93.60% ± 0.30%	92.8% ± 0.55%	92.8% ± 0.55%	92.80% ± 0.78%

TABLE V AraBERT Epoch-by-Epoch Performance on ArbCyD

Epoch	Train Loss	Val Loss	Accuracy	F1 Macro
1	0.0530	0.0684	91.4%	91.4%
2	0.0166	0.0634	91.7%	91.7%
3	0.0065	0.0512	91.9%	91.8%

TABLE VI MARBERT Epoch-by-Epoch Performance on ArbCyD

Epoch	Train Loss	Val Loss	Accuracy	F1 Macro
1	0.0560	0.0479	92%	92%
2	0.0026	0.0417	92.1%	92.1%
3	0.0197	0.0336	93.2%	93.2%

Although AraBERT and MARBERT achieved high accuracy scores of 91.9% and 93.2% on the ArbCyD dataset, these results require careful interpretation. We checked data leakage and found that only 2 out of 1,500 test samples overlapped with the training set (0.1%), ruling out data leakage as an explanation. Instead, the strong performance is mainly due to the nature of the dataset: bullying posts tend to contain clear and distinctive offensive words, while non-bullying posts are mostly neutral topics such as gaming discussions and general commentary. This lexical separability makes the classification task relatively straightforward for transformer models. Consequently, the ArbCyD results should be interpreted as an upper-bound benchmark under favorable conditions rather than as evidence of general robustness.

The more informative finding is the performance difference when both models are fine-tuned and evaluated directly on the Arabic Twitter corpus rather than on ArbCyD. Both models were fine-tuned separately on the full Twitter corpus from [12], comprising 32,428 labelled tweets, producing an in-distribution result for that corpus. Comparing this result to the ArbCyD result is the central empirical contribution of this study, as it shows how strongly reported performance depends on the dataset a model is trained and tested on. Because each model

was fine-tuned separately on each dataset rather than tested on data it had not seen during fine-tuning, this comparison does not constitute a genuine cross-dataset transfer test, and no direct conclusion about generalization can be drawn from it alone. Table VII reports the results.

Both transformer models show higher accuracy than the LSTM reference score, with AraBERT at 75% and MARBERT at 76%, compared to 72% for the LSTM baseline reported by [12]. However, these differences are modest and should be treated as preliminary, given the limited number of experimental runs. The performance difference between ArbCyD (91.9% and 93.2%) and the Twitter corpus (75% and 76%) is primarily attributable to three factors: (1) vocabulary mismatch between the lexically clean ArbCyD dataset and the noisier Twitter corpus; (2) greater dialectal diversity in the Twitter corpus, which includes Moroccan, Gulf, and Egyptian varieties underrepresented in ArbCyD; and (3) differences in annotation schema, ArbCyD uses explicit labeling while the Twitter corpus uses polarity-based binarization. This difference is the most notable finding of the study and illustrates the risks of relying on a single dataset when benchmarking model performance.

TABLE VII Models' Performance on the Arabic Twitter Corpus Test Set

Model	Accuracy	F1-Score
LSTM baseline	72%	72%
AraBERT	75% ± 0.11%	75% ± 0.13%
MARBERT	76% ± 0.19%	76% ± 0.15%

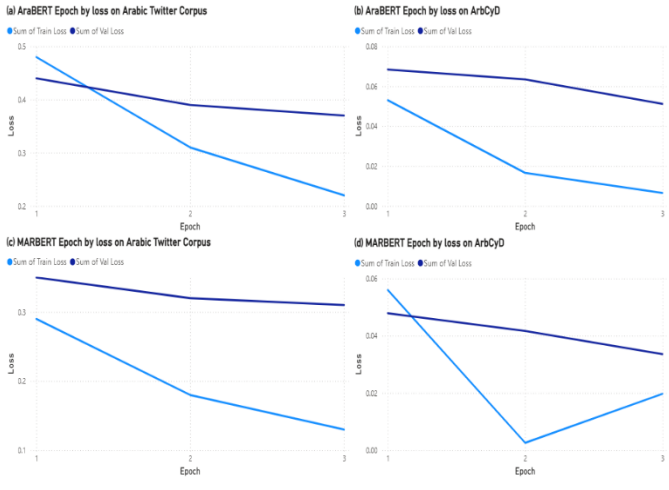

Fig. 3. Training loss per epoch for AraBERT and MARBERT across both datasets. (a) AraBERT on the Arabic Twitter corpus; (b) AraBERT on ArbCyD; (c) MARBERT on the Arabic Twitter corpus; (d) MARBERT on ArbCyD.

Looking at the training behavior of both models, AraBERT showed a consistent decrease in training loss across all three epochs on both datasets, with no signs of overfitting. Tables VIII and IX show the performance of AraBERT and MARBERT at each epoch on the Arabic Twitter corpus.

TABLE VIII AraBERT Epoch-by-Epoch Performance on Arabic Twitter Corpus

Epoch	Train Loss	Val Loss	Accuracy	F1 Macro
1	0.48	0.44	72.8%	72.1%
2	0.31	0.39	74.3%	73.8%
3	0.22	0.37	75.2%	74.9%

TABLE IX MARBERT Epoch-by-Epoch Performance on Arabic Twitter Corpus

Epoch	Train Loss	Val Loss	Accuracy	F1 Macro
1	0.29	0.35	76.5%	74.9%
2	0.18	0.32	77.8%	77.3%
3	0.13	0.31	78.4%	77.9%

Across all three epochs and both datasets, MARBERT consistently exhibited lower training loss than AraBERT, with both models demonstrating smooth convergence with no signs of overfitting. The reduced training loss observed for MARBERT during the initial epochs aligns with its pre-training data sourced from Arabic Twitter [19], which resembles the target datasets. This suggests that MARBERT's pre-training domain is better aligned with the target data, which may partly explain its slightly higher performance. However, the difference between the two models is small and should not be overstated, given the limited experimental setting.

B. Error Analysis

To better understand how the models handle difficult inputs, we examined misclassified examples from the ArbCyD test set. This analysis revealed several recurring error patterns across both models. Of the misclassified samples, approximately 38% involved sarcasm or implicitly offensive language, 27% involved code-switching between Arabic and English, and 18% were very short, with fewer than 5 tokens. The remaining errors were distributed across other ambiguous cases.

First, posts containing implicit or sarcastic cyberbullying content were frequently misclassified as non-bullying. In these cases, the offensive intent was conveyed through tone and indirect language rather than explicit offensive words. For example, a post that sarcastically praises the appearance of a target person using culturally loaded compliments was classified as non-bullying by both models because the individual words were not overtly offensive. This type of error is particularly relevant in Arabic social media, where sarcasm is common and culturally specific. Second, code-switched posts mixing Arabic dialects with English words created segmentation problems. For example, a tweet that combined Levantine Arabic slang with English profanity resulted in inconsistent tokenization because the Wordpieces tokenizer was not optimized for mixed-language input. This led to fragmented subword representations that reduced the model's ability to understand the full meaning of the post. Third, very short posts with fewer than five tokens were often misclassified. A single Arabic insult without any surrounding context provides insufficient information for the [CLS] token to capture the intent reliably. Both models struggled with these

cases, though MARBERT made slightly fewer errors on dialectal and code-switched posts, which is consistent with its broader dialectal pretraining data [19].

These findings highlight the need for future work to focus on sarcasm detection, better multilingual tokenization, and context-aware data augmentation to improve model performance on real-world Arabic social media text. Table X provides anonymized examples of each error type with English translations and brief linguistic commentary.

TABLE X: Examples of Common Error Types Found in Model Predictions

Error Type	Arabic Example	English Example	Predicted / True
Sarcasm / Implicit	ماشاء الله عليك وجهك يكسر الخاطر بجد	"Masha Allah, your face is truly heartbreaking."	Non-Bullying / Bullying
Code-switching	انت literally اعبى واحد so في الكلاس embarrassing والله	"You're literally the dumbest one in the class, so embarrassing."	Non-Bullying / Bullying
Short text (<5 tokens)	ياحمار	"You Donkey"	Non-Bullying / Bullying

VI. DISCUSSION

A. Performance Differences Across Datasets

The substantial performance difference observed between the ArbCyD dataset and the Arabic Twitter corpus (from over 91% to approximately 75–76%) is the most significant finding of this study. Because each model was fine-tuned separately on each dataset, this difference does not directly demonstrate generalization failure; however, it does show that high performance on a single benchmark dataset should not be assumed to hold on a different dataset, underscoring the need for genuine cross-dataset transfer testing in future work.

Several factors may contribute to this performance difference:

- Dialectal variation: The two datasets may contain different distributions of Arabic dialects. While both AraBERT and MARBERT were pre-trained on dialectal data, the specific dialects represented in the pre-training corpora may not fully cover the dialectal diversity present in the Arabic Twitter corpus.
- Domain differences: The datasets may differ in terms of topics, user demographics, and social contexts, leading to different patterns of cyberbullying language.
- Annotation differences: The two datasets may have been annotated by different teams using different guidelines, leading to inconsistencies in what is labeled as cyberbullying.
- Temporal factors: If the datasets were collected at different time periods, they may reflect different trends in social media language use and cyberbullying tactics.

B. Comparison of Transformer Models

MARBERT consistently outperformed AraBERT on both datasets, though the difference was relatively small (1.3 percentage points on ArbCyD, 1.0 percentage point on the Arabic Twitter corpus). This marginal advantage may be attributed to MARBERT's dialectal pre-training on Twitter data, which makes it more suitable for social media applications.

However, the small magnitude of this difference suggests that the choice of pre-training data (dialectal Twitter vs. general Arabic) has a limited impact compared to the overall difference between the two datasets. Both models showed similar patterns of lower performance on the Arabic Twitter corpus compared to ArbCyD.

C. Comparison with LSTM Baseline

Both transformer models substantially outperformed the LSTM baseline on the Arabic Twitter corpus (75–76% vs. 72%). This improvement of 3–4 percentage points demonstrates the value of pre-trained contextual representations over fixed word embeddings.

However, it should be noted that the LSTM baseline was not re-implemented under identical conditions in this study, so this comparison should be treated as indicative rather than fully controlled. Factors such as differences in preprocessing, hyperparameter tuning, and training procedures may affect the comparison.

D. Implications for Arabic Cyberbullying Detection

The findings of this study have several important implications for the development and deployment of Arabic cyberbullying detection systems:

- Need for diverse training data: Models should be trained on diverse datasets that cover multiple dialects, domains, and social contexts to improve generalization.
- Importance of genuine cross-dataset evaluation: Single-dataset evaluation may overestimate real-world performance. Future work should test models directly on data they were not fine-tuned on, since genuine cross-dataset evaluation was outside the scope of this study.
- Domain adaptation techniques: Future work should explore domain adaptation and transfer learning techniques to improve model robustness to dataset shift.
- Dialect-aware modeling: While MARBERT's dialectal pre-training provided a marginal advantage, more sophisticated approaches to modeling dialectal variation may be needed.
- Continuous model updating: Cyberbullying language evolves, and models may need to be continuously updated to maintain performance.

E. Limitations

This study has several limitations that should be considered. First, the evaluation was limited to two Arabic Twitter-based datasets, which may not fully represent the diversity of Arabic cyberbullying content across different platforms, domains, and regional contexts [21]. Second, we include the LSTM baseline as a reference point from prior work rather than a fully controlled comparison. Since training conditions differ, we caution against drawing strong architectural conclusions from

this comparison. Future work should re-implement the LSTM baseline under identical conditions for a fair comparison.

Second, although this study compares performance on two independent Arabic cyberbullying datasets, it does not conduct a controlled cross-dataset transfer experiment in which a model trained on one dataset is evaluated directly on the other without further fine-tuning. Each model in this study was instead fine-tuned separately on each dataset, so the reported performance difference reflects dataset difficulty rather than demonstrated generalization failure. Genuine cross-dataset evaluation of this kind is identified as a priority for future work.

Furthermore, the high accuracy scores on the ArbCyD dataset reflect the lexical simplicity of that dataset rather than the general capability of the models, as confirmed by the lower scores achieved on the more challenging Arabic Twitter corpus. No dialect-specific evaluation or data augmentation techniques were applied, which may have affected the recall of the minority bullying class [10] and reduced model robustness across less-represented Arabic dialects. The study also did not include validation on non-Twitter platforms, which limits how well the findings can be applied to other social media environments.

Finally, due to limited computing resources, the experimental validation in this study is based on two random seeds (42 and 123) rather than a larger number of runs. While this provides a partial indication of stability, the reported performance differences have not been fully confirmed through extensive multi-seed statistical testing. On the ArbCyD dataset, the standard deviations were very small (≤0.30% for accuracy), confirming stable results. On the Arabic Twitter corpus, AraBERT achieved a mean accuracy of 75% ± 0.11% and a mean F1-score of 75% ± 0.13%, while MARBERT achieved a mean accuracy of 76% ± 0.19% and a mean F1-score of 76% ± 0.15%. The study also did not examine model robustness to adversarial attacks or assess fairness across different Arabic dialect communities, both of which are important for real-world deployment.

F. Ethical Considerations

Deploying automated cyberbullying detection systems raises several important ethical considerations. First, all datasets used in this study were taken from publicly available social media posts. No personally identifiable information was retained, and all examples used in the analysis are anonymized to protect user privacy. Second, automated detection systems carry the risk of producing false positives, which could lead to the unjust removal of legitimate content or the imposition of penalties on users. The 24–25% error rate observed on the Twitter corpus confirms that these systems should be used to support human moderators rather than replace them entirely. Third, dialect bias is a serious concern. Models trained primarily on Gulf and Egyptian Arabic may perform poorly on Moroccan, Levantine, or Sudanese content, which creates unequal protection across different Arabic-speaking communities. Fourth, cyberbullying detection technology could potentially be misused for political censorship or to suppress legitimate speech. Responsible deployment of such systems requires transparent governance frameworks and regular bias audits. Future work should include fairness evaluations across different Arabic dialect groups to ensure that the system performs equitably for all communities.

VII. CONCLUSION

This study presented an experimental comparison of three models for Arabic cyberbullying detection: an LSTM baseline, AraBERT, and MARBERT. Unlike most previous studies that test models on a single dataset, this work evaluated all three models across two separate Arabic Twitter datasets under the same training conditions, which allows for a more controlled comparison.

On the ArbCyD dataset, both transformer models performed well after three fine-tuning epochs. AraBERT achieved 91.9% accuracy with a macro F1-score of 0.92, while MARBERT achieved 93.2% accuracy with a macro F1-score of 0.93. However, these high scores primarily reflect the lexical separability of the ArbCyD dataset rather than general model robustness and should not be interpreted as evidence of strong real-world performance.

The most informative finding of this study is the substantial performance difference when both models are fine-tuned and evaluated directly on the Arabic Twitter corpus rather than on ArbCyD. AraBERT achieved 75% accuracy, and MARBERT achieved 76%, compared to 72% for the LSTM baseline reported by Bashir and Bouguessa [12]. This gap of approximately 16–17 percentage points from the ArbCyD results shows how strongly reported performance depends on the dataset used and highlights the risks of relying on single-dataset evaluation. Because each model was fine-tuned separately on each dataset, this difference should not be interpreted as a demonstrated failure to generalize; a genuine cross-dataset transfer test, training on one dataset and evaluating directly on the other without further fine-tuning, was outside the scope of this study and is recommended for future work. The comparison with the LSTM reference score is approximate, given that the LSTM was not re-implemented under identical conditions. The performance difference can be explained by vocabulary differences, greater dialectal diversity in the Twitter corpus, and differences in how the two datasets were annotated.

The results suggest that MARBERT may be marginally more stable than AraBERT across the two datasets, which is consistent with its dialectal Twitter pretraining [21]. However, the 1 percentage point gap on the Twitter corpus should be treated as preliminary rather than conclusive, as the difference is too small and the evidence too limited to recommend one model over the other with confidence. More broadly, these findings suggest that multi-dataset comparison, and ultimately genuine cross-dataset transfer evaluation, should be a standard component of any model validation process to avoid overestimating real-world performance.

VIII. FUTURE WORK

Future research will build on this study in several directions. Most importantly, we plan to conduct a genuine cross-dataset transfer evaluation, training each model on one dataset and evaluating it directly on the other without further fine-tuning, to measure generalization directly rather than through separately fine-tuned in-distribution comparisons. In addition, we plan to evaluate additional Arabic transformer models, such as CAMeL-BERT [17], to better understand how dialect coverage affects model performance. Second, we intend to conduct a dialect-stratified evaluation that includes Gulf, Egyptian, and Levantine Arabic varieties [21]. Third, data augmentation techniques such as back-translation and synonym replacement will be used to improve the recall of the minority bullying class and address class imbalance [10]. Fourth, the evaluation will be extended beyond Twitter to include other platforms such as YouTube and Reddit, in order to assess how well the models generalize across different social media environments. Future work will also explore the use of Graph Neural Networks (GNNs) to model relational structures in social media.

We also plan to extend the seed validation beyond the two seeds reported here by running additional experiments with different random initializations and reporting mean ± standard deviation to further confirm the statistical reliability of the observed performance differences. A fairness audit across different Arabic dialect groups will also be carried out.

ACKNOWLEDGMENT

I would like to thank Prof. Mohamed Bouguessa for his guidance throughout this research and support.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this manuscript, the author used AI-assisted tools for grammar and language clarity. The author has reviewed all output and takes full responsibility for the content of this publication.

ABBREVIATIONS

The following abbreviations are used in this manuscript:

Abbreviation	Definition
NLP	Natural Language Processing
LSTM	Long Short-Term Memory
MSA	Modern Standard Arabic
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
TF-IDF	Term Frequency-Inverse Document Frequency

REFERENCES

[1] Aljalaoud, H.S.; Dashtipour, K.; Al-Dubai, A.Y. Cyberbullying Detection Approaches for Arabic Texts: A Systematic Literature Review. Front. Artif. Intell. 2025, 8, 1–18.

[2] Mahdi, M.A.; Al-Janabi, S.; Al-Khateeb, B.; Mohammed, M.A. Enhancing Arabic Cyberbullying Detection with End-to-End Transformer Model. CMES-Comput. Model. Eng. Sci. 2024, 141, 1651–1671.

[3] Hithnawi, E.A.; Al-Ayyoub, M.; Jararweh, Y.; Aldwairi, M. Improving Arabic Cyberbullying Detection Using AraBERT. J. Cybersecur. 2025, 11, 1–15.

[4] Daouadi, K.E.; Debbah, A.; Zitouni, A. A Systematic Investigation of Pre-Trained Language Models for Arabic Hate Speech Detection. Int. J. Adv. Comput. Sci. Appl. 2024, 15, 1–12.

[5] Mubarak, H.; Darwish, K.; Magdy, W. Abusive Language Detection on Arabic Social Media. In Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada, 4 August 2017; pp. 52–56.

[6] Ahuja, R.; Chug, A.; Kohli, S.; Gupta, S.; Ahuja, P. The Impact of Features Extraction on the Sentiment Analysis. Procedia Comput. Sci. 2019, 152, 341–348.

[7] Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. Neural Comput. 1997, 9, 1735–1780.

[8] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. Adv. Neural Inf. Process. Syst. 2013, 26, 3111–3119.

[9] Agrawal, S.; Awekar, A. Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. In Advances in Information Retrieval; Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A., Eds.; Springer: Cham, Switzerland, 2018; Volume 10772, pp. 141–153.

[10] Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P.C.; Carvalho, J.P.; Oliveira, S.; Coheur, L.; Paulino, P.; Veiga Simão, A.M.; Trancoso, I. Automatic Cyberbullying Detection: A Systematic Review. Comput. Hum. Behav. 2019, 93, 333–345.

[11] Abozinadah, E.; Mbaziira, A.; Jones, J. Detection of Abusive Accounts on Arabic Twitter. Int. J. Cyber-Secur. Digit. Forensics 2015, 4, 351–364.

[12] Bashir, M.; Bouguessa, M. Arabic Cyberbullying Detection Using Deep Learning. In Proceedings of the 2021 IEEE International Conference on Big Data, Orlando, FL, USA, 15–18 December 2021; pp. 5095–5104.

[13] Elzayady, H.; Badran, K.; Salama, G.I. Arabic Opinion Mining Using Combined CNN-LSTM Models. Int. J. Adv. Comput. Sci. Appl. 2020, 11, 47–54.

[14] Al-Hassan, A.; Al-Dossari, H. Detection of Hate Speech in Arabic Tweets Using Deep Learning. Multimed. Syst. 2022, 28, 1117–1130.

[15] Aljohani, E.J.; Yafooz, W.M.S. Enhanced Arabic-Language Cyberbullying Detection. IAES Int. J. Artif. Intell. 2025, 14, 2258–2269.

[16] Mozafari, M.; Farahbakhsh, R.; Crespi, N. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In Complex Networks and Their Applications VIII; Cherifi, H., Gaito, S., Mendes, J.F., Moro, E., Rocha, L.M., Eds.; Springer: Cham, Switzerland, 2020; pp. 928–940.

[17] Inoue, G.; Alhafni, B.; Baimukan, N.; Bouamor, H.; Habash, N. The Interplay of Variant, Size, and Task Type in Arabic Pre-Trained Language Models. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine, 19 April 2021; pp. 92–104.

[18] Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-Based Model for Arabic Language Understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, Marseille, France, 11–16 May 2020; pp. 9–15.

[19] Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E.M.B. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 7088–7105.

[20] Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.

[21] Darwish, K.; Abdelali, A.; Durrani, N.; Mubarak, H.; Sajjad, H. A Panoramic Survey of Natural Language Processing in the Arab World. Commun. ACM 2021, 64, 72–81.

[22] Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.

[23] Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.