Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# A Dialect-Aware Deep Learning Framework for HateSpeech Detection in Yemeni Arabic Social Media Text

**Ahlam Ahmed**
Taiz University
**Murad A. Rassam**
Taiz University
**Dina M. Ibrahim**
d.hussein@qu.edu.sa
Qassim University

**Article**

**Keywords:**

**Posted Date:** June 22nd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9918302/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# A Dialect-Aware Deep Learning Framework for Hate Speech Detection in Yemeni Arabic Social Media Text

**Ahlam Ahmed**[1,+], **Murad A. Rassam**[1,2,+], **and Dina M. Ibrahim**[2,*,+]

[1]Faculty of Engineering and Information Technology, Taiz University, 6803 Taiz, Yemen
[2]Department of Information Technology, College of Computer, Qassim University, Buraydah, 51452, Saudi Arabia
[*]corresponding.d.hussein@qu.edu.sa
[+]these authors contributed equally to this work

## ABSTRACT

Hate speech detection in Arabic social media remains challenging due to dialectal diversity, limited annotated resources, and the prevalence of implicit and context-dependent expressions. While transformer-based models have achieved strong results for Modern Standard Arabic and major dialects, low-resource dialects such as Yemeni Arabic remain underexplored. In this work, we present Ye-HS, a manually annotated dataset of 4,572 Yemeni Arabic tweets and comments collected from the X platform and labeled into hate and non-hate categories. To address dialectal variability, we propose a dialect-aware hybrid framework that integrates a pretrained Arabic transformer (MARBERT) with a dual bidirectional GRU architecture composed of shared and private components enhanced with attention mechanisms. The shared component captures general Arabic sequential patterns, while the private component learns dialect-specific linguistic characteristics. Extensive experiments using stratified cross-validation demonstrate that the proposed model consistently outperforms strong machine learning, deep learning, and transformer-based baselines. Additional evaluations across multiple Arabic dialect datasets confirm the robustness and transferability of the approach. Error analysis further highlights the model's strengths and limitations in handling implicit hate speech and sarcasm. The proposed framework and dataset provide practical resources for Arabic hate speech detection and contribute toward more inclusive and dialect-sensitive language technologies.

## Introduction

Social media has been widely used for various purposes, such as advertising, business, news, etc. The idea of allowing users to post anything at any time on social media has contributed to the prevalence of inappropriate content. As a result, these platforms have become a fertile environment for this type of content. Hate speech is the most common form of destructive content on social media and can come in the form of text, images, or videos[1].

X is a platform where discussions can take place on various topics and issues related to current events, political opinions, and news dissemination, not only by the public but also by the government. This provides freedom to communicate, express opinions, and publish controversial topics, which constitutes an opportunity to spread electronic hate speech. X platform's online hate speech policy defines it as promoting violence against, directly attacking, or threatening others, or any communication that demeans a person or group based on characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, political affiliation, national origin, age, disability, or disease[2]. Hate speech weakens and destroys societies because it spreads hatred and prejudice among its members. If left unaddressed, it can lead to violence and may exacerbate hate crimes. It also leads to dire consequences such as low self-esteem, anxiety, depression, despair, and, in some cases, a lack of motivation to survive, which ultimately leads to the death of the victims[3]. Detecting hate content on social media is a fundamental requirement for social media platforms.

The wide variety of Arabic dialects used by Arab users on social media makes the problem even more complex. In fact, Arabic dialects are numerous and vary not only from country to country but also within the same country, resulting in many words with similar spellings that have different meanings across dialects and regions[4]. X platform aims to prevent aggressive behavior that could cause real-world harm or distress among its users.

Significant research efforts have been made to provide automated solutions for detecting hate speech, ranging from simple machine learning models to more complex deep neural network models[4]. In[5], an ensemble learning method was used to vote on two standard datasets to detect hate speech on X platform, using logistic regression models, decision trees, and support vector machines, achieving a significant improvement in task performance. While in[6] deep learning techniques such as LSTM, BiLSTM, and CNN were used for the same task on a dataset collected from multiple sources, the BiLSTM model was the most accurate for detecting hate speech and capturing the complex nature of the context. In[7], a Pattern-Based Deep Hate Speech Detection (PDHS)-based approach was used to improve representation as a pre-trained transformed language model, to classify

tweets based on POS tag features as a feature architecture, and to incorporate sentiment polarity scores to improve performance. As a result, the evaluation achieved a high-performance score.

Automatically detecting offensive language on Arabic social media is an important step toward regulating content for Arabic-speaking social media users. In recent years, numerous researchers have contributed to numerous studies proposing and developing new methods and techniques to suppress and limit its spread. In[8], deep learning (LSTM) models, CNN+LSTM, GRU, and CNN+GRU, were developed, which outperformed an SVM model using TF-IDF for text representation in multi-class hate classification. In[9], a pre-trained transfer-learning-based model was presented to recognize hate speech in Arabic tweets. The latest techniques were applied to multi-dialectal, monolingual, and multilingual language models. The results demonstrated the superiority of the multi-dialectal language model (AraBERT-X platform) and that tuning modern PLMs improves hate speech detection performance on Arabic tweets, achieving higher accuracy. In[10], a hybrid approach was proposed for detecting hateful and offensive speech in Arabic, combining a pre-trained CNN Attention model for feature extraction, with a Random Forest model optimized using Particle Swarm and Grey Wolf Optimizer for feature classification. The evaluation was conducted on two Arabic datasets classified as hateful, offensive, and normal, and the experimental results showed that CNN Attention performed better when combined with the Random Forest model optimized using the Grey Wolf Optimizer. In[11], models were designed for the automatic detection of hate speech subclasses, using CNN, CNN-LSTM, and BiLSTM-CNN architectures, with skip-gram (SG) embedding, continuous bag-of-words (CBOW) embedding at the word level, and pre-trained multilingual universal set encoder (MUSE) embedding at the sentence level. In different experiments to detect appropriate embeddings, models with linear embedding performed better than other embeddings. The SG-CNN model outperformed the other models in binary classification. In ternary classification, the SG-CNN and SG-BiLSTM-CNN models achieved the best accuracy. In multi-class classification, the SG-CNN and SG-BiLSTM-CNN models achieved the best results, all on the ArHS dataset.

Although studies exist that address the detection of hate speech in Modern Standard Arabic and various mixed and distinct Arabic dialects, to our knowledge, no study has yet focused specifically on detecting hate speech in the Yemeni dialect using a more precise and effective modern approach. This is the primary motivation for this study. In[12], the authors analyzed emotions in the Yemeni dialect using a hybrid machine-learning and deep-learning stacking model, achieving an accuracy of 94.71%. However, it did not directly address hate speech. In[13], a CNN's pre-trained model was used to monitor hate speech related to the COVID-19 pandemic in the Arab region over a specific time, including Yemeni tweets. However, these models did not use modern Transformer models, limiting their ability to capture deeper context and implicit meaning. Furthermore, we are not aware of a Yemeni Arabic dialect dataset that could help the research community generalize their hate speech detection methods to Arabic. Therefore, providing such a dataset would support research into the automatic detection of offensive content and hate speech in several Arabic dialects.

Transformer-based architectures and large-scale language models have advanced Arabic hate speech identification. Arabic-specific transformers like AraBERT, MARBERT, and dialect-aware variations outperform classical machine learning and deep learning in offensive language and hate speech categorization tasks, according to recent studies. However, cross-dialect generalization, implicit hate expressions, sarcasm, code-switching, and the lack of annotated resources for low-resource Arabic dialects remain key issues[14,15]. Recent studies have shown that contextual transformer representations, recurrent neural networks, attention mechanisms, and hybrid learning architectures improve robustness and classification performance[16,17]. These gains aside, Yemeni Arabic is underrepresented in hate speech research, with no published benchmark dataset for Yemeni hate speech recognition. Thus, the Ye-HS dataset and a dialect-aware MARBERT-based hybrid framework that learns shared Arabic linguistic patterns and Yemeni dialect-specific behaviors fill this need.

To this end, this study aims to propose a hybrid transformer and deep learning-based model for detecting Yemeni hate speech to fill the gap identified in the literature. To achieve this aim, the study first develops a Yemeni dialect hate speech dataset from tweets on the X platform and compares the model's performance against datasets from other dialects. This work makes the following contributions:

1. A new Yemeni Arabic hate speech dataset (Ye-HS). We introduce a manually annotated dataset of 4,572 tweets and comments collected from the X platform, addressing the scarcity of publicly documented Yemeni dialect resources for hate speech detection.

2. A dialect-aware hybrid detection framework. We propose a transformer-based hybrid architecture that combines MARBERT with shared and private bidirectional GRU components and attention mechanisms to jointly model general Arabic patterns and dialect-specific linguistic features.

3. Comprehensive evaluation across Arabic dialects. The proposed model is evaluated using stratified cross-validation and benchmarked against strong traditional, deep learning, and transformer-based baselines on multiple Arabic dialect datasets.

4. Error and robustness analysis. We provide a detailed analysis of misclassifications, highlighting challenges such as sarcasm, political discourse, and implicit hate speech in dialectal Arabic.

5. Practical insights for dialect-sensitive NLP. The results demonstrate the effectiveness of hybrid architectures for low-resource dialects and offer guidance for developing robust hate speech detection systems in linguistically diverse settings.

The remainder of this paper is organized as follows: Section 2 reviews and analyzes relevant work to reveal the gaps that the current work should address. Section 3 presents the methodology used in the study and the models employed to achieve the objectives of this work. It also describes the collected dataset, preprocessing techniques, and feature engineering techniques used. It details how the model was constructed and the evaluation measures applied. Section 4 presents the results of implementing and testing the proposed model for detecting hate speech. Section 5 explains how the results were validated and benchmarked against related models, and how errors were analyzed. Section 6 presents the challenges faced and future work needed to extend the proposed model. Section 7 concludes this paper.

## Related Work

In this section, traditional machine learning, deep learning, translational learning, and hybrid models for hate speech detection are reviewed.

### Traditional Machine Learning Models

Traditional machine learning models were the first models adapted for hate speech detection in social media applications. In[18], the authors presented a machine learning-based system for filtering hate speech from Facebook content, and the results showed the effectiveness of the proposed method on a multi-category, emotion-tagged dataset from Facebook and X platform, using a profanity dataset containing 6,000 posts. In[19], the performance of machine learning models developed by Facebook during the cartoon controversy in France was evaluated on a dataset containing 3471 deleted Facebook posts and another dataset containing 1856 posts. The models demonstrated difficulty processing the complex linguistic features of Arabic. In[20], authors used Support Vector Machines (SVM) and Naïve Bayes (NB), with multiple n-gram plots, and term frequency weight (TF) to detect abuse and hate in an L-HSAB dataset in the Levantine dialect. NB outperformed SVM in binary and multi-category classification. In[21], the authors performed lexical content analysis using emojis and feature extraction using n-gram and TF-IDF. They used a simple lexicon-based model (Lex) with a list of potential abusive words and phrases, as well as the SVM and Chance classification models. The SVM was found to perform better on multi-platform datasets for abuse detection. The study[22] used machine learning models with N-gram feature engineering, TF-IDF, and CBOW to detect Saudi hate speech mixed with symbols on 12,356 mixed-symbol tweets and 11,231 Arabic tweets. SGD with TF-IDF achieved 98.18% accuracy on the mixed-symbol dataset.

In[23], the authors used an SVM to classify 16,000 YouTube comments that included foreign words transcribed into the Arabic alphabet and Arabic words transcribed into non-Arabic alphabets, achieving an F1 Score of 82%. In[24], SVM outperformed Naïve Bayes in detecting bullying, achieving an F1 score of 92.7%, on 35,273 tweets in Arabic dialects and 91,431 tweets in English, using a Tweet to Senti Strength feature vector filter based on the SentiStrength mechanism to convert strings into vectors. In[25], the SVM, RBF, Gaussian NB, DT, and RF classifiers were trained on 15 datasets using different feature sets. The RF classifier with TF-IDF and profile features outperformed the other classifiers and feature sets, achieving an accuracy of 91.3% and a mean G of 91.2%. The study[26] proposed previously trained MLP and SVM models for detecting hateful tweets and used sentiment analysis and appropriate Word2Vec embeddings to classify and analyze 100,000 tweets. The SVM classification accuracy was 92%. In[27], the NB classifier outperformed SVM in binary and multi-class classification of the Tunisian dataset (T-HSAB). In[28], the authors used the FastText classifier to propose a learning model based on character n-grams, and it achieved 90% F1 on 1100 tweets, outperforming both the baseline classification using a list of offensive words and the SVM classifier.

In[29], the authors used several machine learning models, and XGBoost with TF-IDF achieved the best performance, with an accuracy of 89.95%, compared to the other models on 9000 tweets categorized as bullying and non-bullying. In[30], they used the LR, RFC, SVM, and MNB machine learning models to classify 46,898 multi-category Instagram comments, and the SVM classifier outperformed, achieving an F1 score of 69% for bullying comments and 85% for positive comments. In[31], several machine learning models (using text representation techniques w2v, TF-IDF, and AraBERT) performed well on a Jordanian dataset stratified by emotion-based categories of hate speech severity. However, these methods fall short in capturing the deeper, more nuanced meaning of the textual context, especially in the diverse, morphologically rich Arabic dialects.

As summarized in Table 1, traditional machine learning approaches such as SVM and Naïve Bayes have been widely applied to hate speech detection, achieving moderate success through handcrafted feature engineering. However, their inability to capture contextual and semantic nuances, particularly in multilingual and dialectal settings, limits their effectiveness—motivating the adoption of deep learning and transformer-based models.

| Paper | Used models | Dataset | The best model | Performance Measures |
|---|---|---|---|---|
| [18] | Linear SVC, Multinomial NB, Bernoulli NB, LR, SGD, SVC, Nu SVC, DT, RF | Hate dataset categorized into hate and non-hate | Linear SVC with TF-IDF using Chi-square feature | Accuracy=82.39%, F1=92.45% |
| [20] | NB, SVM | L-HSAB Multi-category (hateful, offensive, or normal), and binary (offensive and normal) | NB | Precision=90.5<br>Recall=89.0<br>F-measure=89.5<br>Accuracy=90.3 |
| [21] | SVM, Chance Lex | MPOLD classified into (offensive (OF) and non-offensive (NOT) ETCD , L-HSAB, OFD | SVM on MPOLD | F-measure = 0.74 |
| [22] | LR, NB, RF, K-NN, SGD | Arabic dataset Code-mixing dataset All of them are classified as hate and non-hate. | SGD with TF-IDF | Acc=98.21 on Code-Mixing Tweets dataset |
| [23] | SVM | Dataset categorized as offensive, non-offensive, and neutral | SVM | F-measure=82% |
| [24] | SVM, NB | Arabic Tweets English Tweets Facebook database All classified as yes or no | SVM | Precision=93.4<br>Recall=94.1<br>TP=38.1<br>FP=5.7<br>F-Measure=92.7 |
| [25] | SVM, NB, DT, RF | The dataset is classified into “positive” (HS) and “negative” (NOT HS). | RF with TF-IDF Profile features | Accuracy=93.13<br>G-mean=0.9127 |
| [26] | MLP, SVM | Dataset classified as extreme and non extreme Arab Jordanian dataset (AJGT) classified as positive and negative | SVM, MLP | Accuracy= 92%<br>Accuracy= 91% |
| [27] | SVM, NB | T-HSAB Binary and multi-class classifier | NB with (TF2I) | In one class:<br>F-Measure=92.3<br>Accuracy=93.29<br>On multi-class:<br>F-Measure=84.6<br>Accuracy=97.33 |
| [28] | word-1st, SVM, Fast text | Dataset classified as offensive and non offensive | Fast text | F1=90%<br>Accuracy=89.95% |
| [29] | SVM, RF, NB, LR, GBDT, Bagging | Arabic dataset classified into bullying and non-bullying | XGBoost | Accuracy=89.95% |
| [30] | SVM, MNB, LR, RF | Multiclass dataset | SVM | F1= 69% for bullying F1= 86% for insults |
| [31] | SVM, LR, NB, RF, GBDT, and | JHSC | CAMELBERT | F1= 85% for compliments<br>F1= 78% for insults;<br>F2: HS=80.49<br>F1=0.86<br>acc=0.61<br>AUC=0.69 |

**Table 1.** Summary of existing machine learning-based methods employed for hate speech and related topics in the Arabic context

## Deep Learning Models

Deep learning techniques have been used to overcome the limitations of traditional methods. In[2], a model was proposed that integrates Bi-LSTM, CNN, and GRU, with FastText and GloVe included to detect bullying in 266,747 records sorted into 0 and 1. They found that the model with the Adam optimizer and Sigmoid activation function identified instances of bullying in text with an accuracy of 98.83% and an F1 score of 98.13%. In[8], the authors used LTSM, CNN+LTSM, GRU, and CNN+GRU models, compared to the SVM model, on 11,000 tweets. All of them outperformed the SVM model in multi-category hate classification. Adding the CNN layer to the LTSM improved performance, with 72% accuracy, 75% recall, and 73% F1 score. In[11], CNN, CNN-LSTM, and BiLSTM-CNN techniques, with word-level skip-gram and CBOW embeddings and a pre-trained multilingual universal set encoder (MUSE) embedding at the sentence level, were used to detect hate categories. Models with linear embedding performed better than the others. The SG-CNN model achieved the highest accuracy of 81% in binary classification on ArHS.

In[13], a pre-trained CNN model was used to detect hate speech related to the COVID-19 pandemic in 547,554 Arabic tweets.

Each tweet was assigned a value; if the value exceeded 0.5, the speech was considered hate speech, otherwise it was not. The results showed ACCURACY = 83% and F1-SCORE = 79%. In[32], they used SOTA systems and an MTL configuration. To evaluate the (Let-Mi) dataset on misogyny in the Levantine dialect, the results showed that using MTL improves performance. In[33], the authors used CNN, LSTM, and CNN-LSTM models, with pre-trained 300D Skip Gram and word2vec embeddings (CBOW), on 13481 binary-classified tweets and 15486 multi-classified tweets. LSTM outperformed, achieving an accuracy of 95.59% and an F1 score of 96.73% in binary classification and 89% in multi-class classification. In[34], they proposed a 10-layer DRNN-2 model for multi-layer classification and a five-layer hidden DRNN-1 model for binary classification. When comparing performance with machine learning models, DRNN-1 achieved the highest accuracy of 99.73% for binary classification on 4203 comments classified as binary, triple, and seven hate categories. In[35], they combined CNN and LSTM and used pre-trained Word2Vec and AraVec word-embedding models to extract features from 3696 tweets. In the AraVec model, an N-gram model was used in conjunction with a skip-gram model and a CBoW. The proposed approach with AraVec using skip-gram achieved an F1 score (71.688%).

Recent deep learning experiments have shown that recurrent and hybrid neural architectures recognize Arabic hate speech. On an Arabic hate speech dataset,[36] tested CNN, BiLSTM, and a hybrid CNN-BiLSTM architecture. The CNN-BiLSTM model performed best, with 95.2% accuracy and 94.8% F1-score, demonstrating the benefits of local feature extraction and sequential contextual learning.[37] examined BiLSTM-, GRU-, and attention-based architectures on an Arabic social media corpus and found that Attention-BiLSTM had the greatest F1-score of 95.7%. These results show that merging attention processes with recurrent neural networks can improve hate speech detection by focusing on the most informative social media text linguistic patterns.

Table 2 summarizes deep learning-based approaches for hate speech detection, highlighting the evolution from traditional machine learning to representation learning techniques. Models such as CNN, LSTM, and hybrid CNN-BiLSTM architectures demonstrate improved capability in capturing semantic and contextual information, leading to enhanced performance compared to conventional methods. However, the reported evaluation results remain highly dependent on dataset characteristics and often exhibit marginal variation across architectures. Despite these advancements, deep learning models still face limitations in handling implicit, context-dependent, and multilingual hate speech, particularly in low-resource settings. These challenges underscore the need for more advanced architectures, such as transformer-based models, which can better capture long-range dependencies and contextual nuances.

### Transformer-based Models

Several existing studies used transformer models to extract context-rich text representations, especially when dealing with different dialects. In[9], the latest technologies for multilingual, monolingual, and multilingual language models were tested. The multilingual language model (AraBERT-X platform) outperformed traditional machine learning (ML) models, deep learning (DL) models, and monolingual and multilingual language models (PLMs). In[38], the authors used the Marbertv2, Marbert, and Arabert models in MTL tasks for Egyptian, Levantine, Saudi, Gulf, and Algerian Arabic dialects. The model was trained on multiple dialects simultaneously and achieved the highest F1 score of 98% for the Egyptian dialect. However, this study focused on the multiplicity of dialects rather than on improving text representation.

In[39], the authors used the BERT model with WordPiece encoding during training. The performance of the Arabic BERT-Mini technique was evaluated in comparison with SVM, DT, RF, KNN, NB, gradient enhancement, LSTM, CNN-LSTM, and some other works, and the results showed that ABMM delivered high performance with a score of 98% for all performance metrics compared to other models and works. In[40], the following models were used: AraBERT v0.2 Large X platform, CAMELBERT Mix, MARBERTv2, AraELECTRA, and mBERT (Multilingual BERT). The best performance was achieved by AraBERT v0.2 Large X platform, with Macro F1 = 97.52%. In[41], the ARBERT, MARBERT, and TunBERT (Tunisian-specific) models were used, and the best-performing model was TunBERT, with an F1-score of 97

Although transformer models are effective, when combined with deep or automated technologies, or when other technologies are combined, the shared features are leveraged more effectively to improve performance.

The reviewed studies, as indicated in Table 3, consistently show that monolingual Arabic transformer models, especially AraBERT variants fine-tuned on Twitter data (e.g., AraBERT-Twitter v0.2), outperform classical machine learning and multilingual models (e.g., mBERT) in Arabic hate speech detection. These models achieve high F1 scores, ranging from 84% to 98%, with top results (e.g., 97.42% F1) on fine-grained categories such as sexism, religious discrimination, and racism. However, the wide variation in performance is largely due to differences in datasets (dialects, label schemes, class balance), and near-perfect results (98%) may indicate dataset-specific overfitting or limited real-world robustness. Feature engineering is rarely needed with transformers, but challenges remain in handling multiple dialects, code-switching, and implicit hate speech, while computational cost and cross-dialect generalization are often unreported.

Recently, transformer-based designs have maintained their dominance in Arabic hate speech detection research.[17] examined multitask transformer learning across various Arabic dialects and reported significant enhancements in cross-dialect hate

| Paper | Used models | Dataset | The best model | Evaluation results |
|---|---|---|---|---|
| [2] | CNN-BiLSTM-GRU | cyberbullying dataset and classified into 0 (no bullying) and 1 (presence of bullying) | CNN-BiLSTM-GRU with Adam, Sigmoid, ReLU | Accuracy = 98.83%<br>recall = 99.55%<br>precision = 98.15%<br>specificity = 98.09%<br>F1 = 98.13% |
| [8] | LSTM, CNN+LSTM, GRU, CNN+GRU | Multi-category hate speech dataset (Religious, Racism, Sexism, General hate speech, Not hate speech) | CNN+LSTM | Accuracy = 72%<br>Recall = 75%<br>F1 = 73% |
| [11] | CNN, CNN-LSTM, BiLSTM-CNN | AHS, L-HSAB, OSACT4, classified: (hate, non-hate), three categories (hate, offensive, normal), and multiple categories (misogyny, racism, religious discrimination, offensive language, normal) | SG-CNN on two-class AHS; SG-CNN, SG-BiLSTM-CNN on three-class AHS; SG-CNN, SG-BiLSTM-CNN on multiclass AHS | Accuracy = 81%<br>Accuracy = 74%<br>Accuracy = 73% |
| [13] | CNN pre-trained | Data is categorized as hate and non-hate | CNN pre-trained | Accuracy = 83%<br>F1-score = 79% |
| [32] | BOW + TF-IDF + NB, NB, +SVM+ LR, LSTM, BERT, MTL using BERT | Let-Mi. Multiclass and binary classified dataset | MTL using BERT on target classification; Misogyny identification | Accuracy = 88%<br>Accuracy = 90% |
| [33] | CNN, LSTM, CNN-LSTM | Dataset categorized as bullying and non-bullying. Dataset categorized as sexual bullying, animal-related bullying, visual bullying, religious bullying, and psychological bullying, but not bullying | LSTM; CNN | on two classes"<br>Accuracy = 0.9559<br>F1 = 96.73%<br>on 6 classes:<br>F1 = 89%<br>on 6 classes:<br>Accuracy = 78.75% |
| [34] | DT, MLP, LR, RF, NB | A dataset classified binary into negative and positive, tri-classified into negative and positive, hate speech, and the seven categories of anti-religion, racist, gender equality, violent, abusive, insulting/bullying, normal positive, and normal negative | DRNN-1; DRNN-2 | on 3 classes:<br>Accuracy = 99.73%<br>Accuracy = 95.38%<br>on 7 classes:<br>accuracy = 84.14% |
| [35] | CNN+LSTM | Dataset classified as hate or normal | CNN+LSTM with argmax (N-grams & SG) | Accuracy = 66.564<br>Recall = 79.768<br>Precision = 68.965<br>F1-score = 71.688 |
| [36] | CNN, BiLSTM, CNN-BiLSTM | Arabic hate speech dataset | CNN-BiLSTM | Accuracy = 95.2%<br>F1 = 94.8% |
| [37] | BiLSTM, GRU, Attention-BiLSTM | Arabic social media corpus | Attention-BiLSTM | F1 = 95.7% |

**Table 2.** Summary of existing deep learning-based methods employed for hate speech and related topics in the Arabic context

speech classification.[15] performed a thorough assessment of contemporary Arabic transformer models and determined that dialect-aware pretraining is a critical component influencing classification performance. Nonetheless, despite their efficacy, transformer-only systems may continue to have difficulties with implicit hate speech, brief loud social media texts, and dialect-specific idioms. Recent works increasingly integrate transformers with recurrent or attention-based architectures to capture sequential relationships and enhance contextual modeling[16].

| Paper | Used models | Feature engineering | Dataset | The best model | Evaluation results |
|---|---|---|---|---|---|
| [9] | MNB, SVM, CNN, LSTM, Bi-LSTM, PLMs (BERT, XLM, AraBERT, AraBERT-Twitter) | BoW, TF-IDF, WE, Aravec | 2 datasets are classified into (Not-Hate Speech, General HS, Religious HS, Racial HS, Sexism( | PLMs base-arabertv02-twitter | Macro-Weighted = 84.69% Accuracy = 84.66% |
| [38] | NA | NA | GHSD, Algerian dialect toxic dataset (RHSD) Egyptian-Arabic Custom dataset Levantine hate speech dataset | AraBERT | F1 macro = 84.6% |
| [39] | ABMM Linear SVC, SVC, MNB, DT, KNN, BNB, SGD, RF, LSTM, CNN+LTSM | N-Gram encoding, Word Piece | Dataset categorized into (normal, abuse, and hate speech( | ABMM | Accuracy = 0.98 F1_score = 0.98 Recall = 0.98 Precision = 0.98 |
| [40] | AraBERT v0.2 Large Twitter, CAMeLBERT Mix, MARBERTv2, Ara-ELECTRA, mBERT | NA | A dataset classified into Neutral, offensive, Sexism, Religious Discrimination and Racism | AraBERT v0.2 Large Twitter | F1 = 97.42% |
| [41] | TunBERT, AraBERT, MARBERT and AR-BERT | NA | T-HSAB | TunBERT | F1 = 0.97 |
| [17] | AraBERT, MAR-BERT, MARBERTv2 | NA | Multi-dialect Arabic hate speech | MARBERTv2 | Macro F1 = 98.0% |
| [15] | AraBERT, MAR-BERT, ARBERT | NA | Arabic hate speech benchmarks | MARBERT | F1 = 96.8% |

**Table 3.** Summary of existing deep learning-based methods employed for hate speech and related topics in the Arabic context

### Hybrid Models

Hybrid approaches allow users to leverage the advantages of combining multiple technologies to achieve high performance. In[4], the authors proposed CNN, GRU, CNN+GRU, and BERT models using CBOW, pre-trained on word2vec, for Arabic hate speech detection. The performance of the BERT binary classification model was evaluated and compared with SVM and LR classifiers using n-gram features. CNN outperformed the other models, achieving an F1 score of 79% and an AUROC of 89%. In[10], a hybrid approach was proposed that combines a pre-trained CNN Attention model for feature extraction with a random forest model optimized using particle swarms and the Gray Wolf Optimizer to classify the extracted features on the T-HSAB dataset, achieving an accuracy of 97.83%. In[42], the authors used traditional machine learning classifiers, with RBF being the best, and Bert models, with ArabicBERT being the best, and some hybrid models based on ArabicBERT, along with deep learning models (1D-CNN), BiLSTM, D-CNN-BiLSTM1, and BiLSTM-1D-CNN, followed by an RBF model to predict offensive language in separate experiments. The implementation of BiLSTM-RBF ArabicBERT yielded the highest accuracy and F1 score of 98.4% and 98.4%, respectively. In[43], the NileULex polar dictionary, fasttext, AraVec skip-gram, and Mazajak embeddings, and BERT base-multilingual and AraBERT were used to represent tweets. A new fasttext classification model with n-grams as features was trained, along with SVM classifiers with the RBF kernel, AdaBoost, and LR. The AraBERT model performed best, followed by Mazajak/SVM, compared to BERT.

In[44], classification experiments were conducted using traditional learning and deep learning models. The Bi-GRU model achieved the highest performance in classifying DL models of hate speech, with an accuracy of 73.6% and an F1 score of 75.8%. In[45], 5 BERT-based models were trained to assess offensive language, and the Darija RoBERTa-based model using the CNN-LSTM rating was the most effective, achieving 90% accuracy and an F1 score of 85%. In[46], the authors used GNB, LR,

RF, SGD, LSVC, CNN, Multilayer Perception (MLP), LSTM, and Bi-LSTM classifiers. Word2vec and FastText were used with Skip-Gram and CBOW for feature extraction. LSVC, MLP, and BiLSTM performed best, achieving up to 91% accuracy when combined with SG. In[47], a multi-channel deep learning technique based on CNN, BiGRU, and transformer block models was proposed to jointly process inputs for detecting aggressive behavior. The proposed model was compared with several other models (CNN, BiGRU, Transformer, Linear SVC, FILL, LR, RF, and ET) and achieved an accuracy of 87.99%.

In[48], an approach combining a deep random forest and contextualization (CDRF) was proposed for hate speech detection. The AraBERT v02-X platform model outperformed other models on F1, and the proposed CDRF approach achieved the highest F1 score and the highest accuracy among models in other studies. In[49], the MARBERTv2, AraELECTRA, and AraGPT2 models combined with the BiGRU layer were evaluated for accurate detection of Arabic hate speech. MARBERTv2+BiGRU significantly outperformed the other models and some related work, achieving an accuracy of 67.86% and an F1 score of 61.68%. In[50], machine learning models were developed to infer personality traits and investigate the correlation between offensive language and personality. The AraBERT model was proposed to enhance it with personality traits, using the pre-trained Arabic-language model "bert-base-arabertv1". TF-IDF and binary features were used to test the machine learning models, and AraVec 2.0 and binary features were used to test the deep learning models. AraBERT achieved an F1 score of 82.3%, compared to the other models. In[51], models were designed using six machine learning algorithms, and the Ara-BERT and AJGT-BERT transition learning models. BOW and TF-IDF techniques were used, followed by BERT-based contextual text representations for feature extraction. Ara-BERT outperformed with an accuracy of 79%.

In[52], the AraBERTv0.2-X platform was configured to use the BERT model's training data, and the Farasa text hashing function was implemented. In the SVM and LR classifiers, n-grams were used for words and characters, and word-based TF-IDF and character-based counting vectors were used as training features. Five models of each classifier were fitted for each task. The BERT model outperformed all other models in detecting all classes. In[53], the multilingual BERT model, the Arabic BERT model, the ALBERT model, and another model combining CNN and BiLSTM were proposed for detecting Arabic hate speech and offensive language. The Arabic BERT model and the multilingual BERT model achieved the best performance. In[54], they used LR, Linear SVC, RF, Multinomial NB, and DziriBert. The best results were obtained from Linear SVC and RF when using TF-IDF + SMOTE, with Accuracy = 98% and Precision = 1, while DziriBert performed with Accuracy = 82% in the binary classification

In[12], the study analyzed emotions in the Yemeni dialect using a hybrid machine-learning and deep-learning stacking model, achieving an accuracy of 94.71%. However, it did not directly address hate speech.

Table 4 indicates that Hybrid models for Arabic hate speech detection consistently outperform standalone architectures, with the best results (98–99% accuracy/F1) coming from combinations of transformers (especially AraBERT variants) with BiLSTM, GRU, or RBF layers, or from classical ML classifiers (LR, SVM, RF) optimized with SMOTE, GWO, or PSO. However, performance drops sharply when moving from binary offensive classification to fine-grained hate speech types (e.g., from 90.8% to 45.1% F1 in one study), and dialect-specific transformers like DziriBert or DarRoBERTa do not always beat well-tuned traditional models on their own. Low scores on OSACT-5 scores (around 61–62% F1) highlight persistent challenges in detecting implicit, code-switched, or nuanced hate speech. Overall, hybrid approaches offer flexibility, but task complexity, class imbalance, and dialectal diversity remain the primary bottlenecks.

Recent hybrid architectures have shown encouraging outcomes for hate speech detection by integrating contextual language models with sequential learning methods and attention modules.[16] introduced a multi-view hybrid system that combines transformer embeddings with symbolic representations, demonstrating enhanced robustness on multilingual hate speech datasets. Likewise,[17] shown that transformer-based multitask systems may proficiently transfer knowledge of hate speech between dialects. These findings reinforce the emerging agreement that hybrid designs are especially effective for low-resource dialects, as they utilize both semantic contextualization and sequential feature acquisition. The proposed method integrates MARBERT contextual embeddings with shared and private Bi-GRU modules and attention mechanisms to identify both general Arabic hate speech patterns and linguistic features strange to the Yemeni dialect.

Identifying and detecting hate speech in all its forms on social media is crucial and represents the first step toward developing effective countermeasures to combat harmful hate-inciting content online. The digital space necessitates more sophisticated technologies for monitoring and combating online hate speech, considering the linguistic characteristics of Arabic dialects, including Yemeni. The following section presents a proposed methodology for detecting hate speech in Yemeni.

## Proposed Methodology

In this section, a detailed description of the proposed model is provided. Moreover, a detailed specification of the newly collected Yemeni dialect dataset is provided.

| Paper | Used models | Feature engineering | Dataset | The best model | Evaluation results |
|---|---|---|---|---|---|
| [4] | GRU, CNN, CNN+GRU, BERT, SVM, LR | CBOW, n-gram | GHSD (hateful, abusive or normal), RHSD | CNN | Accuracy=0.83<br>F1=0.79<br>Precision=0.81 |
| [10] | CNN Attention + RF+ PSO, CNN Attention + RF+ GWO | MARBERT, TF-IDF, PCA | L-HSAB, T-HSAB (OFF, HS, NORMAL) | CNN Attention + RF+ GWO, PSO | Accuracy=97.83% on T-HSAB |
| [42] | AraBERT, ArabicBERT, XLMRoBERTa, Giga BERT, MBERT, QARiB, 1D-CNN, BiLSTM, 1D-CNN-BiLSTM, SVM, MLP, KNN, NB, RBF | Word2Vec | Multiclass: bullying, insulting, racist, vulgar, non-offensive; Binary:offensive or non-offensive | Arabic BERT-BiLSTM-RBF | Accuracy=98.4%<br>Precision=98.2%<br>Recall=92.8%<br>F1=98.4% |
| [43] | SVM, FastText, AdaBoost, LR, RF, DT, Gaussian NB, Perceptron, Gradient Boosting, BERT, AraBERT | FastText, AraVec, Mazajak | Arabic offensive dataset classified as offensive or clean | AraBERT | Precision=84.6<br>Recall=82.4<br>F1=83.2 |
| [44] | Bi-LSTM, GRU, LSTM, Bi-GRU, CNN, MNB, RF, LG, SGD, Linear SVC | TF-IDF, Word2Vec, CBOW | Algerian dialect dataset categorized into hate, abusive, and bullying | Bi-GRU with FastText | Accuracy=73.6%<br>F1=75.8% |
| [45] | MARBERT, ARBERT, DawiaBERT, DaBERTa, AraRoBERTa, CNN, LSTM | – | Moroccan dialect classified: offensive and non-offensive | DawiaBERT + CNN-LSTM | F1=85%<br>Accuracy=90% |
| [46] | GNB, LR, RF, SGD, LSVC, CNN, MLP, LSTM, BiLSTM | Word2Vec, FastText, CBOW, SG | Hate speech group (Arabic, French, German) classified into hate and non-hate | LSVC, MLP and with SG | Precision=91%<br>Recall=91%<br>F1=91% |
| [47] | Transformer block-BiGRU-CNN | – | Dataset from 3 hate speech groups classified into offensive and non-offensive | Transformer block-BiGRU-CNN | Accuracy=88% |
| [48] | CDRF | AraBERTv02-twitter | Datasets classified into (Non-Hate Speech, General HS, Religious HS, Racial HS, Sexism) | CDRF + AraBERT v02-twitter | F1-score=0.90 |
| [49] | MARBERTv2+BiGRU, AraELECTRA+BiGRU, AraGPT2+BiGRU | – | OSACT-5 classified as clean speech, offensive speech, or hate speech | MARBERTv2+BiGRU | Precision = 67.86%<br>Recall = 59.64%<br>F1-score = 61.68% |
| [50] | RF, ET, DT, SVM, Gradient Boosting, XGBoost, LR, BI-LSTM, GRU, LSTM, CNN, CNN-LSTM, CNN-BILSTM, CNN-GRU, AraBERT | TF-IDF, AraVec2.0, Skip Gram 300D embeddings | Ara Personality, Arabic hate speech dataset | AraBERT | Macro-F1 = 82.3% |
| [51] | Gradient Boosting, K-NN, LR, NB, Passive Aggressive Classifier, SVM, AraBERT, BERT-AJGT | TF-IDF, BoW, AraBERT | Religious Hate Speech Group (RHSA) classified as Hate and Non-Hate | Ara-BERT; BERT-AJGT | Precision = 79%<br>Recall = 79%<br>F1 = 79%<br>Accuracy = 79% |
| [53] | Arabic BERT, ALBERT, Multilingual BERT, Ensemble DL model | wiki-news-300d-1M, crawl-300d-2M | L-HSAB, OSACT | Arabic BERT, Multilingual BERT | on 2 classes:<br>F1 = 90.8%<br>on 7 classes:<br>F1 = 80% |
| [54] | (Linear SVC, Random forest, multinomial NB, LR) with SMOTE, XGBoost | TF-IDF + SMOTE | T-HSAB, Binary and multi-class classifier | LR, Multinomial NB, RF, XGBoost | Acc = 98%<br>Acc = 99.75%<br>Acc = 82% |
| [16] | Transformer + Neuro-Symbolic Learning | NA | Arabic hate speech corpus | Hybrid Framework | F1 = 97.1% |
| [17] | MARBERT + Multi-task Learning | NA | Multi-dialect Arabic datasets | Multi-task MARBERT | F1 = 98.0% |

**Table 4.** Summary of existing hybrid methods employed for hate speech and related topics in the Arabic context

## Proposed Model Architecture

In this paper, a hybrid model based on integrating an efficient transformer structure with a two-way GRU deep learning model is proposed. The MARBERT[55], a pre-trained model on 128 GB of Arabic dialect and Modern Standard Arabic texts, was used as the primary extractor to extract contextual linguistic implications from the raw text. In addition, the Bidirectional Gated Recurrent Unit (BI-GRU), a sequence modeling architecture in deep learning that extends the GRU by processing data in both forward and backward directions, is used for classification. A GRU is a type of recurrent neural network (RNN) designed to handle sequences (text, time series, signals). According to[8], GRUs are simpler and faster to train than LSTMs, which are a more complex type of RNN. Two layers of the Bi-GRU method were used to extract sequential features and enhance representation from the MARBERT outputs. GRU units process data at each time step to produce hidden states that encode the

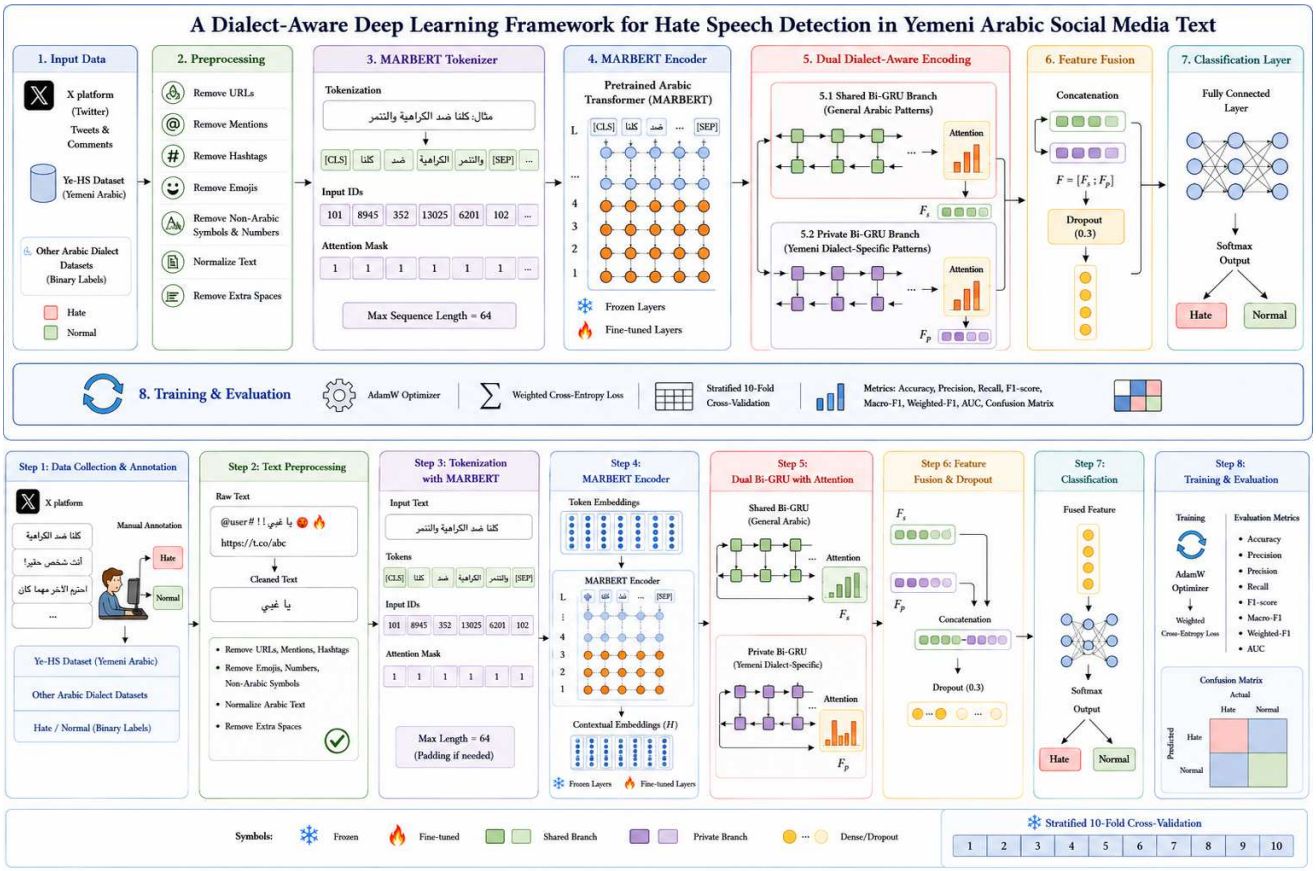



**Figure 1.** The proposed Dialect-Aware Deep Learning Framework for Hate Speech Detection in Yemeni Arabic Social Media Text

entire sentence's information. The attention mechanism was used to extract important features from the Bi-GRU outputs, which were then merged and passed to the classification layer. Figure 1 depicts the proposed model, followed by a description of its phases and steps.

As illustrated in Figure 1 , a hybrid deep learning model for hate speech detection in the Arabic Yemeni dialect on the X platform is introduced. The model consists of several sequential components, which can be described from input to output.

### *Input Preprocessing*

The input is a tweet written in the Arabic Yemeni dialect (or other Arabic dialects for comparison). Raw tweets are collected from the X platform and undergo several preprocessing operations, cleaning the text by removing irrelevant characters, removing links (URLs), and user mentions (@usernames), removing non-Arabic symbols and special characters, and normalizing spaces to ensure consistent tokenization.

### *MarBERT Tokenizer*

After preprocessing, the cleaned text is passed to the MarBERT tokenizer, which converts it into two numerical representations. The first is the input_ids, which constitute a sequence of integer IDs corresponding to each token. The second is the attention_mask, a binary sequence indicating which tokens are real (1) and which are padding (0).

### *MarBERT Pretrained Transformer*

The tokenized input is fed into MarBERT, a pretrained transformer model specifically designed for Arabic. The model is configured as follows:

- Lower layers: Frozen (not trained) to preserve general Arabic linguistic knowledge
- Upper layers: Trainable to adapt to the specific task of hate speech detection in the Yemeni dialect
- Output: The last hidden state, which is a sequence of contextualized vector representations for each input token

#### *Shared Bi-GRU with Attention Layer*

The output from MarBERT is passed into a shared bidirectional GRU (Bi-GRU), which is reponsible of processing the sequence in both forward and backward directions to capture context from both sides. It is called "shared" because it learns general sequential patterns applicable across different Arabic dialects. This layer is followed by an attention layer, which automatically assigns higher weights to the most important tokens in the tweet (e.g., offensive or hateful terms).

#### *Private Bi-GRU with Attention Layer*

In parallel, the same MarBERT output is also passed into a private bidirectional GRU with its own attention layer. This component has its own independent weights (not shared with the first Bi-GRU) and is designed to capture Yemeni dialect-specific features that may differ from other Arabic dialects. Its attention layer focuses on dialectal cues relevant to hate speech detection

#### *Concatenation*

The output vectors from the shared attention layer and the private attention layer are concatenated (joined together) into a single, richer feature vector. This combines general Arabic knowledge with dialect-specific knowledge. vii. Dropout Layer A dropout layer with a rate of 0.3 is applied to the concatenated vector. This randomly disables 30% of the neurons during training, which helps prevent overfitting and improves generalization to unseen tweets.

#### *Classification Layer*

The final vector is passed to a classification layer (a fully connected layer with softmax activation). This layer outputs two probabilities: Hate: Probability that the tweet contains hate speech, and Normal: Probability that the tweet is normal (non-hateful). The model predicts the class with the higher probability. The key innovation of this model is the dual Bi-GRU architecture with shared and private components: (1) The shared Bi-GRU learns patterns that are common across all Arabic dialects. For example, general offensive words or aggressive sentence structures. (2) The private Bi-GRU learns patterns that are unique to the Yemeni dialect. For example, dialect-specific slang, curse words, or expressions that might not be offensive in other dialects but are hateful in the Yemeni context. By combining both, the model can leverage general Arabic knowledge from the pretrained MarBERT, capture universal sequential patterns via the shared Bi-GRU, adapt to the low-resource Yemeni dialect via the private Bi-GRU, and focus on the most important words using attention mechanisms. This makes the model particularly suitable for hate speech detection on dialectal Arabic social media data, where labeled data is scarce and dialectal differences matter.

To provide a structured overview of the proposed framework, Algorithm 1 summarizes the complete workflow of the dialect-aware hate speech detection model. The framework initiates by gathering and preparing Arabic social media postings to remove extraneous elements and standardize textual inputs. The sanitized text is then tokenized and encoded with the MARBERT transformer model to produce contextualized semantic representations. The representations are processed via two complimentary Bidirectional Gated Recurrent Unit (Bi-GRU) branches: a shared branch aimed at identifying broad Arabic hate speech patterns across dialects and a private branch focused on acquiring linguistic characteristics specific to the Yemeni dialect. Attention techniques are utilized to highlight the most salient contextual features derived from both branches. The resultant feature vectors are amalgamated by concatenation, regularized with dropout, and subsequently forwarded to a softmax classification layer to ascertain whether the input text comprises hate speech or standard material. The model parameters are ultimately optimized with the AdamW optimizer and assessed by stratified 10-fold cross-validation employing several performance measures.

As illustrated in Algorithm 1, the proposed framework combines transformer-based contextual representation learning with dialect-aware sequential modeling and attention-based feature selection. The shared-private architecture enables the model to simultaneously learn transferable linguistic knowledge from Arabic dialects and dialect-specific hate speech indicators, while the attention mechanism improves the model's ability to focus on discriminative textual cues. This design enhances the detection of both explicit and implicit hate speech expressions and contributes to robust performance across multiple Arabic dialect datasets, particularly for the low-resource Yemeni dialect considered in this study.

### The Ye-HS Dataset Specification

In this work, a dataset for hate speech on the X platform for the Yemeni dialect called (Ye-HS), containing 4572 tweets and their comments, was created. Collecting and annotating a dataset to train the proposed model for automatic hate speech detection is essential in combating hate speech in Yemen, given the lack of a pre-existing Yemeni dataset. The collected tweets were not limited to a specific topic but encompassed several aspects, including political, religious, social, sports, and sexual. The tweets were annotated as either hate or normal. The classification was based on the following policy:

- Normal: Clean tweets that do not contain any offensive language, hate speech, profanity, insults, threats, or bullying.
- Hate: Tweets that contain offensive and hurtful language, hate speech, profanity, insults, threats, or bullying.

**Algorithm 1** A Dialect-Aware Deep Learning Framework for Hate Speech Detection in Yemeni Arabic Social Media Text

**Require:** Raw Arabic social media dataset $D$
**Require:** Pretrained MARBERT model $M$
**Require:** Maximum sequence length $L = 64$
**Require:** Number of folds $K = 10$
**Ensure:** predicted label $\hat{y}_i \in$ (Hate, Normal)

Collect Yemeni Arabic tweets and comments from the X platform
Manually annotate each sample as *Hate* or *Normal*
Merge and harmonize labels in external Arabic dialect datasets into binary classes
Remove duplicate, empty, or irrelevant samples
**for** each text sample $x_i \in D$ **do**
    Remove URLs, mentions, hashtags, emojis, numbers, and non-Arabic symbols
    Normalize Arabic text and remove extra spaces
    Tokenize $x_i$ using the MARBERT tokenizer
    Generate $input_ids_i$ and $attention_mask_i$ with maximum length $L$
**end for**
Split $D$ using stratified $K$-fold cross-validation
**for** each fold $k = 1$ to $K$ **do**
    Divide data into training set $D^k_{train}$ and testing set $D^k_{test}$
    Initialize MARBERT, shared Bi-GRU, private Bi-GRU, attention layers, dropout, and classifier
    Freeze lower MARBERT layers and fine-tune upper layers
    **for** each training epoch **do**
        **for** each mini-batch $B \subset D^k_{train}$ **do**
            Pass $input_ids$ and $attention_mask$ into MARBERT
            Obtain contextual token embeddings
            Pass $H$ into the shared Bi-GRU branch
            Apply attention to obtain shared Arabic features
            Pass $H$ into the private Bi-GRU branch
            Apply attention to obtain Yemeni dialect-specific features
            Concatenate shared and private representations
            Apply dropout regularization
            Predict class probabilities using softmax:
            Compute weighted cross-entropy loss
            Update trainable parameters using AdamW optimizer
        **end for**
    **end for**
    Evaluate the trained model on $D^k_{test}$
    Compute Accuracy, Precision, Recall, F1-score, Macro-F1, Weighted-F1, AUC, and Confusion Matrix
**end for**
Average evaluation results across all folds
Return final classification performance and predicted labels

Table 5 gives examples of both classes in the Ye-HS dataset.

In addition, four other available datasets in different Arabic dialects were used, namely, the Levantine dialect dataset (L-HSAB)[20], the Tunisian dialect dataset (T-HSAB)[27], the Egyptian dialect dataset (EAHS)[40], and the Levantine dialect misogyny dataset[32]. These four datasets have been manually annotated to align with the existing categories in the Ye-HS: either hate or normal, although they are all unbalanced across categories. The L-HSAB and T-HSAB datasets originally had three categories (hate, abusive, and normal), which were merged into two classes, hate and normal. The dataset also contained three classification columns: the (TARGET) column indicating whether the target was a woman, the MISOGYNY column indicating whether the text contained hate speech, and the CATEGORY column indicating the type of hate speech directed at women. The TARGET and CATEGORY columns were deleted here to standardize the evaluation process and ensure fair benchmarking against the proposed dataset. In the MISOGYNY column, the word "ACTIVE" was replaced with "hate," and the word "NONE" was replaced with "normal" to match the Ye-HS classification. The EAHS dataset, originally classified into five categories, was modified by replacing the NEUTRAL category with normal and the remaining hate categories with hate.

| Tweet text | Classification |
|---|---|
| لعنتي عليك ياست الناس انت امريكي يا راداد وهينكل يمني الاصل منك عند اصحابك الموت لامريكا<br>Damn you, you've made people despair! You're American, Raddad, and Henkel is Yemeni, more authentic than you. Death to America, according to your friends! | Hate |
| يا دحباشي لا تتعب نفسك محد صدقك لو تخرج من جلدك<br>Oh Dahbashi, don't tire yourself out, no one believes you even if you turn your skin off. | Hate |
| هذا ممثل اشرف وانزة منك بكتير لا يحشر نفسة في امور السياسة و الصورة تا كد ان مثل هذاالشخص وغيرة الملايين من طالبين الله في مناطق الجنوب مرحب بهم عالعين والرحب والسعة انته لودخلت عدن باتدخلها مع رولة الريخ وملابس البويا ونشاطك علامنصة لن يكون الافي في دهاليز الضلام ولابايقع لك دعس<br>His representative is far more honorable and dignified than you. He doesn't involve himself in political matters. Rest assured that this person, and millions like him seeking a livelihood in the southern regions, are welcome with open arms. If you enter Aden, you'll enter with paint tools and paint-stained clothes, and your activity on the platform will be confined to the shadows, and you will be beaten. | Hate |
| بالمختصر انت زنبيل<br>In short, you are a follower soldier | Hate |
| الجناح العسكري الاخواني قتلوا النظافة . لتبقى وساختهم ... رحمة الله تغشاها<br>The military wing of the Muslim Brotherhood killed cleanliness so that their filth would remain. | Hate |
| انتهى وقت اللعب الآن وقت جني الاستحقاقات وانتزاعها من أحسن له الحق ومن أساء عليه ان يتقبل<br>The game is over; now is the time to reap the rewards and seize what is rightfully yours. Those who have done good deserve it, and those who have done wrong must accept the consequences. | Normal |
| نتمنى الا تنساقوا إلى تجارب الإدانة، استمروا في إصدار القرارات، ودعوا الحكومة تنفذ<br>We hope you will not be drawn into the cycle of condemnation; continue issuing resolutions, and let the government implement them. | Normal |

**Table 5.** Example tweets classified as Normal and Hate in the Ye-HS dataset

Table 6 summarizes all datasets used in this research and the number of samples in each category.

| Ref. | Dataset | Source | Dialect | Number of hate tweets | Number of normal tweets | Total |
|---|---|---|---|---|---|---|
| Newly Proposed | Ye-HS | X platform | Yemeni | 2821 | 1751 | 4572 |
| [27] | T-HSAB | Various social media platforms | Tunisian | 2207 | 3832 | 6039 |
| [20] | L-HSAB | X platform | Levantine | 2235 | 3111 | 5446 |
| [40] | EAHS | X platform | Egyptian | 6034 | 2135 | 8169 |
| [32] | Let-mi | X platform | Levantine | 2584 | 2654 | 5241 |

**Table 6.** Summary of Datasets Used in This Study

Since raw data are not in a format usable by automated classification models, data preprocessing is crucial for preparing and cleaning the raw data, making it suitable for training the proposed model and improving accuracy and efficiency. This step involves processing sections containing noise, typographical errors, duplicate characters, missing data, or distorted information, replacing, deleting, or modifying them as needed. Techniques used for data cleaning in this work include removing links, mentions, hashtags, and non-Arabic symbols such as English letters, numbers, and emojis, as well as standardizing spacing. Preprocessing techniques are usually used for natural language processing. Here, the MarBERT converter was used to extract features from the tweet content, then a Bi-GRU was used to improve representation by extracting sequential features and learning public and private features, and finally, an attention mechanism was used to attend to important features. A 10-fold cross-validation was used to avoid overfitting and improve classification performance.

## Experiments and Results

The proposed model was applied primarily to the core dataset in our work, which is the Yemeni dialect dataset Ye-HS, which was collected specifically for this work. Then the proposed model was also applied to other datasets in other dialects for benchmarking and evaluating the efficiency and effectiveness of the model. The subsequent sections reports the experimental settings that was adopted and the results on the different datasets.

### *Experimental Settings*

We conducted multiple trials of the proposed model with different settings to identify the optimal settings that would yield strong and reliable results. The final settings for the proposed model, which provide the best performance, are shown in Table 7.

| **Aspect** | **Experiments** |
|---|---|
| Running Platform | [a]Google Colab + Google Drive |
| Framework | [b]PyTorch + [c]HuggingFace Transformers |
| Pretrained Model | [1]MARBERT (UBC-NLP/marbert) |
| Architecture | MARBERT + Shared/Private Bi-GRU + Attention |
| Optimizer | AdamW |
| Learning Rate | $2\ x\ 10^{-5}$ |
| Loss Function | Cross-Entropy Loss (with class weights) |
| Batch Size | 4 |
| Gradient Accumulation | 4 steps |
| Effective Batch Size | 16 |
| Number of Epochs | 5 |
| Max Sequence Length | 64 |
| Hidden Size (GRU) | 256 |
| Number of Folds | 10 (Stratified K-Fold) |
| Regularization | Dropout (0.3) |
| Mixed Precision | FP16 (Autocast + GradScaler) |
| Evaluation Metrics | F1-score, Accuracy, Precision, Recall, Confusion Matrix |

[1] https://huggingface.co/UBC-NLP/MARBERT
[a] https://colab.research.google.com
[b] https://pytorch.org/
[c] https://huggingface.co/transformers/

**Table 7.** Experimental Settings

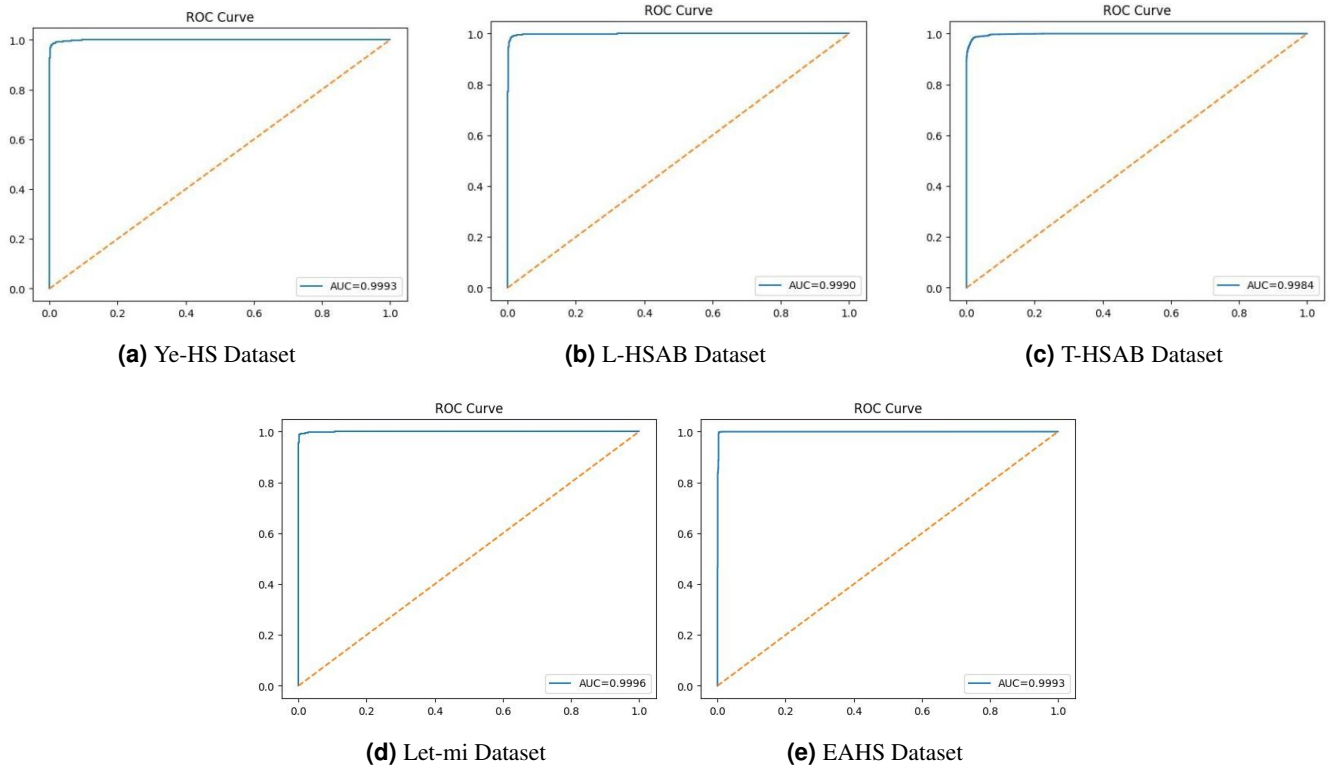
## Results of the Proposed Model

In this section, we present the results of applying the proposed model to the Yemeni dialect dataset (Ye-HS) collected for this study, as well as to other Arabic dialect datasets. Table 8 reports the performance results using the common performance evaluation measures adopted in the literature, including accuracy, precision, recall, and the F1-Score.

| Datasets | Accuracy | precision | Recall | F1-score | F1 Std | F1 Macro | F1 Weighted | AUC |
|---|---|---|---|---|---|---|---|---|
| Ye-HS | **0.986** | **0.987** | **0.991** | **0.989** | **0.026** | **0.985** | **0.986** | **0.999** |
| L-HSAB | 0.987 | 0.983 | 0.985 | 0.984 | 0.037 | 0.987 | 0.988 | 0.999 |
| T-HSAB | 0.980 | 0.969 | 0.978 | 0.974 | 0.069 | 0.979 | 0.980 | 0.998 |
| Let-mi | 0.992 | 0.994 | 0.991 | 0.992 | 0.021 | 0.992 | 0.992 | 0.999 |
| EAHS | 0.998 | 0.998 | 0.999 | 0.999 | 0.003 | 0.997 | 0.998 | 0.999 |

**Table 8.** Performance Evaluation of the Proposed Model on the Ye-HS Dataset and Other Datasets

The proposed model on the Ye-HS dataset achieved an accuracy of 0.986, an F1-score of 0.989, a recall of 0.991, and a precision of 0.987, which is a very high and outstanding performance in detecting Yemeni hate speech. The model, using the EAHS dataset in Egyptian Arabic, achieved an accuracy of 0.998, an F1 Score of 0.999, a recall of 0.999, and a precision of 0.998, the best results achieved with the model proposed in this work. It was followed by an experiment using the Let-mi dataset in the Levantine dialect, which is known for its anti-misogynist stance. This resulted in an Accuracy score of 0.992, an F1-score of 0.992, a Recall score of 0.991, and a Precision score of 0.998. The performance achieved on the L-HSAB dataset in




**(a)** Ye-HS Dataset **(b)** L-HSAB Dataset **(c)** T-HSAB Dataset

**(d)** Let-mi Dataset **(e)** EAHS Dataset

**Figure 2.** Area Under Curve (AUC) of the Proposed Model on the Ye-HS Dataset and Other Datasets.

the Levantine dialect followed, with an Accuracy score of 0.987, an F1-score of 0.984, a Recall score of 0.985, and a Precision score of 0.983. The performance of the proposed model on the Tunisian dialect T-HSAB dataset was relatively weaker than on the other datasets in this work, achieving an Accuracy score of 0.980, an F1-score of 0.974, a Recall score of 0.978, and a Precision score of 0.969. However, all experiments yielded excellent results in detecting hate speech patterns across Yemeni, Egyptian, Levantine, and Tunisian dialects. These results demonstrate the model's high capacity to understand the Yemeni dialect and other Arabic dialects.

## Result Analysis and Discussion

In this work, 10-fold cross-validation was used to ensure consistent results, reduce bias and effects arising from data distribution, and ensure that the evaluation was performed on the entire dataset. In the application to the Ye-HS dataset, we observed a slight performance variation in fold1. However, performance began to stabilize and gradually improve in subsequent folds, approaching 0.99 and 1. This suggests that the first fold contained complex or ambiguous data samples, and that the model generalized better in subsequent folds and was not affected by changes in the training data. This is evidenced by the stability of its performance and the convergence of results. The average F1 score achieved was 0.989 across all folds, with a standard deviation of 0.03, influenced by the first fold. This indicates the overall stability of the model's performance on the Ye-HS dataset, its ability to distinguish between the "hate" and "normal" categories, and a balance between the recall and the precision, avoiding both overprediction and over-caution. We also observe that the model achieved 0.985 in the F1 macro and 0.986 in the F1 weighted. This convergence indicates that the model is not biased towards either class and performs efficiently across both. Furthermore, the area under the curve (AUC) was 0.999 across the various datasets, including Ye-HS, as shown in Figure 2.

This is a very high and optimal value, meaning the model perfectly distinguishes between the two classes and accurately predicts them. In the application to the L-HSAB dataset, we observed an F1 score of 0.984. This indicates a high balance between Precision and Recall, with their values being nearly equal. Consequently, the model is balanced, and its performance is stable. Furthermore, the AUC value of 0.999 indicates the model's strong ability to discriminate between the two classes, as shown in Figure 2.b. However, the F1 std value of 0.037 suggests some fluctuation in performance between folds, particularly in the first fold, where the F1 value was 0.86. It then stabilized in subsequent folds between 0.97 and 1, making the model sensitive to the data distribution. Nevertheless, its overall performance is stable.

In the application to the T-HSAB dataset, we observed that although the F1 score (Mean = 0.974, AUC = 0.998) was high,

indicating the model's ability to discriminate between categories, as shown in Figure 2.b. Some fluctuation in performance was observed. The F1 standard value appeared at 0.069 because the F1 value in the first fold was 0.75, then increased to 0.99 in the subsequent folds. This means the model performed poorly on this dialect compared to other experiments in this study. This explains why it performed slightly worse in some parts due to the variety of dialects and levels of hate in the dataset, and perhaps the heterogeneity of the Tunisian dialect data, while performing strongly in other parts. In the application to the Let-Mi dataset, we observed that the F1 score of 0.992 was high, as was the AUC of 0.999, as shown in Figure 2.d. This indicates a strong ability to distinguish between the two categories almost completely. The Std value of 0.021 suggests that the model rarely makes mistakes. This is likely due to the relatively balanced distribution of categories in this dataset, as well as the clear and direct language used, given that it is specifically a sample of misogynistic. The model performed excellently from the very first fold, starting with an F1 score of 0.91 and then ranging between 0.99 and 1 in subsequent folds. This also reflects the model's sensitivity to data segmentation. In the application to the EAHS dataset, we observed significantly stronger performance than in all previous datasets, achieving an F1 score of 0.998 and an AUC of 0.999, as shown in Figure 2.e. This indicates a high balance between Recall and Precision, enabling the model to distinguish between categories with near-perfect accuracy. The model accurately captures all instances of hate. Furthermore, the Std score of 0.003 confirms consistent performance across all folds, with no impact from data segmentation.

To conclude, experiments with different dialects, the performance of the proposed model starts relatively weak in the first fold, then gradually improves in subsequent folds until it reaches optimal performance. This is because, in the first fold, the model begins to learn patterns and recognize the linguistic complexities of the dialect, understanding the semantic meaning of sentences and the implicit contexts of hate speech, outright hate, and sarcastic, mocking, and derogatory language. Then, in subsequent folds, the model gradually learns until it is fully trained. However, this variation in performance across the first two folds leads to classification errors because the model is still learning, resulting in a higher standard deviation in F1 that affects the overall performance. Nevertheless, it can be said that all the experiments were very successful and achieved high, ideal performance. The optimal results obtained from applying the proposed model are due to its utilization of the hybrid model's features. The model combines MarBERT, which can understand context and the implicit meaning of dialects, with Bi-GRU, which captures sequential word relationships, and an attention mechanism that focuses on the most important features and words. Furthermore, the use of class weights yielded high accuracy and improved feature representation, consequently improving classification across all datasets used in this study. This makes the proposed model suitable for generalization and practical application to the Yemeni dialect and other Arabic dialects. We also applied the t-test to the model results across all datasets used in this study, from the F1 values on the test set, to assess the model's strength. It turned out that there was no significant difference between the results, and because they were all close and performed at the same level for the model, this proves the effectiveness and strength of the proposed model across all Arabic dialects and can be generalized.

### Error Analysis

The samples predicted incorrectly when the proposed model was applied to all datasets across the different Arabic dialects were extracted to analyze the reasons for the errors. We found that the model made mistakes in 61 samples when applied to the Ye-HS, which is the basis of this study. Of these, 25 were false negatives that should be classified as hate, but the model misclassified them as normal. Additionally, 31 were false positives classified as normal, but the model was over-sensitive and misdiagnosed them as hate. Figure 3.a presents the confusion matrix, showing the number of false hate-speech predictions and the number of instances where actual hate speech was not detected on the Ye-HS. Table 9 shows some of the samples where the model made mistakes on the Ye-HS dataset.

We concluded that the incorrect classification of the Ye-HS dataset in the hate category, with FN=25, stems from text ambiguity and the lack of clear contextual information, as some texts may carry more than one possible meaning. Variations in expression within the Yemeni dialect itself, as well as across its local dialectal forms, also contribute to this problem. The model may fail to recognize implicit hate, such as الشماليين ينظرون لكل مسؤول جنوبي نظرت احتقار, or sarcasm, such as الدكتور رشاد فاصولياء, or indirect allusions that require external contextual knowledge, such as خفف العيار وهات كلام منطقي. Furthermore, the model may interpret some sentences as ordinary political discourse rather than hate speech. In addition, long sentences may cause the model to lose the implicit context and reduce its ability to focus on hateful cues embedded within the text. The reasons for the incorrect classification of the Normal category, with FP=31, are mainly related to the presence of samples with a strong tone that appears hateful, such as ميش شغلك انت. Moreover, some words are frequently associated with hateful samples, such as حوثي, إسرائيلي, and عميل. Consequently, the model may become confused and classify sentences containing these words as hateful, even when the overall context is not hate speech. Some samples also fall between the hate and normal categories, forming borderline cases that are difficult for the model to predict correctly, such as ضحكتني وانا ضايع. Finally, because the dataset is imbalanced and the number of hate samples exceeds the number of normal samples, the model tends to classify uncertain cases as hate.

We also found that the model made 116 errors when applied to the L-HSAB. Of these, 48 were false negatives classified as

| Text | True | Pred |
|---|---|---|
| ميش شغلك انت<br>Don't interfere. | normal | hate |
| اخي العزيز والله ان الشماليين ينظرون لكل مسؤول جنوبي نظرت احتقار الجنوب يتسع للجميع<br>My dear brother, I swear by God that northerners look down upon every southern official with contempt; the south is open to everyone. | hate | normal |
| كانت باتتحرر زمان لكن الامريكي والبريطاني والفرنسي اخوان الرافضه وقفوا مع إيران اللضغط على دول التحالف والتحرير جاي لامحاله وبيننا الأيام الحديده<br>Al Hudaydah was about to be liberated, but the Americans, British, and French—the brothers of the Rafidites—sided with Iran to pressure the coalition countries. Liberation is inevitable, and time will tell. | hate | normal |
| انت والامريكي والإسرائيلي<br>You, the American, and the Israeli | normal | Hate |
| الدكتور رشاد فاصولياء<br>Dr. Rashad beans | hate | Normal |
| خفف العيار وهات كلام منطقي<br>Cut down on drugs and speak sensibly. | hate | normal |

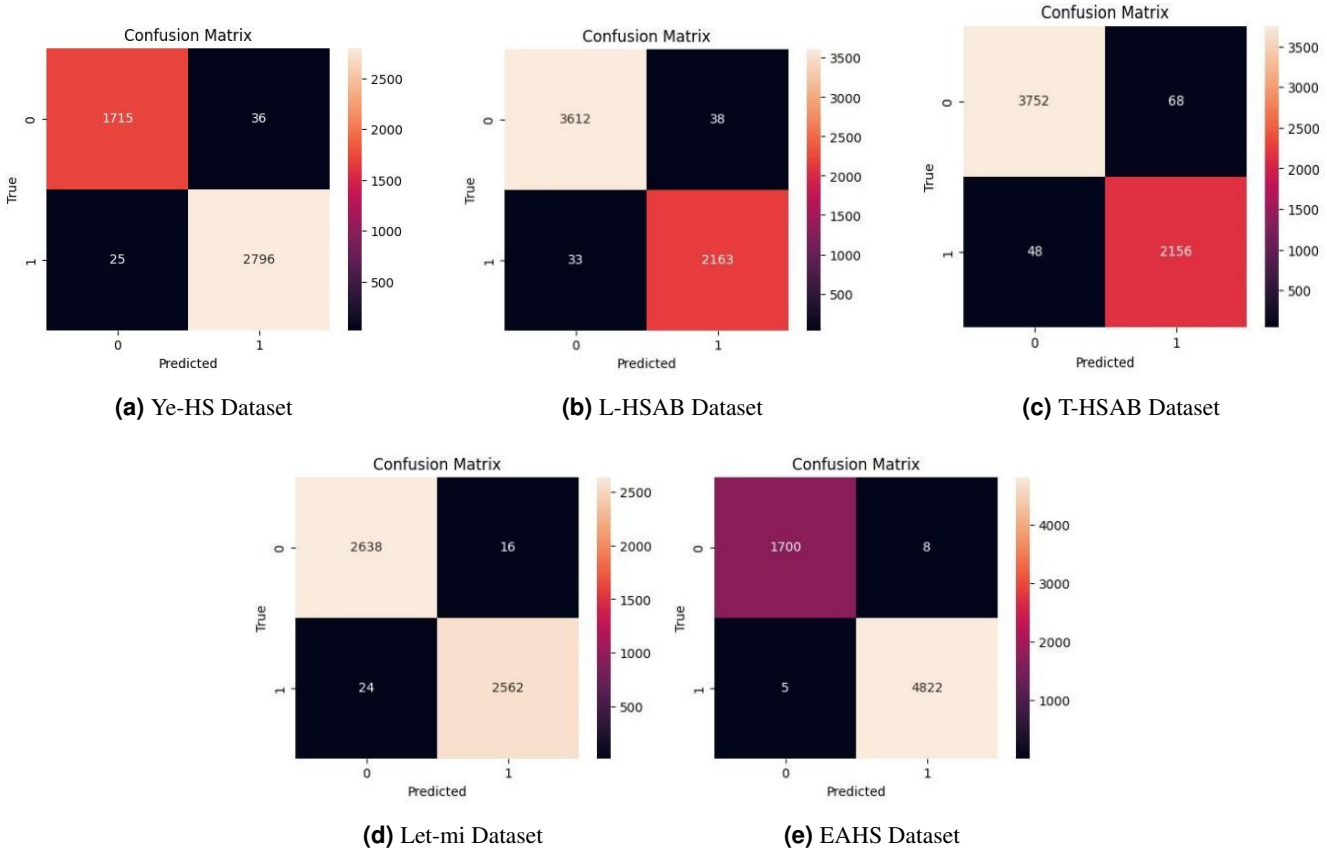
**Table 9.** Samples of some Classification Mistakes Made by the Proposed Model on Ye-HS dataset

hate but incorrectly treated as normal, and 68 were false positives classified as normal but were overly sensitive and treated as hate. Figure 3.b displays the confusion matrix, showing the number of false hate-speech predictions and the number of instances where actual hate speech was not detected on the L-HSAB. The number of samples that the model misidentified when applied to the T-HSAB was 71. Of these, 38 were false negatives classified as hate but were misidentified as normal, and 33 were false positives classified as normal but were oversensitive and misidentified as hate. Figure 3.c shows the confusion matrix, illustrating the number of false hate-speech predictions and the number of instances in which actual hate speech was not detected on the T-HSAB. The model made 40 errors when applied to Let-mi, including 24 false negatives classified as hate but misclassified as normal, and 16 false positives classified as normal but misclassified as hate. Figure 3.d shows the confusion matrix, illustrating the number of false hate-speech predictions and the number of instances where actual hate speech was not detected on Let-mi. The model made 13 errors when applied to EAHS, including 5 false negatives classified as hate but misclassified as normal, and 8 false positives classified as normal but over-sensitive and misclassified as hate. Figure 3.e illustrates the confusion matrix, showing the number of false hate-speech predictions and the number of instances where actual hate speech was not detected on EAHS.

We conclude that the model in the Tunisian dialect dataset has difficulty detecting implicit hate speech, such as الشعب التونسي اخج شعب, sarcasm, and complex and diverse cultural and dialectal expressions, especially since there are rare words that may not be mentioned during the training samples. It also tends to classify neutral content as hate, such as انا وانت مستحيل نتفاهم. However, there were few errors in applying the model to the EAHS dataset, where it mixed criticism and implicit hatred in some samples, such as الشعب الفرنساوي ده زي الفلاحين عندنا and sarcasm such as صباح التوهان على عيونك. Compared to the other datasets, it performed exceptionally well and was very consistent and stable. We can say that the proposed model depends on the impact of FN errors on the recall and of the FP errors on the precision, and that its performance depends on the nature of the diverse dialects present in the dataset, its complexity, ambiguity, and the semantic meaning of the texts.

## Comparison with Existing Works

We reviewed prior studies[13],[27],[32], and[40], which served as the foundation for the datasets used in this work, given their direct relevance to social media text classification. The models proposed in these studies were implemented and evaluated. In addition, we examined recent, high-performing studies[38],[54], and[10] that used the same datasets used in this research. The models reported in these works, particularly those achieving the best performance in hate speech detection, were also implemented and compared against the proposed model. These models collectively served as baseline approaches for evaluating the effectiveness of the proposed framework. During implementation, we followed the configurations and methodologies described in the original studies as closely as possible, preserving their overall structure and experimental setup. In cases where specific details regarding hyperparameters or training settings were unavailable, widely accepted standard configurations were adopted while maintaining consistency with the original methodology. Furthermore, based on[20] and[27], we implemented traditional machine learning




**(a)** Ye-HS Dataset **(b)** L-HSAB Dataset **(c)** T-HSAB Dataset

**(d)** Let-mi Dataset **(e)** EAHS Dataset

**Figure 3.** Confusion Matrices of the Proposed Model on the Ye-HS Dataset and Other Datasets.

models, including Support Vector Machine (SVM) and Naïve Bayes (NB), using various N-gram configurations and TF-IDF feature representations as described in the respective studies. Among these models, the SVM presented in Table 10 achieved the best performance.

| Ref | Model | Dataset | Accuracy | Precision | Recall | F1 Macro | F1 Weighted |
|---|---|---|---|---|---|---|---|
| [20,27] | SVM(uni, bi, tri df: 2) | Ye-HS | 0.6987 | 0.6802 | 0.6613 | 0.6656 | 0.6904 |
| | NB(uni, bi, df: 3) | | 0.7031 | 0.7155 | 0.6354 | 0.6331 | 0.6709 |
| | SVM(uni, bi, df: 2) | T-HSAB | 0.7828 | 0.7773 | 0.7418 | 0.7521 | 0.7754 |
| | NB(uni, bi, tri df: 3) | | 0.7728 | 0.7930 | 0.7110 | 0.7238 | 0.7549 |
| | SVM(uni, df: 2) | L-HSAB | 0.8017 | 0.7958 | 0.7734 | 0.7812 | 0.7978 |
| | NB(uni, df: 3) | | 0.8034 | 0.8231 | 0.7567 | 0.7710 | 0.7924 |
| | SVM(uni, bi, tri df: 3) | Let-mi | 0.8444 | 0.8469 | 0.8439 | 0.8439 | 0.8540 |
| | NB(uni, df: 3) | | 0.8331 | 0.8360 | 0.8325 | 0.8326 | 0.8327 |
| | SVM(uni, bi, df: 2) | EAHS | 0.9009 | 0.8793 | 0.8590 | 0.8683 | 0.8995 |
| | NB(uni, df: 3) | | 0.8678 | 0.8840 | 0.7658 | 0.8012 | 0.8560 |

**Table 10.** Results of Baseline Machine Learning Models on all Datasets

Table 10 shows that NB underperformed SVM across all datasets and n-gram configurations. The results further indicate that no single configuration guarantees optimal performance across all dialects. This confirms that each dataset has a different impact on the model's efficiency. The best performance, represented by the F1 score, is reported on the EAHS dataset, followed by Let-mi and then L-HSAB. However, the models performed poorly on the T-HSAB dataset and further weakly on the Yemeni dialect (Ye-HS).

From the study[32], we selected the best-performing original AraBERT model for identifying the Misogyny class and obtained the results shown in Table 11.

It is shown that the model achieves the best performance on the Egyptian dataset, as in the previous model, followed by the

| Ref. | Model | DATASET | Accuracy | Precision | Recall | F1 (Macro) |
|---|---|---|---|---|---|---|
| 32 | AraBERT | Ye-HS | 0.847 | 0.839 | 0.834 | 0.837 |
| | | T-HSAB | 0.836 | 0.825 | 0.819 | 0.822 |
| | | L-HSAB | 0.902 | 0.894 | 0.899 | 0.896 |
| | | Let-mi | 0.891 | 0.895 | 0.890 | 0.891 |
| | | EAHS | 0.953 | 0.948 | 0.929 | 0.938 |

**Table 11.** Results of the AraBERT model on all Datasets

Let-mi and L-HSAB datasets, while performance decreases on the T-HSAB and Ye-HSAB datasets. Based on the study in[40], the AraBERT v0.2 Large X platform model, which performed best in the original study, was used, and the implementation results are shown in Table 12

| Ref. | Model | Dataset | Validation | | Test | | | |
|---|---|---|---|---|---|---|---|---|
| | | | F1 | Accuracy | Precision | Recall | F1 (Macro) | F1 (Weighted) |
| 40 | AraBERT v0.2 Large X platform | Ye-HS | 0.85 | 0.863 | 0.860 | 0.848 | 0.853 | 0.862 |
| | | T-HSAB | 0.38 | 0.634 | 0.402 | 0.634 | 0.388 | 0.492 |
| | | L-HSAB | 0.918 | 0.923 | 0.923 | 0.923 | 0.918 | 0.923 |
| | | Let-mi | 0.92 | 0.926 | 0.927 | 0.926 | 0.926 | 0.926 |
| | | EAHS | 0.97 | 0.974 | 0.970 | 0.96 | 0.966 | 0.974 |

**Table 12.** Results of the AraBERT v2.0 Model on all Datasets

The results presented in Table 11 showed that AraBERT v2.0 performed best on the Egyptian dataset, followed by Let-mi and L-HSAB, and performed well on Ye-HS compared to the previous models. However, the model performed very poorly on the T-HSAB dataset, indicating a lack of compatibility with this dialect, possibly due to complex linguistic features. Following[38], we implemented AraBERT multitasking (MLT) and single-tasking (STL), which outperformed the original MarBERT and MarBERT V2.0 in the original study for multitasking and single-tasking learning as reported in Table 13.

| Ref. | Model | Dataset | Test | |
|---|---|---|---|---|
| | | | F1 | F1 (Macro) |
| 38 | AraBERT(MTL) | Ye-HS | 0.828 | |
| | | T-HSAB | 0.845 | |
| | | L-HSAB | 0.894 | |
| | | Let-mi | 0.905 | 0.901 |
| | | EAHS | 0.960 | |
| | AraBERT(STL) | All datasets | | 0.883 |

**Table 13.** Results of the AraBERT Single-tasking and Multitasking Model on all Datasets

The AraBERT multitasking model outperformed the single-tasking model and performed well on the Egyptian dialect (EAHS) but poorly on the Yemeni dialect (YE-HS).

Following[54], we applied all traditional machine learning models included in the study, Linear SVC, Multinomial NB, RF, and LR. The results are reported in Table 14. The Synthetic Minority Over-sampling Technique (SMOTE) was used to study the impact of over-sampling on the learning process of these traditional machine learning models.

It is shown that the impact of SMOTE varies across models; some improve with its use, while others are negatively affected and experience a decline in performance. Conversely, some models perform well without SMOTE. The Egyptian dataset achieved the best performance, while the Yemeni dataset (YE-HS) performed the worst. Performance varied across the T-HSAB, L-HSAB, and Let-mi datasets. Overall, the Ye-HS dataset is more challenging according to the reported results. Furthermore, based on the same study, the DziriBert model was implemented, and the results are shown in Table 15.

As indicated in Table 13, the DziriBERT model performed best with the Egyptian dialect, followed by L-HSAB and Let-mi. The performance was lower with T-HSAB and Ye-HS. Following[10], we implemented the model that performed best in the original study: GWO + RF + CNN Attention, after preprocessing with MARBERT + TF-IDF + PCI. The model results are shown in Table 16.

| Ref. | Model | Datasets | Without SMOTE | | | With SMOTE | | |
|---|---|---|---|---|---|---|---|---|
| | | | F1 | Accuracy | Precision | F1 | Accuracy | Precision |
| [54] | LR | Ye-HS | 0.672 | 0.695 | 0.691 | 0.682 | 0.679 | 0.689 |
| | Linear SVC | | 0.707 | 0.709 | 0.706 | 0.676 | 0.673 | 0.681 |
| | Random Forest | | 0.653 | 0.670 | 0.659 | 0.665 | 0.660 | 0.664 |
| | Multinomial NB | | 0.665 | 0.705 | 0.728 | 0.689 | 0.686 | 0.697 |
| | LR | T-HSAB | 0.763 | 0.784 | 0.807 | 0.751 | 0.748 | 0.758 |
| | Linear SVC | | 0.778 | 0.784 | 0.781 | 0.737 | 0.733 | 0.745 |
| | Random Forest | | 0.767 | 0.778 | 0.779 | 0.754 | 0.758 | 0.754 |
| | Multinomial NB | | 0.759 | 0.777 | 0.791 | 0.783 | 0.782 | 0.784 |
| | LR | L-HSAB | 0.782 | 0.798 | 0.819 | 0.780 | 0.779 | 0.781 |
| | Linear SVC | | 0.812 | 0.817 | 0.818 | 0.791 | 0.790 | 0.792 |
| | Random Forest | | 0.789 | 0.801 | 0.814 | 0.770 | 0.783 | 0.790 |
| | Multinomial NB | | 0.782 | 0.798 | 0.819 | 0.819 | 0.819 | 0.820 |
| | LR | Let-mi | 0.807 | 0.807 | 0.810 | 0.810 | 0.810 | 0.812 |
| | Linear SVC | | 0.824 | 0.824 | 0.825 | 0.819 | 0.819 | 0.819 |
| | Random Forest | | 0.793 | 0.793 | 0.794 | 0.797 | 0.797 | 0.798 |
| | Multinomial NB | | 0.814 | 0.814 | 0.814 | 0.811 | 0.811 | 0.811 |
| | LR | EAHS | 0.866 | 0.876 | 0.879 | 0.899 | 0.898 | 0.901 |
| | Linear SVC | | **0.913** | **0.915** | **0.914** | **0.917** | **0.917** | **0.917** |
| | Random Forest | | 0.865 | 0.871 | 0.867 | 0.871 | 0.872 | 0.871 |
| | Multinomial NB | | 0.837 | 0.856 | 0.873 | 0.909 | 0.907 | 0.913 |

**Table 14.** Results of the Traditional Machine Learning Models on all Datasets

| Ref. | Model | DATASET | F1 | Accuracy | Precision |
|---|---|---|---|---|---|
| [54] | DziriBERT | Ye-HS | 0.813 | 0.813 | 0.812 |
| | | T-HSAB | 0.817 | 0.819 | 0.817 |
| | | L-HSAB | 0.893 | 0.894 | 0.893 |
| | | Let-mi | 0.884 | 0.884 | 0.884 |
| | | EAHS | 0.954 | 0.955 | 0.954 |

**Table 15.** Results of the DziriBERT Model on all Datasets

Since the CNN+RF+GWO model mentioned in the study[10] used cross-validation with k=5, we implemented the proposed model in our study with the same settings to perform statistical analysis test fairly.

Table 16 showed that the best performance was on the Egyptian dataset, while it declined on the Ye-HS and T-HSAB datasets. In this model, 5-fold cross-validation was used to support performance stability. However, the model still suffers from sensitivity to the nature of the data, especially in texts with complex dialects.

Overall, the results obtained from the implemented baseline models highlight the importance of tailoring detection approaches to the specific characteristics of the data. They also demonstrate that dialectal diversity and linguistic variations significantly influence the effectiveness of hate speech detection systems. As illustrated in Figure 4, the proposed model surpasses all previous approaches, owing to its enhanced text representation capabilities and its effectiveness in capturing the linguistic complexities and dialectal variations across different Arabic dialects.

Figure 4 illustrates that the proposed model achieved outstanding performance across all datasets according to the standard evaluation metrics, particularly F1 Macro and Accuracy. The F1 Macro scores ranged from 0.979 to 0.997, reflecting the robustness and effectiveness of the proposed approach. In contrast, traditional machine learning models employing various N-GRAM configurations achieved considerably lower F1 Macro scores, ranging between 0.665 and 0.868. Although the application of SMOTE slightly improved the performance of some traditional models by addressing data imbalance, the improvement was limited and fell well short of the proposed model's performance. This limitation can be attributed to conventional machine learning approaches' inability to effectively capture complex linguistic context and semantic relationships.

Transformer-based models, including AraBERT, AraBERT MTL/STL, and AraBERT v0.2 Large for the X platform, demonstrated better performance than traditional machine learning approaches. However, the AraBERT v0.2 Large X platform model performed poorly on the Tunisian dialect. Similarly, the DziriBERT model achieved F1 Macro scores ranging from 0.813 to 0.954, while the hybrid CNN+RF+GWO model produced scores between 0.730 and 0.910. Despite these improvements, the

**Table 16.** Results of the GWO + RF + CNN Attention Model on all Datasets

| Ref. | Model | Dataset | F1-Score | Accuracy | Precision | Recall | F1 (Macro) | F1 (Weighted) |
|---|---|---|---|---|---|---|---|---|
| [10] | CNN Attention RF + GWO | Ye-HS | 0.780 | 0.783 | 0.780 | 0.780 | 0.770 | 0.780 |
| | | T-HSAB | 0.750 | 0.756 | 0.750 | 0.760 | 0.730 | 0.750 |
| | | L-HSAB | 0.840 | 0.841 | 0.840 | 0.840 | 0.830 | 0.840 |
| | | Let-mi | 0.780 | 0.779 | 0.780 | 0.780 | 0.780 | 0.780 |
| | | EAHS | 0.930 | 0.930 | 0.930 | 0.930 | 0.910 | 0.930 |

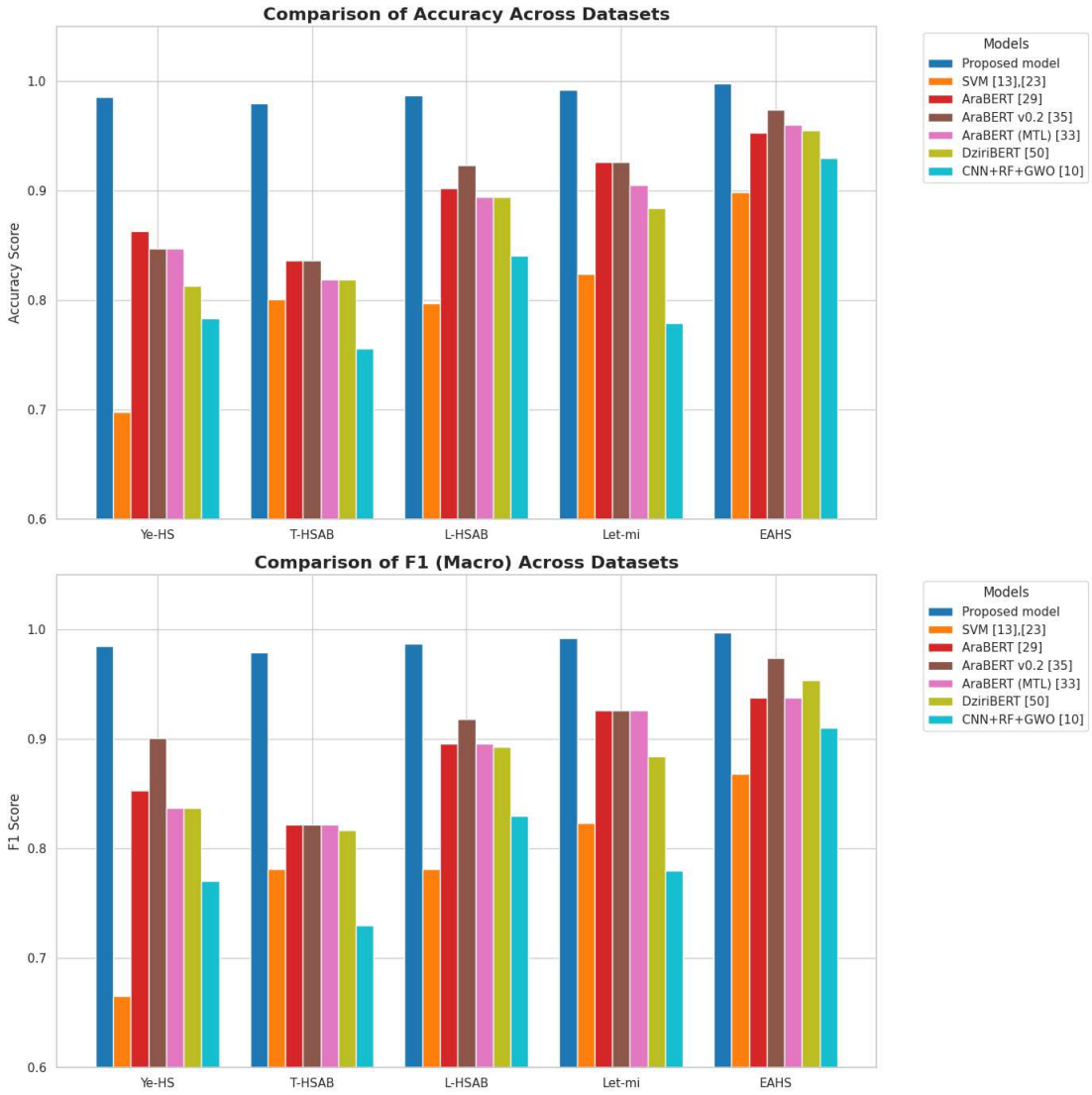



**Figure 4.** Comparison of the Proposed Model with the Best and Recent Existing Models on all Datasets.

proposed model consistently outperformed all comparative models, demonstrating superior capability in capturing linguistic representations and dialect-specific characteristics, and greater flexibility in handling imbalanced data across all Arabic dialects considered in this study.

Moreover, the adoption of 10-fold cross-validation in the proposed framework provides a more reliable and stable evaluation of model performance, thereby enhancing the credibility and robustness of the obtained results compared to models evaluated using a simple train-test split strategy.

## Statistical Significance Tests

A paired t-test was performed on the compared model using the same dataset, with F1 scores calculated for each fold. This process was repeated across all datasets, and the findings, presented in Table 17, reveal statistically significant differences for

every dataset ($p < 0.01$), with particularly high significance for the Let-Mi and Egyptian EAHS datasets. These results provide strong statistical evidence in favor of the proposed model, thereby reinforcing the credibility of the outcomes.

| Dataset | t-value | p-value | Significance |
|---|---|---|---|
| Ye-HS | 7.243 | 0.0019 | Significant ($p<0.01$) |
| T-HSAB | 7.168 | 0.0020 | Significant ($p<0.01$) |
| L-HSAB | 6.654 | 0.0026 | Significant ($p<0.01$) |
| let-mi | 11.697 | 0.00031 | Highly Significant |
| EAHS | 15.125 | 0.00011 | Highly Significant |

**Table 17.** Statistical Significance Test (t-test) between the proposed mode and the model in[10]

Furthermore, a t-test was conducted between the proposed model and the best models from previous works[20,27,32,40,38], and[54], based on F1-macro scores across all datasets, using the same model setup. The t-statistic and p-value were statistically significant in all cases, as shown in Table 16, confirming the superiority of the proposed model in performance over all other models. However, a high variance was observed in one dataset (T_HSAB) for the AraBERT v0.2 Large X platform model, indicating instability in its performance compared to the proposed model. This is obvious from the results shown in Table 18, which indicate the F1 score of 0.388 on the T_HSAB.

| Ref. | Model | t-value | p-value | Significance |
|---|---|---|---|---|
| [20,27] | SVM | 6.316 | 0.0032 | Statistically significant |
| [32] | AraBERT | 6.06 | 0.0037 | Statistically significant |
| [40] | AraBERT v0.2 Large X platform | 1.65 | 0.174 | No significant due to high variance. |
| [38] | AraBERT (MTL) | 4.83 | 0.0085 | Statistically significant |
| [54] | DziriBERT | 4.85 | 0.0083 | Statistically significant |

**Table 18.** Statistical Significance Test (t-test) between the proposed model and the Best and Recent Models on all Datasets

## Challenges and Future Work

Minor performance fluctuations were observed in the proposed model across validation folds for certain datasets, especially those involving Tunisian dialect data. This likely stems from data variability and linguistic noise. The model struggled with detecting implied meanings and indirect forms of hate speech, as well as subtle sarcasm and contempt. At times, it appeared to depend on recognizing specific keywords rather than grasping the broader contextual meaning. Its performance was only moderate when classifying political criticism and statements containing strong opinions or emotionally charged language. Additionally, capturing the linguistic diversity across Arabic dialects remains a significant challenge due to structural differences and the variety of ways a single concept can be expressed. These limitations highlight the need for expanded training data and larger datasets that include more intricate examples of sarcasm, political criticism, emotional expression, and diverse social contexts, particularly across various local dialects. Addressing these needs would improve the model's generalizability and support the development of robust techniques to better understand implicit meanings and indirect intentions in text.

## Conclusion

This study presented a proposed model for detecting hate speech in the Yemeni dialect. It was evaluated on several diverse datasets, encompassing the Yemeni dialect and other Arabic dialects, each with distinct linguistic and thematic patterns. We also conducted trials of models from previous studies on the same datasets as a baseline for comparison. The proposed model outperformed all other models, consistently achieving the highest scores on the core evaluation criteria across all datasets, demonstrating strong generalizability and the ability to handle diverse data. In general, the results confirm that the proposed model, based on MarBERT, Bi-GRU deep learning technology, and the attention mechanism, yields better and more accurate results for detecting Arabic hate speech. It provided an effective, advanced solution for detecting hate speech in the Yemeni dialect and other dialects, with superior performance and greater stability than traditional methods and other modern models.

## References

1. Saleh, H., Alhothali, A. & Moria, K. Detection of hate speech using bert and hate speech word embedding with deep model. *Appl. Artif. Intell.* **37**, 2166719 (2023).

2. Daraghmi, E. Y. *et al.* From text to insight: An integrated cnn-bilstm-gru model for arabic cyberbullying detection. *IEEE Access* **12**, 103504–103519, DOI: 10.1109/ACCESS.2024.3431939 (2024).

3. Hellfeldt, K., López-Romero, L. & Mershed, H. Cyberbullying and psychological well-being in young adolescence: The potential protective mediation effects of social support from family, friends, and teachers. *Int. J. Environ. Res. Public Heal.* **17**, 45 (2020).

4. Alshalan, R. & Al-Khalifa, H. Hate speech detection in saudi x platformsphere: A deep learning approach (2020). Online.

5. Mutanga, R. T., Naicker, N. & Olugbara, O. O. Detecting hate speech on x platform network using ensemble machine learning. *Int. J. Adv. Comput. Sci. Appl.* **13** (2022).

6. Toktarova, A. *et al.* Hate speech detection in social networks using machine learning and deep learning methods. *Int. J. Adv. Comput. Sci. Appl.* **14** (2023).

7. Sharmila, P., Anbananthen, K. S. M., Chelliah, D., Parthasarathy, S. & Kannan, S. Pdhs: Pattern-based deep hate speech detection with improved tweet representation. *IEEE Access* **10**, 105366–105376, DOI: 10.1109/ACCESS.2022.3210177 (2022).

8. Al-Hassan, A. & Al-Dossari, H. Detection of hate speech in arabic tweets using deep learning. *Multimed. Syst.* 1963–1974, DOI: 10.1007/s00530-020-00742-w (2021).

9. Daouadi, K. E., Boualleg, Y. & Guehairia, O. Systematic investigation of recent pre-trained language model for hate speech detection in arabic tweets. *ACM Transactions on Asian Low-Resource Lang. Inf. Process.* DOI: 10.1145/3674970 (2024).

10. Aljohani, A., Alharbe, N., Al Mamlook, R. E. & Khayyat, M. M. A hybrid combination of cnn attention with optimized random forest with grey wolf optimizer to discriminate between arabic hateful, abusive tweets. *J. King Saud Univ. - Comput. Inf. Sci.* **36**, DOI: 10.1016/j.jksuci.2024.101961 (2024).

11. Duwairi, R., Hayajneh, A. & Quwaider, M. A deep learning framework for automatic detection of hate speech embedded in arabic tweets. *Arab. J. for Sci. Eng.* **46**, 4001–4014, DOI: 10.1007/s13369-021-05383-3 (2021).

12. Brihi, A. A. H. & Ghurab, M. A hybrid deep learning-machine learning stacking model for yemeni arabic dialect sentiment analysis. *IEEE Access* **13**, 134160–134171, DOI: 10.1109/ACCESS.2025.3593204 (2025).

13. Alshalan, R., Al-Khalifa, H., Alsaeed, D., Al-Baity, H. & Alshalan, S. Detection of hate speech in covid-19-related tweets in the arab region: Deep learning and topic modeling approach. *J. Med. Internet Res.* **22**, DOI: 10.2196/22609 (2020).

14. Itriq, A. *et al.* Arabic hate speech detection: A comprehensive survey of deep learning approaches. *Inf. Process. Manag.* **62** (2025).

15. Fat'hAlalim, M. *et al.* Evaluating arabic transformer models for hate speech detection across dialects. *Appl. Artif. Intell.* **40** (2026).

16. Mahouachi, S. *et al.* Multi-view hybrid transformer framework for hate speech detection. *Knowledge-Based Syst.* **312** (2026).

17. Abdelsamie, M. *et al.* Multi-task learning for arabic dialect hate speech detection using transformer models. *Expert. Syst. with Appl.* **268** (2026).

18. Omar, A., Mahmoud, T. M., Abd-El-Hafeez, T. & Mahfouz, A. Multi-label arabic text classification in online social networks. *Inf. Syst.* **100**, DOI: 10.1016/j.is.2021.101785 (2021).

19. Omar, A. & Hashem, M. E. An evaluation of the automatic detection of hate speech in social media networks: The case of arabic posts on facebook regarding france's muhammad cartoon controversy (2022).

20. Mulki, H., Haddad, H., Ali, C. B. & Alshabani, H. L-hsab: A levantine x platform dataset for hate speech and abusive language (2019).

21. Chowdhury, S. A. *et al.* A multi-platform arabic news comment dataset for offensive language detection (2020).

22. Alhazmi, A., Mahmud, R., Idris, N., Abo, M. E. M. & Eke, C. I. Code-mixing unveiled: Enhancing the hate speech detection in arabic dialect tweets using machine learning models. *PLoS One* **19**, DOI: 10.1371/journal.pone.0305657 (2024).

23. Alakrot, A., Murray, L. & Nikolov, N. S. Dataset construction for the detection of anti-social behaviour in online communication in arabic. In *Procedia Computer Science*, 174–181, DOI: 10.1016/j.procs.2018.10.473 (2018).

24. Haidar, B., Chamoun, M. & Serhrouchni, A. A multilingual system for cyberbullying detection: Arabic content detection using machine learning. *Adv. Sci. Technol. Eng. Syst.* **2**, 275–284, DOI: 10.25046/aj020634 (2017).

**25.** Aljarah, I. *et al.* Intelligent detection of hate speech in arabic social network: A machine learning approach. *J. Inf. Sci.* **47**, 483–501, DOI: 10.1177/0165551520917651 (2021).

**26.** Mursi, K. T., Alahmadi, M. D., Alsubaei, F. S. & Alghamdi, A. S. Detecting islamic radicalism arabic tweets using natural language processing. *IEEE Access* **10**, 72526–72534, DOI: 10.1109/ACCESS.2022.3188688 (2022).

**27.** Haddad, H., Mulki, H. & Oueslati, A. T-hsab: A tunisian hate speech and abusive dataset. In *Communications in Computer and Information Science*, 251–263, DOI: 10.1007/978-3-030-32959-4_18 (Springer, 2019).

**28.** Mubarak, H. & Darwish, K. Arabic offensive language classification on x platform. In *Social Informatics: 11th International Conference, SocInfo 2019*, 269–276 (Springer, 2019).

**29.** Musleh, D. *et al.* A machine learning approach to cyberbullying detection in arabic tweets. *Comput. Mater. Continua* **80**, 1033–1054, DOI: 10.32604/cmc.2024.048003 (2024).

**30.** Albayari, R. & Abdallah, S. Instagram-based benchmark dataset for cyberbullying detection in arabic text. *Data* **7**, DOI: 10.3390/data7070083 (2022).

**31.** Ahmad, A. *et al.* Hate speech detection in the arabic language: corpus design, construction, and evaluation. *Front. Artif. Intell.* **7**, DOI: 10.3389/frai.2024.1345445 (2024).

**32.** Mulki, H. & Ghanem, B. Let-mi: An arabic levantine x platform dataset for misogynistic language (2021).

**33.** Alkhatib, M., Faisal, A., Alfalasi, F., Shaalan, K. & Mohmed, A. Deep learning approaches for detecting arabic cyberbullying social media. In *Procedia Computer Science*, 278–286, DOI: 10.1016/j.procs.2024.10.201 (2024).

**34.** Al Anezi, F. Y. Arabic hate speech detection using deep recurrent neural networks. *Appl. Sci.* **12**, DOI: 10.3390/app12126010 (2022).

**35.** Faris, H., Aljarah, I., Habib, M. & Castillo, P. A. Hate speech detection using word embedding and deep learning in the arabic language context. In *International Conference on Pattern Recognition Applications and Methods*, 453–460, DOI: 10.5220/0008954004530460 (Science and Technology Publications, Lda, 2020).

**36.** Alam, F. *et al.* Deep learning architectures for arabic hate speech detection. *Neural Comput. Appl.* **37** (2025).

**37.** Husain, T. Attention-based recurrent networks for arabic offensive language detection. *Expert. Syst.* **43** (2026).

**38.** Abdelsamie, M. M., Azab, S. S. & Hefny, H. A. The dialects gap: A multi-task learning approach for enhancing hate speech detection in arabic dialects. *Expert. Syst. with Appl.* **295**, DOI: 10.1016/j.eswa.2025.128584 (2026).

**39.** Almaliki, M., Almars, A. M., Gad, I. & Atlam, E. S. Abmm: Arabic bert-mini model for hate-speech detection on social media. *Electronics* **12**, DOI: 10.3390/electronics12041048 (2023).

**40.** Ahmed, I., Abbas, M., Hatem, R., Ihab, A. & Fahkr, M. W. Fine-tuning arabic pre-trained transformer models for egyptian-arabic dialect offensive language and hate speech detection and classification. In *Proceedings of the 20th Conference on Language Engineering, ESOLEC 2022*, 170–174, DOI: 10.1109/ESOLEC54569.2022.10009167 (Institute of Electrical and Electronics Engineers Inc., 2022).

**41.** Torjmen, R. & Haddar, K. Tunisian dialect hate speech and abuse detection with bert-based models. In *Procedia Computer Science*, 3548–3557, DOI: 10.1016/j.procs.2025.09.480 (Elsevier B.V., 2025).

**42.** Mousa, A., Shahin, I., Nassif, A. B. & Elnagar, A. Detection of arabic offensive language in social media using machine learning models. *Intell. Syst. with Appl.* **22**, DOI: 10.1016/j.iswa.2024.200376 (2024).

**43.** Mubarak, H., Rashed, A., Darwish, K., Samih, Y. & Abdelali, A. Arabic offensive language on x platform: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 110–118 (Kyiv, Ukraine (Online), 2021).

**44.** Mazari, A. C. & Kheddar, H. Deep learning-based analysis of algerian dialect dataset targeted hate speech, offensive language and cyberbullying. *Int. J. Comput. Digit. Syst.* **13**, 965–972, DOI: 10.12785/ijcds/130177 (2023).

**45.** Abdellaoui, I. *et al.* Investigating offensive language detection in a low-resource setting with a robustness perspective. *Big Data Cogn. Comput.* **8**, DOI: 10.3390/bdcc8120170 (2024).

**46.** Guellil, I. *et al.* Detecting hate speech against politicians in arabic community on social media. *Int. J. Web Inf. Syst.* **16**, 295–313, DOI: 10.1108/IJWIS-08-2019-0036 (2020).

**47.** Alotaibi, M., Alotaibi, B. & Razaque, A. A multichannel deep learning framework for cyberbullying detection on social media. *Electronics* **10**, 2664, DOI: 10.3390/electronics10212664 (2021).

**48.** Daouadi, K. E., Boualleg, Y. & Guehairia, O. Deep random forest and arabert for hate speech detection from arabic tweets. *J. Univers. Comput. Sci.* **29**, 1319–1335, DOI: 10.3897/jucs.112604 (2023).

**49.** Bensoltane, R. & Zaki, T. Fine-grained hate speech detection in arabic using transformer-based models. *Int. J. Electr. Comput. Eng.* **14**, 2927–2936, DOI: 10.11591/ijece.v14i3.pp2927-2936 (2024).

**50.** Elzayady, H., Mohamed, M. S., Badran, K. M. & Salama, G. I. A hybrid approach based on personality traits for hate speech detection in arabic social media. *Int. J. Electr. Comput. Eng.* **13**, 1979–1988, DOI: 10.11591/ijece.v13i2.pp1979-1988 (2023).

**51.** Masadeh, M., Davanager, H. J. & Muaad, A. Y. A novel machine learning-based framework for detecting religious arabic hatred speech in social networks. Online.

**52.** Althobaiti, M. J. Bert-based approach to arabic hate speech and offensive language detection in x platform: Exploiting emojis and sentiment analysis. Online.

**53.** Al-Jarrah, H., Al-Smadi, M., Hammad, M. & Shannaq, F. Using deep learning techniques to detect hate and abusive language in arabic tweets. *Int. J. Intell. Eng. Syst.* **17**, 553–569, DOI: 10.22266/ijies2024.1031.43 (2024).

**54.** Al-Jarrah, H., Al-Smadi, M., Hammad, M. & Shannaq, F. Using deep learning techniques to detect hate and abusive language in arabic tweets. *Int. J. Intell. Eng. Syst.* **17**, 553–569, DOI: 10.22266/ijies2024.1031.43 (2024).

**55.** Abdul-Mageed, M., Elmadany, A., Moatez, E. & Nagoudi, B. Arbert & marbert: Deep bidirectional transformers for arabic. Online.

## Acknowledgements

The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2026).

## Author contributions statement

A.A. conceived the experiment(s); A.A. and M.A.R. conducted the experiment(s); M.A.R. and D.M.I. analyzed the results. All authors reviewed the manuscript.

## Funding

The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2026).

## Data availability:

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

## Competing interests

The authors declare no competing interests.