KFUPM
Journal of Undergraduate Research International

Diamond Open Access 48

Article Computer Science

Building and Evaluating an Arabic Social-Media Dataset for Deepfake Text Detection

Budoor Alshehri, Amal Sunba, and Tarek Helmy*

Cite https://doi.org/10.64589/juri/221315

Submitted: November 23, 2025 **Revised:** February 25, 2026 **Accepted:** March 01, 2026

ABSTRACT

Rapid advances in large language models have enabled the creation of sophisticated deepfake text, which poses a significant threat to social media information integrity. Arabic deepfake text detection is underexplored owing to the morphological complexity of the language and its diverse regional dialects. Studies in this domain often rely on narrow datasets or focus exclusively on Modern Standard Arabic and fail to capture the informal linguistic nuances prevalent on platforms like X. This study addresses these gaps by compiling a high-quality Arabic dataset spanning six major dialects and seven distinct domains, providing a representative benchmark for real-world environments. Real Arabic texts were collected from X, deepfake texts were generated via a prompt-based approach using GPT-4o-mini across multiple deception styles, and a pretrained BERT-base-uncased model was fine-tuned for binary classification using the Hugging Face Transformers framework. Experimental results validated the high quality of the proposed dataset, The model achieved training, validation, and test accuracies of 93.7%, 93.2%, and 92.5%, respectively, with an F1 score of 92.5%. A significant finding of this study was the variation in performance across different data sources. The highest accuracy was observed in entertainment content from YouTube, suggesting that the model effectively identifies the stylistic and expressive markers inherent in synthetic conversational Arabic. Our contributions include the curation of an authentic multi-dialectal corpus and validation of transformer-based architectures tailored for identifying AI-manipulated text. This study provides a benchmark for the Arabic natural language processing community.

Keywords: deepfake detection, Arabic text, machine learning, social media

1. INTRODUCTION

The rapid evolution of large language models (LLMs) has ushered in an era of synthetic text being increasingly indistinguishable from that written by humans. Although these advances offer transformative benefits in natural language processing (NLP), they also facilitate the creation of highly realistic deepfake texts, which are frequently used to spread misinformation, conduct large-scale phishing, and manipulate public discourse, especially across social-media platforms [1,11]. Recent findings indicate that although LLMs excel at generating coherent text, they introduce significant ethical risks by enabling the production of deceptive content, which can deceive even human experts [11,12]. Despite the global prevalence of this threat, existing detection frameworks are overwhelmingly optimized for English and Indo-European languages, often overlooking the linguistic complexity and vulnerability of Arabic [12].

This study is motivated by several fundamental challenges in the detection of deceptive Arabic content. The primary obstacle is the inherent morphological complexity of the language—Arabic is characterized by a root-and-pattern system where a single semantic root can generate numerous forms with distinct meanings, making it one of the most intricate languages. Therefore, there is a persistent shortage of standardized and inclusive datasets that capture the linguistic diversity of Arabic, including the variability between Arabic dialects commonly used on social media [12].

The complexity of Arabic is mirrored in other morphologically rich and right-to-left languages such as Urdu, where standard, unsupervised models often struggle without specialized feature extraction and syntactic rules [13]. Just as document clustering in Urdu requires specific handling of complex syntax—where the absence of Latin-script features such as capitalization complicates standard NLP approaches—Arabic detection requires models that consider its unique agglutinative nature and intricate orthography [12,13].

The absence of comprehensive, multi-dialectal Arabic datasets and the lack of detection models tailored to Arabic's morphological complexity represent critical gaps in the literature. This study addresses these gaps through three steps: collecting diverse data from social media, generating synthetic counterparts using LLM, and implementing rigorous quality control to address the morphological framework.

Figure 1. Workflow of the proposed approach

2. RELATED WORK

Although detecting machine-generated text in Arabic is critical, it is an emerging research field. Current literature can be mainly categorized into three main directions: rumor and fake news detection, deep learning approaches for synthetic text, and development of large-scale Arabic corpora.

Early research focused on rumors rather than AI-generated text. Alzanin and Azmi[4] used semi-supervised and unsupervised expectation maximization to detect rumors in Arabic on Twitter. Although pioneering, this approach targeted misinformation patterns rather than LLM linguistic markers. Similarly, Awajan[1] explored shallow learning techniques for fake news detection, emphasizing the importance of platform-specific features on X. Qasem et al.[7] expanded this into ensemble learning specifically for COVID-19 rumors, demonstrating that domain-specific models achieve high accuracy but often fail to generalize to broader contexts.

Recent studies have shifted toward identifying deepfake or machine-generated text. Sadiq et al.[2] leveraged FastText embeddings combined with deep learning to identify synthetic tweets, providing a baseline for how machine-generated Arabic differs from human prose. For data-collection methodology, Fagni et al.[3] established the TweepFake dataset; although focused on English, it provided a blueprint for using application programming interfaces (APIs) and heuristic searches to build deepfake-specific corpora, which were subsequently adapted by Arabic researchers.

The introduction of AraBERT by Antoun et al.[8] revolutionized the field by providing a robust transformer-based encoder for Arabic. Building on this, Alshammari et al.[11] recently proposed a dedicated detector for AI-generated Arabic text using an encoder-based transformer architecture, specifically designed to distinguish between human-written text and LLM output. To evaluate these complex models, Swaminathan and Tantri[10] recently emphasized the use of advanced metrics based on confusion matrices to ensure that performance is not skewed by imbalanced datasets.

Despite these technical leaps, the progress in this field is hindered by lack diverse data. Nofal[5] and Seddik[6] provided large-scale datasets through web scraping and APIs; however, these are largely restricted to Modern Standard Arabic (MSA) and Egyptian dialects. As noted by Alayba[12] in a 2025 comprehensive review, the primary challenge remains "Diglossia"—the vast difference between MSA and regional dialects. Most datasets focus on news and political content, leaving a notable gap in social and creative domains, where machine-generated text and deepfakes are increasingly prevalen. These limitations—narrow dialect coverage, domain imbalance, and models inadequate to Arabic's morphological complexity—collectively motivate the present study. This study aims to address these challenges by constructing a benchmark dataset spanning six Arabic dialects and seven domains, and by evaluating a transformer-based model for deepfake text detection. The key contributions of this work include the curation of an authentic multi-dialectal corpus collected from X, the generation of deepfake texts using multiple deception styles, and the fine-tuning of a BERT-based classifier.

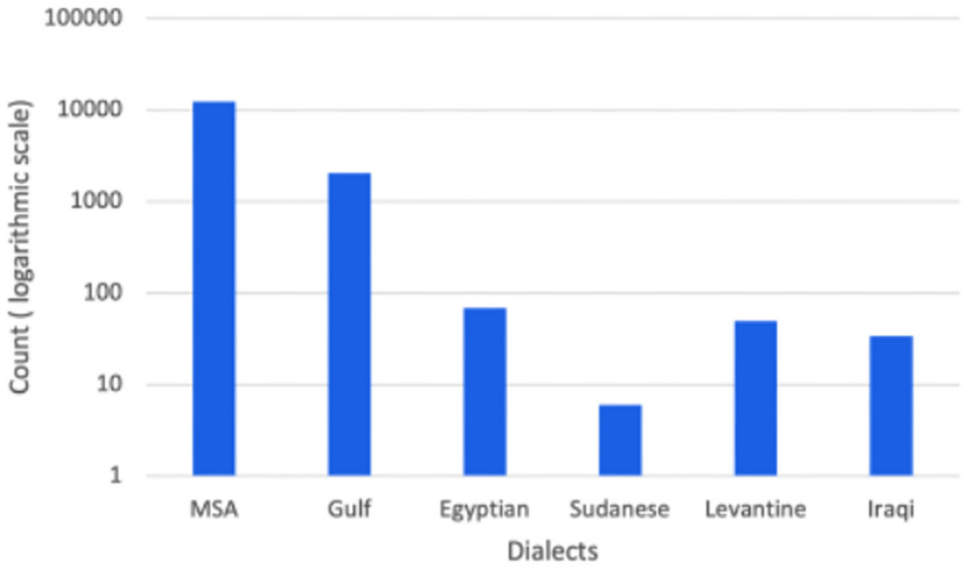

Figure 2. Distribution of the dataset across dialects

Table 1. *Attributes of the dataset*

Attribute	Description
Text	Target text
Type	Tweet, Reply
Platform Type	Social media
Source	X (formerly Twitter)
Dialect	MSA, Gulf, Egyptian, Levantine, Iraqi, and Sudanese
Domain	Health, Politics, Education, Ads, Sports, Religion, and Economy
Class	Real or Deepfake

3. METHODOLOGY

This section explains the methodology followed in this study to construct and validate an Arabic deepfake-detection dataset, as shown in Figure 1.

3.1. Real Data Collection. We gathered real Arabic text from X and compiled a representative dataset using a wide range of linguistic and thematic variations across Arabic dialects. Public posts and replies were searched and collected using a combination of hashtag queries, keyword searches, automated scraping, and manual reviews. The search targeted diverse domains, including news, opinions, e-commerce, and general discussions. The text was collection between June 2022 and November 2022, focusing on posts shared before the public release of ChatGPT to minimize the likelihood of AI-generated content. No private or personal identifiable information was collected; all data were obtained from publicly accessible sources according to data usage and content access policies of X. The resulting dataset reflects natural Arabic discourse across dialects and topics. Table 1 shows the key attributes of the dataset.

We categorized the data by dialect and domain using a combination of a prompt-based method and manual reviews. The texts already had preliminary topic labels based on the collection criteria; however, the dialects were not labeled. We used a prompt-based classification approach powered by the GPT-4o-mini LLM. The instructions were to output a dialect label per text and select the dominant dialect in case of mixed usage. The prompt explicitly listed the targeted dialects and enforced consistent formatting. At the end, an Arabic speaker manually reviewed each domain and dialect label to resolve ambiguous cases such as domain-overlapping or dialectally mixed texts.

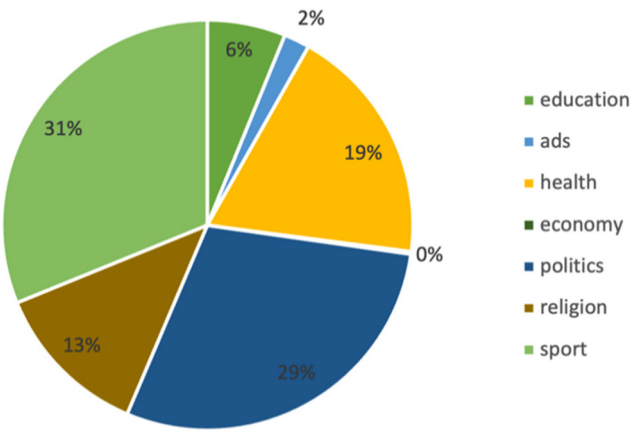

Figure 3. Distribution of the dataset across topics

The dataset comprises six Arabic dialects—MSA, Gulf, Egyptian, Sudanese, Levantine, and Iraqi—as shown in Figure 2. It also spans seven domains—health, politics, education, online advertisements, sports, religion, and economy. The initial dataset consisted of 21,000 texts; however, after cleaning and filtering, the final dataset comprised 14,403 real texts. Raw data were preprocessed to ensure quality; this included duplicate removal to prevent model overfitting, empty-text removal (such as tweets containing only hashtags), and text normalization to remove URLs, usernames, emojis, and hashtags. The clean dataset was used as the foundation for deepfake generation. To better understand the structure of the dataset, Figure 3 illustrates the distribution of text samples across topics. Domains such as sports and politics exhibited higher representation, reflecting their popularity on social media.

3.2. Deepfake Data Generation. To simulate realistic misinformation, deepfake texts were generated via a prompt-based approach using the OpenAI API. Text was generated using GPT-4o-mini with the following parameters: temperature = 0.9, top-p = 1.0 (default), and max-tokens = ~4096 (default). This stage aimed to generate deceptive text that is similar to real-world misinformation while preserving the linguistic and thematic attributes of the original text. The prompt was designed to mimic user-generated posts across different domains. Each deepfake text was generated based on the corresponding real post, ensuring that it followed the same topic, tone, length, and dialect. The generation process employed multiple deception styles, including exaggeration, omission, mixed truth, and contradiction, which are common types of misinformation found on X. The final dataset contained 28,806 real and fake texts. Table 2 shows one sample of real text and its corresponding deepfake, with the type of deception used.

3.3. Model Architecture and Training. To detect deceptive Arabic text effectively, we adopted a transfer-learning approach using a pretrained BERT-base-uncased model as the baseline architecture, which is a stacked bidirectional transformer encoder[8]. This model was chosen for its deep bidirectional understanding of language context. Although this model was originally trained by Google on large-scale English-language data, its architecture can be effectively generalized to other languages after fine-tuning. The model was fine-tuned for a binary classification task to label each text sample as real or fake.

As illustrated in Figure 4, the model was fine-tuned for a binary classification task to label each text sample as real (1) or fake (0). Fine-tuning was conducted in the Hugging Face Transformers framework to achieve state-of-the-art performance. We used the AdamW optimizer, which decouples weight decay from gradient updates and helps to prevent overfitting—a crucial factor given the linguistic complexity of Arabic and limited dataset size. The learning rate was set to 2e-5, a value empirically shown to yield stable convergence in BERT-based fine-tuning without catastrophic forgetting of pretrained weights[9]. A binary cross-entropy loss function was applied since the output space was

Table 2. *Real vs. deepfake Arabic text sample*

Real text	Deepfake text	Deception type
أزمة التملك السكني تعكس فشل سياسة السلطة بسبب القمع	أزمة التملك السكني تعكس نجاح سياسة السلطة في تحسين الأوضاع بسبب الدعم الكبير الذي تقدمه للمواطنين.	Contradiction

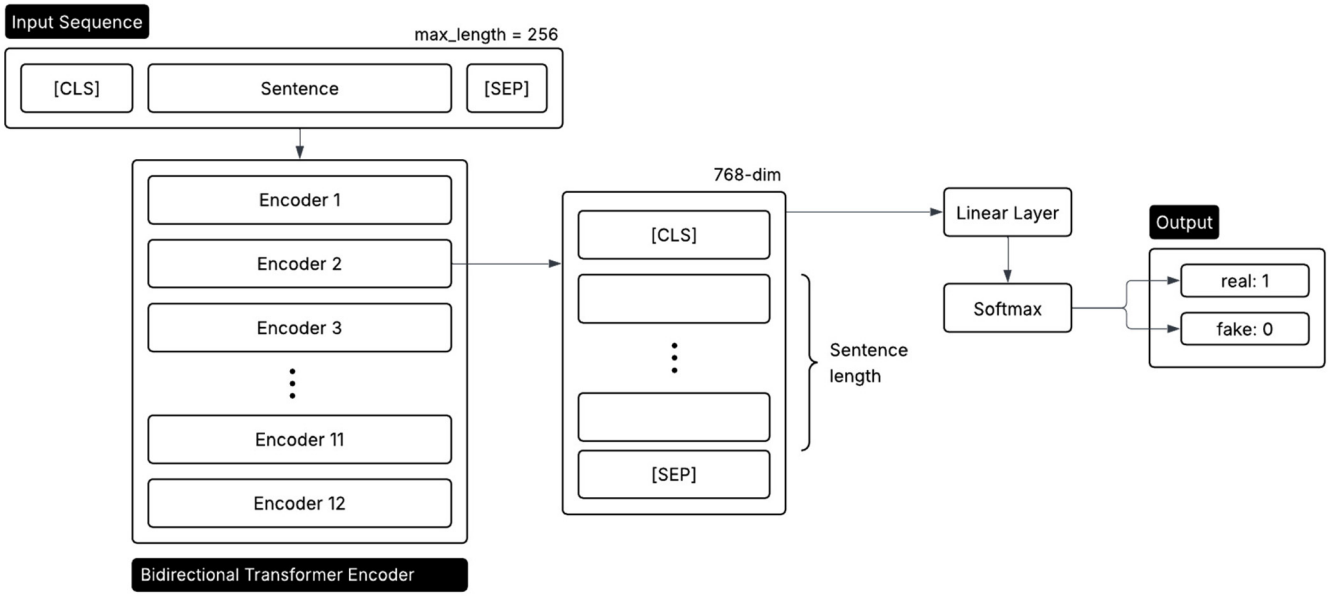

Figure 4. BERT-base-uncased architecture for Arabic deceptive-text detection

Table 3. *Training configuration*

Framework	Hugging Face Transformers
Optimizer	AdamW
Learning Rate	2×10^{-5}
Batch Size	16
Epochs	3
Loss Function	Binary cross-entropy
Label Encoding	{fake: 0, real: 1}

restricted to two mutually exclusive classes: fake = 0 or real = 1. All training configurations are shown in Table 3. The training was conducted on Google Colab using a high-performance T4 GPU environment, which enabled parallel computations, thus reducing training time. The model was trained on Arabic text samples, each labeled with its corresponding class (fake or real).

3.4. Evaluation Metrics and Validation. The model was evaluated using four metrics—accuracy, precision, recall, and F1-score. A confusion matrix was generated, which is a performance-evaluation tool used in classification tasks to visualize model predictions compared to actual values[10]. For our binary classification problem, the matrix was structured as shown in Table 4, where

Table 4. *Confusion matrix structure*

		Predicted	
		Real (1)	Fake (0)
Actual	Real (1)	True Positive (TP)	False Positive (FP)
	Fake (0)	False Negative (FN)	True Negative (TN)

- TP (True Positives): Real tweets correctly identified as real.
- TN (True Negatives): Fake tweets correctly identified as fake.
- FP (False Positives): Fake tweets incorrectly predicted as real (also called a "Type I error").
- FN (False Negatives): Real tweets incorrectly predicted as fake (also called a "Type II error").

Accuracy provides an overall measure of correctness, whereas precision and recall offer deeper insights into the tendencies of Type I and Type II errors, where the cost of false negatives can be high.

To ensure reliable performance assessment, a stratified train-validation-test split was applied, maintaining an approximately 1:1 class ratio across all subsets. This helped preserve the class balance and prevent data leakage between splits. The data were split into 75% training, 15% testing, and 15% validation. Cross-domain validation was performed using four datasets, each representing a different data domain, to further assess the generalization of the model across varying linguistic and contextual settings.

4. RESULTS AND DISCUSSION

The fine-tuned BERT model performed well across all evaluation sets, as summarized in Table 5, demonstrating that transfer learning can be effective for detecting deception in Arabic text. The model achieved training, validation, and test accuracies of 93.7%, 93.2%, and 92.5%, respectively, with an F1 score of 92.5%. These

Table 5. *Evaluation Metrics*

Subset	Loss	Accuracy	Precision	Recall	F1
Training	0.1377	0.937236	0.937235	0.937236	0.937235
Validation	0.186762	0.932534	0.932561	0.932534	0.932525
Test	0.212243	0.925264	0.92553	0.925264	0.925248

Table 6. *Cross-domain validation results*

Sources	Accuracy	Precision (weighted)	Recall (weighted)	F1 (weighted)
entertainment	**0.849417**	**0.857303**	**0.849417**	**0.848581**
governmental	0.569704	0.691781	0.569704	0.495685
news	0.499614	0.494677	0.499614	0.351474
e-commerce	0.454203	0.493177	0.454203	0.414901

results across datasets indicate effective generalization and ability to capture linguistic cues associated with deception. The high performance confirms that even with a transformer-based architecture pretrained on English data, the model can adapt to Arabic linguistics through fine-tuning. Moreover, the close alignment between training and testing metrics suggests minimal overfitting, indicating that the model successfully learned the underlying patterns in the data. The learning curve presented in Figure 5 shows that all evaluation metrics increased steadily and uniformly over the three epochs. This indicates stable model learning and improvement. All scores had effectively converged by Epoch 2, suggesting that the model achieved high performance with minimal trade-offs between precision and recall.

A confusion matrix for the test set (Figure 6) shows a balanced distribution of true positives and true negatives with relatively few misclassifications, confirming the effectiveness of the model in distinguishing between deceptive and authentic Arabic text.

To further evaluate the generalization and robustness of the mode, a cross-domain validation was conducted across four content domains—entertainment, government, news, and e-commerce—as shown in Table 6. The entertainment domain achieved the highest performance, with 84.9% accuracy and 0.85 F1 score, reflecting the ability of the model to detect deepfakes in conversational and emotionally expressive text. Other domains achieved significantly lower accuracy, such as e-commerce (45.4%).

1This variation is attributed to the linguistic diversity and nature of the content in different domains. The data for the entertainment domain were obtained from YouTube. The comments and posts in X and YouTube typically use informal and expressive Arabic styles with identifiable stylistic patterns. In contrast, governmental and news data typically contain formal and factual language with limited emotional tone, making it hard for the model to distinguish between fake and real text. e-commerce demonstrated the lowest performance, likely due to the nature of

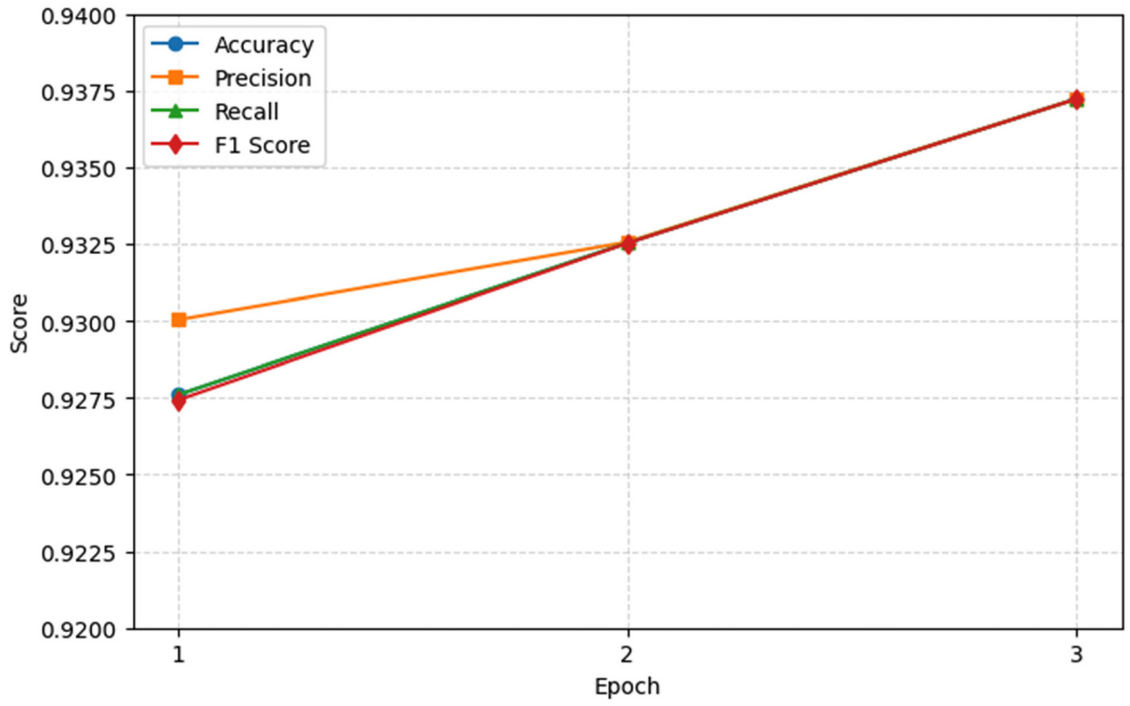

Figure 5. Learning curve across epochs

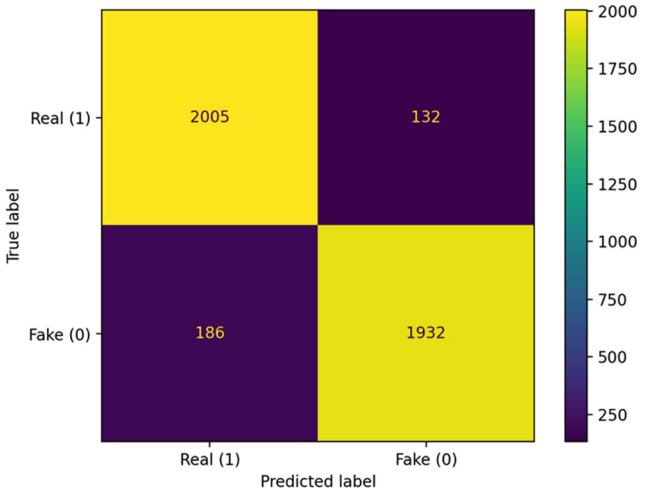

Figure 6. Confusion matrix for the testing dataset

the data. Content from e-commerce platforms such as Amazon consists of short product titles, often mixed with English phrases, and lack contextual depth. In comparison, data from X are contextually rich and linguistically varied. Hence, the model, which was fine-tuned on X data, could perform well on a similar type of expressive content. This indicates that the model can capture stylistic and lexical markers of deception more effectively in subjective and expressive contexts than in objective or factual ones, as shown in Figure 7.

A closer examination of the confusion matrices for the cross-domain validation in Figure 8 further clarifies this variation. The matrices for the news and e-commerce datasets illustrate the degradation in performance, showing predictions that are nearly random (approximately a 50/50 split between correct and incorrect classifications for each class).

The fine-tuned BERT model represents a significant advancement over existing benchmarks that typically focus on MSA in isolated news domains [1,2]. By successfully identifying deceptive markers across six dialects and seven domains, this approach demonstrates robustness superior to that of traditional models that require manual feature extraction for the complex Arabic morphology [4,13]. Furthermore, the 92.5% F1-score of the model highlights the need for fine-tuning encoder-based architectures to overcome the performance failures observed in general-purpose tools such as GPTZero when processing intricate Arabic text [11,12].

One limitation was the reduced performance in code-mixed and English-mixed Arabic text. For example, tweets containing nonstandard spellings or repetitive expressions that were challenging for the tokenizer to handle reduced the classification reliability. Furthermore, dialect categorization presented some challenges, especially with Gulf Arabic. Owing to annotation difficulties, all Gulf variants, such as Kuwaiti and Bahraini, were grouped under Gulf. Consequently, the dataset comprises 28,806 samples, which is relatively limited for training and evaluating deep-learning models. The limited size of the dataset may constrain the generalizability and robustness of the model. Despite these limitations, the dataset represents a valuable step toward building a comprehensive Arabic deepfake-detection dataset spanning diverse dialects and domains.

5. CONCLUSIONS

This study investigated the detection of AI-generated Arabic text on social media by developing a comprehensive dataset spanning multiple domains and regional dialects. Experimental results validated the high quality of the proposed dataset, with the trained model achieving a robust 92% accuracy and a 92.5% F1 score. A significant finding of this study was the variation in performance across different data sources. The highest accuracy was observed in entertainment content from YouTube, suggesting that the model effectively identifies the stylistic and expressive mark-

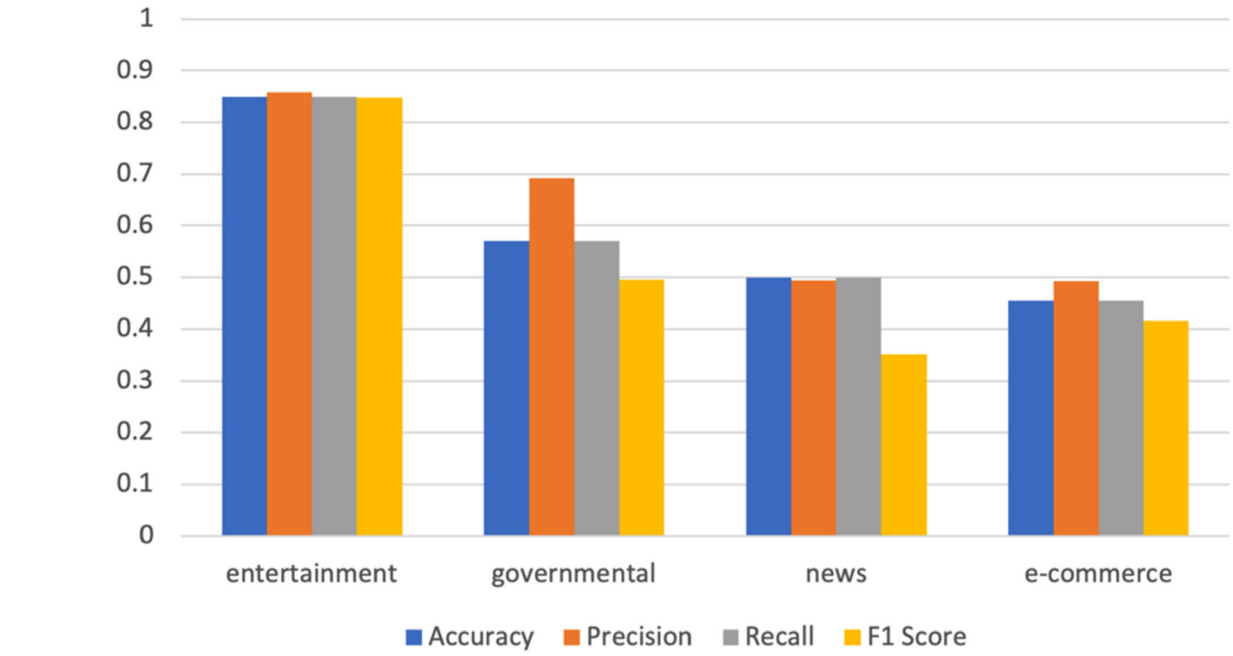

Figure 7. Performance comparisons across domains

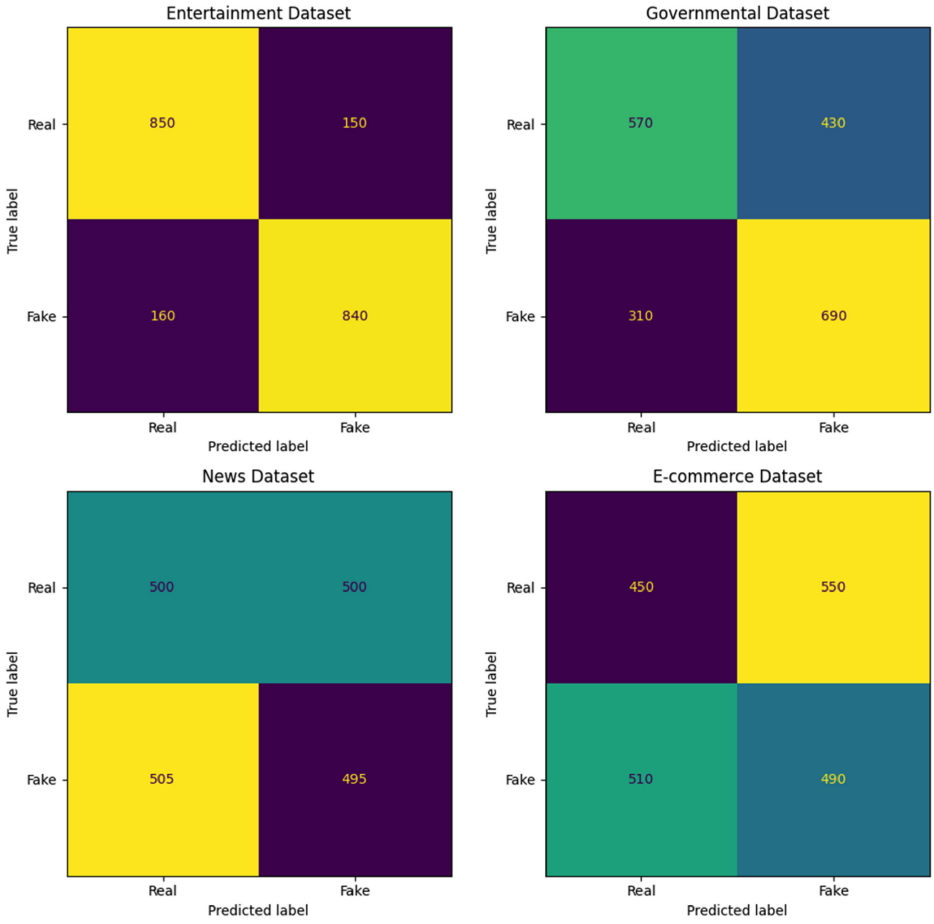

Figure 8. Confusion matrices on testing data for the four datasets

ers inherent in synthetic conversational Arabic. This highlights the critical need for incorporating diverse data sources, including social-media platforms and official governmental records, to ensure that detection frameworks can capture a broad spectrum of linguistic patterns and generalize successfully to unseen data. Although the use of BERT-base-uncased posed certain semantic constraints for the Arabic language, the success of transfer learning demonstrated that the dataset can support resilient and effective detection models. Consequently, this study provides a benchmark for the Arabic NLP community. Future work will focus on further expanding the dataset to include broader dialectal variations and specialized topics.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Budoor Alshehri – *Department of Information & Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*; 0009-0000-9661-6261
Email: s202177210@kfupm.edu.sa

Author

Amal Sunba – *Research collaborator, Department of Information and Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*; 0009-0002-4712-2104
Email: g202203780@kfupm.edu.sa

Corresponding Author

Tarek Helmy – *Research Mentor, Professor, Department of Information and Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*; 0000-0001-9683-682X
Email: helmy@kfupm.edu.sa

ACKNOWLEDGEMENTS

This study was conducted under a collaboration between the General Authority for Defense Development (GAAD) and King Fahd University of Petroleum & Minerals (KFUPM) under agreement number CSSP2643, Saudi Arabia.

REFERENCES

(1) Awajan, A. Enhancing Arabic fake news detection for Twitter's social media platform using shallow learning techniques. *Journal of Theoretical and Applied Information Technology*. 101, 5 (2023).

(2) Sadiq, S., Aljrees, T. & Ullah, S. Deepfake detection on social media: Leveraging deep learning and FastText embeddings for identifying machine-generated tweets. *IEEE Access* 11, (2023). https://doi.org/10.1109/ACCESS.2023.3308515

(3) Fagni, T., Falchi, F., Gambini, M., Martella, A. & Tesconi, M. TweepFake: About detecting deepfake tweets. *PLoS ONE* 16, 5 (2021). https://doi.org/10.1371/journal.pone.0251415

(4) Alzanin, S. M. & Azmi, A. M. Rumor detection in Arabic tweets using semi-supervised and unsupervised expectation–maximization. *Knowledge-Based Systems*. 185, (2019). https://doi.org/10.1016/j.knosys.2019.104945

(5) Nofal, M. 2.5+ million rows Egyptian datasets collection. Kaggle (2023). https://www.kaggle.com/datasets/mostafanofal/two-million-rows-egyptian-datasets

(6) Seddik, F. Arabic viral tweets (عربي). Kaggle (2022). https://www.kaggle.com/datasets/fahdseddik/arabic-viral-tweets

(7) Qasem, S. N., Al-Sarem, M. & Saeed, F. An ensemble learning-based approach for detecting and tracking COVID-19 rumors. Computers, Materials & Continua. 70, 1721–1747 (2022). https://doi.org/10.32604/cmc.2022.018972

(8) Antoun, W., Baly, F. & Hajj, H. AraBERT: Transformer-based model for Arabic language understanding. *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4)*, (2020). https://doi.org/10.48550/arXiv.2003.00104

(9) Baharuddin, F. & Naufal, M. F. Fine-tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy. *Journal of Information Systems Engineering and Business Intelligence*. 9, 2 (2023).

(10) Swaminathan, S. & Tantri, B. R. Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, 27, 4023–4031 (2024). https://doi.org/10.53555/AJBR.v27i4S.4345

(11) Alshammari, H., El-Sayed, A., & Elleithy, K. (2024). AI-Generated Text Detector for Arabic Language Using Encoder-Based Transformer Architecture. *Big Data and Cognitive Computing*, 8(3), 32. https://doi.org/10.3390/bdcc8030032

(12) Alayba, A. M. (2025). Arabic Natural Language Processing (NLP): A Comprehensive Review of Challenges, Techniques, and Emerging Trends. *Computers*, 14(11), 497. https://doi.org/10.3390/computers14110497

(13) Amin, A., Rana, T. A., Mian, N. A., Iqbal, M. W., Khalid, A., Alyas, T., & Tubishat, M. (2020). TOP-Rank: A Novel Unsupervised Approach for Topic Prediction Using Keyphrase Extraction for Urdu Documents. *IEEE Access*, 8, 212675-212686.