Arabic Tweets Classification using MARBERT For Spam Detection

Hanan Abu Kwaider
dept of Computer Engineering of Mersin University
Mersin, Türkiye
hanan.abuquader@gmail.com

Large-language models (LLMs) have revolutionized natural language processing (NLP) by achieving state-of-the-art performance across various tasks. This paper explores the use of MARBERT, a pre-trained transformer model designed specifically for Arabic, to classify Tweets as Spam or Ham. The unique challenges posed by the Arabic language, including diglossia, dialectal variations, and morphological complexity, make spam detection a particularly demanding task. To address these challenges, a publicly available dataset of 132,421 labeled Arabic Tweets was used, split into training and testing subsets at an 80%-20% ratio, and evaluated using a 5-fold cross-validation strategy. MARBERT demonstrated strong classification performance, achieving an overall accuracy of 99.54%, with F1-scores of 99.4% for Spam and 99.6% for Ham. These results validate MARBERT's effectiveness in capturing the linguistic nuances of Arabic text, including its ability to handle both formal and dialectal varieties. This study highlights MARBERT's applicability in social media spam detection, showcasing its potential as a reliable tool for real-time monitoring and cybersecurity applications. Furthermore, the research contributes to advancing Arabic NLP by providing a comprehensive evaluation of MARBERT for spam detection and emphasizing the importance of task-specific datasets in improving performance. The findings underscore the significance of leveraging language-specific LLMs for underrepresented languages in NLP, opening avenues for future research in Arabic content analysis.

Keywords—MARBERT, Arabic Tweets Classification, Spam Detection, Natural Language Processing.

I. Introduction

Large-language models (LLMs) have significantly advanced natural language processing (NLP) by achieving state-of-the-art performance across various linguistic tasks. These models, trained on massive datasets with billions of parameters, are designed to learn and generate complex linguistic patterns, enabling them to handle a wide range of applications, from text classification to language generation. Recent efforts in LLM research have focused on developing models that cater to specific languages and dialects, addressing linguistic and cultural nuances that general-purpose multilingual models often overlook [1, 2].

Arabic, the fifth most spoken language in the world, poses unique challenges and opportunities for NLP due to its complex linguistic characteristics [2]. The language is characterized by diglossia, where the formal written form of Arabic, including Classical Arabic and Modern Standard Arabic, differs substantially from spoken regional dialects. Furthermore, Arabic exhibits significant dialectal diversity and morphological complexity, requiring language models to handle not only variations in vocabulary but also intricate grammatical structures. These factors make tasks such as text classification and spam detection particularly challenging [2].

Spam detection in social media platforms such as Twitter is a critical NLP application that requires robust language understanding to differentiate between spam (unwanted or irrelevant content) and legitimate messages. MARBERT, a pre-trained transformer model designed specifically for Arabic, addresses these challenges by leveraging contextualized embeddings trained on diverse Arabic text, encompassing both formal and dialectal variations [1]. In this study, we fine-tune MARBERT on a publicly available dataset of 132,421 labeled Arabic Tweets [3], focusing on the binary classification of Tweets as Spam or Ham. The dataset is balanced across classes, and an 80%-20% train-test split, coupled with 5-fold cross-validation, is used to ensure robust evaluation.

This work highlights MARBERT's ability to capture the linguistic nuances of Arabic, making it a reliable tool for spam detection. Furthermore, this study contributes to advancing Arabic NLP by addressing a critical application area and emphasizing the importance of tailored models and high-quality datasets.

II. Background and related work

A. Arabic Large-Language Model

The development of Arabic large-language models (LLMs) has significantly advanced natural language processing (NLP) capabilities for Arabic, addressing challenges such as diglossia, dialectal diversity, and rich morphology. Monolingual models like MARBERT, which is pre-trained on both Modern Standard Arabic (MSA) and dialectal Arabic, have proven highly effective for tasks involving informal and noisy data, such as social media text. In comparison, multilingual models like mBERT and XLM-R, while capable of processing Arabic, often underperform due to limited representation of Arabic-specific features [1, 2]. MARBERT's ability to handle diverse text makes it well-suited for downstream tasks like spam detection, sentiment analysis, and dialect identification [2].

B. Spam Detection in Arabic Text

Spam detection is a critical NLP task, especially in social media analysis, where differentiating between legitimate content and spam is essential. Various approaches have been proposed for Arabic spam detection, ranging from classical machine learning algorithms to deep learning models.

In [4], the authors proposed a data augmentation method combined with machine learning algorithms to enhance Arabic spam detection on Twitter. They achieved an F1-score of 89% and an overall accuracy of 92%. Another study [5] examined Arabic spam reviews in Facebook comments, using a dataset of 3,000 comments. The authors evaluated classifiers such as Decision Trees, k-Nearest Neighbors (kNN), Support Vector Machines (SVM), and Naïve Bayes, reporting that the Decision Tree classifier outperformed the others, achieving an accuracy of 92.63%.

Further, [6] utilized a dataset of 3,503 tweets with Word2Vec embeddings and machine learning classifiers, including Naïve Bayes, Decision Trees, and SVM. The highest accuracy of 87.33% was achieved using SVM with the skip-gram Word2Vec technique. In [7], a Recurrent Neural Network (RNN) architecture with Gated Recurrent Units (GRU) and pre-trained word embeddings was used to detect Arabic religious hate speech, achieving an Area Under the Receiver Operating Characteristic (AUROC) of 0.84. Finally, [8] proposed an ensemble approach for detecting spam in Arabic opinion texts, achieving a detection accuracy of 95.25%.

III. METHODOLOGY

A. Dataset

The dataset used in this study is the publicly available Arabic Spam and Ham Tweets Dataset [3], which consists of 13,240 labeled tweets, including 1,941 spam tweets and 11,299 ham tweets, as shown in Table 1. The dataset was curated to provide a comprehensive benchmark for spam detection in Arabic social media.

Table 1. Data Count

Tweets	*Count*
Spam	1941
Ham	11299
Total	13240

1) *Ham Tweets*

Ham tweets were collected from prominent, verified Twitter accounts such as Arabiya, Emarat Al Youm, and Sky News Arabia. These accounts were selected to ensure the quality and authenticity of the ham data. The collected tweets were manually verified to remove any instances of spam or mislabeled content.

2) *Spam Tweets*

Spam tweets were gathered by querying Twitter using specific Arabic spam keywords. To enhance the dataset's reliability, the top 10 accounts with high spam percentages were identified and crawled. The crawled data was then inspected manually to ensure that only spam tweets were retained, with any ham tweets incorrectly included in the list removed.

3) *Preprocessing*

To ensure the quality of the dataset:

- Duplicate Tweets: All duplicate tweets were removed from both the ham and spam datasets.
- Spam Content Warning: The spam data may contain inappropriate language, as these tweets reflect typical spam content in Arabic.

This dataset provides a balanced and realistic representation of Arabic Twitter content, making it an excellent resource for evaluating spam detection models like MARBERT.

B. *Model Architecture*

This study utilizes MARBERT, a monolingual Arabic large-language model (LLM) based on the BERT architecture. MARBERT has been pre-trained on a diverse corpus of Arabic text, encompassing both Modern Standard Arabic (MSA) and dialectal Arabic. Its ability to understand both formal and informal text makes it particularly suitable for spam detection tasks [2].

1) *Fine-Tuning MARBERT*

The fine-tuning process was conducted on the Arabic Spam and Ham Tweets Dataset described in Section 3.1. The implementation leveraged the Hugging Face Transformers library for training MARBERT on the binary classification task, where tweets were labeled as 1 for spam and 0 for ham.

2) *Training Configuration*

The fine-tuning process was configured as follows:

- Tokenizer and Model Initialization: MARBERT's pre-trained tokenizer was used to tokenize tweets, while the model was initialized for a binary classification task with two labels (spam and ham).
- Cross-Validation: A 5-fold cross-validation approach was employed, with the dataset split into 80% training and 20% validation sets for each fold. This ensured robust evaluation across multiple data partitions.
- Training Hyperparameters:
  - Batch size: 8 for both training and evaluation.
  - Learning rate: 2×10^{-5}
  - Number of epochs: 3.
  - Weight decay: 0.01.
  - Mixed precision training (FP16): Enabled to accelerate computations.
  - Evaluation strategy: Performance evaluation after each epoch.

3) *Hardware Environment*

The fine-tuning process was conducted in a GPU-enabled environment to handle the computational demands of training MARBERT on large-scale data efficiently.

C. *Evaluation Metrics*

To evaluate the performance of MARBERT on the spam detection task, a range of standard classification metrics was employed. These metrics were calculated for each fold in the 5-fold cross-validation and aggregated to assess the model's overall performance.

1) Accuracy: Accuracy measures the percentage of correctly classified tweets (spam or ham) out of the total number of tweets. It is calculated as:

$$\text{Accuracy} = \frac{Total\ Samples}{True\ Positives + True\ Negatives} \quad (1)$$

2) Precision: Precision evaluates the proportion of correctly classified spam tweets out of all tweets predicted as spam. It indicates the model's ability to avoid false positives and is defined as:

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (2)$$

3) Recall: Recall, also known as sensitivity, measures the proportion of correctly classified spam tweets out of all actual spam tweets. It assesses the model's ability to detect spam and is given by:

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

4) F1-Score: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between these two measures. It is defined as:

$$\text{F1-Score} = 2. \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

5) Confusion Matrix: The confusion matrix was used to provide a detailed breakdown of the model's predictions, indicating the number of:

- True Positives (TP): Spam tweets correctly identified as spam.
- True Negatives (TN): Ham tweets correctly identified as ham.
- False Positives (FP): Ham tweets incorrectly classified as spam.
- False Negatives (FN): Spam tweets incorrectly classified as ham.

6) Evaluation Process: Performance metrics were calculated on the validation set for each fold in the cross-validation process, and the final results were averaged across all folds to ensure robustness.

IV. Results and Discussions

A) Results

The performance of MARBERT on the spam detection task was evaluated using the metrics described in Section 3.3. The results, averaged across all folds of the 5-fold cross-validation, are summarized in Table 2.

Table 2 . OVERALL PERFORMANCE

Metric	*Class 0 (Ham)*	*Class 1 (Spam)*	*Overall*
Precision	0.9943	0.9963	-
Recall	0.9950	0.9957	-
F1-Score	0.9947	0.9960	-
Accuracy	-	-	0.9954

Table 3 presents the confusion matrix summarizing MARBERT's predictions for spam and ham tweets.

Table 3. CONFUSION MATRIX

Predicted \ Actual	*Ham (Class 0)*	*Class 1 (Spam)*
Ham (Class 0)	11,189	56
Spam (Class 1)	64	14,851

These results demonstrate MARBERT's ability to effectively distinguish between spam and ham tweets, achieving high precision, recall, and F1-scores for both classes.

B) Discussion

The results indicate that MARBERT performs exceptionally well on the binary classification task of spam detection, achieving an overall accuracy of 99.54% and high F1-scores for both classes. The following observations can be made:

Class Imbalance Handling: Despite the dataset's imbalance (1,941 spam tweets vs. 11,299 ham tweets), MARBERT demonstrated strong performance across both classes, with minimal disparity between precision and recall values.

Spam Detection Strength: The model's precision and recall for the spam class (0.9963 and 0.9957, respectively) highlight its ability to identify spam tweets accurately, minimizing false positives and false negatives.
Real-World Applicability: The low misclassification rates (56 false positives and 64 false negatives) suggest that MARBERT can be effectively deployed in real-world scenarios, such as automated moderation systems for social media platforms.

Comparison with Prior Work: Compared to earlier studies that employed traditional machine learning or word embedding techniques [5]-[8], MARBERT demonstrates significant improvements in accuracy and F1-score. For instance, prior work using Word2Vec and SVM achieved an accuracy of 87.33% [5], while ensemble methods achieved 95.25% [7]. MARBERT's performance surpasses these benchmarks, underscoring the advantages of transformer-based models for Arabic spam detection.

These findings validate the effectiveness of MARBERT's pre-trained embeddings and its ability to handle noisy and dialect-rich Arabic social media text.

V. RESULTS AND DISCUSSIONS

This study applied MARBERT, a monolingual Arabic large-language model, to the task of spam detection on Arabic tweets. By fine-tuning MARBERT on the Arabic Spam and Ham Tweets Dataset, the model achieved an overall accuracy of 99.54%, with high F1-scores for both spam (0.9960) and ham (0.9947). The results highlight MARBERT's effectiveness in handling the linguistic complexity and noise of Arabic social media text.

Future work could explore extending this approach to other social media platforms or spam domains, such as fake review detection. Expanding the binary classification task to multi-class spam detection and deploying MARBERT in real-time systems would provide valuable insights into its efficiency and scalability. These directions can help broaden MARBERT's applications in Arabic natural language processing.

Acknowledgments

The author would like to express their gratitude to Assoc. Prof. Dr. Çiğdem Aci for her guidance and support throughout the duration of this course, which greatly contributed to the successful completion of this research. The author also acknowledge the creators of the Arabic Spam and Ham Tweets Dataset for making their dataset publicly available, which was instrumental in conducting this study.

References

[1] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7088–7105, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.acl-long.551. doi: 10.18653/v1/2021.acl-long.551

[2] M. Mashaabi, S. Al-Khalifa, and H. Al-Khalifa, "A Survey of Large Language Models for Arabic Language and its Dialects,", Oct. 2024. [Online]. Available: https://arxiv.org/pdf/2410.20238

[3] S. Kaddoura and S. Henno, "Dataset of Arabic spam and ham tweets," *Data in Brief*, vol. 52, 2024, Art. no. 109904. [Online]. Available: https://doi.org/10.1016/j.dib.2023.109904

[4] S. Al-Khalifa, "Enhancing Detection of Arabic Social Spam Using Data Augmentation and Machine Learning,". [Online]. Available: https://www.researchgate.net/publication/365304571

[5] A. I. Ahmed, "Detecting Arabic Spam Reviews in Social Networks Based on Classification Algorithms," *Proceedings of the ACM International Conference on Web Intelligence*, 2021. doi: 10.1145/3476115. [Online]. Available: https://dl.acm.org/doi/10.1145/3476115

[6] A. Khater et al., "Detection of Arabic Spam Tweets Using Word Embedding and Machine Learning," *2021 IEEE International Conference on Computer and Information Sciences (ICCOINS)*, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/8855747

[7] M. A. Hussein et al., "Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere," *2021 IEEE International Conference on Computer and Information Sciences (ICCOINS)*, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/8508247

[8] F. Rashid et al., "An Ensemble Approach for Spam Detection in Arabic Opinion Texts," [Online]. Available: https://www.researchgate.net/publication/336471271_An_Ensemble_Approach_for_Spam_Detection_in_Arabic_Opinion_Texts