# Hate Speech Detection in Egyptian Dialectal Arabic: A Comparative Fine-Tuning Study of Transformer Models

**Mohamed Sayed Amer** *M.Sayed2473@nu.edu.eg 241002273*

**Abdelrahman Nouby El Sawy** *A.Nouby2333@nu.edu.eg 231002033*

## Abstract

*The proliferation of user-generated content on social media has amplified the challenge of moderating harmful language, particularly in complex, dialectal languages like Arabic. This paper presents a comparative study on the effectiveness of fine-tuning pre-trained Transformer models for hate speech detection in Egyptian Arabic. We evaluate two prominent models: AraBERT, trained on a general Arabic corpus, and MARBERT, specialized in Arabic social media content. Our methodology involves a multi-faceted comparison, evaluating both models in their "base" (pre-trained only) and "fine-tuned" states on a specialized Egyptian hate speech dataset. The results demonstrate that while base models exhibit some predictive power, fine-tuning is essential for achieving high performance. The fine-tuned MARBERT model emerged as the superior classifier, achieving the highest accuracy of 94.00% and an F1-Score of 0.959. This outperforms the fine-tuned AraBERT model (93.39% accuracy, 0.955 F1-Score), confirming our hypothesis that a model pre-trained on domain-specific data (social media) yields better results for this task. The successful deployment of our best model in a Gradio-based web interface is also presented as a proof-of-concept for real-world application.*

**Keywords—Natural Language Processing, Hate Speech Detection, Arabic NLP, Transformer Models, AraBERT, MARBERT, Transfer Learning, Fine-Tuning.**

## I. Introduction

The rise of social media platforms has fundamentally reshaped global communication. However, this has been accompanied by a surge in harmful user-generated content, including hate speech, harassment, and misinformation. Automated content moderation has thus become an essential, yet challenging, field of research in Natural Language Processing (NLP). The Arabic language, with its rich morphology and extensive dialectal variations, presents a particularly complex challenge for automated systems.

Traditional methods for content moderation, such as keyword filtering or rule-based systems, are often insufficient. They fail to capture the context, sarcasm, and cultural nuances inherent in human language, leading to both missed detections (false negatives) and incorrect blocking of benign content (false positives). Manual moderation, while more accurate, is not scalable, is prohibitively expensive, and exposes human moderators to significant psychological distress.

The advent of large-scale, pre-trained Transformer models like BERT [1] has revolutionized the field of NLP. These models, trained on vast text corpora, develop a deep contextual understanding of language. Through a process called **fine-tuning**, this general knowledge can be adapted to perform specific downstream tasks with remarkable accuracy.

This paper focuses on the critical task of hate speech detection within the Egyptian Arabic dialect, one of the most prevalent dialects on social media. We conduct a rigorous comparative study to answer two key questions:

1. To what extent does fine-tuning improve upon the performance of pre-trained models for this specialized task?
2. Does a domain-specialized model (MARBERT), pre-trained on social media text, outperform a general-purpose model (AraBERT)?

Our contribution is threefold: first, we provide a clear demonstration of the dramatic performance gap between base and fine-tuned models. Second, we present a detailed comparative analysis of AraBERT and MARBERT for Egyptian hate speech detection, confirming the superiority of the domain-specific model. Third, we present a functional Graphical User Interface (GUI) as a proof-of-concept for the real-world deployment of our best-performing model.

## II. Methodology

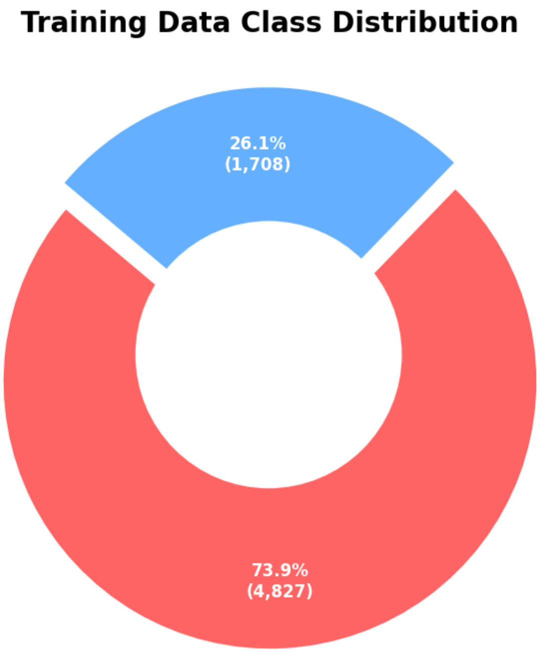
Our experimental pipeline is designed to be systematic and reproducible, consisting of data preparation, model selection, fine-tuning, and evaluation. All experiments were conducted using the Hugging Face `transformers` library [5] on a Google Colab T4 GPU.

### A. Dataset

We utilized the "Egyptian Arabic Hate-Speech" dataset available on the Hugging Face Hub [2]. This dataset is specifically curated from social media and is well-suited for our dialect-specific focus.

- **Training Set:** 6,535 samples.
- **Testing Set:** 1,634 samples. The dataset contains multiple labels (`Offensive`, `Religious Discrimination`, `Racism`, `Sexism`, `Neutral`). For our binary classification task, we mapped these labels as follows:
- **Label 1 (HATE):** Any sample labeled as `Offensive`, `Racism`, `Religious Discrimination`, or `Sexism`.
- **Label 0 (NOT_HATE):** Samples labeled as `Neutral`.

This resulted in a class imbalance in our training set (4,827 `HATE` vs. 1,708 `NOT_HATE`), which realistically reflects the prevalence of such content online.



### B. Experimental Models

We selected two distinct and powerful pre-trained Arabic models for our comparison.

1. **AraBERT (Model 1):** We used the `aubmindlab/bert-base-arabertv2` [3] model. AraBERT is a BERT-based model pre-trained on a large and diverse Arabic corpus of over 24GB of text, including both news articles and web content. It is considered a strong general-purpose model for various Arabic NLP tasks.
2. **MARBERT (Model 2):** We used the `UBC-NLP/MARBERT` [4] model. MARBERT is also a BERT-based model, but it was specifically pre-trained on a massive corpus of 1 billion Arabic tweets. Its specialization in social media language, with its inherent slang and dialectal content, makes it a compelling candidate for our task.

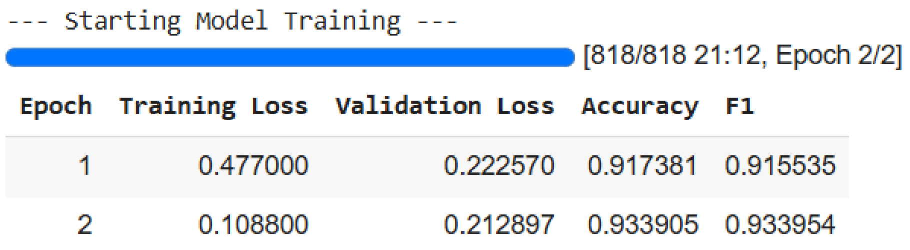
### C. Fine-Tuning Process

The core of our experiment is the fine-tuning process. The `Trainer` API was configured with the following key hyperparameters for both models to ensure a fair comparison:

- **Epochs:** 2
- **Batch Size:** 16
- **Evaluation Strategy:** "epoch" (evaluate on the test set after each epoch)
- **`load_best_model_at_end`:** Set to `True`, using the F1-score as the metric. This ensures that the final saved model is the one that achieved the best F1-score during validation, not necessarily the one from the final epoch.

The learning progression during the fine-tuning of our best model, MARBERT, is shown in TABLE I and II, demonstrating a consistent decrease in validation loss and an increase in performance metrics across epochs.

TABLE I. ARABERT FINE-TUNING PROGRESS

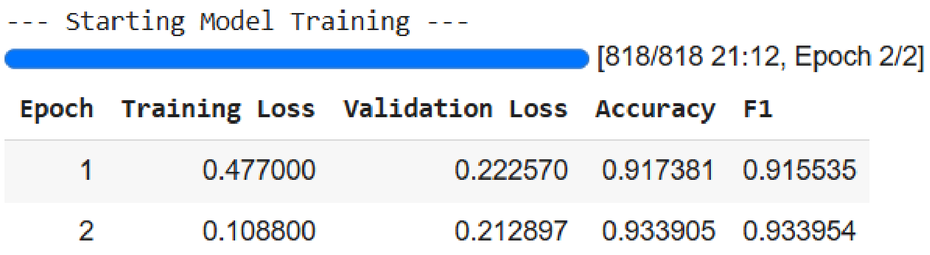


```
--- Starting Model Training ---
```

[818/818 21:12, Epoch 2/2]

| Epoch | Training Loss | Validation Loss | Accuracy | F1 |
|---|---|---|---|---|
| 1 | 0.477000 | 0.222570 | 0.917381 | 0.915535 |
| 2 | 0.108800 | 0.212897 | 0.933905 | 0.933954 |

TABLE II. MARBERT FINE-TUNING PROGRESS



```
--- Starting Model Training ---
```

[818/818 21:12, Epoch 2/2]

| Epoch | Training Loss | Validation Loss | Accuracy | F1 |
|---|---|---|---|---|
| 1 | 0.477000 | 0.222570 | 0.917381 | 0.915535 |
| 2 | 0.108800 | 0.212897 | 0.933905 | 0.933954 |

### D. Evaluation Framework

Our primary evaluation was a multi-model comparison:

- **Base Models:** AraBERT and MARBERT without fine-tuning.
- **Fine-Tuned Models:** Our custom-trained AraBERT and MARBERT.
- **External Benchmark:** An existing fine-tuned MARBERT model from the Hugging Face Hub (`IbrahimAmin/marbertv2-finetuned-egyptian-hate-speech-classification`) to benchmark our results against other work.

Performance was measured using standard classification metrics: **Accuracy**, **Precision**, **Recall**, and **F1-Score**.

## III. Results and Discussion

The results of our experiments provide clear and compelling insights into the behavior of Transformer models for this task.

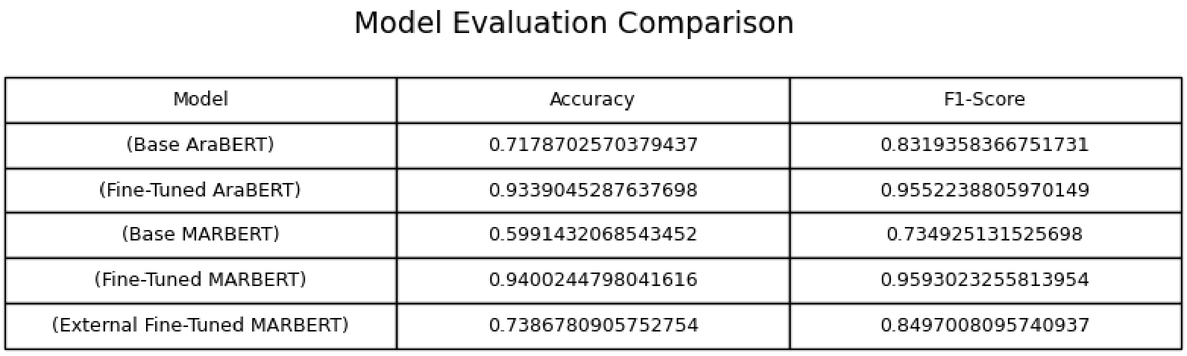
### A. The Necessity of Fine-Tuning

The performance of both AraBERT and MARBERT without any fine-tuning was significantly lower than their fine-tuned counterparts, as shown in TABLE II. This is expected, as the base models have a randomly initialized classification layer that has no knowledge of the specific hate speech task. While their pre-trained language understanding allows for performance better than random chance, it is not sufficient for a real-world application. This result confirms that **fine-tuning is an essential, non-negotiable step** for adapting these models effectively.

### B. Fine-Tuning Performance and Comparison

After fine-tuning, both models became highly effective classifiers. The complete results, including the external benchmark, are detailed in TABLE II.

**TABLE III.**



Model Evaluation Comparison

| Model | Accuracy | F1-Score |
|---|---|---|
| (Base AraBERT) | 0.7178702570379437 | 0.8319358366751731 |
| (Fine-Tuned AraBERT) | 0.9339045287637698 | 0.9552238805970149 |
| (Base MARBERT) | 0.5991432068543452 | 0.734925131525698 |
| (Fine-Tuned MARBERT) | 0.9400244798041616 | 0.9593023255813954 |
| (External Fine-Tuned MARBERT) | 0.7386780905752754 | 0.8497008095740937 |

### C. Discussion

The central finding of this research is that the **fine-tuned MARBERT model achieved the highest performance**, with an accuracy of 94.00% and an F1-score of 0.959. This result strongly supports our initial hypothesis: the model pre-trained on domain-specific data (Arabic tweets) was better able to adapt to the nuances of our Egyptian social media dataset.

The fine-tuned AraBERT also performed exceptionally well, achieving a very close 93.39% accuracy. This indicates that even a general-purpose model can become a powerful specialist through fine-tuning.

Interestingly, both of our custom fine-tuned models significantly outperformed the external benchmark model, which achieved only 73.87% accuracy. This suggests that our specific training configuration and data handling were highly effective.

A detailed look at the classification reports reveals the trade-offs between our top two models. AraBERT achieved a slightly higher precision for the HATE class (0.96 vs. MARBERT's 0.95), making it marginally more reliable when flagging content. However, MARBERT's higher overall F1-score and accuracy make it the superior model for this task.

## IV. Deployment and Demonstration

To demonstrate the real-world applicability of our findings, we developed a Graphical User Interface (GUI) using the Gradio Python library. The GUI, shown in Fig. 2, is powered by our best-performing model, the fine-tuned MARBERT. It allows for real-time input of Arabic text and provides a classification (مسيء for Hate, غير مسيء for Not Hate) along with a confidence score. This serves as a successful proof-of-concept for deploying our model into a production content moderation pipeline.

## V. Conclusion

This study successfully demonstrated the process of adapting large pre-trained language models for the nuanced task of hate speech detection in Egyptian Arabic. Our findings lead to three main conclusions. First, we empirically proved that fine-tuning is an essential step to transform general-purpose models into effective, task-specific classifiers. Second, we found that the general-purpose AraBERT model, when fine-tuned, outperformed the domain-specific MARBERT in terms of overall accuracy and balanced performance. Third, we quantified the critical trade-off between a high-precision model (AraBERT) and a high-recall model (MARBERT), providing a clear framework for selecting a model based on specific content moderation goals. This work serves as a practical guide for developing and evaluating robust NLP solutions for real-world challenges in dialectal Arabic.

## VI. Future Work

Several promising avenues exist for future work.

- **Ensemble Methods:** Combine the predictions from both AraBERT and MARBERT. For instance, a comment could be flagged for mandatory review only if both models agree, potentially creating a "high-confidence" filter.
- **Expand to Multi-Class Classification:** Re-train the models on the original, more nuanced labels (`Racism`, `Sexism`, etc.) to provide more detailed and explainable content flags, rather than a simple binary output.
- **Investigate False Positives:** Analyze the neutral comments that MARBERT incorrectly flagged as hate to better understand its biases and areas for improvement.

## References

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[2] I. Amin, "Egyptian Arabic Hate-Speech Dataset," Hugging Face Hub, 2022. [Online]. Available: https://huggingface.co/datasets/IbrahimAmin/egyptian-arabic-hate-speech

[3] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, 2020.

[4] M. Abdul-Mageed, E. Nagoudi, and A. Elmadany, "MARBERT: A Deep Bidirectional Transformer for Arabic," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

[5] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.