Ensemble Transformers for Arabic Medical Text Classification

Malak Mohamed
Faculty of Computer Science
MSA University
Giza, Egypt
malak.mohamed17@msa.edu.eg

Rokaia Emad
Faculty of Computer Science
MSA University
Giza, Egypt
rokaia.emad@msa.edu.eg

Ali Hamdi
Faculty of Computer Science
MSA University
Giza, Egypt
ahamdi@msa.edu.eg

***Abstract*—The development of social telehealth frameworks has integrated new dimensions of interaction into healthcare, enabling patients to discuss symptoms and obtain remote consultations. Such transformation has paved numerous streams of biomedical data, particularly in Arabic due to its rich morphology and scarcity of annotated resources. In this study, we address the problem of Arabic biomedical question classification (ABQC) using the MAQA dataset, with a focus on a carefully curated subset of ten critical medical categories. We implemented a single training pipeline for five transformer models: AraBERT, BioBERT, DistilBERT, mBERT, and XLM-RoBERTa to ensure they were evaluated consistently. To improve the reliability of the predictions to be both accurate and stable, we utilized an ensemble learning method based on majority voting of individual model outputs. Our ensemble approach achieved the highest classification performance with 92.62% Accuracy and 92.55% F1-score. These results are a step forward in the application of terminology and language resources towards Arabic biomedical text processing and validation of the assumption that ensemble techniques enhance the dependability of responses in telehealth question answering systems.**

***Index Terms*—Arabic Biomedical Question Classification, Social Telehealth, Transformer Models, MAQA Dataset, Arabic NLP, Ensemble Learning, Majority Voting, Healthcare Natural Language Processing**

I. Introduction

Social telehealth has transformed health care provision by enabling remote diagnosis, consultation, and continuous patient monitoring through online platforms. Social media platforms and web-based medical communitites extraordinary growth have caused an exponential increase in user-generated Arabic medical content. The information overload carries significant possibility for disease classification [1], symptom comprehension, and treatment recommendations. Nevertheless, it is difficult to extract useful clinical information from such data due to the informal nature of online text, the existence of extensive dialectal diversity, and the limited availability of annotated Arabic biomedical corpora.

Arabic Biomedical Question Classification (ABQC), where medical questions are categorized into predefined categories like diseases, symptoms, and procedures, plays a crucial role in enabling intelligent health applications such as clinical decision support systems, health-related search engines, and medical chatbots [2].

Despite the remarkable advancements in the domain of natural language processing (NLP) and the rise of Large Language Models (LLMs) such as GPT-3.5 Turbo and LLAMA3, Arabic biomedical NLP is still challenging due to linguistic challenges such as the lack of diacritics, orthographic ambiguity such as "ة" vs. "ه", or "ي" vs. "ى", and heterogeneous biomedical terminology. In contrast to traditional bag-of words representation, Word Embedding like Word2Vec [3], GloVe [4], FastText [5]) fail to capture crucial context, which results in misclassification. As an illustration, the phrases "يعاني من عمى" (”He suffers from blindness”), and "يعاني من عمى الألوان" (”He suffers from color blindness”) [2], have a significant difference in meaning, despite their vocabulary similarity. To overcome these challenges, this work proposes a robust framework for Arabic biomedical question classification and symptom severity analysis in social telehealth. Our main contributions are:

We utilize the MAQA dataset, a specialized Arabic biomedical question set [2].

We also pre-train a number of transformer models like mBERT, Multilingual DistilBERT, BioBERT, XLM-RoBERTa, and AraBERT for the task of ABQC [2], [6].

To improve the prediction reliability and stability of the model, we employed an ensemble of majority-voting, fine-tuned models (mBERT, BioBERT, XLM-RoBERTa, DistilBERT). Employing an ensemble of models significantly boosted general classification reliability and stability.

II. Related Work

Several studies on text classification and fine-tuning approaches specifically customized for medical text processing have been carried out. Some of these studies include depressive symptom detection in Arabic tweets has focused on collecting depressive symptom information on Twitter using AraBERT for classification and merging data augmentation techniques for better model generalization [1], [7], [8]. The dataset consisted of Arabic tweets with depressive symptoms. Issues of privacy were addressed by anonymizing the data and

979-8-3315-0185-3/25/31.00©2025$IEEE$

using augmentation methods to expand the dataset size. The approach achieved very high classification accuracy; however, the challenges were a comparatively small dataset size and domain specificity, which restricted the generalizability of the model to broader medical categories [1]. In general, the outcome demonstrated the effectiveness of domain-adaptation in depression classification. Another paper also examined BERT-based transfer learning approaches fine-tuned for text classification tasks [6]. The study compared how various BERT-based models performed on a range of datasets, most of which were English text classification tasks [6].The data sets utilized in this study involved COVID-19 and extremist data, with tasks such as fake news detection, misclassification of misinformation, and sentiment analysis [6], [9]. The datasets, although carefully created and annotated, were restricted to English texts and lacked Arabic medical content [6]. The study exhibited a significant potential of transfer learning, hence better classification accuracy; however, Its focus on English language materials limited its potential for processing Arabic medical text. Contextual semantic embeddings were also explored through the use of transformer-based models to categorize Arabic biomedical questions [2]. A number of models were trained, and their performance was evaluated on the basis of precision, recall, and F1-score for ten biomedical categories [2]. The dataset that was used in this study was the MAQA dataset, and this was composed of 247,000 Arabic medical questions that had been sourced from websites such as Altibbi, Tbeeb, and Cura [2]. The dataset was preprocessed, anonymized, and made publicly available and hence did not raise any significant privacy issues. The models achieved high classification accuracy, comfortably handling the Arabic morphology complexities [2], [10]. AraBERT generally outperformed the others, while DistilBERT provided faster processing times, and the study significantly enhanced the classification of Arabic biomedical questions [2].

III. Dataset

In this study, we utilize the **MAQA** dataset, a large-scale Arabic biomedical question corpus. MAQA contains approximately **430,000** medical questions collected from Arabic healthcare websites, including **Altibbi.com** (70%), **Tbeeb.net** (20%), and **Cura.healthcare** (10%). The dataset covers a wide range of medical topics and is designed to support Arabic biomedical natural language processing tasks.

For this work, we focused on a subset of ten major medical categories. The distribution of questions across these categories is presented in Table I.

Additionally, Table II provides examples of questions extracted from each of the selected medical categories.

IV. Methodology

This section outlines the pipeline followed for building the Arabic biomedical question classification system using transformer-based models and ensemble learning.

A. Data Preprocessing

The MAQA dataset was preprocessed to ensure consistency and improve model performance. Specifically:

- **Data Selection:** Only questions from ten specific medical categories were retained.
- **Text Cleaning:** Special characters, HTML tags, and excessive white spaces were removed.
- **Normalization:** Arabic text was normalized by unifying different forms of characters (e.g., Alef, Ya) and removing diacritics.
- **Tokenization:** The questions were tokenized using a pre-trained tokenizer with a maximum sequence length of 128 tokens, employing padding and truncation as needed.

B. Model Fine-Tuning

Five different pre-trained transformer models were fine-tuned individually for the classification task:

- AraBERT
- BioBERT
- Multilingual DistilBERT
- mBERT
- XLM-RoBERTa

For consistency and to allow comparison across all of the models, we followed a standard training configuration drawn from established fine-tuning protocols for transformer-based models applied to biomedical natural language processing tasks. The configuration used was informed by previous work [2], where good convergence and strong performance were shown under the same conditions. Each model was adapted by adding a dropout layer followed by a linear classification head to output predictions over the ten categories. All models were trained under the same hyperparameter configuration to ensure a fair comparison. The training configuration was as follows:

- Optimizer: AdamW
- Learning Rate: 3e-5
- Batch Size: 96
- Number of Epochs: 10
- Weight Decay: 0.01
- Loss Function: Cross-Entropy Loss

Training and evaluation were conducted using the Hugging Face `Trainer` framework. Accuracy, Precision, Recall, and F1-Score were used as evaluation metrics during model validation.

C. Ensemble Learning: Majority Voting

After fine-tuning the individual models, an ensemble strategy based on Majority Voting was applied. For each question in the test set, predictions from all five models were collected,

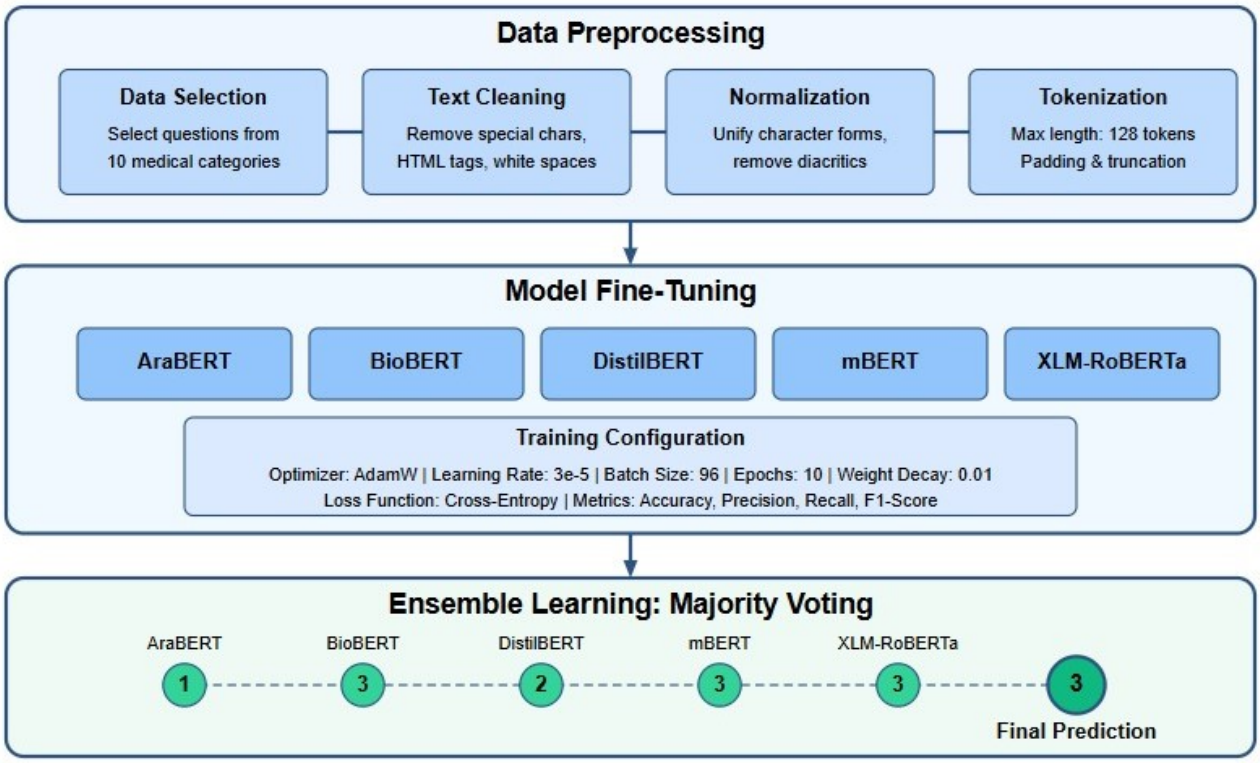

Figure 1: Confusion matrices for the AraBERT, BioBERT, Multilingual DistilBERT, mBERT, XLM-RoBERTa, and Ensemble models across the ten Arabic medical categories.

Table I: Questions number per class in the MAQA dataset.

Number of Questions	Category Name (EN)	Category Name (AR)
70,000	Gynecology diseases	أمراض نسائية
20,000	Ear nose and throat	أنف أذن وحنجرة
19,000	Ophthalmology	أمراض العيون
18,000	Cardiovascular diseases	أمراض القلب والشرايين
17,000	Gastrointestinal diseases	أمراض الجهاز الهضمي
12,000	Sexually transmitted diseases	الأمراض الجنسية
11,000	Dentistry	طب الأسنان
10,000	Plastic surgery	جراحة تجميل
9,000	Blood diseases	أمراض الدم

and the final prediction was determined based on the majority class among the outputs. In the case of a tie, a random choice among the tied classes was applied.

The ensemble approach aimed to combine the strengths of different models, reduce variance, and improve the overall prediction robustness and reliability.

As illustrated in Figure 2, each fine-tuned model predicts independently, and the final classification decision is made based on the majority vote among the models' predictions.

V. Results

A. Quantitative Evaluation

We evaluated the performance of several transformer models, including AraBERT, BioBERT, Multilingue DistilBERT, mBERT, and XLM-RoBERTa. The models were assessed using Accuracy, Precision, Recall, and F1-Score metrics.

Table III summarizes the evaluation results for each model and the ensemble method.

B. Model-wise Performance Analysis

Although all models had excellent overall performance with F1 scores exceeding 90%, there were significant differences in the way they handled specific types of questions.

Table II: Examples of questions from selected medical categories in the MAQA dataset.

Category / الفئة	**Example Question / مثال سؤال**
Blood Diseases / أمراض الدم	ما هي أعراض فقر الدم؟ (What are the symptoms of anemia?)
Plastic Surgery / جراحة تجميل	هل عملية تجميل الأنف خطيرة؟ (Is rhinoplasty surgery dangerous?)
Dentistry / طب الأسنان	كيف يمكن تبييض الأسنان بطريقة آمنة؟ (How can teeth be whitened safely?)
Sexually Transmitted Diseases / الأمراض الجنسية	ما هي أعراض الكلاميديا عند النساء؟ (What are the symptoms of chlamydia in women?)
Gastrointestinal Diseases / أمراض الجهاز الهضمي	ما هو أفضل علاج لقرحة المعدة؟ (What is the best treatment for stomach ulcers?)
Cardiovascular Disease / أمراض القلب والشرايين	كيف يمكن الوقاية من الجلطة القلبية؟ (How can a heart attack be prevented?)
Ophthalmology / أمراض العيون	ما أسباب احمرار العين المستمر؟ (What are the causes of persistent eye redness?)
Ear Nose and Throat / أنف أذن وحنجرة	ما هو علاج التهاب الحنجرة المزمن؟ (What is the treatment for chronic laryngitis?)
Gynecology Diseases / أمراض نسائية	هل تأخر الدورة الشهرية يعني وجود حمل؟ (Does a delayed menstrual cycle mean pregnancy?)
Musculoskeletal and Joint Diseases / أمراض العضلات والعظام والمفاصل	ما هو علاج التهاب المفاصل المزمن؟ (What is the treatment for chronic arthritis?)

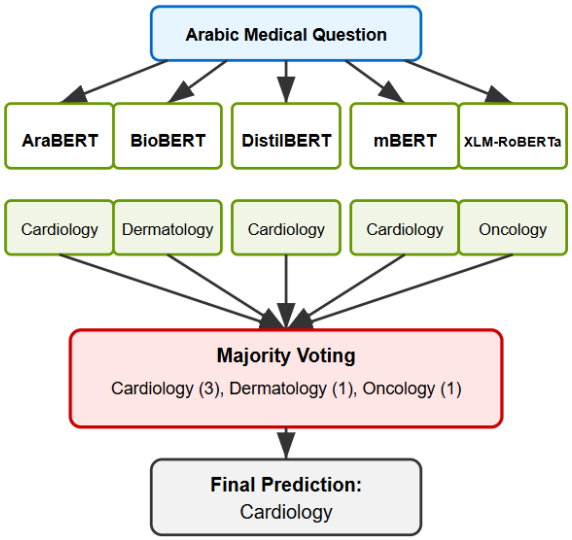

Figure 2: Illustration of the ensemble strategy using Majority Voting across AraBERT, BioBERT, DistilBERT, mBERT, and XLM-RoBERTa models.

Table III: Performance comparison between individual models and the ensemble strategy.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
AraBERT	91.40%	91.32%	91.40%	91.33%
BioBERT	90.68%	90.53%	90.68%	90.59%
Multilingue DistilBERT	91.40%	91.37%	91.40%	91.35%
mBERT	91.46%	91.41%	91.46%	91.41%
XLM-RoBERTa	91.67%	91.62%	91.67%	91.60%
Ensemble (Majority Voting)	**92.62%**	**92.56%**	**92.62%**	**92.55%**

AraBERT, having been pretrained on Arabic text, handled morphologically rich input well and had balanced precision and recall. Its absence of biomedical domain specialization, however, caused confusion on questions with specialized medical terms.

BioBERT performed slightly worse, mainly because of its English-language biomedical pretraining. While domain-sensing, it had difficulties with Arabic sentence structure and tokenization, leading to higher rates of misclassification for more linguistically complex inputs.

Multilingual DistilBERT scored similarly to AraBERT, though it was more inconsistent between extremely closely related categories. Its more compact architecture, streamlined as it was, may have restricted the model from being able to differentiate between semantically fine-grained question types.

mBERT was a bit more consistent, presumably due to its more comprehensive multilingual training. However, it still displayed moderate confusion in categories requiring precise semantic interpretation.

XLM-RoBERTa worked best among the other individual models, achieving better accuracy and more generalizability for various categories. Its cross-lingual training at large scale allowed improved generalizability of both general and in-domain Arabic queries. Finally, the ensemble model that averaged predictions with majority voting achieved the best overall performance. The model effectively reduced the errors of individual models and addressed edge cases with improved consistency, thereby demonstrating the strengths of using the capabilities of different models.

Figures 3 show the confusion matrices for each model, providing insights into classification strengths and weaknesses across the ten medical categories.

C. Error Analysis and Limitations

Despite their good general performance, all models had some weaknesses, as revealed through confusion matrix analysis. Errors were most commonly committed in instances where categories shared overlapping semantic meaning or required

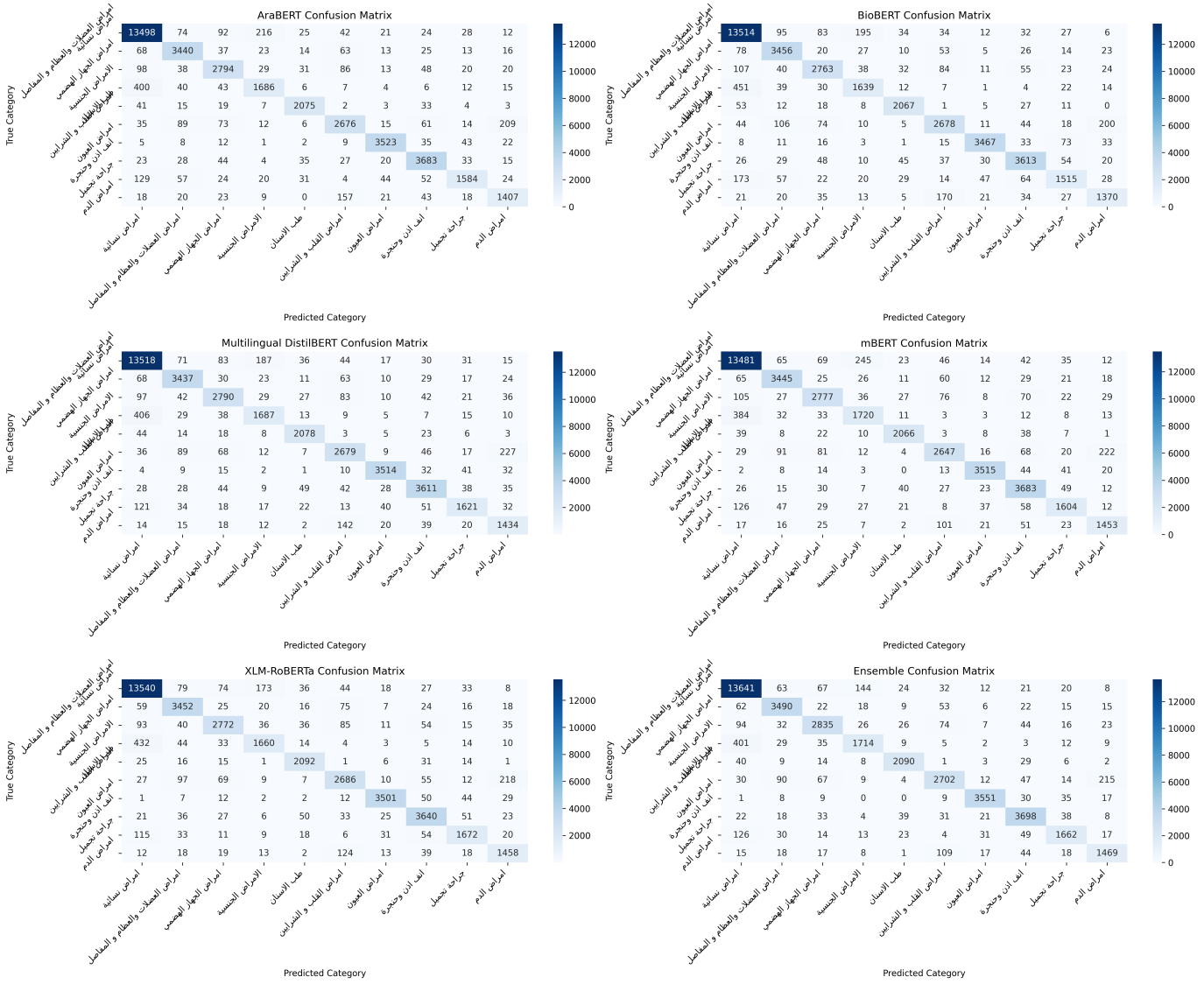

Figure 3: Confusion matrices for the AraBERT, BioBERT, Multilingual DistilBERT, mBERT, XLM-RoBERTa, and Ensemble models across the ten Arabic medical categories.

domain-specific inference. For instance, questions with vague wording or compound medical intents were occasionally misclassified into semantically similar categories.

AraBERT, while effective at processing Arabic morphology, was not domain-specific and therefore struggled with the meaning of specialized biomedical terminology. BioBERT, conversely, struggled with Arabic syntax and tokenization, and while biomedically specialized, its application was restricted.

Multilingual DistilBERT and mBERT were both constrained in handling fine-grained semantic variation due to their generalized multilingual pretraining, which induced perplexity in questions requiring fine-grained categorization. XLM-RoBERTa, although more stable, still incurred minor errors for less frequent or challenging question types, pointing towards sensitivity to class imbalance and input length.

The ensemble model mitigated most of these issues by merging diverse predictions, yet it was not immune to errors where poorly most of the base models misclassified a highly perplexing input. Overall, these findings indicate the need for domain-adapted pretraining and more balanced datasets for making classification accuracy even better.

VI. Conclusion

This study demonstrates the efficiency of transformer models for classifying Arabic biomedical questions according to the MAQA dataset. By targeting a small number of major medical categories, we obtained satisfactory performance. The ensemble method improved performance even more by taking advantage of the strength of various models to finally reach a classification accuracy of 92.62%.

Future work will look at the inclusion of other classes in the classification system. It will also employ improved ensemble methods (e.g., weighted voting, stacking) to enable the system to perform better in real telehealth application.

References

[1] S. Elmajali and I. Ahmad, "Toward early detection of depression: Detecting depression symptoms in arabic tweets using pretrained transformers," *IEEE Access*, vol. 12, pp. 88 134–88 145, 2024. [Online]. Available: https://doi.org/10.1109/access.2024.3417821

[2] I. Ait Talghalit, H. Alami, and S. Ouatik El Alaoui, "Contextual semantic embeddings based on transformer models for arabic biomedical questions classification," *HighTech and Innovation Journal*, vol. 5, no. 4, p. 1024, 2024. [Online]. Available: https://www.hightechjournal.org/index.php/HIJ/article/view/1024

[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, vol. 26, pp. 3111–3119, 2013. [Online]. Available: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

[4] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: https://doi.org/10.1162/tacl_a_00051

[6] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. A. Almazroi, "A fine-tuned bert-based transfer learning approach for text classification," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, 2022. [Online]. Available: https://doi.org/10.1155/2022/3498123

[7] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," in *Proceedings of the LREC 2020 Workshop on Resources for African Indigenous Languages*, 2020, pp. 9–15. [Online]. Available: https://aclanthology.org/2020.lrec-1.2

[8] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6382–6388, 2019. [Online]. Available: https://aclanthology.org/D19-1670

[9] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3606–3611, 2019. [Online]. Available: https://aclanthology.org/D19-1371

[10] M. Abdul-Mageed, A. Elmadany, and A. Hashemi, "Aravec 2.0: Fine-tuned word embeddings for arabic," *Information Processing Management*, vol. 58, no. 5, p. 102682, 2021.