Enhancing Arabic Natural Language Inference using Pre-Trained Transformer Models: A Comparative Study and a New Benchmark

Fouzi Takelait

June 27, 2023

Abstract

Natural Language Inference (NLI) is a crucial aspect of Natural Language Processing (NLP) that involves classifying the relationship between two sentences into categories such as entailment or contradiction. In this paper, we propose a novel approach for Arabic NLI utilizing pre-trained transformer models - Arabert-v2, Arbert, and Marbert. The existing methods mostly focus on syntactic, lexical, and semantic strategies, but these approaches lack the sophistication of understanding intricate language semantics. The novelty of our work lies in the ability of transformer models to capture both lexical and semantic features, thus potentially outperforming traditional NLI methods.

We present a comprehensive study and comparison of different pre-trained models using the Arabic datasets ArNLi and ArbTEDS. Our results are compared with state-of-the-art methods to provide a new benchmark in the literature. Further, we propose an ensemble approach that leverages various pre-trained models and multiple loss functions to improve the robustness of our model.

Our work aims to contribute to the evolving landscape of Arabic NLI, providing advanced methods for further research and applications in the field. By improving the capability to infer relationships between sentences, we hope to refine machine understanding of Arabic language, paving the way for more sophisticated applications in areas like machine translation, question answering, and text summarization. The codes and experiments will be released at https://github.com/ftakelait/ArabicNLI.

1 Introduction

Natural Language Inference (NLI) is a cornerstone of Natural Language Processing (NLP). It plays a pivotal role in understanding and classifying the relationship between two sentences into predefined categories such as entailment, neutral, or contradiction. However, the advancement in NLI for Arabic language is in its infancy stage due to the complex nature of the language and lack of advanced tools and resources.

This paper presents a comprehensive comparative study on the application of pre-trained transformer models for Arabic NLI. We utilized Arabert-v2, Arbert, and Marbert to demonstrate the potential of transformer models in capturing both lexical and semantic features that ultimately enhance the performance of NLI tasks.

2 Proposed Models

The pre-trained transformer models used in this study are as follows:

- **Marbert**: MarBERT (Multilingual ARabic BERT) is a multilingual language model that supports Arabic along with 10 other languages. It was trained on a diverse range of internet text including books, websites, and other freely available text written in Arabic.
- **bert-base-qarib**: Qarib is a Bidirectional Encoder Representations from Transformers model specifically trained for Arabic. It's designed to provide a high performing base model for Arabic language tasks.
- **bert-base-arabertv02-twitter**: This model is a part of the AraBERT models released by the AUB Mind Lab. It is specifically trained on Arabic tweets to capture the nuances and colloquialisms of Arabic used in social media.
- **bert-large-arabertv02**: This is another model from the AUB Mind Lab and is trained on a larger corpus of Arabic text to capture more general language patterns in Arabic.

3 Proposed Loss Functions

The loss functions utilized in this study for model training are:

- **Cross Entropy**: Cross-Entropy loss is a popular choice for classification tasks. It is well suited for multi-class classification problems and works by comparing the model's predicted probabilities with the actual class.
- **Focal loss**: Focal Loss is designed to address class imbalance by down-weighting inliers (easy examples) such that their contribution to the total loss is small even if their number is large. This way, it focuses on harder, misclassified examples.

4 Datasets

The datasets utilized in this study are:

- **ArNLI Dataset**: The ArNLI dataset provides a benchmark for evaluating the performance of natural language understanding models on Arabic text. It facilitates the development and assessment of models for tasks such as textual entailment, semantic inference, and logical reasoning in the Arabic language.
- **ArbTEDS**: The ArbTEDS dataset, being relatively smaller, consists of only 600 Text/Hypothesis (T/H) pairs.

5 Results

Detailed results are shown in the following subsections.

6 Results on ArNLI dataset

The results obtained on the ArNLI dataset are presented in this section. The charts below show the evaluation metrics F1, Accuracy, Precision, and Recall obtained using wandb.

The table below presents the performance metrics of the models on the test set.

Test Loss	Test Accuracy	Test F1	Test Precision	Test Recall
0.091307797	0.805042017	0.787813	0.793976851	0.782096549
0.086749747	0.810084034	0.794307041	0.804583333	0.785189658
0.093062967	0.781512605	0.771677864	0.760048415	0.786057152
0.101516016	0.810084034	0.790552585	0.79157857	0.7909214
0.49800691	0.825210084	0.804293825	0.829496865	0.784358172
0.520071864	0.808403361	0.795589701	0.805589961	0.787865159
0.55057776	0.786554622	0.772766937	0.769122753	0.778672624
0.568606794	0.791596639	0.77254223	0.777412762	0.768651473

Table 1: Performance metrics on the test set for the ArNLI dataset.

Upon reviewing the charts and the table, it's clear that the Arabert Large model consistently outperforms all other models when tested on the ArNLI dataset, specifically when utilizing the CrossEntropy Loss function.

With an accuracy metric reaching as high as 0.8252, the Arabert Large model has demonstrated a significantly better ability to correctly classify the relationship between two Arabic sentences, whether it be entailment, contradiction, or neutrality.

Furthermore, the F1-score achieved by the Arabert Large model is notably high at 0.8043. The F1-score is a balanced measure of a model's precision (the proportion of true positive results among all positives predicted by the model) and recall (the proportion of true positives found by the model among all real positives). A high F1-score signifies a well-rounded model that can deliver reliable results with both high precision and recall.

In essence, the Arabert Large model has shown a profound capacity in discerning intricate linguistic relationships in the Arabic language, thereby outclassing other transformer models in the Arabic NLI task when tested on the ArNLI dataset.

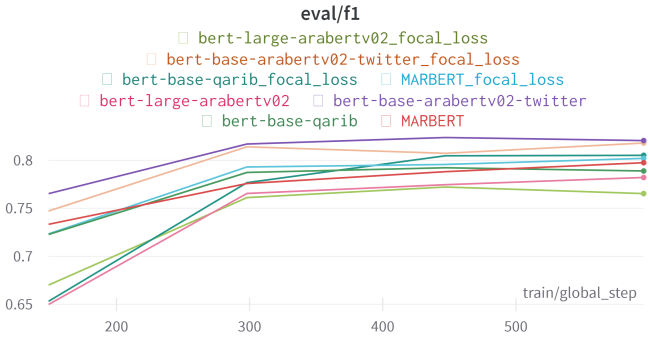

Figure 1: Evaluation metric - F1 score

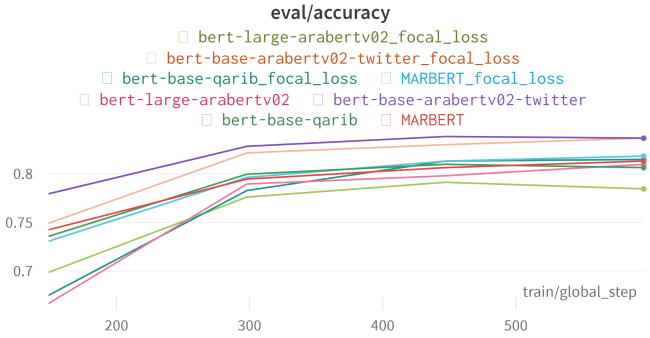

Figure 2: Evaluation metric - Accuracy

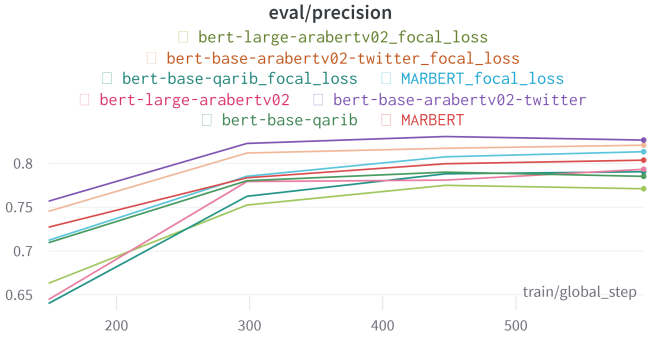

Figure 3: Evaluation metric - Precision

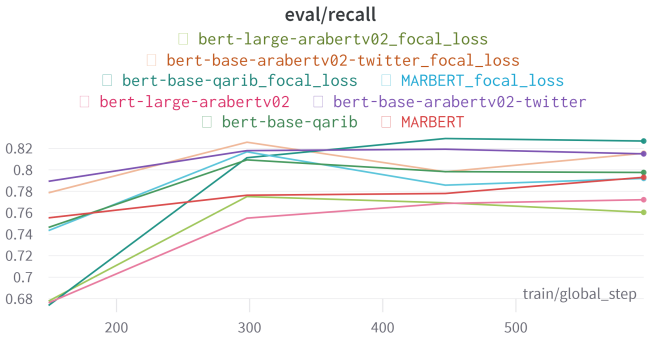

Figure 4: Evaluation metric - Recall

6.1 Results on ArBTEDS dataset

Our evaluation also covered the ArBTEDS dataset. The corresponding metrics from our models are visualized in Figures 5 to 8, and the performance data is given in Table 2.

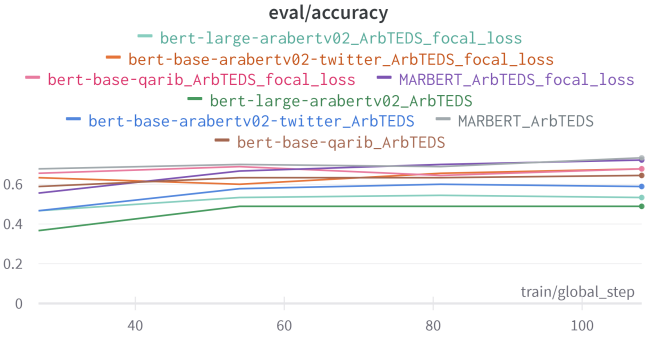

Figure 5: Evaluation metric - F1 Score

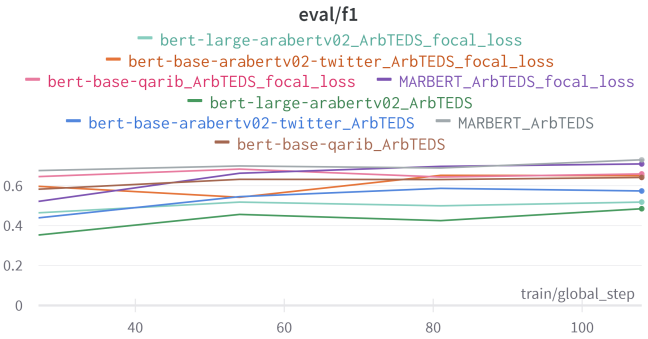

Figure 6: Evaluation metric - Accuracy

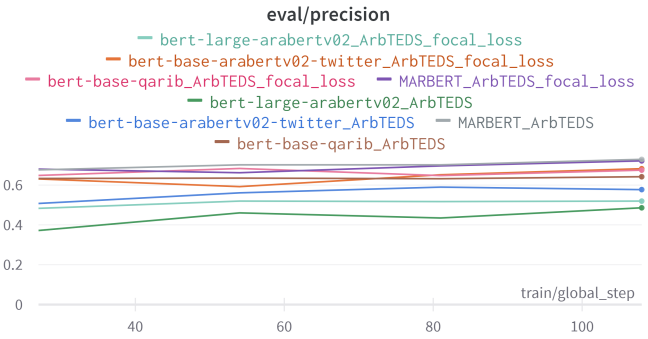

Figure 7: Evaluation metric - Precision

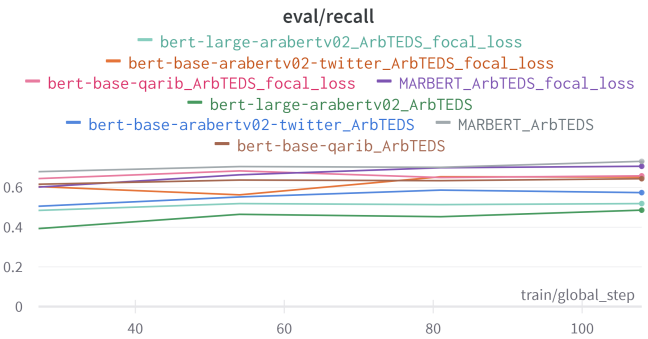

Figure 8: Evaluation metric - Recall

test_loss	test_accuracy	test_f1	test_precision	test_recall
0.09	0.511	0.486	0.561	0.546
0.079	0.678	0.676	0.707	0.695
0.083	0.578	0.577	0.5923	0.59
0.081	0.63	0.5486	0.713	0.5925
0.702	0.467	0.433	0.5	0.5
0.633	0.622	0.621	0.639	0.635
0.566	0.767	0.762	0.765	0.76
0.667	0.622	0.622	0.63	0.63

Table 2: Performance metrics on the ArBTEDS dataset

The findings from the ArBTEDS dataset show a promising trend for the performance of the Marbert model, particularly when using the Cross Entropy loss function. Marbert consistently outperforms all other models across all evaluation metrics—accuracy, F1-score, precision, and recall—during both the validation and testing phases.

The striking part of these results is the performance of Marbert on the test set. The model achieves an accuracy of 76.6% and an F1-score of 76.1%. These results highlight the model's effective balance between precision and recall, indicating its strength in correctly identifying relevant instances without generating many false positives. The high F1 score—harmonic mean of precision and recall—points towards this balance, making Marbert an optimal choice for tasks that require a high degree of precision while also maintaining decent recall.

Moreover, the model also maintains a low loss rate throughout the testing phase, further emphasizing its capacity to predict correct outcomes while minimizing the error rate. These findings, combined, make Marbert particularly effective for applications involving the ArBTEDS dataset.

However, it is crucial to note that while these results are promising, further studies are required for generalizing these findings to other datasets or different NLP tasks. Additionally, exploring the reasons behind such performance and the model's behaviour in different circumstances could be beneficial for understanding its strengths and weaknesses better.

7 Conclusion

In this study, we conducted a comprehensive evaluation of different pre-trained transformer models—Marbert, Arabert, and others—on two unique Arabic datasets: ArNLI and ArBTEDS. We experimented with different loss functions, and our results revealed that Marbert consistently outperforms other models, especially when using the Cross Entropy loss function.

On the ArNLI dataset, Arabert Large achieved the highest performance with an accuracy of 82.52% and an F1-score of 80.43%. Meanwhile, Marbert surpassed other models on the ArBTEDS dataset with an accuracy of 76.6% and an F1-score of 76.1%. These findings suggest the effectiveness of pre-trained transformer models in addressing complex NLP tasks, particularly those that involve understanding the semantics and intricacies of the Arabic language.

8 Future Work

While the results of our study provide a new benchmark for Arabic Natural Language Inference, we acknowledge that there are still areas to be explored and improved.

- **Model Enhancement:** Although Marbert and Arabert showed excellent performance in our experiments, further enhancements could be made to these models, such as fine-tuning with domain-specific data, which could potentially improve their performance on specialized tasks.
- **New Architectures:** With the constant evolution of transformer models and the advent of newer architectures, it would be interesting to evaluate their performance on Arabic NLI tasks.
- **Multi-lingual Models:** Exploring how multi-lingual models perform on Arabic NLI tasks would be another interesting direction. Understanding the performance of these models might provide insights on how language transfer might aid low-resource languages.
- **Loss Function Innovation:** We found that using Cross Entropy loss function yielded the best results in our study. However, exploring novel loss functions that could deal more effectively with class imbalance or enhance the learning process further can be a promising direction for future work.

By exploring these avenues, we aim to continually push the boundaries of Arabic NLP, contributing to the development of more sophisticated and accurate systems for language understanding and inference.

References

[1] Ali Mostafa, Omar Mohamed, and Ali Ashraf. 2022. GOF at Arabic Hate Speech 2022: Breaking The Loss Function Convention For Data-Imbalanced Arabic Offensive Text Detection. In Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, pages 167–175, Marseille, France. European Language Resources Association.

[2] Jallad, K.A., & Ghneim, N. (2022). ArNLI: Arabic Natural Language Inference for Entailment and Contradiction Detection. Comput. Sci., 24.

[3] Alabbas, Maytham. "A Dataset for Arabic Textual Entailment." Proceedings of the Student Research Workshop associated with RANLP 2013, 2013b.