# KNOWLEDGE DISTILLATION OF BERT LANGUAGE MODEL ON THE ARABIC LANGUAGE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The absence of good Arabic language models led to significant setbacks in the Arabic language related tasks and lag with respect to robustness and accuracy. While a pre-trained version of BERT on Arabic language is available, a smaller distilled version could be proven to be highly scalable. In this research paper, we propose the development of a smaller and more efficient version of BERT, known as DistilBERT for the Arabic language for the pursuit of achieving comparable results with significantly less computational resources. Employing knowledge distillation to create a compact model allows for wider implementation, even in areas with limited computational resources. Ultimately, this project aims to break down language barriers, bring greater inclusivity and improve the accessibility of the Arabic language in NLP applications worldwide. This project serves as a starting point for further research and investigation of the performance of the Arabic DistilBERT model across various NLP tasks.

## 1 INTRODUCTION

Despite the prevalence of the Arabic language worldwide, it remains one of the most underrepresented languages in the field of natural language processing (NLP). Arabic has a rich and complex morphology distinct from other languages, which poses a significant challenge for NLP models and tasks thus restricting their potential and limiting their scope.
On the contrary, the burgeoning field of natural language processing (NLP) has seen considerable progress in recent years. One such model is BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), a transformer-based architecture that has shown remarkable results in enhancing the fine-tuning approaches of various NLP tasks, However, the high computational cost and large corpus required for its pre-training have limited the availability of BERT to English and other high-resource languages. Even though a multilingual BERT model was trained, a BERT model pre-trained specifically for the Arabic language (Antoun et al., 2020) was developed to address the suboptimal performance achieved by the multilingual BERT for the Arabic language. Nonetheless, operating the Arabic BERT under resource constraints remains challenging for many real-world applications. To overcome this limitation, we propose employing the knowledge distillation technique to distill the knowledge learned by the large Arabic BERT into a smaller more efficient model, DistilBERT (Sanh et al., 2019) for the Arabic Language, while maintaining its performance. We aim to maximize the performance of the Arabic DistilBERT for Question Answering and to identify any challenges or limitations in using this model for this task. Through this work, we provide a valuable resource for Arabic speakers seeking information through natural language interactions. Our ultimate goal is to facilitate inclusivity and accessibility of the Arabic language in NLP technologies thereby breaking down language barriers and promoting cross-cultural communication.

## 2 METHOD

In this paper, we suggest developing an Arabic DistilBERT Model using the knowledge distilled from an Arabic BERT Model. Initially, we've introduced additional pre-processing to the data used before training the model. The dataset used in this research was carefully selected to be public, large, relatively new with modern language, correct grammar and annotated correctly. As a result, we merged a few datasets, including OSCAR 2019 (Open Super-large Crawled Aggregated CoRpus), Wikipedia Corpus, Arabic BERT Corpus and ARCD ( Arabic Language Comprehension Data) in

Table 1: Comparing the performance of the models on Question Answering (QA) task

| $Model$ | $QA(F1)$ |
|---|---|
| Arabic DistilBERT | 43.47 |
| Arabic BERT-base | 43.76 |
| DistilBERT | 63.72 |

text format. Aside from removing duplication, a pre-processing script was used to process the dataset according to techniques such as removing URLs, English words, emoticons, symbols flags, pictographs, transport and map symbols and unicodes. The results was a collection of over 90 million sentences of text. The goal was to create a balanced final version of the dataset with no duplication and suitable sentence length.
For the training stage, we employed the Knowledge Distillation technique, which is a compression technique used to reproduce the behavior of the larger model into a smaller model using a triple loss function 2. We used a pre-trained Arabic BERT model which has a general understanding of the context of the Arabic Language. The distilled model is then trained on the smaller corpus of the previously pre-processed Arabic dataset, leveraging on the knowledge transferred from the pre-trained Arabic BERT model.

$$L = \alpha * L_{distill} + \beta * L_{training} + \gamma * L_{cosine} \tag{1}$$

## 3 Experimental Setup

In this study, we used a BERT-large model for Arabic Language as the teacher model. To develop the DistilBERT model, we initialized it from the final weights of the Arabic BERT model. The configurations of the student were modified to match those of the teacher, including the transformer layer's hidden state size set to $1024$, the attention heads to $16$, the intermediate size to $4096$ and vocabulary size to $32000$. The pre-training was done using $4$ Tesla V100 GPUs, each of 32GB RAM. The pre-training process was carried out with a batch size of $16$, a learning rate of $3\mathrm{e}{-5}$, and maximum sequence length of $512$ running on 3 epochs.

## 4 Results

Being constrained to the resources available and the large data size, the model training is currently ongoing. To showcase some results, we obtained a checkpoint from the model upon the completion of the first epoch. Table 1 shows the experimental results of testing the performance of Arabic DistilBERT model on Question Answering task, compared against the performance of Arabic BERT-base model and English DistilBERT model. The results indicate that the Arabic DistilBERT achieved comparable performance to that of the Arabic BERT-base model. Moreover, It's generally expected that the Arabic DistilBERT model to be of lower accuracy than its English counterpart given the complexity of the Arabic language and the size of the training data available. These findings provide evidence that the Arabic DistilBERT model can achieve competitive results on Question Answering tasks while requiring fewer computational resources.

## 5 Conclusion and Future Work

We present Arabic DistilBERT, a language model that meets similar performance levels to its larger counterpart with significantly less computational resources. We showed that the model has the potential to notably contribute to the development of Arabic language-related tasks, improve its accuracy and robustness and thereby acts as a valuable addition to the natural language processing toolkit for the Arabic language. For future work, we aim to test the effectiveness of using the Arabic DistilBERT model across various NLP tasks.

## URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

## REFERENCES

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. arxiv 2020. *arXiv preprint arXiv:2003.00104*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

## A APPENDIX

Table 2: Examples from Arabic DistilBERT Model fine-tuned on Question Answering

| Context | Question | Answer | Score |
|---|---|---|---|
| عجائب الدنيا السبع الجديدة، هو مشروع استثماري أطلقه منتج أفلام كندي من أصل سويسري يدعى برنارد فيبر عام ١٩٩٩، وذلك من خلال شركة ربحية أنشأها باسم مؤسسة العالم المفتوح الجديد ( و اوپن وارلد ثرپرتـ■ن). | من هو صاحب مشروع عجائب الدنيا السبع؟ | برنارد فيبر | 0.953 |
| على جانبي الحدود الثقافية بين أوروبا الجرمانية واللاتينية، وفي بلجيكا لغتين رئيسيتين هما: الهولندية المتكلمين بها (حوالي ٥٩ % )، ومعظمهم الفلمنكية، والمتكلمين بالفرنسية (حوالي ٤١ % )، ومعظمهم الولونيين، بالإضافة إلى مجموعة صغيرة من المتحدثين بالألمانية. اثنين من أكبر المناطق في بلجيكا هي المنطقة الناطقة باللغة الهولندية الفلاندر في الشمال والمنطقة الناطقة بالفرنسية جنوب والونيا. في إقليم العاصمة بروكسل، ثنائية اللغة رسميا، هو جيب الناطقة بالفرنسية في الغالب داخل الإقليم الفلمنكي.* | بلجيكيا لها لغتين رئيسيتين. اي من اللغتين يستخدمونه اكثر؟ | الهولندية | 0.836 |
| شهد حكم بشار الأسد بدأ موجة احتجاجات بدأت في ١٥ آذار عام ٢٠١١ ولا زالت رافعة شعارات ضد القمع والفساد وكبت الحريات ومطالبة بإسقاط النظام البعثي الذي استخدم ضدها الأسلحة الثقيلة وقوات الشبيحة بحسب منظميها في تحد غير مسبوق لحكم بشار الأسد وحزب البعث السوري ومتأثرة بموجة الاحتجاجات العارمة التي اندلعت في الوطن العربي أواخر عام ٢٠١٠، في حين أعلنت الحكومة السورية أن هذه الحوادث من تنفيذ متشددين وإرهابيين من شأنهم زعزعة الأمن القومي وإقامة إمارة إسلامية في بعض أجزاء البلاد. وقد نتج عن قمع هذا الحراك الشعبي الآلاف من الضحايا والجرحى والمعتقلين، بالإضافة إلى الخسائر في صفوف القوات المسلحة. | متى بدات الموجة الاحتجاجية فى عهد بشار؟ | ١٥ آذار عام ٢٠١١ | 0.958 |