Benchmarking Transformer-based Language Models for Arabic Sentiment and Sarcasm Detection

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proceedings of the Sixth Arabic Natural Language Processing Workshop

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Benchmarking Transformer-based Language Models for Arabic Sentiment and Sarcasm Detection

Ibrahim Abu Farha¹ and Walid Magdy¹,²

¹School of Informatics, The University of Edinburgh, Edinburgh, United Kingdom
²The Alan Turing Institute, London, United Kingdom
i.abufarha@ed.ac.uk, wmagdy@inf.ed.ac.uk

Abstract

The introduction of transformer-based language models has been a revolutionary step for natural language processing (NLP) research. These models, such as BERT, GPT and ELECTRA, led to state-of-the-art performance in many NLP tasks. Most of these models were initially developed for English and other languages followed later. Recently, several Arabic-specific models started emerging. However, there are limited direct comparisons between these models. In this paper, we evaluate the performance of 24 of these models on Arabic sentiment and sarcasm detection. Our results show that the models achieving the best performance are those that are trained on only Arabic data, including dialectal Arabic, and use a larger number of parameters, such as the recently released MARBERT. However, we noticed that AraELECTRA is one of the top performing models while being much more efficient in its computational cost. Finally, the experiments on AraGPT2 variants showed low performance compared to BERT models, which indicates that it might not be suitable for classification tasks.

1 Introduction

In recent years, the development of contextualised language representations led to a revolution in the natural language processing (NLP) field. Early work on representing language started with pre-trained word representations such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2017). However, these word representations were static and did not rely on the context, in which they appear. Other works tried to address this issue with contextualised word representations such as ELMo (Peters et al., 2018). Howard and Ruder (2018) proposed ULMFiT, which is a method for transfer-learning that can be applied to any task in NLP. The introduction of BERT (Devlin et al., 2019) revolutionised the research on NLP. BERT is based on complicated neural network models, namely transformers. The utilisation of BERT led to achieving state-of-the-art results on many NLP tasks such as sentence completion, question answering and sentiment analysis. The advantage of BERT and similar models is that they are pre-trained on large amounts of data, which the model utilises to learn a representation of the language. Following BERT, many other transformer-based language models were released such as ELECTRA (Clark et al., 2020), GPT-1/2/3 (Radford et al., 2019; Brown et al., 2020) and T5 (Raffel et al., 2020). These models tried to improve the performance of BERT through some slight modifications on the training objective such as RoBERTa (Liu et al., 2019). Or completely changing the training methodology such as ELECTRA (Clark et al., 2020). These models were mostly specific for English with the exception of some multilingual ones such as the multilingual BERT (mBERT) (Devlin et al., 2019). Researchers on other languages followed the trend and released other language-specific models such as CamamBERT (Martin et al., 2020) for French, PhoBERT (Nguyen and Tuan Nguyen, 2020) for Vienamese, FinBERT (Virtanen et al., 2019) for Finnish, BERTje (de Vries et al., 2019) for Dutch and others.

AraBERT (Antoun et al., 2020) was the first Arabic-specific transformer-based language model. The introduction of AraBERT helped improving the performance in many Arabic NLP tasks. Recently, a large set of transformer-based Arabic language models has been released. These include BERT based models such as the new large version of AraBERT (Antoun et al., 2020), QARIb (Chowdhury et al., 2020), ARBERT/MARBERT (Abdul-Mageed et al., 2020). Also, Arabic variants of other models were released such as AraGPT2
(Antoun et al., 2021b), AraELECTRA (Antoun et al., 2021a) and Arabic ALBERT (KUIS-AI-Lab). These models vary in their architectures, sizes and the nature of their training data. While most of these models were trained on modern standard Arabic (MSA) data; some of them, such as MARBERT, included dialectal Arabic in their training data.

Most of these models were evaluated on a small set of Arabic NLP tasks and without any direct comparison with other models. Thus, there is no clear measure of the effectiveness of one of these models on a specific task compared to the others. Hence, in this paper, we aim to provide a comparative study of the performance of all of the recently introduced Arabic language models for the well-studied Arabic sentiment analysis (SA) task and the emerging Arabic sarcasm detection task.

In this paper, we provide a rigorous comparison of the effectiveness of 24 recently-released Arabic language models. These models were evaluated for the tasks of Arabic sentiment analysis and sarcasm detection. We test these models on the newly released ArSarcasm-v2 dataset (Abu Farha et al., 2021), which was released along with the shared task on Arabic sarcasm detection. The experiments show the following: First, the models trained on dialectal Arabic are the most effective to handle the tasks under study, where the best model MARBERT achieved an $F_{PN}$ of 0.724 on the SA task and F1-sarcastic of 0.584 on the sarcasm detection task. Second, language-specific models achieve higher results than the multilingual ones. Third, the training procedure of ELECTRA is better than the other models, as AraELECTRA (Antoun et al., 2021a) achieved results close to larger models while being smaller and more efficient. Fourth, AraGPT2 is not suited for classifications tasks as it performed worse than all the other models.

2 Related Work

2.1 Arabic Language Models

The introduction of Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2019) led to a revolution in the NLP world. Since then, many other models have been released such as ELECTRA (Clark et al., 2020), GPT-1/2/3 (Radford et al., 2019; Brown et al., 2020) and RoBERTa (Liu et al., 2019). Those models helped achieving state-of-the-art results on different tasks such as sentiment analysis, named entity recognition (NER), sentence completion and others. However, those models were trained mostly on English data while others included data from other languages such as the multilingual BERT (Devlin et al., 2019). Recently, Arabic NLP researchers started training Arabic variants of these models such as the works of (Antoun et al., 2020, 2021a; Chowdhury et al., 2020; Abdul-Mageed et al., 2020). AraBERT (v0.1/v1) (Antoun et al., 2020) was built using the same architecture as BERT-base (Devlin et al., 2019). AraBERT was trained using a combination of different Arabic news corpora. The authors utilised Farasa (Abdelali et al., 2016) for the pre-processing and segmentation, then they trained a SentencePiece tokenizer (Kudo, 2018) on the segmented text with a vocabulary of 60K subword tokens. Recently, the authors released AraBERT (v0.2/v2), which was trained on a larger dataset of 77GB of text. The authors trained two variants of AraBERT based on BERT-base and BERT-large architectures. AraBERT was evaluated on three tasks: named entity recognition (NER), question answering and sentiment analysis. In another work, the authors released AraELECTRA (Antoun et al., 2021a), which is trained using the same architecture and procedure used to train the original ELECTRA model (Clark et al., 2020). AraELECTRA was trained using the same preprocessing used with AraBERT(v2) and using the same data. AraELECTRA was tested only for question answering task, where it achieved state-of-the-art results on multiple datasets. In (Antoun et al., 2021b), the authors released AraGPT2, which is based on the original GPT2 (Radford et al., 2019) architecture, this model was also trained on the same data used for AraELECTRA and AraBERT(v2). Since GPT2 is trained using causal language modelling objective, the authors did not test the model on any datasets and relied on the perplexity reported during training.

In (Lan et al., 2020a), the authors pre-trained a customised bilingual BERT, GigaBERT, that is designed specifically for Arabic NLP and English-to-Arabic zero-shot transfer learning. The training data was around 13M news articles collected from different sources. They also augmented their data with code-switched samples to improve the cross-lingual performance. GigaBERT was evaluated on multiple NLP tasks such as: NER, part of speech (POS) tagging, relation extraction and argument role labelling. In (Abdul-Mageed et al., 2020), the authors proposed two new Arabic specific BERT
models, ARBERT and MARBERT. For ARBERT, they used BERT-base architecture and 61 GB of text as training data. The training data for ARBERT is mostly in modern standard Arabic (MSA) with a small portion in Egyptian dialect. For MARBERT, the authors aimed to improve the model’s ability to handle dialectal Arabic. They enriched their training data through adding a set of randomly sampled 1B Arabic tweets. The final training dataset was around 128 GB of text, 50% of which are tweets. Abdul-Mageed et al. (2020) provide an extensive evaluation of their models on many tasks such as sentiment analysis, dialect identification, NER, and others. In (Chowdhury et al., 2020), the authors introduced a new Arabic BERT (QARiB). In their work, the authors tried to improve the performance of the model through diversifying the training data. In their experiments, they show that a BERT model trained on a mixture of formal and informal data has much better generalization power compared to BERT models that are trained on formal text only. QARiB was evaluated only on a text categorization task.

From the previous summary, it is noticeable that there is a lack of direct comparison between these new language models. This, in turn, raises the question about the effectiveness of each of them against the others. In this work, we offer a direct comparison between all the new Arabic language models on two classifications tasks, sentiment analysis and sarcasm detection.

2.2 Arabic Sentiment and Sarcasm Classification

Arabic sentiment analysis (SA) has been under the researchers’ radar for a while. Early work on Arabic SA such as (Abdul-Mageed et al., 2011; Abbasi et al., 2008), focused on modern standard Arabic (MSA). Later, attention started moving towards dialects such as the work of (Mourad and Darwish, 2013), where the authors introduced an expandable Arabic sentiment lexicon along with a corpus of tweets. Other works aimed to create datasets such as the works of (Kiritchenko et al., 2016; Rosenthal et al., 2017; Elmadany et al., 2018). Regarding sentiment analysis systems, there were many attempts such as the works of (El-Beltagy et al., 2017; Al-Smadi et al., 2019; Abdulla et al., 2013; Alayba et al., 2018; Abu Farha and Magdy, 2019). Abu Farha and Magdy (2021) provide a thorough comparative analysis of the available SA approaches. In their work, they compared a large variety of models on three benchmark datasets. Their analysis shows that deep learning models combined with word embeddings achieve much better performance compared to classical machine learning models, such as SVMs. However, their experiments show that the utilisation of transformer-based language models achieves better results than deep learning models. They show that a fine-tuned AraBERT (Antoun et al., 2020) model outperforms all existing classical and deep learning models on all the three benchmark datasets they examined (Abu Farha and Magdy, 2021). AraBERT achieved $F_{PN}$ scores of 0.69 and 0.92 on SemEval-2017 (Rosenthal et al., 2017) and ArSAS (Elmadany et al., 2018) datasets respectively. Abdul-Mageed et al. (2020) tested their new language models on many SA datasets such as SemEval-2017, ArSAS and ArSarcasm (Abu Farha and Magdy, 2020) datasets. The best model (MARBERT) achieved $F_{PN}$ scores of 0.710, 0.930, and 0.715 on the three datasets respectively.

Unlike Arabic SA, the work on Arabic sarcasm is scarce and limited to a few attempts. The earliest work on Arabic sarcasm/irony is (Karoui et al., 2017), where the authors created a dataset of Arabic tweets, which they collected using a set of political keywords. They filtered sarcastic content using distant supervision, where they used the Arabic equivalent of #sarcasm. The dataset contains 5,479 tweets, 1,733 of which are sarcastic/ironic. In their work, the authors utilised various features to experiment with their data. These features include punctuation marks, emoticons, quotations, opposition words, sentiment features, shifters features and contextual clues. They experimented with various classifiers such as SVM, Naive Bayes, Logistic Regression, Linear Regression. Random Forest was the best model, where it achieved an F1-score of 0.73. The authors of (Ghanem et al., 2019) organised a shared task competition for Arabic irony detection. They collected their data using distant supervision and used similar Arabic hashtags to the ones in (Karoui et al., 2017). In addition, they manually annotated a subset of tweets, which were sampled from ironic and non-ironic sets. The first place was (Khalifa and Hussein, 2019), where they achieved an F1-score of 0.85. In their work, they utilised a set of features that include word n-grams, topic modelling features, sentiment features, statistical features and word embeddings. They experimented with multiple classifiers such as BiLSTM,
Random Forest, XGBoost.

Abbes et al. (2020) created a corpus of ironic tweets, namely DAICT, which contains around 5,000 sarcastic/ironic tweets. In (Abu Farha and Magdy, 2020), the authors proposed ArSarcasm dataset for sarcasm detection, which contains 10,547 tweets, 1,682 of which are sarcastic. They created their data through the reannotation of previous Twitter sentiment dataset. In their work, they provided a baseline model, which was based on a BiLSTM and achieved an F1-score of 0.46 on the sarcastic class. Recently, Abdul-Mageed et al. (2020) created a set of BERT-based models and tested them on the ArSarcasm dataset, where the best model (MARBERT) achieved a macro F1 of 0.76. Those results are not directly comparable with the baseline in (Abu Farha and Magdy, 2020), as their official metric is F1-score over the sarcastic class.

A new version of ArSarcasm dataset was released, namely ArSarcasm-v2 (Abu Farha et al., 2021). This dataset is an extension of the original ArSarcasm dataset (Abu Farha and Magdy, 2020), where the authors annotated an additional 5,000 tweets and provided a new train/test split. In this work, we compare the performance of all existing Arabic transformer-based LMs on the sentiment and sarcasm detection tasks of the ArSarcasm-v2 dataset to have a comprehensive report of their performance on such tasks.

3 Experimental Setup

3.1 Dataset

In the experiments, we use ArSarcasm-v2 dataset which was released along with the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic (Abu Farha et al., 2021). ArSarcasm-v2 is an extension of the original ArSarcasm dataset (Abu Farha and Magdy, 2020). The authors extended the dataset through annotating more data for sarcasm, sentiment and dialect. The authors collected the new data through random sampling from Twitter. To ensure the presence of sarcastic tweets, the authors utilised portions of DAICT corpus (Abbes et al., 2020), which contains around 5,000 tweets, most of which are sarcastic. To ensure annotation consistency, the authors re-annotated the portions from DAICT along with the newly collected tweets. The authors used Appen\(^1\) crowd-sourcing platform for the annotation. Each tweet was annotated for sarcasm, sentiment and dialect. The final dataset consists of 15,548 tweets divided into 12,548 training tweets and 3,000 testing tweets. Table 1 provides the detailed statistics of the dataset.

3.2 Models

This section goes over the models used in the experiments. Each of these models has been fine-tuned for sentiment classification and sarcasm detection. The fine-tuning is done through adding a fully connected layer on top of the pre-trained model. After that, each model is fine-tuned for the specific task. In this work, we compare the following models:

- **Bi-LSTM**: a baseline model based on a BiLSTM combined with Mazajak word embeddings (Abu Farha and Magdy, 2019).

- **mBERT**: multilingual BERT provided by (Devlin et al., 2019). This model is based on BERT-base architecture and was trained on data from the Wikipedia dumps of 104 languages.

- **GigaBERT**: provided by (Lan et al., 2020a). The model was trained on a large set of Arabic news articles. The training data was augmented with English translations to improve cross lingual performance.

- **XLM-RoBERTa (XLM-R)**: multilingual extension of the original RoBERTa model Liu et al. (2020) provided by (Conneau et al., 2020). We use two variants of this model, XLM-R-base and XLM-R-large.

- **AraBERT**: Arabic-specific BERT provided by (Antoun et al., 2020). We use all the versions of AraBERT (v0.1/v0.2). AraBERT (v0.2) models are trained on more data compared to AraBERT (v0.1). We experiment with all the variants of these models (base and large) and the models with and without Farasa (Abdelali et al., 2016) pre-segmentation. AraBERT (v0.1) was trained on 23GB of text while AraBERT (v0.2) was trained on 77GB of text.

- **AraELECTRA**: Arabic-specific ELECTRA provided by (Antoun et al., 2021a). ELECTRA contains two modules, a generator and a discriminator. Usually, the discriminator is taken and fine-tuned for downstream tasks. In

\(^1\)https://appen.com
## Table 1: Statistics of training and testing datasets, showing the number of examples for both sarcasm detection and sentiment analysis tasks.

<table>
<thead>
<tr>
<th>Set</th>
<th>Sarcasm</th>
<th>Sentiment</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sarcastic</td>
<td>Non-sarcastic</td>
<td>Positive</td>
</tr>
<tr>
<td>Training</td>
<td>2,168</td>
<td>10,380</td>
<td>2,180</td>
</tr>
<tr>
<td>Testing</td>
<td>821</td>
<td>2,179</td>
<td>575</td>
</tr>
<tr>
<td>Total</td>
<td>2,989</td>
<td>12,559</td>
<td>2,577</td>
</tr>
</tbody>
</table>

In this work, we experiment with both the generator and the discriminator. AraELECTRA was trained on the same 77GB of text used for AraBERT.

- **Arabic BERT**: provided by (Safaya et al., 2020). The model was trained on 95GB of text from the Arabic version of the unshuffled OSCAR corpus (Ortiz Suárez et al., 2020) and the Arabic Wikipedia. The model is available in two variants based on the number of parameters (base and large).

- **Arabic ALBERT**: provided by (KUIS-AI-Lab). An Arabic version of ALBERT (Lan et al., 2020b). This model was trained on data from the Arabic version of the unshuffled OSCAR corpus (Ortiz Suárez et al., 2020) and the Arabic Wikipedia. There are three variants of this model based on the number of parameters (base, large, xlarge).

- **ARBERT/MARBERT**: provided by (Abdul-Mageed et al., 2020). These models are based on the BERT-base and trained on a set of books and news articles. ARBERT was trained on 66GB of text only from news articles. MARBERT was trained on a larger dataset (128 GB), 50% of which is tweets. The variation in MARBERT’s training data gives it the ability to better handle the variations in dialectal Arabic, which is very useful to the tasks in this paper.

- **QARiB**: provided by (Chowdhury et al., 2020). This model was trained on various sources of data including news articles and tweets.

- **AraGPT2**: Arabic-specific GPT2 provided by (Antoun et al., 2021b). AraGPT2 is a stacked transformer-decoder model trained using the causal language modelling objective. The model was trained on 77GB of Arabic text (same as AraELECTRA and AraBERT).

AraGPT2 comes in four variants: AraGPT2-base, AraGPT2-medium, AraGPT2-large and AraGPT2-mega. We experiment with the base, medium and large variants.

A summary of details about these models and their variants is shown in Table 3.

### 3.3 Hyper-parameters and evaluation

In the experiments, we relied on the implementation provided by HuggingFace’s Transformers library (Wolf et al., 2019). We used the provided AutoModelForSequenceClassification which matches each model to the proper implementation. We trained the models for 5 epochs with a learning rate of $5 \times 10^{-6}$. The maximum sequence length was set to 128 tokens. For AraBERT experiments, we used the provided pipeline, which uses Farasa (Abdelali et al., 2016) segmentation for some models. For the BiLSTM model, we used 128 hidden units, Rectified Linear Unit (ReLU), and Adam (Kingma and Ba, 2015) optimiser with a learning rate of 0.0001.

For evaluation, we used the official metrics used in the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic (Abu Farha et al., 2021). For the sarcasm detection task, F1-score of the sarcastic class is the main metric. For sentiment analysis, the macro average of the F1-scores of the positive and negative classes ($F_{PN}$) is the main metric. This metric, $F_{PN}$, has been adopted as the standard metric for sentiment classification in SemEval tasks (Rosenthal et al., 2017).

### 4 Results

#### 4.1 Models Effectiveness for Sentiment and Sarcasm Detection

Table 2 shows the results achieved by all the models on the dataset for both sentiment analysis and sarcasm detection tasks. The first five rows include all the experiments in this work were conducted on a server machine with a 32 core CPU, 512GB RAM and a Quadro RTX 6000 GPU (24GB GPU Memory).
Table 2: Results achieved by all models on sentiment analysis and sarcasm detection tasks.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sentiment Analysis</th>
<th>Sarcasm Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Accuracy</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>0.623</td>
<td>0.671</td>
</tr>
<tr>
<td>mBERT</td>
<td>0.570</td>
<td>0.624</td>
</tr>
<tr>
<td>GigaBERT</td>
<td>0.625</td>
<td>0.662</td>
</tr>
<tr>
<td>XLM-R-base</td>
<td>0.605</td>
<td>0.643</td>
</tr>
<tr>
<td>XLM-R-large</td>
<td>0.641</td>
<td>0.678</td>
</tr>
<tr>
<td>AraBERT-base (v01)</td>
<td>0.630</td>
<td>0.670</td>
</tr>
<tr>
<td>AraBERT-base (v1)</td>
<td>0.638</td>
<td>0.677</td>
</tr>
<tr>
<td>AraBERT-base (v02)</td>
<td>0.654</td>
<td>0.686</td>
</tr>
<tr>
<td>AraBERT-base (v2)</td>
<td>0.651</td>
<td>0.690</td>
</tr>
<tr>
<td>AraBERT-large (v02)</td>
<td>0.659</td>
<td>0.695</td>
</tr>
<tr>
<td>AraBERT-large (v2)</td>
<td>0.660</td>
<td>0.700</td>
</tr>
<tr>
<td>AraELECTRA (discriminator)</td>
<td>0.649</td>
<td>0.687</td>
</tr>
<tr>
<td>AraELECTRA (generator)</td>
<td>0.604</td>
<td>0.648</td>
</tr>
<tr>
<td>Arabic BERT-base</td>
<td>0.627</td>
<td>0.668</td>
</tr>
<tr>
<td>Arabic BERT-large</td>
<td>0.648</td>
<td>0.678</td>
</tr>
<tr>
<td>Arabic ALBERT-base</td>
<td>0.600</td>
<td>0.653</td>
</tr>
<tr>
<td>Arabic ALBERT-large</td>
<td>0.603</td>
<td>0.657</td>
</tr>
<tr>
<td>Arabic ALBERT-xlarge</td>
<td>0.623</td>
<td>0.674</td>
</tr>
<tr>
<td>MARBERT</td>
<td>0.664</td>
<td>0.693</td>
</tr>
<tr>
<td>ARBERT</td>
<td>0.642</td>
<td>0.673</td>
</tr>
<tr>
<td>QARiB</td>
<td>0.661</td>
<td>0.688</td>
</tr>
<tr>
<td>AraGPT2-base</td>
<td>0.594</td>
<td>0.647</td>
</tr>
<tr>
<td>AraGPT2-medium</td>
<td>0.602</td>
<td>0.649</td>
</tr>
<tr>
<td>AraGPT2-large</td>
<td>0.562</td>
<td>0.612</td>
</tr>
</tbody>
</table>

The BiLSTM baseline along with the multilingual BERT models. The rest of the table contains the results achieved by Arabic-specific language models. As can be seen in Table 2, most models, including the BiLSTM baseline, are achieving good results on the sentiment analysis task. On the sarcasm detection task, which is more challenging, the use of Arabic-specific language models provides a large boost in performance. The overall best model is MARBERT, which is a BERT model trained on 128GB of textual data, 50% of which is dialectal Arabic. MARBERT achieved an F_{PN} score of 0.724 on the sentiment analysis task and an F1-score (sarcastic class) of 0.584. AraBERT-large (v2) achieved similar performance on SA, while slightly falling behind on the sarcasm detection task.

From Table 2, it is noticeable that larger models tend to achieve higher results, which is due to their larger representational power. The large variants of AraBERT are achieving higher than the other smaller models. Additionally, the nature of the training data has a significant effect on the performance. Models such as MARBERT and QARiB were trained on a mixture of MSA and dialectal Arabic. It is noticeable that the performance of these models is better than other similar or even larger models such as Arabic BERT-large and Arabic ALBERT-large. Nevertheless, it is also noticeable the applied preprocessing has an effect where similar models achieve different results. For example, AraBERT-base and Arabic BERT are based on the same architecture and trained on similar data, but AraBERT achieves higher results. The preprocessing used for AraBERT includes the removal of non-Arabic words, while Arabic BERT data has some inline non-Arabic words. Additionally, it is clear that the training methodology is an important factor in a model’s performance. Models trained with masked language modelling (MLM), BERT variants, or replaced token detection (RTD), ELECTRA, are better for classification tasks. AraGPT2 is trained using causal language modelling objective, which is quite useful for sentence completion and language generation tasks, but it seems that it is not as effective for classification tasks. Finally, it is clear that monolingual models achieve higher scores than the multilingual ones such as mBERT, XLM-R, and GigaBERT.

4.2 Models Computational Cost

The development of transformer-based language models embarked the war to develop larger and larger models with billions of parameters. This raised the question of the computational cost, the
accessibility of such models and the time needed
to train and fine tune them. The largest model we
experimented with is AraGPT2-Mega, which we
were not able to fine-tune on the aforementioned
machine due to lack of memory. Figure 1 shows
the time needed to fine-tune each of the models
for one epoch with the same batch size. From the
figure, it is noticeable that the time is monotonically
increasing with the number of parameters within
the model. AraGPT2-large (792M parameters) is
the slowest while the fastest is Arabic ALBERT-
base.

When considering both the time and perfor-
mance, AraELECTRA is one of the fastest and
lightest models to fine-tune, yet it achieves results
closer to other larger models. Thus, it could be
the choice when dealing with limited resources. In
our experiments, AraELECTRA achieved results
closer to larger models (AraBERT-large), while it
consumes half the space in the memory and can be
fine-tuned much faster.

5 Summary of Findings

From the experimentation of the existing
transformer-based LMs on Arabic sentiment and
sarcasm detection, we can learn the following:

• Existing Arabic specific models are bet-
ter than multilingual ones. In the ex-
periments, Arabic-specific models achieved
higher results than their multilingual equiv-
alents. Larger multilingual models achieve
similar scores to smaller monolinguals.

• Large models (based on the number of param-
eters) achieve better results than small ones,
when trained on similar data. This is due to
the larger representational power.

• Models that included social media text in the
training data, achieve better results on these
tasks, which also contain dialectal Arabic. En-
riching the training data with social media
data would increase the model’s ability to han-
dle dialectal Arabic, which is crucial for many
NLP tasks.

• ELECTRA is more efficient while still being
effective. The training procedure of ELEC-
TRA produces better representation than other
models. AraELECTRA achieved similar re-
sults to AraBERT that has double the number
of parameters.

• For classification tasks, BERT is better than
GPT2. BERT is trained using masked lan-
guage modelling (MLM) objective, while
GPT2 is trained using causal language mod-
ing. Since all the variants of AraGPT2
achieved lower scores compared to other mod-
els, we can conclude that the representation
learnt by BERT/ELECTRA is better suited for
classification.

Table 3 summarises all our findings on these
models, including their details, performance, archi-
tecture, number of parameters, training data size
and source.

We hope that this benchmark study would guide
future research in Arabic sentiment and sarcasm
detection, and generally in Arabic NLP tasks in-
cluding the creation of new language resources.

6 Conclusion and Future Work

In this paper, we provided a comprehensive bench-
mark to the 24 existing transformed-based lan-
guage models that support Arabic on two NLP
Table 3: Summary of the results achieved by each model on the sarcasm detection and sentiment analysis tasks. The table includes details about the model variant, architecture, training data size, training data nature, number of parameters, time needed to fine-tune for one epoch (batch size=4). In text nature types are: tweets (T), Wikipedia (W), news (N), OSCAR corpus (OC), Common Crawl (CC), subtitles (S), and books (B). (CS) stands for code-switching, and (*) indicates that Farasa segmentation is applied to the text. Finally, (N/A) indicates that information is not available.

For future work, one of the immediate studies that is essential for Arabic NLP researchers is to extend a more comprehensive benchmark study on the effectiveness of these model on other Arabic NLP tasks, including other classification tasks, text generation, and information extraction tasks. Such work would be an excellent guide for researchers in the field and will provide insights on the required efforts for improving these language models on different NLP tasks.

7 Acknowledgement

This work was partially supported by the Defence and Security Programme at the Alan Turing Institute, funded by the UK Government.

References


Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In ICLR.


Samhah R. El-Beltagy, Mona El Kalamawy, and Abu Bakr Soliman. 2017. NileTMRG at SemEval-
Taku Kudo. 2018. Subword regularization: Improving


KUIS-AI-Lab. *Arabic-albert*.


