Prediction of Author's Profile Basing on Fine-Tuning BERT Model

Bassem Bsir[1,2], Nabil Khoufi[3], Mounir Zrigui[1,2]
[1] ISITCom, University of Sousse, 4011 Hammam Sousse,
[2] Laboratory in Algebra, Numbers Theory and Intelligent Systems, University of Monastir, Monastir, Tunisia
[3] ANLP Research Group, FSEGS, Sfax, Tunisia
E-mail : Bsir.bassem@yahoo.fr, nabil.khoufi@outlook.com, mounir.zrigui@fsm.rnu.tn

Keywords: BERT, Author profiling (AP), fine tuning, PAN 2018 Corpus dataset, NLP, Transformer-model, deep learning, Self-attention Transformers.

Received: Februar 5, 2023

The task of author profiling consists in specifying the infer-demographic features of the social networks' users by studying their published content or the interactions between them. In the literature, many research works were conducted to enhance the accuracy of the techniques used in this process. In fact, the existing methods can be divided into two types: simple linear models and complex deep neural network models. Among them, the transformer-based model exhibited the highest efficiency in NLP analysis in several languages (English, German, French, Turk, Arabic, etc.). Despite their good performance, these approaches do not cover author profiling analysis and, thus, should be further enhanced. So, we propose in this paper a new deep learning strategy by training a customized transformer-model to learn the optimal features of our dataset. In this direction, we fine-tune the model by using the transfer learning approach to improve the results with random initialization. We have achieved about 79% of accuracy by modifying model to apply the retraining process using PAN 2018 authorship dataset.

Povzetek: Članek predstavlja novo metodo za napovedovanje avtorjevega profila, ki temelji na modelu Fine-Tuning BERT.

1 Introduction

As defined by [6], author profiling (AP) is a Natural Language Processing (NLP) research domain that aims at deducing social-demographic data on the author or user of a specific application or software service. It consists first in extracting automatically, from the text, information showing the authors' gender, age and other demographic features. These data are used in several fields such as in forensics, security and marketing.

In the last decades, the main methods utilized in Natural Language Processing (NLP) are deep neural networks relying on Transformers. As instance of these techniques, Self-attention Transformers and, particularly, the self-supervised-trained variants, also called BERT (Bidirectional Encoder Representations from Transformers) models [26], showed high performance in several tasks such as text classification [18], Sentiment Analysis [14], question answering [38], natural language inference [45][40], etc. In fact, these novel methods have revolutionized NLP tasks by dropping the recurrent part and only keeping attention mechanisms.

Indeed, transformers-based pre-trained language models, such as OpenAI GPT [3], BERT [26], RoBERTa [41], have proven their good performance in learning language representation by employing huge quantity of unlabeled data [4]. Nevertheless, their training is often performed on large monolingual English corpora or on multi-lingual corpora involving more than one hundred languages. Recent study has demonstrated that the performance of the fine-tuning from multi-lingual models is almost similar to that of monolingual models for low resource languages [1].

Despite the wide use of the afore-mentioned methods, their accuracy for Arabic Author profiling should be further enhanced, particularly in tokenization level in the task of data processing. For this reason and as no previous has focused on the identification of the gender of the author form Arabic texts published on social networks using these models, we examine the efficacy of several multilingual models for AP tasks. Then we choose to fine-tuned model. We are focusing on the Ara-BERTv2-base model in order to change its parameters and search for the most suitable ones for the gender identification task.

The present manuscript is structured as follows. Section II presents the works deal-ing with author profiling. In Section III, we depict the introduced approach, the employed datasets, and the training details. Section IV shows a comparative study of the obtained findings with those obtained in the stat-of- the art and discusses the experimental results. We end the paper with a short conclusion.

2 State of the art

Several approaches and methods have been recently developed and applied in AP.

We can classify these approaches into two categories. The first category includes traditional machine learning methods [2]. The second category includes deep learning techniques. [11] [37] [14].

Traditional machine learning methods have been explored by researchers for the task of gender prediction in author profiling. Indeed, Poulston et al. in 2017 used the genism Python library for LDA topic extraction with SVM classifiers. Their results proved that the topic models are useful in developing author's profiling systems. Argamon et al. in 2012 analyzed an analogous sample taken from the BNC consist-ing of fiction and non-fiction documents. Their corpus includes 604 texts equally divided by genre and controlled for authorial origin for a total size of 25 million words. Their anal ysis consists in a frequency count of basic and most frequent function words, part-of speech tags and part-of-speech two-grams and three-grams. The counts were processed by a machine-learning algorithm used to classify the texts according to the author's gender. They obtained an accuracy of 80%.

In 2017, Martinc et al based on the corpus collected from Twitter text written by four different languages (Arabic, English, Portuguese and Spanish), they obtained 70.02 by using logistic regression by combining character, word POS n-grams, emo-ji's, senti ments, character flood in gland lists of words per variety in PAN 2017 competition.

González-Gallardo et al. predicted the gender, age and personality traits of Twitter users. They accounted stylistic features represented by character N grams and POS N-grams to classify tweets. They applied Support Vector Machine (SVM) with a linear kernel called LinearSVC and obtained 83.46% for gender detection [16].

While these methods have shown some success, they are often limited by the quality of the features used and the complexity of the task, which can lead to lower accuracy compared to deep learning methods.

In the last three years have been many recent modeling improvements on NLP tasks. These models have largely focused on building separate models for each language or for a small group of related languages. However, Transfer Learning from large-scale pre-trained models in Natural Language Processing (NLP) becomes more prevalent they often have several hundred million parameters and current research on pre-trained models indicates that training even larger models still leads to better perfor-mances on NLP tasks [5][41].

Indeed, Devlin and Chang proved that the main challenge in NLP consists of the small quantity of the training data [26]. To deal with this issue, they have suggested transformer-based models trained on huge unlabeled datasets (e.g., Wikipedia's dataset). The authors were able to apply the pretrained models on smaller datasets without the need for developing training models from scratch. De-spite the fact that the proposed technique provided high accuracy in executing various NLP tasks [26][29], "fine-tuning" should be performed on the pretrained models before being applied on smaller datasets. As example of the pretrained models, we can mention the bidirectional en-coder representations from transformers (BERT) characterized by its bidirectionality.

In 2017, Vaswani et al introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. It's designed to pre-train deep bidirectional representations from unlabeled text by joint-ly conditioning on both right and left context in all layers. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5, MultiNLI accuracy to 86.7%, SQuAD v1.1 question answering Test F1 to 93.2 and SQuAD v2.0 Test F1 to 83.1[28]. Unlike Radford et al. [3], which uses unidirectional language models for pre-training, BERT uses masked language models to enable pretrained deep bidirectional repre-sentations. It's also reducing the need for many heavily engineered task specific architectures. BERT is the first finetuning based representation model that achieves state-of-the-art performance on a large suite of sentence-level and token-level tasks, outperforming much task-specific architecture [3].

Ai, M. proposed the tasks of Russian news event detection. They present datasets for the Russian news event clustering, headline selection, and headline generation tasks along with baselines. Authors demonstrated the successful models were classifica-tion-based BERT models. However, it turns out clustering embeddings can be almost as effective when trained with correct pooling and loss function [2].

Rangel et al. in 2021 explored the use of the BERT language model for author pro-filing in multiple languages. The authors found that the BERT model achieved high accuracy rates for gender and age prediction in several languages. they achieved an accuracy of 96.4% for gender prediction and 77.5% for age prediction on the English dataset, and an accuracy of 92.3% for gender prediction and 62.5% for age predic-tion on the Spanish dataset. The study also evaluated author profiling in French, Portuguese, and Italian, achieving similarly high accuracy rates. [10].

In the same year, 2021, other study used a combination of n-gram-based features and a random forest classifier to predict the gender and age of authors was presented by Khader and Al-Ani. The results showed that the approach achieved high accuracy rates, particularly for gender prediction. Indeed, for gender prediction, the approach achieved an accuracy rate of 97.9%, while for age prediction, the accuracy rate was 91.3%.

In 2019, Victor SANH et al. Show that it is possible to reach similar performances on many downstream-tasks using much smaller language models pre-trained with knowledge distillation, resulting in models that are lighter and faster at inference time. It is possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster [18].

For instance, while processing the word bank (which have two meanings (financial institution or the shore of a river), the BERT model analyzes all words in the sen-tence at both valences and produces a score showing the best representation of the meaning of the words in a specific context.

The main objective of this research work is to study the impact of the common pre-processing methods utilized to determine the author’s age and gender in case of using the pretrained model called BERT.

The following section presents first the existing research works based on the preprocessing techniques applied in author profiling. Then, the implementation of the five considered cases of the preprocessing methods and the different steps of each con-ducted experiment is detailed. Subsequently, the findings obtained in the experiments are described and the impacts of each preprocessing method on the accuracy of the model in predicting the authors’ gender are discussed. Finally, in the conclusion, we show briefly the important results of the current work study and highlight the directions of our future work.

[13] employed the NN model with GRU to determine the writer’s by examining Facebook's and Twitter posts. The used NNP model input was prepro-cessed and divided into two layers: embedding layer and stylometric features extrac-tion phases. In fact, the embedding layer output was linked to a bidirectional GRU layer and then, to an activation layer. However, the stylometric features were nor-malized and attached directly to the same activation layer utilized after the GRU layer. The authors compared the obtained findings, which are inferior to the best result in PAN'AP (2017), to the best findings provided by Basile et al. in PAN' AP (2017).

In 2017, Estruch et al. enhanced an early fusion model, which was based on perform-ing fusion after the decision level single source classification. The authors achieved 91% GI accuracy on an English dataset in Singapore retrieved from Foursquare, Instagram and Twitter.

The approach developed by Sebastian Sierra and al. in 2018 was applied to assess the authors’ gender employing multi-modal information (texts and images). The multi-modal representation was learned using GMUs. Indeed, accuracy rates equal to 0.74 and 0.81 were obtained in the multi-modal scenario for the test partition for English, Spanish and Arabic, respectively.

Moreover, the gold standard data was translated by Veenhoven et al in 2018 into the language of interest. Bi-LSTM and CNN architectures were also utilized to solve the GI problem by considering PAN-AP (2018) dataset. By considering the RNN, the highest obtained GI accuracy was equal to 79.3%, 80.4% and 74.9% for English, Spanish and Arabic languages, respectively.

The deep learning approach introduced by Yasuhide Miuraand et al. (2017) provided the best result when applied on the Portuguese language. The authors used the Re-current Neural Networks (RNN) for words and Convolutional Neural Networks (CNN) for characters. Therefore, they obtained two representations of various levels for a single message. The representations were, then, classified according to the writers’ gender by employing attention mechanism, max-pooling layer and fully connected layer. More precisely, the word embeddings layer was first trained by the skip-gram. On the other hand, in the character embedding layer, weights were arbi-trarily initialized using the uniform distribution.

In 2013 [20], 2014 [21] and 2015 [22] PAN competition, the age and gender profiling was performed

by analyz-ing the English and Spanish datasets and applying the traditional supervised machine learning approaches, namely Logistic Regression, Random Forest, SVMs, etc. The objective of PAN competition organized in in 2016 [23] consists in validating the robustness of techniques from the cross-genre perspective. The obtained results showed that SVMs were the dominant paradigm. Then, F. Rangel et al in 2017 added, in 2017, two more languages (Arabic and Portuguese) to the dataset. Although SVMs were selected by several participants, deep neural networks (i.e., Windowed Recurrent Convolutional Neural Network as an extension of the Recur-rent Convolutional Neural Network) attained the state-of-the-art performance in terms of gender identification.

Among the PAN 2017 tasks, we cite the gender identification from Twitter texts. Concerning the Arabic language, the best model relied on representing the text as a vector including the combinations of character, word and POS n-grams with emojis, character flooding, and sentiment words. Besides, logistic regression was employed to train the classifier [34]. Approaches for predicting an AP can be broadly categorized into three types of methods as shown in table1.

Since the task of determining an author's profile can be seen as a classification task, we can benefit from pre-trained models. Indeed, pre-trained language models like BERT, GPT, ELMo, etc., capture extensive linguistic knowledge from large amounts of textual data. These models can be fine-tuned for specific authorship profiling tasks. on one side, by employing pre-trained models, transfer learning enables the transfer of general language and contextual knowledge to more specific author attribution tasks. This can enhance the models' ability to grasp subtle characteristics of an author's writing style. On the other side, Transfer learning is particularly useful when specific datasets for author attribution are limited. By fine-tuning pre-trained models on smaller datasets, better performance can be achieved with fewer specific data.

3 The Proposed Approach

The present work presents a fine tune model Approach. More specifically, we build an Arabic pretrained model based on the Ara-BERTv2-large model which is an improved version of BERT model [26] To design the proposed model, multi-lingual transformer models trained on large corpus, were fine tuned. The general architecture of this model is shown in Figure 1.

Table 1: Three types of methods for authorship identification task

Approaches	**Example of features**	**Example of authors and Results**
stylometry methods	The total number of characters The number of capitalized letters Character N-Grams The ratio of capital letters to total number of characters The ratio of white-space characters to total number of characters	Corney & al 2002 described an investigation of authorship gender attribution mining from e-mail text documents. They obtained 70.2 % precision rate for gender detection. Koppel and Peneebaker, analyzed a corpus of 71,000 blogs incorporating almost 300 million words. They obtained 43.8% and 86% for age and gender accuracy prediction, respectively (Schler & al 2006). In 2016 Bilan & al 2016, built a Cross-genre Author Profiling System (CAPS). Their system attained 74.36% accuracy for gender identification.
Content-based methods	Frequency of Function words The Number of contraction words Frequency of punctuations Stopwords The proportion ratio of singular to plural nouns and proper nouns and pronouns	Busger et al., 2016 obteined 0.5575 accuracy for gender identification in English data PAN 2016 competition. Dichiu & al 2016, applied SVM classifier and neural network on TF-IDF and verbosity features. Their results are almost similar to those provided in Bayot & al 2016. They got 61.5% gender accuracy and 41.03% age accuracy.
Deep learning models	- subword character embedding (word - n-gram embeddings - GloVe - FastText - ELMo (Embeddings from Language Models)	Nils Schaetti et al., 2017 used TF-IDF and a Deep-Learning model based on Convolutional Neural Networks. They obtained 0.66%, 0.73%, 0.81% and 0.57% of accuracy in the test partition for English, Spanish, Portuguese and Arabic respectively in PAN 2017 competition. Salvador et al., 2017 generated embeddings of the authors' text based on subword character n-grams. They got 0.7919 for gender identification in PAN 2017 competition[1]. Victor SANH et al, 2019 showed that it is possible to reach similar performances on many downstream-tasks using much smaller language models pre-trained with knowledge distillation, resulting in models that are lighter and faster at inference time. It is possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster.

The PERT tokenizer, trained by the WordPiece tokenization, was used to split the input text into a list of tokens. Such division reveals that means that a word can be broken down into several sub-words. The BERT vector assigned to a word is a function of the entire sentence. Therefore, a word can have different vectors according to the contexts in which they are used. There are different built-in tokenizers. The basic one is character tokenizer. However, the pretrained Arabic BERT utilizes a word-by-word tokenizer.

3.1 Corpus

The dataset, collected from Twitter, is part of the author's profiling task of PAN@CLEF 2018. For each tweet collection, Arabic texts are composed of tweets written by 2400 authors: 100 tweets per authors. Four varieties of the Arabic language were used in this corpus: Egypt, Gulf, Levantine and Maghrebi.

3.2 Pre-training model

Ara-BERTv2-large was trained to learn the distributed representation from the unlabeled texts by jointly conditioning on the left and right contexts of a certain token. The models were trained for 10 epochs with learning rate 1e-5 employing cross entropy loss criterion and Adam optimization algorithm. 32 samples were utilized in each mini batch, except when this did not fit in memory. During the training phase, a sequence of fixed length was used and padding or truncate was applied when necessary. The sequence length consists of 30, 100. Besides, all model parameters were fine-tuned during training, i.e., no layer was kept frozen. The model with the best validation set performance was evaluated on the test dataset.

[1] https://pan.webis.de/clef17/pan17-web/

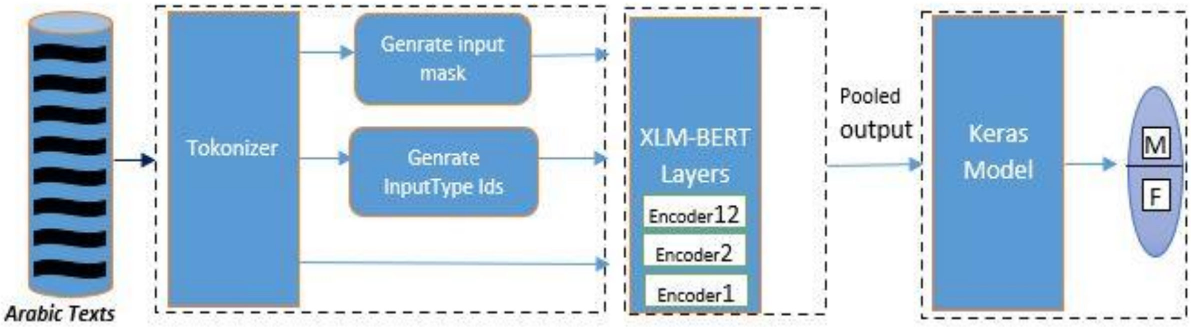

Figure 1 : The architecture of XLM-RoBERTa Approach.

3.3 Regularization of hyperparameters

This section details the fine-tuning of hyperparameters. The selected values and hyperparameters were determined through various tests, considering only those values that demonstrated optimal performance for the introduced model.

During the pretraining phase, we employed an Adam optimizer with a learning rate of 1e-8, a batch size of 64, a maximum sequence length of 512, and a masking probability of 15%. Additionally, a dropout rate of 0.1 was utilized to prevent model overfitting. All models were trained using a batch size of 16 and 5 epochs.

We leveraged Python within the Google Colab environment, a cloud service by Alphabet Inc., for implementing deep learning algorithms. Colab provides access to accelerated cloud tensor processing units (TPUs) developed by Google, each boast-ing up to 180 teraflops of computation power and high-bandwidth memory on a sin-gle board. In this study, Colab Pro was used, providing virtual machines (VMs) with doubled memory compared to standard Colab VMs.

To ensure uniform input sizes for Ara-BERTv2-large, we set a maximum sentence length of 128. Inputs were adjusted by padding and truncating until all sequences reached this length, employing the "pad_sequences" Python function with the "post" value for both padding and truncation, ensuring these operations occurred at the end of the sequences.

4 Experimentation

4.1 Preprocessing

We apply our preprocessing function before training/testing on any dataset. We used the library farasapy for segmentation, stemming, Part Of Speech tagging (POS tagging) and diacritization. Also use the unpreprocess function to reverse the preprocessing changes, by fixing the spacing around non alphabetical characters, and also de-segmenting if the model selected need pre-segmentation.

```
output_text = " و+ لن نبالغ إذا قل +نا : إن' هاتف ' أو ' كمبيوتر ال+ مكتب '
في زمن +نا هذا ضروري"

arabert_prep.unpreprocess(output_text)
```

Figure 2: Example of the unpreprocess function.

4.2 Results

We approached this phase in 2 steps. The first serves as a pre-selection step. In this step, we choose the best-performing model for gender detection, in order to compare it in a second step with the XLM-RoBERTa model. Indeed, AraBERT comes in 6 variants. Each variant of AraBERT is pre-trained on a large cor-pus of Arabic text and can be fine-tuned on specific downstream tasks. In order to explore the most adopted variant for authorship detection task, we experiment with the six variants presented in the table below. We rely on the PAN 2018 corpus da-taset to test these models.

For hyperparameters setting, we used the same parameters in the different experiments for all the models: a peak learning rate to 1 × 10−5, maximum sequence length 128 tokens, batch size 64, 10 training epochs. Two objective functions are used during the language model pretraining step. The bidirectional nature ensures that the model can effectively make use of both past and future tokens for this. The second objective is the next sentence prediction (NSP) task.

The results show that Ara-BERTv2-large achieved the highest accuracy, 79.7%, while Ara-BERTv2-base and Ara-BERTv1-base followed closely with an accuracy rate of 78.1%. Our experiments demonstrate that Ara-BERTv2-base, the largest model with 345 million parameters, is the most effective for the gender detection task in the Arabic language.

Upon exploring the obtained results, we notice that the difference in accuracy between the different tested models does not exceed 1%. These models, despite being trained using different pre-training objectives, such as masked language for Ara-BERTv2-large or masked language combined with next sentence prediction objectives for Ara-BERTv0.1-base, are highly effective for NLP tasks and outperform ML algorithms or N-Gram models.

Table 2: Performance of different models for gender identification task

Model	**HuggingFace Model**	**Size (MB/Params)**	**Pre-Segmentation**	**DataSet (Sentences/Size/nWords)**	**Accuracy**
AraBERT v0.2-Twitter-base	bert-base-arabertv02-twitter	543MB / 136M	No	Same as v02 + 60M Multi-Dialect Tweets	0,763
AraBERT v0.2-Twitter-large	bert-large-arabertv02-twitter	1.38G / 371M	No	Same as v02 + 60M Multi-Dialect Tweets	0,771
AraBERT v0.2-base	bert-base-arabertv02	543MB / 136M	No	200M / 77GB / 8.6B	0,767
AraBERT v0.2-large	bert-large-arabertv02	1.38G / 371M	No	200M / 77GB / 8.6B	0,765
AraBERT v2-base	bert-base-arabertv2	543MB / 136M	Yes	200M / 77GB / 8.6B	0,781
AraBERT v2-large	bert-large-arabertv2	1.38G / 371M	Yes	200M / 77GB / 8.6B	0,797
AraBERT v0.1-base	bert-base-arabertv01	543MB / 136M	No	77M / 23GB / 2.7B	0,771
AraBERT v1-base	bert-base-arabert	543MB / 136M	Yes	77M / 23GB / 2.7B	0,781

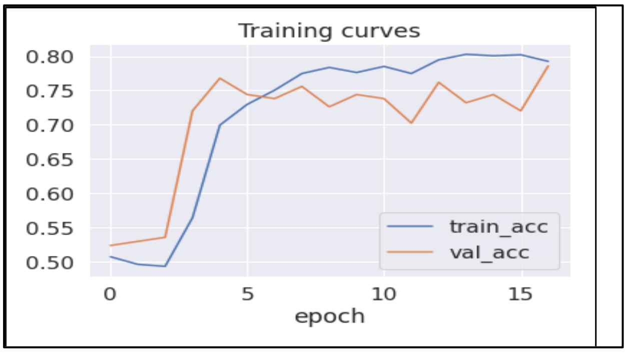

Figure 3: Accuracy of Ara-BERTv2-large model.

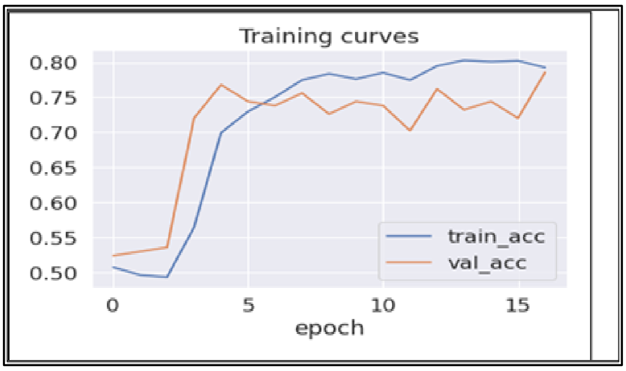

Figure 4: Fine tuning of Ara-BERTv2-large model.

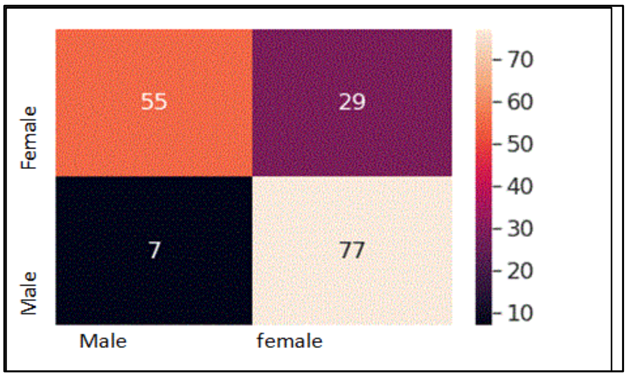

Figure 4 : Confusion matrix of the trained Ara-BERTv2-large

In terms of the adopted model's generalization, we will then compare it with XLM-RoBERTa, which is a multilingual model. This comparison can provide in-sights into the effectiveness of the Ara-BERTv2-large model for Arabic language tasks and whether a more general multilingual model is more suitable for the task at hand.

This comparison can also help researchers and practitioners to choose the most appropriate model for their specific use case and language. Indeed, XLM-RoBERTa is the multilingual variant of RoBERTa trained with a multilingual MLM on one hundred languages, with more than two terabytes of filtered Common Crawl data.XLM-RoBERTa showed its superiority over BERT by its trainability on larger datasets, using larger vocabulary as well as on longer sequences with larger batches in some cases. NSP task was removed and only MLM loss was used for pretraining. XLM-RoBERTa exhibited impressive performance in several multilingual NLP tasks and can perform comparably to monolingual language models (A. Conneau and al, 2019).

To fine-tune XLM-RoBERTa model for gender detection, we run fine-tuning experiments on single GPUs using Transformers software. Then, we fixed the peak learning rate to 1 × 10−5, maximum sequence length 128 tokens, batch size 64, 15 training epochs. We were also experimenting with other hyperparameters setting, but this one gave consistently the best results across all models and datasets. After each training epoch, we evaluated the model on the development dataset (if not pre-sent in the dataset, we randomly held out 10% of the training samples as a develop-ment data) and at the end, we used the best model for evaluation on the test dataset.

Model	Accuracy	Recall	F1-Score
XLM-RoBERTa	0,768	0,75	0,84
Ara-BERTv2-large	0,797	0,76	0,82

Table 3 : Performance of Ara-BERTv2-large and XLM-RoBERTa

4.3 Discussion

Our fine-tuned Ara-BERTv2-large model has achieved higher accuracy (79.7%) on the test data set compared to XLM-RoBERTa (76.8%), as shown in Figure 3 and Table 2. This confirms the suggestion that a model specifically designed for a particular language can perform better than a more general multilingual model in tasks related to that language. Indeed, even when compared to the other models trained specifically on the Arabic language, as shown in the table1, we can see that XLM-RoBERTa remains less competitive and only outperforms 3 out of 8 models.

The reason for this is found in the scale of the Ara-BERTv2-large model, encompassing an increased number of parameters, layers, and a larger hidden size. Consequently, this heightened capacity enables it to capture finer and more intricate patterns, particularly in the context of social media platforms and Arabic datasets, which often include dialectal variations.

As a baseline, we used three results obtained in PAN@CLEF as a baseline method to assess our technique and show its efficiency. Those obtained by applying deep learn-ing method in PAN@CLEF2017: the result of [21] based on using BRNN Gated Recurrent Unit and [24] relying on CNN architecture as well as the best results of gender identification obtained in PAN@CLEF2017 [27].

Kodiyanand et al. in 2017, utilizing GRU, achieving a 71.50% result, and another by Miura et al. in 2017, employing CNN and obtaining a 76.44%.

The results from the conducted experiments demonstrated that the developed Ara-BERTv2-large achieved state-of-the-art performance on Arabic datasets. This vali-dates our initial hypothesis and underscores the effectiveness of pre-trained models for the Arabic language, particularly when dealing with dialectal variations. It emphasizes how these models, with their sophisticated architecture and extensive training on Arabic text, can comprehend the intricate nuances and complexities within the language. This success reinforces the value and potential of deep learning techniques specifically designed for Arabic NLP tasks, showcasing their ability to achieve state-of-the-art results in handling various linguistic challenges present in Arabic datasets.

5 Conclusions

The novelty of this work consists in conducting the Arabic author profiling experi-ments by focusing on the gender of the social networks' users. In this task, the trans-fer learning methods were utilized, for the first time, on the Arabic language.

Three deep learning models were applied with words embeddings for the prediction of Twitter Arabic authors' gender.

The experimental results revealed that the suggested model, called XLM-RoBERTa and used as a contextual embedding technique outperforms the models Ara-BERTv2-base. To sum up, deep learning techniques are not very efficient in detecting the authors' profile and, more precisely, his/her age and gender. As future work, we will explore and enhance the performance of deep learning approaches in author's profiling by augmenting the size of the training set, using different tuning parameters, and employing various types of word embeddings).

References

[1] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116, 2020. https://doi.org/10.18653/v1/2020.acl-main.747

[2] Ai, M.: BERT for Russian news clustering. In: Proceedings of the International Conference "Dialogue 2021", p. 6. Moscow, Russia.2021. https://doi.org/10.28995/2075-7182-2021-20-385-390

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," URL https://s3- us-west-2. amazonaws.com/openai-assets/research covers/languageunsupervised/language_ understanding_paper. pdf, 2018.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[6] Alvarez-Carmona, M. A., Lopez-Monroy, A. P., Montes- ´ y Gomez, M., Villase ´ nor-Pineda, L., and Meza, I. ˜ (2016). Evaluating topic-based representations for author profiling in social media. In Montes y Gomez, M.

[7] Antoun, Wissam, Fady Baly, and Hazem Hajj. "Arabert: Transformer-based model for arabic language understanding." arXiv preprint arXiv:2003.00104 (2020).

[8] Bernard, G.: Resources to compute TF-IDF weightings on press articles and tweets (2022). https://doi.org/10.5281/zenodo.6610406.2022

[9] Bassem, B., & Zrigui, M. (2020). Gender identification: a comparative study of deep learning architectures. In Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 2 (pp. 792-800). Springer International Publishing. https://doi.org/10.1007/978-3-030-16660-1_77

[10] Butt, S., Ashraf, N., Sidorov, G., & Gelbukh, A. F. (2021, September). Sexism Identification using BERT and Data Augmentation-EXIST2021. In IberLEF@ SEPLN (pp. 381-389).2021.

[11] Bsir, B., & Zrigui, M. (2018). Enhancing deep learning gender identification with gated recurrent units architecture in social text. Computación y Sistemas, 22(3), 757-766. https://doi.org/10.13053/cys-22-3-3036

[12] Bsir, B., & Zrigui, M. (2018). Bidirectional LSTM for author gender identification. In Computational Collective Intelligence: 10th International Conference, ICCCI 2018, Bristol, UK, September 5-7, 2018, Proceedings, Part I 10 (pp. 393-402). Springer International Publishing.

[13] Bsir, B., & Zrigui, M. (2019). Document model with attention bidirectional recurrent network for gender identification. In Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part I 15 (pp. 621-631). Springer International Publishing. https://doi.org/10.1007/978-3-030-20521-8_51

[14] Catelli, R., Pelosi, S., & Esposito, M. (2022). Lexicon-based vs. Bert-based sentiment analysis: A comparative study in Italian. Electronics, 11(3), 374.2022.https://doi.org/10.3390/electronics11030374

[15] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune BERT for text classification?” in China National Conference on Chinese Computational Linguistics. Springer, 2019, pp. 194–206.

[16] González-Gallardo, C. E., Montes, A., Sierra, G., Núnez-Juárez, J. A., Salinas-López, A. J., & Ek, J. (2015, September). Tweets Classification using Corpus Dependent Tags, Character and POS N-grams. In CLEF (working notes).2015

[17] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? In China National Conference on Chinese Computational Linguistics, pages 194–206. Springer, 2019.

[18] DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter / Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf // arXiv preprint arXiv:1910.01108. — 2019.

[19] Estruch, C. P., Paredes, R. Rosso, P., Learning Multimodal Gender Profile using Neural Networks. Rec Adv Nat Language Process 2017; Varna: Bulgaria, pp: 577-582. Available from: http://users.dsic.upv.es/~prosso/resources/PerezEtAl_RANLP17.pdf. https://doi.org/10.26615/978-954-452-049-6_075

[20] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. “Overview of the author profiling task at PAN 2013”. CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 2013.

[21] F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, and W. Daelemans. “Overview of the 2nd author profiling task at PAN 2014”. CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 2014.

[22] F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans. “Overview of the 3rd author profiling task at PAN 2015”. CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 2015.

[23] F. Rangel, M. Francisco, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, Martin, and B. Stein. “Overview of the 4th author profiling task at PAN 2016: Cross-Genre Evaluations”. Working Notes Papers of the CLEF 2016 Evaluation Labs, 2016.

[24] F. Rangel, M. Francisco, P. Rosso, M. Potthast, and B. Stein. “Overview of the 5th author profiling task at PAN 2017: gender and language variety identification in Twitter”. Working Notes Papers of the CLEF 2017 Evaluation Labs, 2017.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. 2019.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[27] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[28] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, no. Mlm, pp. 4171–4186, 2019.

[29] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 1, pp. 328–339, 2018, doi: 10.18653/v1/p18-1031. https://doi.org/10.18653/v1/p18-1031

[30] Haffar N., Ayadi R., Hkiri E., Zrigui M. (2021) Temporal Ordering of Events via Deep Neural Networks. In: Lladós J., Lopresti D., Uchida S. (eds) Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science, vol 12822. Springer, Cham. https://doi.org/10.1007/978-3-030-86331-9_49.

[31] Abdellaoui, H., & Zrigui, M. (2018). Using tweets and emojis to build tead: an Arabic dataset for sentiment analysis. Computación y Sistemas, 22(3), 777-786. https://doi.org/10.13053/cys-22-3-3031

[32] Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and Practical BERT Models for Sequence Labeling. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3632–3636, Hong Kong, China. Association for Computational Linguistics.2019.

[33] Kim, J., Aum, J., Lee, S., Jang, Y., Park, E., Choi, D.: FibVID: comprehensive fake news diffusion dataset during the COVID-19 period. Telemat. Inform. 64, 101688 (2021). https://doi.org/10.1016/j.tele.2021.101688.2022.

[34] Matej Martinc, Iza Skrjanec, Katja Zupan, and Senja Pollak. Pan 2017: Author profiling-gender ˇ and language variety prediction. Cappellato et al.[13], 2017.

[35] Martinc, M., Škrjanec, I., Zupan, K., & Pollak, S. Pan 2017: Author profilinggender and language variety prediction. CLEF (Working Notes) 2017. CEUR Workshop Pro ceedings 1866, CEUR-WS.org (2017).

[36] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," arXiv preprint arXiv:1508.07909, 2015.

[37] Suman, C., Kumar, P., Saha, S., & Bhattacharyya, P. (2019, December). Gender Age and Dialect Recognition using Tweets in a Deep Learning Framework-Notebook for FIRE 2019. In FIRE (Working Notes) (pp. 160-166).

[38] S. Garg, T. Vu, and A. Moschitti, "TANDA: Transfer and adapt pre-trained transformer models for answer sentence selection," arXiv preprint arXiv:1911.04118, 2019.

[39] Schlicht, I. B., & de Paula, A. F. M. (2021). Unified and multilingual author profiling for detecting haters. arXiv preprint arXiv:2109.09233.

[40] Velankar, A., Patil, H., & Joshi, R. (2022, November). Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi. In Artificial Neural Networks in Pattern Recognition: 10th IAPR TC3 Workshop, ANNPR 2022, Dubai, United Arab Emirates, November 24–26, 2022, Proceedings (pp. 121-128). Cham: Springer International Publishing.2002. https://doi.org/10.1007/978-3-031-20650-4_10

[41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," CoRR, vol. abs/1907.11692, 2019.

[42] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. arXiv:1906.08237 [cs] (2020)

[43] Yasuhide Miura, Tomoki Taniguchi, Motoki Taniguchi, and Tomoko Ohkuma. Author profiling with word+ character neural attention network. Cappellato et al.[13].2017.

[44] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, "Semantics-aware BERT for language understanding," arXiv preprint arXiv:1909.02209, 2019. https://doi.org/10.1609/aaai.v34i05.6510

[45] Zrigui, M., Ayadi, R., Mars, M., & Maraoui, M. (2012). Arabic text classification framework based on latent dirichlet allocation. Journal of computing and information technology, 20(2), 125-140. https://doi.org/10.2498/cit.1001770