A transformer fine-tuning strategy for text dialect identification

Mohammad Ali Humayun1 · Hayati Yassin1 · Junaid Shuja2 · Abdullah Alourani3 · Pg Emeroylariffion Abas1

Received: 5 March 2022 / Accepted: 11 October 2022 / Published online: 15 November 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract
Online medical consultation can significantly improve the efficiency of primary health care. Recently, many online medical question–answer services have been developed that connect the patients with relevant medical consultants based on their questions. Considering the linguistic variety in their question, social background identification of patients can improve the referral system by selecting a medical consultant with a similar social origin for efficient communication. This paper has proposed a novel fine-tuning strategy for the pre-trained transformers to identify the social origin of text authors. When fused with the existing adapter model, the proposed methods achieve an overall accuracy of 53.96% for the Arabic dialect identification task on the Nuanced Arabic Dialect Identification (NADI) dataset. The overall accuracy is 0.54% higher than the previous best for the same dataset, which establishes the utility of custom fine-tuning strategies for pre-trained transformer models.

Keywords Text classification · Author profiling · Dialect identification · Arabic language

1 Introduction

Recently online medical question-answering services have gained widespread popularity. These websites connect patients to specialist medical practitioners without needing physical visits. Subsequently, these services reduce the overhead of clinic visits and make the first level of medical support significantly more efficient [1]. Online medical answers can be a practical first step for further diagnosis and referrals for relevant physical visits if required. These services have become remarkably more relevant after the introduction of the social distancing policies due to the COVID-19 pandemic. An essential challenge for such medical question-answering services is the automatic categorisation of questions based on the medical speciality for notifying the relevant practitioners for answers. Since patients are generally unfamiliar with the categorisations of medical specialities, and even those with the familiarity might not be sure about the most appropriate medical field based on their symptoms. Hence, the answering services rely on automatic classification models trained using machine learning techniques [2].

The automatic question categorisation makes the service much more efficient. It removes the requirement or manual efforts, knowledge, and skills required to analyse each question and forward it to the right doctor with the relevant medical specialities. However, the existing question classification models only consider medical specialities for question categorisation and disregard the social background of patients. Selecting a doctor with a social background similar to that of the patients can significantly improve the effectiveness and convenience of patient-doctor interaction by removing socio-linguistic barriers. Moreover, recent medical research has proven that information about a patient’s social background can impact the diagnosis and effectiveness of treatment options.

This research aims at identifying the native country of patients by considering the linguistic varieties in their text. Hence, the task can be seen as country-specific dialect classification or author profiling from Arabic text. Since not enough medical questions datasets are available in Arabic, a recent social-media Arabic text dataset has been
considered for the dialect classification task. The dataset is well-researched as it has been presented as an Arabic dialect identification challenge in 2021 [3]. Multiple state-of-the-art text classification models have reported results on the dataset. It is shown that our proposed method achieves higher accuracy than the previous best on the dataset.

Most successful contemporary author profiling techniques utilise transformer models [4]. The transformer architecture was initially tested for auto-regressive sequence to sequence mapping tasks, such as machine translation, and comprises an encoder and a decoder. Each encoder block within the transformer can be further divided into two sub-blocks: the attention and feed-forward sub-blocks. The attention sub-block consists of multiple weights concatenated kernels for query, key, and value, which are then followed by a fully connected layer as the output of the attention sub-block. This output is fed to a feed-forward network with two fully connected layers: one intermediate and one final-output layer for the encoder. The decoder blocks are similar to the encoder blocks, except for an additional attention layer that considers the final output from the encoder stack along with the decoder attention output. For sequence mapping tasks, the encoder block is fed with input text sequence while the decoder considers the shifted output sequence already predicted by the model as input. Devlin et al. [5] presented a bidirectional encoder representations from transformer (BERT) model based on transformer encoders and demonstrated its usage for multiple downstream tasks. The basic version of the BERT architecture has a sequence of 12 different encoder blocks.

Before training on a labelled dataset of interest, the BERT models are usually pre-trained on a large unlabelled text dataset for multiple unsupervised tasks. The pre-trained models can then be fine-tuned using the labelled dataset for supervised classification with minimal changes to its parameters. The pre-training helps the model learn the text structure utilising massive unlabelled text data scraped from the web. The unsupervised pre-training has been proven very effective for a wide range of classification tasks [6].

Two unsupervised tasks have been used to pre-train the BERT model. The first task was the masked word prediction with a softmax classifier for the entire vocabulary as the final classification layer. Random words from the input sentence were masked, with the training objective of predicting the masked words. The second pre-training task was the next sentence prediction. The final layer was replaced with a single-unit binary classification layer that identifies if the second sentence in a pair of sentences is the following consecutive sentence of the first sentence or not. A unique separator token separates the two sentences of the pairs at the transformer input.

The pre-trained BERT models can be used for downstream tasks by fine-tuning or feature-based strategies. With fine-tuning, the final classification layer is replaced for the downstream task, and the rest of the model initialised with the pre-trained weights is tuned by end-to-end training. Consequently, the fine-tuning updates the pre-trained weights of all the transformer layers according to the downstream classification task. On the other hand, the feature-based strategy transforms the input text into a representation at the output of the encoder-stack in the pre-trained BERT model by forward-pass. These text embeddings may then be used as inputs for training and testing with a separate custom downstream classifier [5, 7, 8].

Recently, a novel strategy has been proposed to adapt the transformers for downstream tasks without complete fine-tuning or utilising the embeddings for another classifier [9]. The proposed technique adds new, small adapter layers within the transformer encoder blocks. During fine-tuning for the downstream tasks, only these adapter layers are trained while freezing other transformer layers to their pre-trained weights. The method proposes to add two fully connected adapter layers for each encoder block, one between attention and feed-forward sub-blocks and the other after the feed-forward sub-block as the encoder output.

Many variants of BERT for the Arabic language are available, and they have been pre-trained on Arabic text datasets of different nature. Among the Arabic BERT models, the MARBERT has been multi-dialectal pre-trained on Twitter data and has been shown to give better results than other pre-trained models when fine-tuned for author profiling [10]. The adapter model has recently been used by AlKhamsi et al. [11] to enhance the MARBERT and achieved the highest score for the Arabic dialect classification challenge. Their model fuses the results by adapter-based tuning and complete fine-tuning of a pre-trained MARBERT [3].

Despite the effectiveness of the adapter model, it needs additional layers to be embedded in the pre-trained model, which requires complex implementation. Moreover, bias toward dominating class has also been reported as a significant issue [11]. In this paper, we propose to improve the simple fine-tuning of the pre-trained BERT models without adding adapters by training only the intermediate, fully connected layers while retaining the pre-trained weights of all other layers. Moreover, to utilise the pre-trained model without fine-tuning, the model proposes concatenating the output of all the transformer hidden layers. Another classifier is trained over these concatenated representations for the classification dataset. Fusion of predictions by the
The paper is organised as follows. The next section reviews recent literature on the progress of linguistic features used for author profiling. Relevant datasets, recent challenges, and associated models for Arabic text dialect identification are reviewed, with brief descriptions of models that have reported superior accuracies for the task. The literature section is followed by a section on the proposed method, explaining the proposed technique in this paper and the experimental setup for its evaluation. The results give a performance comparison of the proposed technique for the Arabic dialect identification task compared to the prior models. Finally, the conclusion lists key findings from this work and their implications in the area.

2 Literature review

Identification of the social background of authors through their writing styles and speakers from their accents has attracted significant interest from the research community due to its wide applications ranging from forensic investigations and speech recognition to informed marketing and authorship attribution [12–14]. This section briefly reviews the history of techniques used for author profiling and then focuses on current transformer-based models, specifically their application for Arabic dialect profiling challenges.

2.1 Traditional author profiling

The oldest techniques for author profiling use a stylometric approach by analysing statistics of text style, including lengths, types of words, and sentences [15–17]. Purely stylometric techniques are then surpassed by content-based lexical features [18, 19]. The most famous examples of lexical features are vector space models, such as ‘Term-Frequency Inverse-Document-Frequency’ (TF-IDF), which represents each sentence as a sparse vector indicating the number of appearances of each term in that sentence normalised by appearances throughout the dataset [20]. Terms refer to individual words or n-grams for words in the dataset vocabulary. Sequences of word embeddings then replace the vector space models learned through training. Most common word embeddings capture semantic information of words and are trained for unsupervised tasks on a large unlabelled dataset. Word2Vec [21] and Glove [22] are the two most well-known semantic word embeddings. Word2Vec embeddings are learned by minimising geometric distance in embedding space between semantically similar words. Words having similar neighbours are considered semantically similar. Glove embeddings also use a similar training objective. However, the dot products for embeddings represent semantic similarity for corresponding words.

2.2 Transformer models

More recently, transformer-based embeddings have outperformed other techniques by demonstrating superior results for most natural language processing tasks [4, 23]. Transformers utilise attention mechanisms and capture words’ context and their semantic information.

Many variants of transformers are available and have been used for a variety of text classification tasks. The models are pre-trained on large amounts of unlabelled text data, with the models varying in their architectures and pre-training methodology [5, 24–26]. Most prominent among the transformer-based models are the BERT models [5] [27, 28]. Pre-trained BERT models have been successfully applied for various text classification problems, including author identification, author profiling, and dialect classification tasks [28–32].

The BERT models are available for many languages, with many BERT models, which have been trained with a variety of unlabelled Arabic text datasets, available for the Arabic language. These include QCRI Arabic and Dialectal BERT (QARiB) [7], Arabic BERT (AraBERT) [8], and MARBERT [10]. Most of these models have been trained on web-scraped datasets, except for MARBERT, which has been trained explicitly on Arabic tweets and is particularly suitable for Twitter datasets. The MARBERT pre-training dataset has a vocabulary size of 100 k and 15.6 B tokens and uses the BERT-based transformer architecture.

2.3 Arabic author profiling challenges

Many datasets and challenges have been arranged for author profiling from Arabic text, as well as, particularly for the task of Arabic language variety or dialect identification. These challenges aim to categorise Arabic text into dialect categories with different granularities, ranging from binary classification into dialectical or standard Arabic to classification into regional dialects. The dialects are associated with different regions, such as Egyptian, Gulf, Levantine, or North-West African, and more fine-grained country or province-level author classification [33–36].

More recently, Nuanced Arabic Dialect Identification 2020 (NADI 2020) and NADI 2021 [3] challenges have presented datasets with even broader coverage and multiple classification tasks, such as identifying provinces and countries from standard as well as dialectical Arabic text.

Many researchers have proposed various methods, specifically for the challenges, by utilising a wide variety of techniques, including n-gram TF-IDF, semantic embeddings, and transformers [37–39]. Transformer-based
classification models have been shown to outperform other techniques, with the most prominent among the transformer-based submissions being the MARBERT model. The MARBERT model has been fine-tuned for the downstream dialect classification task to achieve an accuracy of 51.9% [10]. However, the best accuracy has been demonstrated by Khamisi et al. [11]. They have adapted the MARBERT model by utilising the adapter mechanism proposed initially by Houlsby et al. [9] to fine-tune the transformers. The adapter-based fine-tuning of the MARBERT model has demonstrated benchmark accuracy for Arabic dialect identification. In this paper, the final classification result by Khamisi et al. [11] is reported using an ensemble of a MARBERT model tuned using adapters and another MARBERT model, fine-tuned simply for the complete architecture. Individual accuracies by the adapter and simple fine-tuning models have been reported at 52.48% and 51.02%, respectively. In contrast, the ensemble of both MARBERT models achieves 53.42% accuracy on the development set for the country-level dialect identification.

While the MARBERT [10] model was fine-tuned simply for the NADI challenge, Khamisi et al. [11] added another pre-trained MARBERT tuned with the adapter as an ensemble with the simply fine-tuned model. This paper proposes to enhance the ensemble of classifiers by Khamisi et al. [11] by replicating the adapter-based MARBERT model but replacing the simple fine-tuned model with the newly proposed fine-tuning strategy. Moreover, a third feature-based custom classifier fed with concatenated features extracted from all layers of pre-trained MARBERT is also added to the ensemble.

3 The proposed method

The proposed classification model for dialect classification comprises three individual classifiers, with an overall output given by simply averaging the outputs from the three classifiers. Figure 1 represents the proposed classification model. The three individual classification models, which constitute the proposed ensemble, are:

1. The MARBERT model is fine-tuned using the proposed strategy in this paper, which is accomplished by training only selected layers of the transformer.
2. The MARBERT model, fine-tuned with the adapter mechanism, as implemented by Khamisi et al. [11].
3. Standalone classifier using concatenated pre-trained MARBERT representations without fine-tuning.

Simple averaging is performed on the outputs from the three classification models, with the model referred to as the ensemble model. Averaging the three different components is expected to improve the model’s generalisation. The novel fine-tuning strategy employed in the first component of the ensemble also improves the generalisation as it helps to retain most of the parameters learned during unsupervised pre-training on much larger datasets. The adapter model used as the second component has achieved the best results for the particular task. Similarly, the standalone classifier as the third component, using encoder embeddings as input, performs better when fed with representations from multiple encoder layers.

All three individual classifiers considered are built on the MARBERT model [10]. It is a publicly available transformer trained on a large dataset of Arabic tweets. The MARBERT model comprises 12 encoder blocks, with an output of the preceding encoder block serving as input to the following encoder block. The output of the MARBERT model is effectively the output of the 12th encoder block. The MARBERT model utilises 768-dimensional embeddings at every layer.

3.1 Fine-tuning strategy

The proposed fine-tuning model wholly follows the MARBERT model in terms of architecture. Similar to the adapter model, a classification head with C target classes has been added to the fine-tuned model, taking input from the 768-dimensional embedding of the last encoder block. The output classification layer allows the fine-tuned model to be used for the classification task. A fine-tuning strategy has been proposed for the encoder model. The tuning strategy proposes freezing all the sub-layers within the encoder blocks, except for the intermediate fully connected layers within the feed-forward sub-block of each encoder block. Figure 2 represents an encoder block of the proposed fine-tuned model. Every intermediate layer of all 12 encoder blocks is fine-tuned during training while freezing all other layers with the pre-trained weights of the original
MARBERT model. Freezing most layers allows the retention of the rich pre-trained embedded knowledge on most of the MARBERT layers. Still, it allows flexibility to adapt specifically to the classification task by allowing fine-tuning of weights of the intermediate fully connected layers. Additionally, the fine-tuning strategy also avoids overfitting.

### 3.2 Adapter

The adapter model follows the MARBERT model as a basis, but has been adapted by embedding two additional layers, called the adapter, at each encoder block of the transformer. One of the adapters is added after the attention sub-block, and the other adapter, after the feed-forward sub-block. Figure 3 represents the adapter model, with details of the model given by Khamisi et al. [11]. A classification head with C target classes corresponding to the dialect categories has been added to the model. During fine-tuning, all the layers are frozen, except for the two adapter layers of the encoders. Again, this allows retention of rich pre-trained embedded knowledge in the original MARBERT while allowing the model to adapt to the specific classification task via fine-tuning the adapter layers.

---

**Fig. 2** Proposed fine-tuning strategy

**Fig. 3** Adapter tuning [11]

**Fig. 4** Feature-based method
3.3 Feature-based model

The final individual classifier is feature-based and is given in Fig. 4. It comprises 12 pre-trained encoder blocks of the original MARBERT model, with 768-dimensional embedding from each encoder block serving as input to the next consecutive encoder block. As opposed to the previous individual classifiers, the encoders in the proposed feature-based model are not fine-tuned to the classification task and retain the pre-training weights of the original MARBERT model. The MARBERT model without the classification head is used to transform the text token sequences into their respective embeddings. The fix-length token sequences have been forward-passed to the model, and outputs by each of the 12 encoder blocks are captured for only the first, sentence-level token. Consequently, each sentence is transformed into 12 different 768-dimensional embeddings at encoder outputs. These 768-dimensional embeddings outputs of each encoder block are then concatenated to provide a single 768 × 12-dimensional vector output, with the transformer architecture discarded after that. The single 768 × 12-dimensional vector output is then fed to a single fully connected softmax classifier with C target classes, corresponding to the dialect categories of the classification task. The output of the softmax classifier provides output to the overall proposed feature-based model.

4 Results and discussions

The proposed method has been evaluated for Arabic text dialect identification using the NADI 2021 dataset [3] for the country-level identification task. Many methods have been proposed for the classification task, including state-of-the-art models such as the adapter approach. Only the training and development subsets from the NADI 2021 dataset have been utilised for training and testing the models, respectively, as the organisers only provide true labels for these two subsets. The training and development subsets consist of 21,000 and 5,000 text samples, respectively, with each sample annotated with one of the 21 countries, i.e. C = 21 classes. These annotated samples represent the Twitter text of the respective dialectical Arabic selected by considering the origin and location history of the users.

Input text has been tokenised with the MARBERT tokeniser with padding and truncation to the fixed length of 30 tokens per sample. The three individual classifier models were then trained for the classification task on the training dataset. Both the adapter and proposed fine-tuning models have been trained end-to-end. As has been previously discussed, the original MARBERT layers have been frozen in the adapter model, except for the additional adapter layers at every encoder block. Similarly, all MARBERT layers have been frozen in the fine-tuned model, except the intermediate fully connected layer in the feed-forward network of the encoder block. In contrast, only the single 768 × 12-dimensional vector output from the proposed feature-based model has been used to train the classification head.

4.1 Overall accuracy

In the first instance, the performances of the three individual classifier models, the adapter, proposed fine-tuning, and proposed feature-based models, and their ensembles have been compared on the dialect identification task. Table 1 lists the accuracies, macro-average of precision, and recall achieved by the three individual classifiers and their ensembles on the development dataset. It can be seen that the adapter model gives the best accuracy among the individual classifiers, with 52.26% accuracy. However, averaging the outputs from the adapter, proposed fine-tuning, and proposed feature-based models via simple averaging provides the best accuracy of 53.96%. Higher accuracy by the average output indicates that the different classifiers capture complementary information.

Precision is the ratio of true predictions to total predictions for a particular class, while recall is the ratio of true predictions to total test samples of that class. Macro-averaging simply averages total precision and recall for all classes with equal weightage. As can be seen, the ensemble of the three individual classifiers achieves precision and recalls of 0.4335 and 0.3113, respectively. It is noteworthy that among the individual classifiers, the proposed fine-tuning model achieves the highest precision but the lowest recall among all the individual classifiers. Moreover, the final ensemble reports considerably higher precision than the adapter model, with a recall value similar to the adapter model.

From the literature, the best accuracy for the country-level identification task on the NADI2021 dataset has been reported using the adapter method/fine-tuning ensemble for MARBERT [11], and then, followed by fine-tuning for MARBERT [10], achieving 53.42% and 51.90% accuracies, respectively, on the development set. All the individual classifiers used in this paper have also utilised the original MARBERT as their basis. The adapter model used in this paper has been replicated using the same architecture provided by Khamisi et al. [11] using multiple training initialisation seeds but without vertical attention. Table 2 represents the classification accuracy of the proposed ensemble model compared to the two previous best models for the same dataset. The proposed ensemble model
achieves a classification accuracy of 53.96%, which is 0.54% higher than the adapter and fine-tuning ensemble reported by Khamisi et al. [11].

The proposed ensemble model comprises three individual classifiers: the adapter, the proposed fine-tuning, and the proposed feature-based models. Fine-tuning just the intermediate layer while freezing the rest of the layers, in the case of the proposed fine-tuning model, helps to retain useful information learned on large pre-training datasets and prevents overfitting to the smaller fine-tuning training set. Moreover, output embeddings from the different encoder blocks of the pre-trained model represent different levels of complementary information useful to train the downstream classifier without fine-tuning the transformer. The 768-dimensional sentence-level representations by all the 12 encoder blocks have been down-projected to two-dimensional using principal component analysis (PCA) for visualisation. Figure 5 presents the scatter plots for the two down-projected PCA components of the 12 encoder representations for the complete development set samples. Samples for each of the 21 countries have been represented with a unique colour.

4.2 Country-wise classification metrics

Figure 6 presents the confusion matrix for the final predictions by the proposed ensemble model, with rows and columns indicating actual and predicted classes of the samples, respectively. Consequently, each row–column index-pair represents the rounded-off percentage of samples for that row-country (true-class) predicted as the column-country (predicted-class), with the diagonal representing percentage of correct predictions. The confusion matrix shows that the proposed ensemble model achieves the best results for Egypt and Saudi Arabia with

| Model                        | Accuracy % | Precision (macro) | Recall (macro) |
|------------------------------|------------|-------------------|----------------|
| Adapter model                | 52.26      | 0.3491            | 0.3136         |
| Proposed fine-tuning model   | 51.4       | 0.4217            | 0.2762         |
| Proposed Feature-based model | 50.6       | 0.3919            | 0.2790         |
| Proposed Ensemble            | 53.96      | 0.4335            | 0.3113         |

Table 1 Classification accuracy by the components of the proposed method and their ensemble

Table 2 Classification accuracy comparison with previous best models

| Model                          | Accuracy % |
|--------------------------------|------------|
| Fine-tuning for MARBERT [10]   | 51.9       |
| Ensemble with adapter for MARBERT [11] | 53.42     |
| Proposed ensemble              | 53.96      |

Figure 5 Scatter plots of the first and second PCA components at the different encoder layers of the feature-based model

Figure 6 Confusion matrix of the proposed ensemble classification model

93% and 71% accuracies, respectively. Bahrain, Djibouti, and Somalia achieve 0% accuracy, with the most significant confusion reported for Somalia, with 57% of the samples wrongly predicted as being from Saudi Arabia.
The higher accuracy for Egypt and Saudi Arabia can be attributed to the imbalance in the NADI dataset [3]. The dataset contains 4220 training and 1041 development samples for Egyptian dialect and 2110 training and 520 development samples for Saudi Arabian dialect. These are much higher as compared to the rest of the countries.

Table 3 lists the class-wise precision and recall scores, with the number of test samples for each class. Similarly, Egypt has the highest recall value, followed by Saudi Arabia, with 0.932 and 0.713, respectively. Again, these can be attributed to the higher number of samples in the dataset. On the other hand, Mauritania has the highest precision value of 0.727, followed by Egypt and Morocco with 0.719 and 0.707, respectively. Kuwait, Sudan, Oman, and Saudi Arabia have the lowest recall values of 0.328, 0.372, 0.399, and 0.399, respectively. Precision for Saudi Arabia is very low despite its higher recall value due to many misidentifications from other countries as being from Saudi Arabia.

5 Conclusion

Most text classification tasks have achieved the best results by fine-tuning the pre-trained transformer models. This paper proposes a novel technique of fine-tuning a pre-trained transformer model by freezing all, except the intermediate fully connected layers of each encoder block, before training on a particular classification task. The proposed fine-tuning model effectively retains valuable information from the original MARBERT model, which has been pre-trained on a huge Arabic dataset while at the same time adapting to the smaller training dataset. Moreover, the paper has also illustrated the effectiveness of concatenating the output representations from all hidden-layer encoder blocks to train a custom downstream classifier. Fusion of the proposed classifiers with the state-of-the-art adapter-based classification achieves a classification accuracy of 53.96% for identifying the authors’ origin from a span of 21 countries, superseding the previous best model by fusing adapter with simple fine-tuning of the same pre-trained BERT model, with a reported accuracy of 53.42%. Although improvement in the classification accuracy is marginal, the results are significant as the previous best model has already reported quite a high accuracy as compared to many other models. Precision and recall of the proposed ensembles are 0.4335 and 0.3113, respectively. Country-specific confusion matrix has indicated that the model performs considerably better for countries with more text samples in the dataset. Hence, obtaining a more balanced dataset is critical to achieve balanced performances across all the countries.

In the future, we plan to test our proposed algorithm for other languages to establish its utility across text in different languages. Moreover, the model can be used with acoustic models for speaker profiling from spontaneous speech.

Table 3  Country-wise precision and recall

| Country       | Precision | Recall | Test samples |
|---------------|-----------|--------|--------------|
| Algeria       | 0.582     | 0.579  | 430          |
| Bahrain       | 0         | 0      | 52           |
| Djibouti      | 0         | 0      | 27           |
| Egypt         | 0.719     | 0.932  | 1041         |
| Iraq          | 0.583     | 0.66   | 664          |
| Jordan        | 0.533     | 0.154  | 104          |
| Kuwait        | 0.328     | 0.21   | 105          |
| Lebanon       | 0.477     | 0.134  | 157          |
| Libya         | 0.591     | 0.548  | 314          |
| Mauritania    | 0.727     | 0.151  | 53           |
| Morocco       | 0.707     | 0.198  | 207          |
| Oman          | 0.399     | 0.448  | 355          |
| Palestine     | 0.471     | 0.231  | 104          |
| Qatar         | 0.5       | 0.019  | 52           |
| Saudi Arabia  | 0.399     | 0.713  | 520          |
| Somalia       | 0         | 0      | 49           |
| Sudan         | 0.372     | 0.66   | 53           |
| Syria         | 0.267     | 0.252  | 278          |
| Tunisia       | 0.612     | 0.237  | 173          |
| UAE           | 0.382     | 0.318  | 157          |
| Yemen         | 0.455     | 0.095  | 105          |

Declarations

Conflict of interest  The authors have no conflict of interest with the presented research.

Data availability  The dataset used for this study is obtained from the text dialect profiling challenge and is publicly available at https://sites.google.com/view/nadi-shared-task

Human and animal rights  This article does not contain any studies with human participants or animals performed by any authors.

References

1. Cao YG, Liu F, Simpson P, Antieau L, Bennett A, Cimino JJ, Ely J, Yu H (2011) AskHERMES: An online question answering system for complex clinical questions. J Biomed Inform. https://doi.org/10.1016/j.jbi.2011.01.004
2. Chen CW, Tseng SP, Kuan TW, Wang JF (2020) Outpatient text classification using attention-based bidirectional LSTM for robot-assisted servicing in hospital. Inf. https://doi.org/10.3390/info11020106
3. Abdul-Mageed M, Zhang C, Elmadany A, Bouamor H, Habash N (2021) NADI 2021: The Second Nuanced Arabic Dialect Identification Shared Task, ArXiv Prepr. arXiv:2103. http://arxiv.org/abs/2103.08466.

4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst. 5999–6009.

5. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL HLT 2019. In: 2019 Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, 1 (2019) 4171–4186.

6. Humayun MA, Yassin H, Abas PE (2021) Spatial position constraint for unsupervised learning of speech representations. PeerJ Comput Sci 7:1–24. https://doi.org/10.7717/PEERJ-CS.650

7. Abdelali A, Hassan S, Mubarak H, Darwish K, Samih Y (2021) Pre-training BERT on Arabic tweets: practical considerations. http://arxiv.org/abs/2102.10684

8. Antoun W, Baly F, Haji H (2020) AraBERT: Transformer-based model for arabic language understanding. http://arxiv.org/abs/2003.00104.

9. Houlsby N, Giurgiu A, Jastrzbczki S, Morrone B, de Laroussilhe P & MARBERT: deep bidirectional transformers for Arabic, transfer learning for NLP. In: 36th international conference on computational linguistics: human language technologies. 1 (2020) pp 7088–7105. https://doi.org/10.18653/v1/2021.acl-long.551.

10. Abdul-Mageed M, Elmadany A, Nagoudi EMB (2021) ARBERT Neural Computing and Applications (2023) 35:6115–6124 6123

11. AlKhamissi B, Gabr M, ElNokrashy M, Essam K (2021) Authorship attribution and verification with many authors and limited data. Artif Intell Conf. https://doi.org/10.1017/S1351324920000601

12. Humayun MA, Yassin H, Abas PE (2022) Native language identification for Indian-speakers by an ensemble of phoneme-specific, and text-independent convolutions. Speech Commun. https://doi.org/10.1016/j.specom.2022.03.007

13. Björklund J, Zechn er N (2017) Syntactic methods for topic-independent authorship attribution. Nat Lang Eng. https://doi.org/10.1017/S1351324917000249

14. Al-Yahya M (2018) Stylometric analysis of classical Arabic texts for genre detection. Electron Libr. https://doi.org/10.1016/EL-11-2017-0236

15. Abbasi A, Chen H (2008) Writeprints: a stylometric approach to dependent authorship attribution. Nat Lang Eng. https://doi.org/10.1017/S1351324917000249

16. Neal T, Sundararajan K, Fatima A, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst. 5999–6009.

17. Luyckx K, Daelemans W (2008) Authorship attribution and deception detection in Arabic shared task at FIRE. In: Pervasive health applications and technologies for pervasive computing technologies for healthcare. pp 7–9. https://doi.org/10.1017/S1351324917000249

18. Al-Yahya M (2018) Stylometric analysis of classical Arabic texts for genre detection. Electron Libr. https://doi.org/10.1016/EL-11-2017-0236

19. Abbasi A, Chen H (2008) Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans Inf Syst. https://doi.org/10.1145/1344411.1344413

20. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18:613–620. https://doi.org/10.1145/361219.361220

21. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: 1st international conference on learning representations. ICLR 2013 – workshop track proceedings

22. Pennington J, Socher R, Manning CD (2014) GloVe: Global vectors for word representation, in: EMNLP 2014 – proceedings of the 2014 conference on empirical methods in natural language processing. pp 1532–1543. https://doi.org/10.3115/v1/d14-1162

23. Wang C, Banko M (2021) Practical transformer-based multilingual text classification. pp 121–129. https://doi.org/10.18653/v1/2021.naacl-industry.16

24. Radford A, Narasimhan T, Salimans T, Sutskever I (2018) [GPT-1] Improving Language Understanding by Generative Pre-Training, in: Preprint. pp. 1–12.

25. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) GPT-3. NeurIPS. 2020-Decem

26. Dale R (2021) GPT-3: What’s it good for? Nat Lang Eng. https://doi.org/10.1017/S1351324920000601

27. F. Rangel, P. Rosso, A. Charfi, W. Zaghouani, B. Ghanem, J. Sánchez-Junquera (2019) On the author profiling and deception detection in Arabic shared text at FIRE. In: Pervasive health applications and technologies for pervasive computing technologies for healthcare. pp 7–9. https://doi.org/10.1017/S1351324917000249

28. Gao Z, Feng A, Song X, Wu X (2019) Target-dependent sentiment classification with BERT. IEEE Access 7:154290–154299. https://doi.org/10.1109/ACCESS.2019.2946594

29. Sun C, Qiu X, Xu Y, Huang X (2019) How to Fine-Tune BERT for Text Classification?. In: Lecture notes in computer science (Including its subservice lecture notes in artificial intelligence and lecture notes in bioinformatics). https://doi.org/10.1007/978-3-030-32381-3_16

30. Sun C, Qiu X, Xu Y, Huang X (2019) How to Fine-Tune BERT for Text Classification?. In: Lecture notes in computer science (Including its subservice lecture notes in artificial intelligence and lecture notes in bioinformatics). https://doi.org/10.1007/978-3-030-32381-3_16

31. Zhang C, Abdul-Mageed M (2019) BERT-based Arabic social media author profiling. In: CEUR workshop proceedings. pp 84–91

32. Suman C, Naman A, Saha S, Bhattacharyya P (2021) A multi-modal author profiling system for tweets. IEEE Trans Comput Soc. https://doi.org/10.1109/TCSS.2021.3082942

33. W. Zaghouni, A. Charfi, AraP-Tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In: Lr. 2018: eleventh international conference on language resources and evaluation. pp 694–700

34. Zaghouni W, Charfi A (2018) Guidelines and annotation framework for arabic author profiling, CoRR.abs/1808.0

35. F. Rangel, P. Rosso, A. Charfi, W. Zaghouni, B. Ghanem, J. Sánchez-Junquera (2019) On the author profiling and deception detection in Arabic shared text at FIRE. In: Pervasive health applications and technologies for pervasive computing technologies for healthcare. pp 7–9. https://doi.org/10.1017/S1351324917000249

36. Talalha B, Farhan W, Alatkrouri A, Al-Natsheh H (2019) Mawodo3 AI at MADAR shared task: Arabic tweet dialect identification. pp 239–243. https://doi.org/10.18653/v1/w19-4629

37. El Mekki A, Alami A, Alami H, Khoumsi A, Berrada I (2020) Weighted combination of BERT and N-GRAM features for Nuanced Arabic Dialect Identification. In: Proceedings fifth Arabic national language processing work, pp. 268–274. https://www.aclweb.org/anthology/2020.wanlp-1.27
38. Wadhawan A (2021) Dialect Identification in Nuanced Arabic Tweets Using Farasa Segmentation and AraBERT. In: Proceedings sixth Arabic national language processing work. http://arxiv.org/abs/2102.09749
39. Lichouri M, Abbas M, Lounnas K, Benaziz B, Zitouni A (2021) Arabic dialect identification based on a weighted concatenation of TF-IDF features. In: Proceedings sixth Arabic national language processing work. pp 282–286. https://www.aclweb.org/anthology/2021.wanlp-1.33

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.