IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 29, No. 6, December, 2024, pp. 2209-2218
Journal homepage: http://iieta.org/journals/isi

Assessment of LSTM, ARABERT and Prompt-Based Learning for Gender Author Profiling in Modern Standard Arabic Language

Asmaa Mansour Khoudja[1*], Mourad Loukam[2], Fatma Zohra Belkredim[1]

[1] LME Laboratory, Hassiba Benbouali University of Chlef, Chlef 02000, Algeria
[2] LIA Laboratory, Faculty of Exact Sciences and Informatics, Hassiba Benbouali University of Chlef, Chlef 02000, Algeria

Corresponding Author Email: a.mansourkhoudja@univ-chlef.dz

https://doi.org/10.18280/isi.290611

Received: 21 May 2024
Revised: 4 September 2024
Accepted: 25 September 2024
Available online: 25 December 2024

Keywords:
ARABERT, author profiling, deep learning, gender profiling, LSTM, Modern Standard Arabic Language (MSA), prompt based learning

ABSTRACT

Author Profiling aims to extract persons' characteristics (gender, age…) from their writings. This emerging field of NLP poses great challenges for all languages in general and, in particular, for the Modern Standard Arabic Language. This paper presents an assessment study of three state-of-the-art approaches used for gender author profiling, namely, LSTM, ARABERT, and Prompt-Based learning. Using a rich dataset created for this task, our research investigates the effectiveness of these methods in gender identification. Our findings indicate that the ARABERT method obtained the highest scores in terms of accuracy, ranging from 84.6% to 92.4%, and Prompt-Based learning performed competitively compared to ARABERT, with accuracy increasing from 84% to 92.3%. However, while LSTM also showed progress across all batches, it still consistently performed worse than the other two models and reached an accuracy of only 78.5%.

1. INTRODUCTION

Over the past few decades**,** the growth of data on the Internet aroused wide interest in extracting relevant information from this massive amount of data. For textual data (blogs, posts, opinions, articles, …), it is deemed necessary to develop software tools for analysing texts and profiling their authors in order to extract useful characteristics which can help us to understand their personalities and motivations. Thus, Author profiling (AP) has emerged as a crucial field, at the crossroads of Natural Language Processing (NLP) and Text Mining, offering various methods and techniques for identifying and uncovering the personal and demographic characteristics of the authors from their written texts, which could include their gender [1], age, political beliefs, or their mother tongue [2].

Author Profiling is attracting a great deal of interest since many years, due to its wide range of applications, such as: in security and crime investigation, AP can be used to shortlist potential suspects by analysing a given text [3]. In marketing, AP can be useful for devising commercial strategies to target specific groups of consumers based on their age and gender [4]. According to a recent study [5], Author Profiling has been a rapidly growing research area in computer Science.

In this research, we focus on gender profiling, also called gender identification, and applied to the modern standard Arabic (MSA) language. In practical terms, this involves detecting the gender of an author (male or female) from a text written by this author.

Author profiling and its sub-topic "gender profiling" pose great challenges for all languages and, in particular, for MSA. This is due to several reasons: ambiguities of all kinds that are frequently found in texts, linguistic expressions that lead to gender identification are far from obvious, wide cultural differences used in gender expressions, the lack of consistent training sets needed to train computer models to be developed for gender profiling.

Over time, several approaches have been explored to solve NLP tasks. These include statistical models, learning-based models, deep learning, etc. But the emergence of large language models (LLM) and transformers has been an important step in this evolution [6, 7]. Indeed, BERT introduced the notions of pre-training and fine-tuning, which have proved to be a very interesting alternative for solving many NLP problems [8]. More recently, a new approach called 'prompt-based learning' has been introduced [9-11]. This technique tries to make the data fit the pre-trained model by providing clear instructions or prompts to overcome the biggest issue in NLP, which is the need for large amounts of supervised data. This could be a very promising solution for low-resource languages such as MSA.

In this study, we therefore considered it interesting to investigate three main different models for dealing with the problem of gender identification in MSA: LSTM (based on deep learning), ARABERT and Prompt based learning. The aim is to find out which model is best suited to solving this problem.

This study deals with Author Gender Profiling based on texts written in MSA. It is worth noting that Arabic is today one of the five most widely spoken languages in the world. But unfortunately, Arabic language processing is still lagging behind in the fields of AP and NLP in general. This is mainly due to the lack of resources and also to the fact that Arabic is a rich and complex language.

Despite the importance of Gender Profiling for MSA,

research in this field is limited due to the lack of a required consistent data set for training effective models. To get around this difficulty, in this study, we have built and collected more than 10,000 labelled Arabic text data from different resources and applied the three models mentioned above: LSTM, ARABERT and Prompt-Based learning. The goal is to find out which is the most efficient for gender profiling.

This study focuses on addressing the following research questions:

(1) To what extent are the three methods (LSTM, ARABERT, and Prompt-based learning) suitable for solving the problem of gender profiling in MSA?

(2) How should each method be configured for this purpose?

(3) What is the most effective method?

(4) How does the performance of each method evolve according to the size of the training dataset?

The paper is organised as follows:

In Section 2, Author Profiling is introduced with a review of works that have been done in this field, for English, and in particular for MSA. Section 3 is devoted to an overview of the three methods investigated in this study for gender AP: LSTM, ARABERT and Prompt-based learning. Section 4 describes the assessment of the studied methods: data and performance measures used in the evaluation. The results obtained and their discussion are detailed in Section 5. Section 6 outlines the limitations of our study. Finally, a conclusion is given, summarising the work done and providing guidelines for future work.

2. RELATED WORK IN GENDER AUTHOR PROFILING

In this research, we focus on gender identification, which is a sub-topic of Author Profiling. There has been a great deal of interest in this field of research. A lot of progress has been made in AP, but it still a topical issue. Gender AP is the analysis of people's writing in an attempt to predict the gender (male/female) of the author.

There are numerous objectives of gender AP, for instance this task could be useful in domains such as marketing where it can help marketers for targeting specific consumers, based on their gender, to offer them products tailored to their needs. In this case, the gender is predicted from comments left by customers on the websites of commercial companies. In forensic domain, AP can be an interesting tool to use during investigations. By predicting the gender of an anonymous author through his or her writing, it can help to consider him or her as a potential crime-suspect.

Several works have been carried out in the field of AP treating English. For instance, Mamgain et al. [12] used a range of machine learning classification models including Bag-of-Words (BOW), Logistic Regression, Random Forest, LSTM-CNN for the task of gender and language variety prediction. The results showed good performance for gender profiling.

Despite the progress made in this field, it still faces numerous difficulties. One of these main challenges in gender profiling is the lack of training data, as it is hard to acquire a large and diverse labelled dataset of texts. Furthermore, the gender expression can vary from culture to culture, which can make the task even more difficult to solve. Another challenge is to extract features that can accurately detect the gender of the authors from the writing style, owing to the differences in writing styles between males and females that are often subtle and difficult to classify. For example, in the English language the gender markers such as pronouns and first names are ambiguous. Indeed, the pronouns "they" and "their", "I" and "my" can refer to both genders.

For the Arabic language, the gender is mostly morphologically marked. For instance, the equivalent of "I am happy" in Arabic is "أنا سعيد" or "أنا سعيدة" where happy is morphologically identified as either masculine or feminine. Using these features can help in detecting gender dissimilarity in a precise way. However, sometimes it is not easy to predict the gender when the person is writing in the first person with the pronoun 'I'. For example, let us examine in Arabic the following sentence: (حصلت على شهادة البكالوريا من المدرسة الثانوية الجديدة للبنات) when translated to English it becomes (I obtained my baccalaureate degree from the new secondary girls' school). We can only assume that the author is female just because she attended a school which was exclusively for girls [5].

Ameer et al. [13] proposed a content-based approach for author's age group and gender detection. They used a different set of features, such as syntactic n-grams, traditional n-grams, and the combination of word n-grams and character n-grams. They used multiple classifiers. The achieved accuracy of 73% signifies that the combination of word n-grams using the SMO classifier can produce good results.

Mechti et al. [14] presented an approach that discriminated between age and gender through four steps: the calculation of words' occurrences, the selection of classes, and the creation of ARFF2 files. They achieved the highest accuracy in the competition with 36.7%.

López-Monroy et al. [15] approached the gender and age classification task with the idea of second order attributes where, they focused on sub-profiles and building document vectors of the targeted profiles. They achieved the highest accuracy with more than 70%, using data from different social media domains.

In 2018, Takahashi et al. [16] developed a method called TIFNN that fuses images and texts using a Recurrent Neural Network (RNN) for texts and a Convolutional Neural Network (CNN) for images, to identify gender. Their method ranked first in the competition with an average accuracy of 81.9%. Similarly, using image and text fusion for the gender classification task [17] proposed a bi-directional GRU neural architecture. They concluded from the obtained results that weighted attention performs better for gender prediction tasks.

Alongside these works, Ouni et al. [18] proposed two new models in the task to distinguish between bots and humans on Twitter, then to identify the gender of human users. The first model they used was the CNN with topic-based semantic feature extraction from tweets, and the second was a random forests classifier. The results showed the effectiveness of their system in terms of accuracy, precision, recall, F1-score and G-mean. For exactly the same task [19] addressed the problem differently. For the Gender detection they used SVM architecture. They achieved an accuracy of over 80% for the gender detection task.

The main objective of the work presented by Ikae and Savoy [20] was to check if the same model always gives the best results when using similar corpora under the same conditions in gender classification. Thus, they used 10 different classifiers and proposed a 2-stage feature selection to reduce the feature size to a few hundred terms without any significant change in the performance level compared to other approaches using all the attributes. They concluded that neural network or

random forest along with feature reduction produced the highest results.

However, when it comes to Arabic, we can only mention a few related works in the field of AP, because Arabic is in fact a challenging language due to its rich and complex nature (phonology, orthography, morphology, and syntax).

Alsmearat et al. [1] approached the problem by checking whether female authors write with more emotions than male authors. The results showed that there is no evidence that this case is true.

In 2016, AlSukhni and Alequr [21] detected the gender of authors in Arabic with the BOW method along with the name of the authors of the tweets and the total words for each tweet. They found out that using the name as a feature can help in improving the accuracy up to 98%.

Bsir and Zrigui [22] explored gender identification among Arabic authors using two corpora: the PAN-competition corpus and the Facebook corpus. The proposed technique, which combines stylistic models, word embedding, and GRUs, showed effectiveness with a considerable 79% accuracy when compared to the best performance in PAN 2017 using the same corpus. Notably, the bi-directional recurrent neural networks performed well. This demonstrates that neural network models are capable of handling NLP tasks, such as text classification and sentiment analysis.

Nayel [23] Defined the strategies and experiments that have been utilized in the system development of two tasks: author profiling in Arabic tweets and deception detection in Arabic texts. The main experiments relied on classical machine learning, specifically Linear classifiers, SVM and Multilayer Perceptron classifiers. In addition, the author used Bag-of-Words with a range of n-grams for feature extraction. He outperformed other teams with one of his submissions for the second task.

Mubarak et al. [24] worked on gender analysis on Arabic tweets in which they performed an intensive evaluation of the differences between male and female users, where, they studied the differences in user engagement, and topics of interest. Furthermore, they proposed a method using features extraction and the SVM classifier, to infer gender by utilizing usernames, tweets, profile pictures, and networks of friends. They had to manually annotate gender and locations for over 160K Twitter accounts associated with user location. Their method initially achieved impressive results of 82.4%.

In another paper, Zhang and Abdul-Mageed [25] employed BERT models to detect age, gender, and dialect for the Arabic APDA shared task under different data conditions. Their ideal model achieved 81.67% accuracy in gender detection task.

Furthermore, Alzahrani and Jololian [26] used different pre-processing techniques to see how they affect the gender profiling of authors when using the BERT Model. They discovered that the highest accuracy of 86.67% was achieved without applying data pre-processing techniques.

To summarize the literature, different approaches have been used in gender identification. These include features extraction like gender-specific linguistic features, bag-of-words (BOW), etc. Also, various machine learning models have been explored, including Multilayer Perceptron, SVM, and Linear Classifier. Furthermore, recent research has demonstrated the success of BERT models, with accuracy of over 80%.

Unfortunately, we couldn't find any related work to English/Arabic gender classification using the most recent paradigm in NLP, Prompt-Based earning.

The following sections discuss in detail our contribution to Arabic gender AP.

3. THE STUDIED METHODS

NLP has evolved over time, step by step. Starting with simple linguistic models and focusing on using rule-based systems to interpret and process language. After that, statistical methods advanced the field using probabilistic models. The field later experienced a new transformation through machine learning, using different approaches such as SVM, Naïve Bayes, Linear classifier, and Multilayer Perceptron in a variety of tasks like text classification, text generation, and named entity recognition. Allowing systems to learn patterns and use manual feature extraction, such as style-based features, for example character-based features that contained properties like the number of the total words, number of words in each sentence, vocabulary richness, etc.

NLP continued to evolve using advanced techniques called deep learning, which automated feature engineering [27]. These techniques used neural networks such as recurrent neural networks (RNNs) and long short-term memory (LSTM), which allowed models to automatically learn complicated patterns directly from unprocessed data. Furthermore, as of 2013, word embedding was introduced by Mikolov et al. [28] these techniques represent words as vectors and then can be utilized to enhance the efficiency of the used model.

In 2017, The transformer models have revolutionized the field of NLP [29], using large language models (LLMs) such as BERT and GPT.

BERT (Bidirectional Encoder Representations from Transformers) for instance is a machine learning language representation model in NLP developed in 2018 by Google [8]. Its rise caused a huge shift in the NLP domain. BERT structure is almost similar to Transformer. For instance, Transformer contains two mechanisms one that takes the input and understands it as a whole (it learns the whole sequences of a text at once, both left and right), and it is called an encoder and another that predicts the output for the specific task and it is called a decoder [29]. While Transformers use these two mechanisms BERT only uses the encoder part because BERT's goal is to generate a language model, so only the encoder mechanism is necessary. Also, BERT uses masked language model where it hides a word in the input sentence then makes the algorithm predict the hidden word based in the context.

Unlike BERT, GPT (Generative Pre-trained Transformer) is based on a transformer decoder to produce texts that are indistinguishable from human written texts [30]. It is trained on almost 45 terabytes of unlabelled dataset. The 1st generation of the generative pre-trained language model (GPT-1) was proposed in 2018 by the OpenAI team. Based on GPT1, GPT2 was introduced a year later with some improvement to the model structure with additional training data [31]. As of June 2020, GPT-3 appeared as the third successor, where the parameters and the data scale were expanded further. Thus, GPT3 was able to learn the patterns and structures of natural language because the model is trained on a massive dataset. This final model can be used for a wide range of applications, including translation, Chatbots, and human-like text generation which is one of its essential features such as generating poetry fluently or even code generation. It can also adapt to new tasks like text classification, sentiment analysis and many others with

minimal additional training. Recently, GPT4 was introduced as the latest successor of GPT. Given its expanded general knowledge and problem-solving skills, it can generate safer and more accurate responses.

Finally, Prompt-based learning appeared in 2021 with a strategy that allows us to use the same model for different tasks without the need to re-train on huge labelled datasets. The basic idea of Prompt-based learning is to insert clear instructions to the input example. This is what is called prompt [9]. For example, when classifying a sentiment on a movie review, we can use this template to modify the primary input: "Input-Text" It was "Mask". Then a verbalizer is used to define the class of the input example. Sometimes a set of label words are used to define the class. For instance, to define the positive class we can use words like: great, good, nice, etc. and bad or terrible for the negative class [32]. After that a prompt model is used to combine the previous steps with the PLM (pre-trained language models) [32]. There exist many PLMs that can be used in Prompt learning for example Masked LMs like BERT, or Encoder Decoder PLMs like T5 and BART. Finally, the training phase is performed. The main complication of this new strategy is the choice of the right prompt, it is very important to choose the accurate prompt to solve a specific task.

This technique is presumed that in the long run, it could be a really useful solution against the problems facing the Standard Arabic language in NLP. That is why we decided to test our collected data using this technique.

To explore the best possible solution to address the task of gender AP, with a focus on MSA, we settled on applying three of the most successful paradigms used in NLP tasks, namely: LSTM, ARABERT from BERT, and Prompt-based learning using ARABERT.

Recent research discussed in the related work, has demonstrated the success of LSTM and BERT models in gender prediction. Prompt-based method also has proven to be adaptable and effective across other range of applications in NLP, such as text classification [10] and dialogue systems [11]. Its use in Arabic gender prediction or in Arabic NLP in general, however, is still unexplored. That is why we believe that our study fills a critical gap in the literature and contributes to advancing the field by becoming the first to apply Prompt-based for gender identification in Arabic.

3.1 The first method LSTM

LSTM is a type of RNN that have demonstrated effectiveness in addressing a range of NLP tasks, particularly in processing long sequential textual data. The strengths and limitations of LSTM are summarised below.

LSTM Strengths:

- It can capture information over long sequences.
- It is effective in tasks requiring sequence prediction.
- It manages effectively temporal relationships between sequences.

LSTM Limitations:

- Requires high processing power, particularly when handling large datasets and long sequences comparing to newer models such as transformers.
- May not function properly with complex language structures and requires careful tuning.

To implement our LSTM in our work, the model first used tokenization to convert text to integers then, into lists of values based on these tokens (sequences). After that to have some level of equality in size, these sequences were passed to pad sequences to be padded. The next step was embedding where similar words were gathered as vectors. Furthermore, two 128-unit LSTM layers that capture long-term dependency were added. And following the first and the second LSTM layers, two dropout layers (50% rate) were applied for regularization. Finally, the output was passed into an output layer for the gender prediction probability.

Figure 1 summarizes the model architecture:

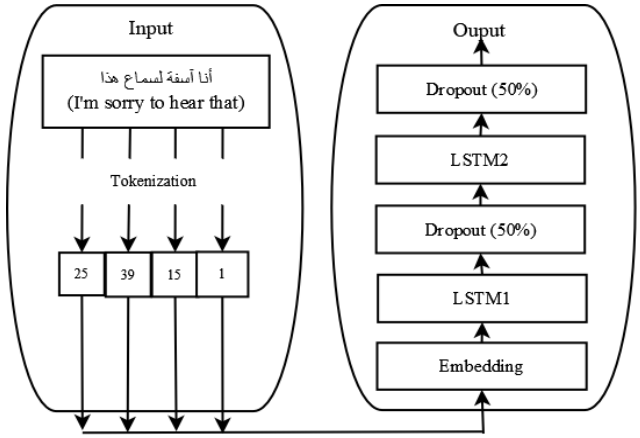

Figure 1. LSTM model architecture

Table 1 summarises the model's compilation and training parameters:

Table 1. LSTM's training parameters

Compilation Parameter	Value
Optimizer	Adam
Loss function	Binary crossentropy
Metrics	Accuracy
Number of epochs	5
Validation split	20%

Note: We used the early stopping function so that when the validation loss stops improving, the training stops to prevent overfitting.

3.2 The second method ARABERT

Arabic BERT [33] a pre-trained monolingual model based on BERT, developed by a team of research at the American University of Beirut in 2020. It is the largest pre-trained language model for processing Arabic language. The strengths and limitations of ARABERT are summarised below.

ARABERT Strengths:

- It understands context effectively through bidirectional context modeling.
- It is designed for Arabic NLP tasks, having been pre-trained on a large Arabic corpus. (70 million MSA and dialectal)

ARABERT Limitations:

- It requires significant computational power to run and train.
- It requires large datasets for good performance.
- Due to its complex architecture, it can be time-consuming and requires higher operational costs.

Like BERT, ARABERT relies on the Transformer model architecture to take advantage of hidden language modelling, which predicts hidden tokens in the input sequence, and the configurations used in ARABERT are the same as BERT-base architecture which contains 12 encoder blocks, 12 attention

heads, 512 sequence length, and approximately 110M parameters. So, for the pre-training phase, ARABERT uses a masked language model where it hides word in the input-sentence then makes the algorithm predict the hidden word based on the context [33]. For the Fine-tuning phase, the embedding of the first token (CLS token) is used. Then, a simple dense layer is added to get the predicted output classes. Figure 2 summarises the steps explained above.

Table 2 summarises the model's compilation and training parameters.

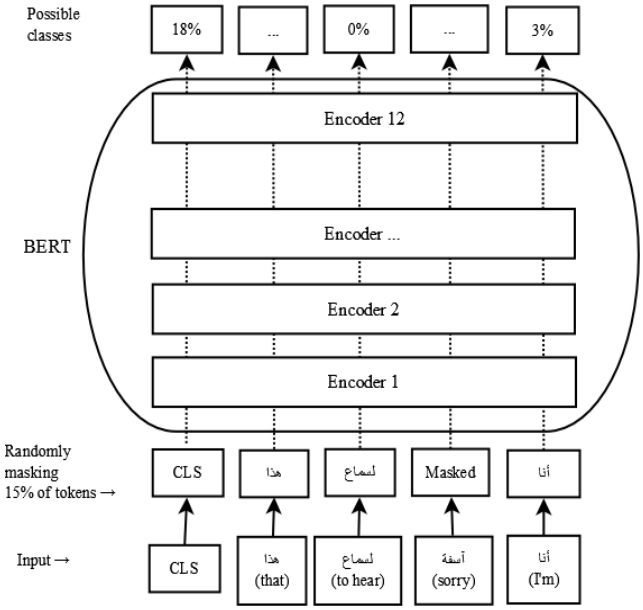

Figure 2. ARABERT model architecture

Table 2. ARABERT's training parameters

Compilation Parameter	Value
The ARABERT model	BERT-base-ARABERTv02-twitter
Optimizer	Adam
Loss function	CrossEntropyLoss
Learning Rate	2e-5
Metrics	Accuracy
Number of epochs	5
Validation split	20%
Seed	25

Note: We used ARABERT instead of a multilingual BERT [8] that supports other languages such as Arabic because it is proven to be more accurate.

3.3 The third method prompt-based learning

The third and last technique was Prompt-Based learning using OpenPrompt [32] a library that provides a framework created in 2021 by a team of researchers, to deploy the Prompt-learning technique.

As mentioned before, Prompt-based learning is a recent technique that involves providing the model with clear instructions or prompts into the input text [9], which can improve usefulness and the accuracy of the output of the model while remaining interpretable.

Model's architecture:

First, to implement our model, the choice of the PLM in our method landed on ARABERT pre-trained language model. Thus, ARABERT tokenizer was used to tokenize texts.

Next, to construct prompt, we designed a textual template that modified the primary input text from our data as follows "Input_Text" this is "Mask". Where, "MASK" is the section where the model needs to analyse and predict the gender.

We then used the PLM (ARABERT) to understand the meaning of the text by analysing the relationships between words, using the tokenized input and template as input.

Furthermore, we constructed the verbalizer using [Male, Female] as label words. The verbalizer took the output of the PLM processing phase and helped in translating the model's predictions task.

Finally, we trained and evaluated the model over multiple epochs.

The strengths and limitations of Prompt based learning are summarised below.

Prompt based Strengths:

- Adaptable and capable of handling various types of tasks.
- Highly efficient, particularly with limited data.
- Easy to use without requiring the model to be whole retrained.

Limitations:

- The prompts' design will determine the success of the models.
- Arabic has had limited research using prompt method, thus it hasn't been tested as much.

Algorithm 1 summarises the general Prompt-Based learning strategy and explains the steps above:

Algorithm 1: *Prompt-based model training*

```
1.  Function predict_gender(text):
2.  /* Preprocessing & Tokenization */
3.  prompt ← combine_tokens(tokenized_text, tokenized_template);
    /* e.g., "Input_Text" this is "Mask" */
4.  model_output ← ARABERT(prompt) ;
5.          /* The choice of PLM */
6.  /* Verbalizer (using a predefined threshold) */
7.  threshold ← 0.5;    /* This value can be adjusted */
8.  if model_output[0] > threshold then
9.  predicted_gender ← "Male";
10. else
11. predicted_gender ← "Female";
12. end if
13. return predicted_gender;
```

Table 3 summarises the model's compilation and training parameters:

Table 3. Prompt-based's training parameters

Compilation Parameter	Value
The PLM model	BERT-base-ARABERTv02-twitter
Template	{"placeholder":"text_a"} this is {"mask"}
verbalizer	[Male, Female]
Optimizer	Adam
Loss function	CrossEntropyLoss
Learning Rate	PLM: 1e-5, Additional model parameters: 1e-4
Number of epochs	5
Validation split	20 %
Seed	25

4. EVALUATION

4.1 Collected data

Due to the lack of gender-labelled Modern Standard Arabic data (MSA) on the web, we have created a new dataset for our gender profiling task.

For that we had to build and collect 10,000 texts of MSA labelled data (Male/Female) from 3 different resources:

- Pan 2018 corpus [34],
- The Arabic Parallel Gender Corpus 2.0 [35] obtained from the English/Arabic Open-Subtitles,
- and the last resource is the result of a work done by the authors of this paper. It is a corpus of opinions collected from university students.

4.2 The first data resource

The first resource that we collected the data from was PAN2018 [34]. Where, PAN is short for (Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection). PAN is a workshop that organizes shared tasks since 2010, highlighting mostly gender and age identification.

The data extracted from PAN 2018 was made of tweets where, Twitter users were labelled according to their gender.

Unfortunately, the Arabic dataset was not completely in MSA, and our task specifically focuses on using MSA. So, we had to translate most of the tweets manually. Finally, we reached almost 1,000 tweets written in MSA. Table 4 explains how the tweets were translated:

Table 4. Examples of tweets translated from dialect to MSA

Dialectal Arabic Tweet	Label	Rewritten Tweet in MSA	English Translation
كيف الحال دكتور، لو فاضي تشيك على المقطع و تعطيني رايك فيه.. اسف على الازعاج	Male	كيف حالك يا دكتور؟ إذا كان لديك وقت فأرجو أن تتفضل بتفقد المقطع وأخبرني برأيك فيه.. اسف على الازعاج	How are you, doctor? If you have time, please take a look at the clip and tell me your opinion on it? Sorry for the inconvenience.
وصلني مسج تم استلام طلبك جهاز رسيفر وسيتم التوصيل خلال يومين وانا مو طالب شي المسج واصله بالخطأ	Male	وصلتني رسالة تم استلام طلبك جهاز رسيفر وسيتم التوصيل خلال يومين وانا لم اطلب شيء الرسالة وصلت بالخطأ	I received a message that your order has been received, a receiver device, and it will be delivered within two days, and I did not order anything. The message arrived by mistake
حتى انا مافهمتش واش قاعد يشوف السيد هذا .... هدف صحيح واين يشوف في الخطا ؟!	Female	حتى انا لم افهم ما يرى هذا الرجل .... هدف صحيح اين رأى الخطأ؟!	Even I did not understand what this man sees.... a right scored goal and where did he see the foul?!
مليت كل يوم نبي انفكر ايش انطيب معاش عندي حتى أفكار جديدة	Female	مللت كل يوم أفكر ماذا اطبخ ليس لدي اي أفكار جديدة	I'm bored every day thinking what to cook I don't have any new ideas

4.3 The Second data resource

The Second data resource was collected from the Parallel corpus for gender identification. This resource involves first and second grammatical persons– I and You. [35] Obtained it from the English Arabic Open Subtitles 2018 dataset [36].

This MSA dataset is full with conversational texts in MSA, first and second person subject [I, we, and you]. It had sentences in which they included: Female talking to Female, Female to Male, Female to anonymous, and group of Females (plural) to Male, etc. and the same for the Male labelled sentences, because they have used the gender mirroring technique. Where, they transform a text originally written with female/male grammar, into its reversed grammatical gender.

Therefore, they assigned a two-letter label to every text in the data, indicating the gender of the first and second person, respectively. For example, FM stands for (female to male). There are four possible possibilities for each letter in the label: F (female), M (male), N (non-existent), which indicates the individual is speaking to no one in particular, or B (invariant or ambiguous). Thus, each sentence will be labelled as one of the 16 possible label combinations: BB, FB, MB, BF, BM, BN, NB, NN, FN, MN, NF, NM, MM, FM, MF, or FF. For more details check [35].

Thus, we tried to re-annotate most of the labels of these texts to only male and female labels, leaving only the texts with the first person "I" as the author and removing all the duplicates as well. For example, if the text was labelled as FB, MB, FN, MN, MM, FM, MF, or FF, we changed it to female or male labels only. However, if the texts were labelled as BM, BF, BN, NB, NN, BB, NF, and NM, we removed them from the corpus.

Finally, we were left with only 8,000 English-Arabic annotated pairs (out of 52,000 total), and the data had already been cleaned. The current state of the data is described in Table 5.

Table 5. The Arabic parallel gender data's outward form after modification

English Translated Text	Original Arabic Text	Label	Text Re-inflection	Re-inflection Label
If I told you something strange, would you think I am crazy	إذا اخبرتك بشيء غريب، هل ستعتقدين أنني مجنون	Male	إذا اخبرتك بشيء غريب، هل ستعتقدين أنني مجنونة	Female
I'm afraid you'll be unlucky their sir	أنا خائف أنك ستكون غير محظوظ هناك سيدي	Male	أنا خائفة أنك ستكون غير محظوظ هناك سيدي	Female
I am a free man and you are a free man	أنني رجل متحرر وأنت رجل متحرر	Male	أنني امرأة متحررة وأنت رجل متحرر	Female
I'm going with you dad	أنا ذاهب معك يا أبي	Male	أنا ذاهبة معك يا أبي	Female

4.4 The third data resource

The third data resource was built for the specific task of gender profiling. We have built this data from a questionnaire. In this dataset, we have collected the opinions and impressions

of university students about their personal experiences in high school or university.

This questionnaire contained six questions in which people were asked to specify their gender to facilitate the labelling process, and each question should be answered in four proper sentences at most. We finally ended up with 1,000 texts.

Unfortunately, the data was not very balanced because females responded to our Google Forms more than males did.

The six questions are shown in Tables 6 and 7, with examples of female and male students' answers.

Table 6. A table showing a female student's answers to the questionnaire

Question in Arabic	Translated Question	Answer in Arabic	Translated Answer
ما هي أهم المواد التي كنت تحب دراستها في مرحلة الثانوية ؟ لماذا ؟	What were the most important subjects that you liked to study in high school? Why?	مادة هندسة الطرائق لأنني كنت جد متعلقة بأستاذتي لهاذا كنت ممتازة فيها واحوز على أفضل النقاط	Process Engineering because I was very attached to my teacher, so I was excellent in it and I got the best points
لخِّص أحسن ذكرياتك في الثانوية التي درست بها.	Summarize your best memories of the high school you attended.	أحسن ذكرياتي في الثانوية هي اوقات الفراغ لأنني كنت سعيدة واستمتع بوقتي مع زميلاتي وأنسى انني كنت حزينة	My best memories in high school are break times because I was happy and I enjoyed my time with my classmates and forget that I was sad
عبّر عن شعورك و أنت تحوز على شهادة البكالوريا.	Express how you felt when you succeeded in the baccalaureate degree.	شعرت كأنني فراشة خفيفة الوزن أكاد اطير وجد سعيدة	I felt like a lightweight butterfly, almost flying, and so happy
هل كان التخصص الذي أرسلت إليه في الجامعة مناسبا لك ؟. لماذا ؟	Was the major you were sent to at the university suitable for you? Why?	نعم انا جد راضية على اختياري فمن خلاله اعمل حرة ولا انشغل عن أطفالي كأنني ربة بيت	Yes, I am very satisfied with my choice, through which I work freely and do not get busy to take care of my children as if I am a housewife
عبّر عن طموحاتك المهنية بعد الجامعة.	Express your career aspirations after college.	من أهم طموحاتي ان أصبح مستقلة ماديا وناجحة في مجال دراستي و اذا حالفني الحظ أصبح أستاذة.	One of my most important ambitions is to become financially independent and successful in my field of study, and if I am lucky, to become a teacher.
عبّر عن مبادرة تمنيت القيام بها في يوم ما.	Express an initiative that you would like to take one day.	من أهمها ان أكون انسانة متطوعة لكل ما فيه الخير لمجتمعي وغيري واساعد بما أستطيع واكون مشاركة ومساهمة إلخ.	One of the most important is to become a volunteer for all that is good for my community and others and to help as I can and to be a participant and a contributor, etc.

Table 7. A table showing a male student's answers to the questionnaire

Question in Arabic	Translated Question	Answer in Arabic	Translated Answer
ما هي أهم المواد التي كنت تحب دراستها في مرحلة الثانوية ؟ لماذا ؟	What were the most important subjects that you liked to study in high school? Why?	العلوم الطبيعية و الفيزياء: لأنني أحب علوم المادة / الأدب العربي و اللغة الإنجليزية: لأنني أحب الكتابة و الشعر.	Natural sciences and physics: because I love material sciences / Arabic literature and the English language: because I love writing and poetry.
لخِّص أحسن ذكرياتك في الثانوية التي درست بها.	Summarize your best memories of the high school you attended.	أحسن الذكريات كانت في التجمع و الصحبة مع الأصدقاء و الزملاء.	The best memories were gathering and company with friends and colleagues.
عبّر عن شعورك و أنت تحوز على شهادة البكالوريا.	Express how you felt when you succeeded in the baccalaureate degree.	عدم الرضى نوعا ما ، لعدم حصولي على المعدل الذي أردته.	Somewhat dissatisfied for not getting the grade that I wanted.
هل كان التخصص الذي أرسلت إليه في الجامعة مناسبا لك ؟. لماذا ؟	Was the major you were sent to at the university suitable for you? Why?	نعم ، لأنني مهتم بالبرمجة و الحاسوب منذ نعومة أظافري.	Yes, because I have been interested in programming and computers since my childhood.
عبّر عن طموحاتك المهنية بعد الجامعة.	Express your career aspirations after college.	العمل كمهندس برمجيات مستقل و إنشاء شركتي الخاصة.	Working as a freelance software engineer and starting my own company.
عبّر عن مبادرة تمنيت القيام بها في يوم ما.	Express an initiative that you would like to take one day.	تكوين فريق عمل من مبرمجين مختصين في مجلات مختلفة لتغطية مساحة عمل واسعة.	Forming a team of programmers specialized in different fields to cover a wide work area.

Furthermore, we balanced the overall corpus and we ended up with a merged corpus of 10,000 out of 13,000 from 3 different resources.

We then, decided to carry out the evaluation (comparison of the three methods) on a corpus of progressive size. Making 3 or 4 corpora of progressive size for example: 2,500, 5,000, 7,500, 10,000 to make the comparison on these corpora.

The idea was to see if the size of the corpus influences the performance so that each time we enlarge the data, we check if the results improve. The same pre-processing technique of the data was used in all models. We also wrote a program to alternatively put a text-Male, then a text-Female, from the different collected resources (Pan, Open Subtitle and Google Forms questionnaire). So, we dispensed with the use of the function responsible for the shuffling because our program now will ensure that each time, we take a sub-corpus (2,500, 5,000, 7,500, 10,000) it will also be shuffled in a balanced way and the Training (80 %) and Test (20%) parts will be well divided and balanced as well.

Table 8. The statistics of the used dataset

Data	Language	Corpus Size	Male	Female
PAN	MSA	1,000	500	500
Open Subtitle corpus	MSA	8,000	4,000	4,000
Google Forms	MSA	1,000	500	500
The overall corpus (PAN+ Subtitle corpus+ Google Forms)	MSA	10,000	5,000	5,000

Table 8 presents the statistics and distribution of the dataset used.

4.5 Performance measures

To compare the performance of the three different methods, we used the accuracy measure. The accuracy is a performance measure that shows how often the model predicts correctly. It is defined as the proportion of the number of correct predictions to the total number of the predictions generated by the model. Accuracy = (number of correct predictions) / (total number of predictions).

Thus, this measure allowed us to evaluate the performance of the three methods and compare them based on how precise their predictions and their ability to approach the target.

5. RESULTS AND DISCUSSION

After 5 epochs, we compared the performance of three methods (LSTM, Prompt-based, and ARABERT) on four batches of data sets with different corpus sizes using accuracy as a performance measure.

The overall measurement results are summarized in Table 9 and shown graphically in Figure 3.

Table 9. Performance of the three models in terms of Accuracy

Corpus Size	DNN Accuracy	Prompt Based Accuracy	ARABERT Accuracy
2,500	55.7%	84.0%	84.6%
5,000	68.0%	89.3%	91.8%
7,500	72.3%	93.0%	92.6%
10,000	78.5%	92.3%	92.4%

From Table 9 we can see that as the corpus size increases, the accuracy of the three methods also increases. This is a logical result for learning-based models in general: for those models, better performance is often associated with larger training data sets.

Also, from the table, we can notice that prompt-based learning is very competitive compared to ARABERT, with accuracy increasing from 84% to 92.3% and from 84.6% to 92.4%, respectively. While the LSTM lags behind with the lowest accuracy, from 55.7% to 78.5%.

Even though, ARABERT slightly outperforms the Prompt-based model, it maintains the lead over all epochs. Leaving the LSTM model significantly behind.

This emphasizes the limitations of traditional DL models for gender classification compared to PLMs, which are more suitable for this task, particularly with limited training data. Because the PLMs are already trained on large amounts of semi-supervised texts, such as Wikipedia texts or online books, which allows them to train and build many more models for different supervised learning tasks.

The results are also illustrated in Figure 3.

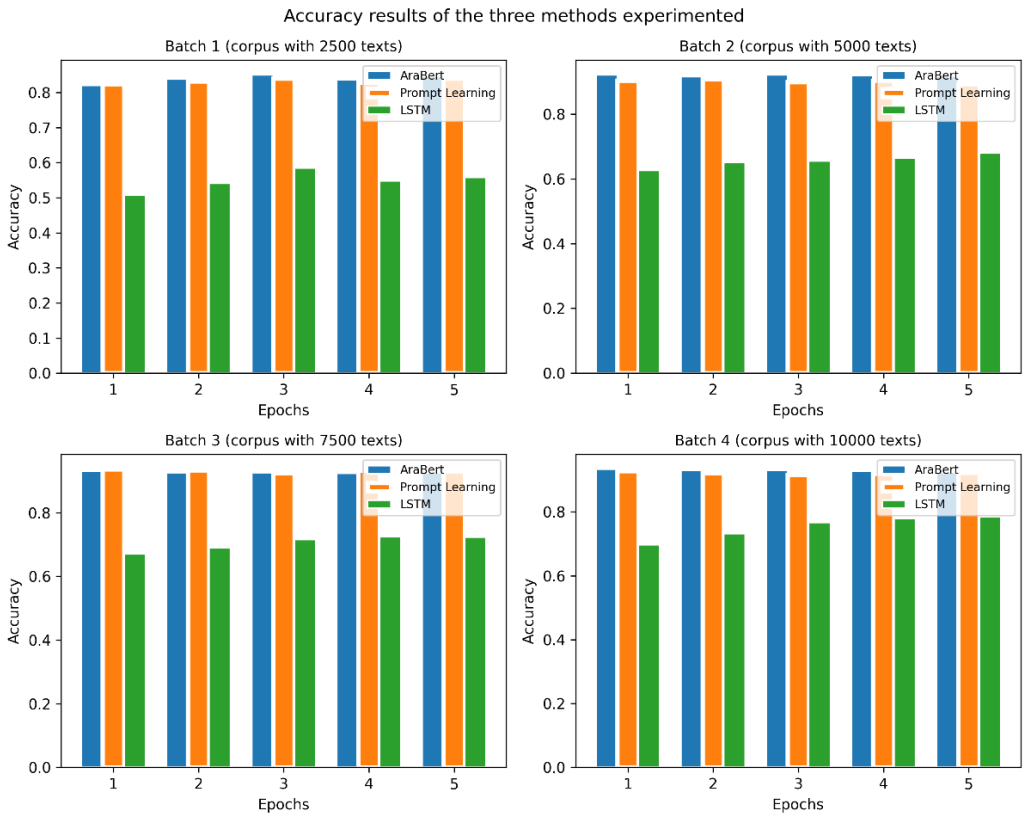

Figure 3. Accuracy results of the three methods experimented

From the chart, we can see that ARABERT outperforms the other methods and shows better results each time the data increases. Its accuracy keeps improving until it reaches an impressive 92.4% on the largest batch size. This demonstrates that ARABERT is the most effective method for gender author profiling in MSA.

Prompt-based learning performs steadily and starts remotely close to ARABERT throughout all batches and epochs and achieves almost a similar impressive accuracy of 92.3%. While, LSTM also shows progress across all batches, it still consistently performs worse than the other two models and reaches an accuracy of only 78.5%.

This gap in results between LSTM and the other two models could be due to the limited training data. Because the PLMs are pre-trained on large amounts of semi-supervised texts, therefore, they can be adapted to solve specific tasks using smaller datasets.

To sum up, the chart clearly compares the three approaches across all batches and epochs, emphasizing ARABERT's consistently superior performance. However, throughout all batches and epochs, there is barely a difference in accuracy between ARABERT and Prompt-based learning.

Overall, the chart indicates that ARABERT gives the highest accuracy, exceeding 90%. We can deduce that ARABERT can be effective for Arabic gender identification task. Additionally, the results of prompt-based learning suggest a potential for improvement with more carefully designed templates and verbalizers for this model.

Furthermore, investigating this problem using other PLMs such as GPT in the Prompt-based model could significantly enhance the results.

6. LIMITATIONS AND FUTURE WORK

This study encounters challenges related to Arabic, such as linguistic ambiguities, cultural differences in gender representation, and the difficulty of gathering a large, diverse labelled corpus.

In order to address these challenges, our next study using Prompt-based learning will be performed on other dedicated models to Arabic, such as ARAGPT that can probably better handle linguistic ambiguities.

Furthermore, we suggest enlarging the current data by splitting the long paragraphs using splitting techniques such as words referring to the end of a sentence, like conjunctions or adverbial phrases and collect more diverse datasets.

7. CONCLUSIONS

The present study was designed to determine whether the recent paradigms of NLP, such as LSTM, ARABERT, and Prompt-based learning are effective in solving the gender profiling task in MSA. The findings of this research suggest that even with the success and outperformance of ARABERT with an accuracy of 92.4%, Prompt-based learning showed competitiveness and performed almost similarly to ARABERT with an accuracy of 92.3%. These promising results show the advantages of PLMs, which are better suited for this task, particularly with limited training data, over traditional DL models such as LSTM. Also, this suggests the potential for improvement in Prompt-based with more carefully designed templates and verbalizers.

REFERENCES

[1] Alsmearat, K., Shehab, M., Al-Ayyoub, M., Al-Shalabi, R., Kanaan, G. (2015). Emotion analysis of Arabic articles and its impact on identifying the author's gender. In 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech, Morocco, pp. 1-6. https://doi.org/10.1109/AICCSA.2015.7507196

[2] Tetreault, J., Blanchard, D., Cahill, A. (2013). A report on the first native language identification shared task. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 48-57.

[3] Horan, C., Saiedian, H. (2021). Cyber crime investigation: Landscape, challenges, and future research directions. Journal of Cybersecurity and Privacy, 1(4): 580-596. https://doi.org/10.3390/jcp1040029

[4] Sotelo, A.F., Gómez-Adorno, H., Esquivel-Flores, O., Bel-Enguix, G. (2020). Gender identification in social media using transfer learning. Pattern Recognition, pp. 293-303. https://doi.org/10.1007/978-3-030-49076-8_28

[5] Mansour Khoudja, A., Loukam, M., Belkredim, F.Z. (2021). Towards author profiling from modern standard Arabic texts: A review. In Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, pp. 745-753. https://doi.org/10.1007/978-981-16-2377-6_69

[6] Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D.F., Chao, L.S. (2019). Learning deep transformer models for machine translation. arXiv preprint arXiv:1906.01787. https://doi.org/10.48550/arXiv.1906.01787

[7] Zhang, H., Song, H., Li, S., Zhou, M., Song, D. (2023). A survey of controllable text generation using transformer-based pre-trained language models. ACM Computing Surveys, 56(3): 1-37. https://doi.org/10.1145/3617680

[8] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[9] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9): 1-35. https://doi.org/10.1145/3560815

[10] Mayer, C.W., Ludwig, S., Brandt, S. (2023). Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models. Journal of Research on Technology in Education, 55(1): 125-141. https://doi.org/10.1080/15391523.2022.2142872

[11] Madotto, A., Lin, Z., Winata, G.I., Fung, P. (2021). Few-shot bot: Prompt-based learning for dialogue systems. arXiv preprint arXiv:2110.08118. https://doi.org/10.48550/arXiv.2110.08118

[12] Mamgain, S., Balabantaray, R.C., Das, A.K. (2019). Author profiling: Prediction of gender and language variety from document. In 2019 International Conference on Information Technology (ICIT), Bhubaneswar, India, pp. 473-477. https://doi.org/10.1109/ICIT48102.2019.00089

[13] Ameer, I., Sidorov, G., Nawab, R.M.A. (2019). Author profiling for age and gender using combinations of features of various types. Journal of Intelligent & Fuzzy

Systems, 36(5): 4833-4843. https://doi.org/10.3233/JIFS-179031
[14] Mechti, S., Jaoua, M., Belguith, L.H., Faiz, R. (2013). Author profiling using style-based features. Notebook Papers of CLEF2. https://downloads.webis.de/pan/publications/papers/mechti_2013.pdf.
[15] López-Monroy, A.P., Montes-y-Gómez, M., Escalante, H.J., Pineda, L.V. (2014). Using intra-profile information for author profiling. In CLEF (Working Notes), pp. 1116-1120. https://downloads.webis.de/pan/publications/papers/lopezmonroy_2014.pdf.
[16] Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., Ohkuma, T. (2018). Text and Image Synergy with Feature Cross Technique for Gender Identification: Notebook for PAN at CLEF 2018. In CLEF (Working Notes). https://pdfs.semanticscholar.org/c6b3/9cf3d580d810c951a56150a4568b6d9070a1.pdf.
[17] Suman, C., Chaudhary, R.S., Saha, S., Bhattacharyya, P. (2022). An attention based multi-modal gender identification system for social media users. Multimedia Tools and Applications, 81: 27033–27055 https://doi.org/10.1007/s11042-021-11256-6
[18] Ouni, S., Fkih, F., Omri, M.N. (2023). Novel semantic and statistic features-based author profiling approach. Journal of Ambient Intelligence and Humanized Computing, 14(9): 12807-12823. https://doi.org/10.1007/s12652-022-04198-w
[19] Bacciu, A., La Morgia, M., Mei, A., Nemmi, E.N., Neri, V., Stefa, J. (2019). Bot and gender detection of twitter accounts using distortion and LSA. In CLEF (Working Notes). https://www.researchgate.net/profile/Massimo-La-Morgia/publication/334591709_Bot_and_Gender_Detection_of_Twitter_Accounts_Using_Distortion_and_LSA_Notebook_for_PAN_at_CLEF_2019/links/5d332f9b92851cd0467640fe/Bot-and-Gender-Detection-of-Twitter-Accounts-Using-Distortion-and-LSA-Notebook-for-PAN-at-CLEF-2019.pdf.
[20] Ikae, C., Savoy, J. (2022). Gender identification on Twitter. Journal of the Association for Information Science and Technology, 73(1): 58-69. https://doi.org/10.1002/asi.24541
[21] AlSukhni, E., Alequr, Q. (2016). Investigating the use of machine learning algorithms in detecting gender of the arabic tweet author. International Journal of Advanced Computer Science and Applications, 7(7): 319-328.
[22] 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
[23] Nayel, H.A. (2020). A new approach for author profiling and identification of deception in texts. Journal of Computational Linguistics, 11(2): 73-79. https://doi.org/10.6025/jcl/2020/11/2/73-79
[24] Mubarak, H., Chowdhury, S.A., Alam, F. (2022). Arabgend: Gender analysis and inference on arabic twitter. arXiv preprint arXiv: 2203.00271. https://doi.org/10.48550/arXiv.2203.00271
[25] Zhang, C., Abdul-Mageed, M. (2019). BERT-based Arabic social media author profiling. arXiv preprint arXiv: 1909.04181. https://doi.org/10.48550/arXiv.1909.04181
[26] Alzahrani, E., Jololian, L. (2021). How different text-preprocessing techniques using the bert model affect the gender profiling of authors. arXiv preprint arXiv:2109.13890. https://doi.org/10.5121/csit.2021.111501
[27] Subramanian, V. (2018). Deep Learning with PyTorch: A practical approach to building neural network models using PyTorch. Packt Publishing Ltd.
[28] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781. https://doi.org/10.48550/arXiv.1301.3781
[29] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1706.03762
[30] Floridi, L., Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30: 681-694. https://doi.org/10.1007/s11023-020-09548-1
[31] Zhang, M., Li, J. (2021). A commentary of GPT-3 in MIT technology review 2021. Fundamental Research, 1(6): 831-833. https://doi.org/10.1016/j.fmre.2021.11.011
[32] Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H. T., Sun, M. (2021). Openprompt: An open-source framework for prompt-learning. arXiv preprint arXiv:2111.01998. https://doi.org/10.48550/arXiv.2111.01998
[33] Antoun, W., Baly, F., Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104. https://doi.org/10.48550/arXiv.2003.00104
[34] Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B. (2018). Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. Working Notes Papers of the CLEF. https://downloads.webis.de/pan/publications/papers/rangel_2018.pdf.
[35] Alhafni, B., Habash, N., Bouamor, H. (2021). The Arabic parallel gender corpus 2.0: Extensions and analyses. arXiv preprint arXiv:2110.09216. https://doi.org/10.48550/arXiv.2110.09216
[36] Lison, P., Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. Vis Innførsel, 923-929. http://urn.nb.no/URN:NBN:no-54046