Discover Artificial Intelligence

Research

ExaAUAC: Arabic Twitter user age prediction corpus based on language and metadata features

Reyhaneh Sadeghi[1] · **Ahmad Akbari**[1] · **Mohammad Mehdi Jaziriyan**[1]

Received: 23 January 2024 / Accepted: 12 June 2024
Published online: 08 July 2024

Abstract
Twitter is a rich resource for analyzing the contents of social media and extracting the age groups of users can be beneficial for recommender systems, marketing and advertising. Age detection task is an aspect of demographic information of users. In this study a large-scale corpus of Arabic Twitter users including 181k user profiles with diverse age groups consisting of −18, 18–24, 25–34, 35–49, 50–64, +65 is presented. The corpus is created by four methods: (1) collecting publicly available birthday announcement tweets using the Twitter Search application programming interface, (2) augmenting data, (3) fetching verified accounts, and (4) manual annotation. To have a best age detection model on the presented corpus, different evaluations are tested to find the model with highest accuracy and efficiency. Number of tweets, regression vs. classification, using metadata of users and tweets, using LSTM+CNN model vs. BERT are some parts of examinations done. Presented methodology is based on language and metadata features and final model is fine-tuned with BERT on 70k users and evaluated on 8200 manually annotated users. We show that our best model, compared with LSTM+CNN model and BERT-based similar model yields an improvement of up to 9% in F1-score and increment of 5% in accuracy, respectively. The model achieved macro-averaged F1-score of 44 on six age groups, and F1-score of 58 on three age groups of −25, 25–34, +35. The link of our proposed data is provided here: www.github.com/exaco/ExaAUAC.

Keywords Author profiling · Twitter age extraction · Age detection

1 Introduction

With the increasing access to the internet, social media (Twitter,[1] Instagram,...) have become a part of everyday life; people of different ages use these platforms, and their usage is on the rise. Twitter especially is a huge source of textual content and information that people share their daily life events and opinions on certain topics and discuss their interests. However, there are some benefits and drawbacks associated with social media usage among all age groups, especially for those in their younger ages. There are benefits such as access to recent news and information, and drawbacks such as some potentially inappropriate content for younger users.

[1] At a time of doing this research Twitter was not transformed to X. So in this paper we use the term "Twitter".

✉ Reyhaneh Sadeghi, r.sadeghi@exalab.co; Ahmad Akbari, ahmadakbari@ut.ac.ir; Mohammad Mehdi Jaziriyan, m_jaziryan@modares.ac.ir | [1]Exa Company, Tehran, Iran.

Analyzing the contents of social media and extracting the age of users can be beneficial for multiple applications such as demographic information of users, interests of the age groups, recommender systems, marketing and advertising, etc. The advent of machine learning-based and deep learning-based methods helped tremendously and gave the pace to machines to learn the multiple varieties of sociolinguistic features and find patterns on the data and tag millions of users that can take a long time for humans to assess social media profiles.

Author profiling is the task of determining the characteristics of a profile on social media based on their network, textual and linguistic features expressed on their profiles. Age detection Schler et al. [20] is one of the important characteristics of a profile, besides gender Schler et al. [20], education Juola et al. [8], personality Pennebaker et al. [14] and etc, where the age of the user is determined based on text attributes. Many PAN competitions were done on author profiling of twitter users between the years of 2013–2019 for many different languages (English, Arabic, Spanish,...) Rangel et al. [16]. The shared task of Arabic Author Profiling and Deception Detection (APDA) Rangel et al. [18], Zhang et al. [22] was done by PAN organizers to profile age, gender and language of Arabic Twitter users. Many participants participated, and among the results, multiple machine learning-based and deep learning-based works were done, and machine learning in those tasks achieved the best results. In one hand, Author profiling in the Arabic language is rarely researched, and few papers are proposed despite 27 million daily tweets in Arabic. On the other hand, Arabic is a very complex language with many different dialects in many countries that make processing the linguistic feature difficult. Recently, BERT-based language models Devlin et al. [6] made a substantial improvement in natural language processing and natural language understanding tasks, which learn complex linguistic features from text.

In this work we propose a new corpus on Arabic Twitter user profiles for age detection task, and use language models to evaluate the proposed corpus. Our contribution to this work are as follows:

- Proposing a new corpus on Arabic Twitter user profiles for age detection task,
- Properly utilizing textual content besides metadata features to make model learn the content and achieve F1-score of 44 on six age group classes, and 58 on three age group classes

In the following section, related works in age classification task is presented; Sect. 3, expresses our proposed corpus, data gathering methods and tagger agreements. Section 4 presents experiments and the suggested model, and in Sect. 5, the result of the model is discussed, and finally in the last section, we conclude our work.

2 Related works

Numerous studies have collected datasets through user self-reporting and have conducted research on these data. Authors of the article Zhang et al. [22], used self-reported age of 2970 Arabic Twitter users and in a similar way Antonio et al. [10], gathered 3184 Twitter users. In the article by Chamberlain [4], the authors examined Twitter users who communicated in English, Spanish, French, and Portuguese. Employing regular expressions to scrutinize user information, they identified 133,000 users with a certain age. In addition, this article has extracted the graph of people's relationships. In Pandya et al. [13], three datasets were employed, with one of them being in Dutch, consisting of 2150 active users. Similarly, in Pandya et al. [12], the dataset for Dutch language users comprises 1365 users, while the dataset for English language users includes 2745 users. Additionally, two English datasets, comprising 1794 and 1074 users, were also utilized in this article. Culotta et al. [5] extracted the data of 1532 websites and reporting the demographic information of their users, such as the age group, gender, income, education, race and political orientation of the people. Then, they checked the information and found 1066 Twitter accounts corresponding to the above pages and tagged the users who followed these accounts with the obtained distribution. Finally, 46649 users with more than 100 followers were extracted for final usage and 9 M tweets were extracted from the followers of mentioned users. Authors of the article Klein et al. [9], developed a model to detect the age of Twitter users by using 1000 tagged accounts. The age of 1000 Twitter users was verified by the taggers according to their self-reported age.

The Chamberlain [4], presents a text-independent method based on the accounts they follow to predict the age of Twitter users. Antonio et al. [10] trained six statistical models, and logistic regression yielded the highest accuracy in detecting the age of users. Best output obtained by using the linguistic information of tweets and user metadata. In addition to the usual features in metadata and texts for age detection, hashtag content and URLs are also exploited in age detection Pandya et al.

Table 1 Summary of related works

Dataset	Classes	Count	Verified	Acc/F1	Measurement	Language
Klein et al. [9]	Age number	1000	Yes	91	F1	EN
Antonio et al. [10]	13–17 18–25 25+	3184	No	74	Accuracy	EN
Pandya et al. [12]	−18 18–30 30–40 40+	1066	Yes	61	Accuracy	EN
Pandya et al. [13]	−17 18–40 40+	1794 1074	No	81 86	F1	EN EN
Pandya et al. [13]	−20 20–40 40+	2150	No	82	F1	De
Zhang et al. [22]	−25 25–34 35+	2970	No	54.72	Accuracy	AR
Chamberlain [4]	10 classes 3 classes	133000	No	$\frac{31}{86}$	Accuracy	Multi lingual
Mubarak et al. [11]	2	6000	No	94.4	Accuracy	AR

[13]. The content of the URLs and hashtags helps to detect the interest of the user, and the detection of interest also helps to detect the age. So, it can be a substantial feature. The methodology employed in this article involves identifying 1000 tweets within approximately a 10-day timeframe surrounding the user's tweet. Subsequently, it extracts the words with the highest co-occurrences in these tweets, utilizing them as representatives of the corresponding hashtag. This article has used the CNN method and obtained promising accuracies in the Dutch and English datasets. In the article Pandya et al. [12], a CNN model has also been used to detect age, and in addition to the existing tweets and metadata, a linguistic model has also been used to understand hashtags and URLs to be used in the model. various methods such as SVM, MultiBERT, and AraBERT were examined and tested in Mubarak et al. [11], to predict the age of users and obtained the best result with *username+user description+tweet inputs* In Zhang et al. [22] They presented a two-layer method based on BERT. In this method, the user's age, gender, and accent are tagged at the tweet-level first, and then the final tag is determined at the user-level by aggregating the obtained tags.

Model reported in Klein et al. [9] recognizes the user's age with the F1-score of 91 if the user has reported his/her age. Antonio et al. [10] has achieved an accuracy of 74. The goal of Mubarak et al. [11] is to identify individuals who share adult-related content on Twitter. In pursuit of this goal, they introduced a two-class classification model with accuracy of 94.4.in this Article, a dataset comprising 6000 tweets was manually labeled into mature and non-mature classes. Additionally, tweets not associated with individuals generating mature content were categorized as non-mature.The results of Culotta et al. [5] claimed that the user's demographic information can be obtained with 73% accuracy only by using the user's following list. This number is reported to be 79% using text, and by using text and users' followers, it is possible to reach 81% accuracy in detecting the user's demographic information. The accuracy of the presented model for detecting the age of users is 61%. The age classes in Zhang et al. [22] are (under 25, between 25 and 34, and over 35). This article has achieved 54.72 accuracy on its data. 10 age classes are used in Chamberlain [4] and the length of the intervals between each class is determined depending on the frequency of that age group. this Article has achieved 31% accuracy on the test data. However, when evaluated on the 3-class dataset, the model obtained an accuracy of 86%.

The number of classes and the selected range for each class have significantly impacted the accuracy of the models presented in the studied articles. Studies that have categorized age into only two classes have achieved an accuracy of 94 [11], whereas those utilizing ten classes for age detection have only reached an accuracy of 31 Chamberlain [4]. Table 1 provides a summary of the existing proposed research datasets. Given the limited size of existing datasets and the necessity for larger datasets to enhance accuracy in age detection, we opted to systematically gather and annotate the age of Twitter users. With this initiative, we aim to provide a comprehensive dataset that can be used to accurately identify the age of Arab Twitter users.

Table 2 Corpus statistics

	Self-report	Augment under 18	Verified accounts	Manual annotation
Train	62k	6k	1000	1100
Test	–	–	–	8200
Total	165k	6k	1000	9300

Table 3 Distribution of train and test data

	−18	18–24	25–34	35–49	50–64	+65	Total
Train	10k	15.3k	17.9k	17.2k	9k	664	70k
Test	743	2684	3502	1163	165	29	8200

3 The corpus

In this section, we present corpus statistics, then go through different methods applied for data gathering and creating the corpus.

3.1 Corpus statistics

In this research, we use the PAN 2014 Rangel et al. [17] age tagging scheme consisting of 18–24, 25–34, 34–49, 50–65, and +65 classes. We also add -18 class to make the corpus include all varieties of different age classes. These classes are finer-grained because they include multiple age groups, such as teenagers, young adults, adults, middle age adults and old-age adults. As age increases, a person's concerns, responsibilities and needs change, which will be reflected on the person's behavior, speeches and writings. So we chose these classes to properly model the differences in all age groups with similar concerns and interests.

Presented corpus is called ExaAUAC (Exa Arabic User Age detection Corpus). The primary advantage of ExaAUAC is the large number of labeled accounts. The distribution has a maximum of 50 tweets for each user, and so each tweet is distributed with a corresponding user id. As such, in total, the distributed training data has 70k users. The official task test set contains 8200 users. There are 6 Classes of labels according to most used classes among related articles: −18, 18–24, 25–34, 35–49, 50–64, +65. The age of users is calculated based on the year 2021, and are considered in the relevant classes of labels. Also, Twitter API v1.1 was used in 2021 to gather the user accounts.

Table 2 presents number of total users were extracted from different methods and the distribution of train and test data from gathering resources. The label distribution of train and test data is mentioned in Table 3. To observe data balance in classes of labels, we used duplicated users in the class of 50–64.

3.2 Data gathering methods

We employed four independent methods to collect data for the age detection corpus.

3.2.1 Data gathering via self-report

Unfortunately, Twitter API does not provide a way to access the age of profiles. To solve this problem, we implemented a web crawler using the selenium package to extract the age of profiles depicted on their pages where they reported themselves. Among 10 M Arabic profiles, 1.65 percent of them had reported their age. After extracting the age and screen

name of the profile, their 50 recent tweets were extracted using the twitter API. Result in totally 165k users with their ages from this method were presented, and 62k users of them were used as train data.

3.2.2 Augment data of under 18

To achieve data balance in 6 classes of the corpus, data in class of −18 needed to be augmented. we used labeled data from other classes in the corpus and searched if there were any tweets from the user account when he/she was under 18. For example we have an account with a reported age of 32 in year of 2021. we searched tweets from this user that are at least for 15 years ago to ensure the extracted tweets were made during the age of under 18. we extract these tweets and report them for a new account with the age of under 18. With this manner, we extracted 6000 accounts and it was added to the corpus to make it balanced.

3.2.3 Data gathering by verified account

The other method that we used for data gathering is using information from verified twitter accounts. Due to the existence of information about celebrities or popular people on Wikipedia, we extract 14k unique verified accounts from twitter and use them to extract the birth date of them from related Wikipedia pages automatically. By this method only the birth date of 1800 users was available on Wikipedia. Afterward we pass these accounts from a bot detector to filter human results. Finally 1000 accounts were extracted in this manner.

3.2.4 Manual annotation

In this study, the guideline for annotators is proposed, which focuses on identifying age indicators within a user's tweets and bio. Annotators trained to identify explicit mentions of birthdays, along with social activities that are often age-specific. For example, for younger adults, this includes events like starting university, taking an exam, and etc. which typically occurs in a defined age range. Additionally, the guideline encompass identifying age clues within the user's bio, capturing any self-reported age information they may choose to share. In instances where age cannot be conclusively determined from the available data, annotators are instructed to use a "Cannot be specified" tag. By integrating these strategies, our guideline aims to enhance the precision of age tagging in social media research, thereby improving the reliability of demographic analyses derived from Twitter data.

Out of 31k annotated accounts by two Arabic annotators, 22k tags were common among two annotators. From common tags, 7k were unknown, 5500 were bot and just 9300 accounts were usable with their ages.

3.3 Tagger agreements of manually annotation method

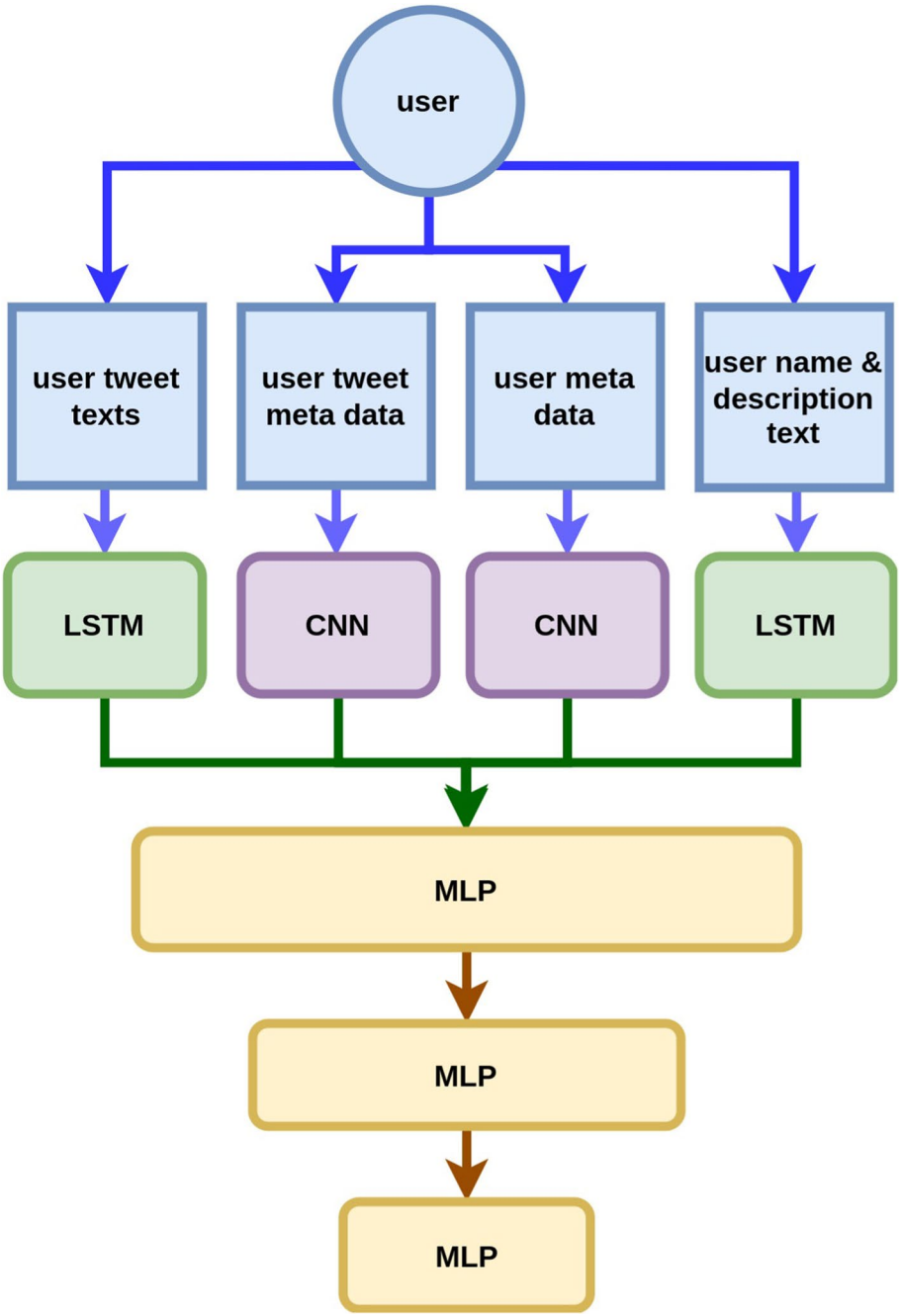
From 31k annotated accounts, each account was annotated twice randomly by two of five Arabic annotators. For the computation of inter-annotator reliability, we removed accounts with only one tag, so we reached tagger agreement of κ=0.6449 (substantial agreement) Fleiss Kappa Fleiss [7] with a "Cannot be Specified" label, and without that, we reached kappa agreement of κ=0.7345 which highlights the complexity of assigning age tags to profiles through manual human annotations.

To assess the annotators, first, annotators were tested on 100 profiles and five best annotators with the highest accuracy selected. After choosing the proper annotators, we hold bi-weekly sessions to observe their differences in tagging and mistakes to help them understand what we expect and consequently increase the data quality.

4 Experiments

Proposing a new corpus needs some models for evaluating and assessing the corpus on them, so we constructed two models on age detection task: (1) LSTM + CNN model (2) BERT-based model. Here we first review the LSTM + CNN model and then explain the proposed model with BERT.

Fig. 1 LSTM+CNN model



4.1 LSTM + CNN model

This model is constructed with a combination of LSTMs and CNNs. The architecture of model is shown in Fig. 1. There are four different feature vectors that model's inputs include: user tweet texts, tweet metadata, user metadata and username and description text. Description of each feature is explained below:

- User tweet texts: In this feature vector, we concatenate all the tweet texts of each user and then utilize ELMO Peters et al. [15] model to vectorize it. Additionally, we append one more feature to the ELMO-processed vector named as *retweeted or not*, therefore, its shape for one user become: (1,2000,1025)[2]

[2] 2000 means that we just used 2000 words for every 200 tweets of a user.

Table 4 Tweet metadata features

Tweet metadata	
Time	The time of post sharing
hashtag_num	Number of hashtag in text
mention_num	Number of mention in text
retweet_count	Number of retweet
favorite_count	Number of favorite
count_emoji	Number of emoji in text
cout_sw	Number of stop words in text
sent_len	Length of text (split by word)
Media	Post has media or not
reBarfo	Retweet/follower
reBarfr	Retweet/friends
foBarfa	Follower/favorite
stBarfa	Status/favorite
frBarfa	Friends/favorite
tweet_ratio	Mean of tweet per days

Table 5 User metadata features

User metadata	
sent_len_des	Length of description text
cout_sw_des	Number of stop words in description text
count_emoji_des	Number of emoji in description text
sent_len_name	Length of name
followers_count	Number of follower
friends_count	Number of friends
favourites_count	Number of favorites
statuses_count	Number of total post
Red	The code of red in RGB
Green	The code of green in RGB
Blue	The code of blue in RGB
profile_image_url	User has profile image or not
profile_banner_url	User has header profile image or not
foBarfr	Follower/friends
foBarfa	Follower/favorites
foBarst	Follower/status
frBarfa	Friends/favorites
frBarst	Friends/status
faBarst	Favorites/status

- Tweet metadata: In this vector we have 15 different features that extract from each tweet. All of these features are depicted in Table 4.
- User metadata: This vector consists of 19 different features that are extracted from each user. All of these features are presented in Table 5.
- Username and description text: In this feature vector we append description and username and then vectorized it using ELMO and then average it on axis = 0, therefore, its shape for one user become: (1,1024).

Table 6 Ablation study

	Classification layer	F1-score (%)
User meta	Linear	40
Tweet meta + user meta	Linear	41
User meta	MLP	43
User meta(only image detection result)	MLP	43.7

4.2 Proposed model

To achieve the best model, we experiment with different cases such as number of tweets per user, classification versus regression model, feeding chunks of tweets or single tweet per user and Using Bio, wordCloud and user metadata. The core model in these experiments is BERT that is fine-tuned on presented corpus.

- Number of tweets per user: In order to choose the best number of tweets per user, we did some experiments by adjusting different numbers of tweets per user and fed it to the model. We did it with 50 and 100 tweets per user, the results showed the same F1-score. So to decrease the time of training procedure and increase the speed of inference we choose the last 50 tweets per user.
- Regression vs. classification: There are two models for implementation: regression and classification with six classes. During training and inference, the regression model employs a regression approach. Due to the different evaluation methods used for regression and classification models, we map the output of the regression model's inference to one of the six classes for comparison in the test phase. Experiments showed that the classification method achieved a 4% increase in F1-score compared to the regression method.
- Single tweet vs. chunk of tweets per user: There are two cases: single tweet per user labeled by its age and fed it to the model and the other, make chunks of tweets per user with its label. Due to limitation of sequence length of BERT with 512 tokens, we used the maximum capacity of BERT and concatenated the number of tweets that can be fit to 512 tokens and made this as one chunk. Between each tweet, the special character of separator of BERT is used. Finally we can have 4–5 chunks per user consisting of 50 tweets with its label of user age and feed it to the model. BERT was fine-tuned in these 2 cases and the case with chunks had the best result with 5% improvement in F1-score.
- Using Bio and wordCloud: According to the article Mubarak et al. [11], user description is critical information that benefits machine learning with more accuracy. It's shown that combining metadata with linguistics has the best throughput than either approach alone.

  We used user description at the beginning of every chunk. Furthermore, we added 30 most common words among up to last 50 tweets per user. This could make the model learn some common words that have important information from related age categories, as this article Culotta et al. [5] says every age category uses some words due to their favorites in tweets. So this could help the model to detect favorites and most common words of literature specific to every age group.
- Using user metadata: To experiment use cases of adding tweet metadata or user metadata we consider different cases as shown on Table 6. Our criterion is a model without using any metadata and the results were compared to that model. In this case chunks of tweets texts fed to the BERT and output of BERT model is concat with user or tweet metadata values and pass from a classification layer.

  Two different classification layers including linear and MLP (Multilayer perceptron) were experimented. User metadata and tweet metadata features that are used in this case were proposed in the LSTM+CNN model Sect. 4.1. Last experiment in this part used only the output of image detection of user profile image. Image detection part using OpenCV Bradski [3] Deep Learning models.

  It is evident from the results, using metadata does not have any significant increase on F1-score and we withdraw it from the final model to increase inference time.

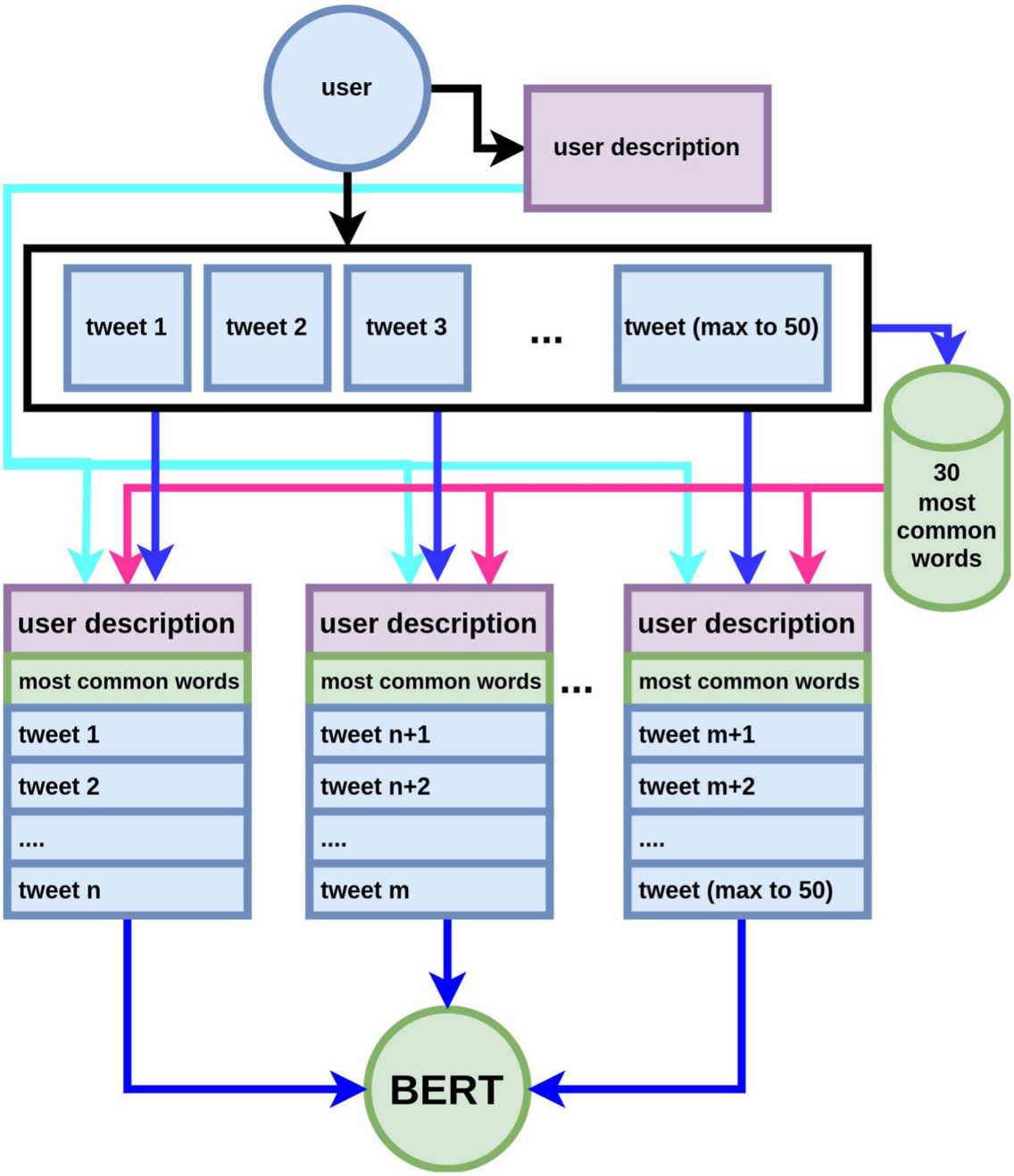


Fig. 2 Model architecture

4.3 Model architecture

We present a new input structure to feed the BERT, as in Fig. 2 shows, for a given user, the tweet text for up to the last 50 tweets were pooled together. We used BERT tokenizer to tokenize the tweets. Afterward, we use chunk methods to cover up to the last 50 tweets. Also user description and 30 most common words among the last 50 tweets are added to the first of every chunk. According to BERT's limit of 512 tokens, we could not feed 50 tweets in one chunk. So for every user there are at most 4 to 5 chunks with the same label of age for feeding to BERT. Finally at inference we use the most predicted label of all chunks of user as an output of the model. In our experiments, we first train the model at the tweet-level and then do the predictions at the user-level. Experiment at the tweet-level means each chunk of tweets

Table 7 Result on test data

	Num of classes	Acc	F1-score	Precision	Recall
BERT model fine-tuned on ExaAUAC	6	0.58	0.44	0.43	0.45
	3 (−25, 25–34, +35)	0.59	0.58	0.60	0.56

Table 8 Model comparison

Models	F1-score
LSTM + CNN model	0.35
Presented model with BERT	0.44

with its label is considered to be predicted and evaluated, but at the user-level, all 4–5 chunks of the user tweets are considered and the most common predicted label is assumed for the evaluation.

5 Results

In this section, an evaluation method is used to measure ExaAUAC corpus accuracy. We chose a BERT model to fine-tune, cause it offers higher accuracy in classification tasks at the time of preparing the article. Different Arabic BERT models were investigated like Arabic-BERT Safaya et al. [19], AraBERT Antoun et al. [2], Multi-dialect-Arabic-BERT Talafha et al. [21] and MARBERT Abdul-Mageed et al. [1]. So, Multi-dialect-Arabic-BERT Talafha et al. [21] language model is selected among them, which is a BERT-based model that is pre-trained on 10 M Arabic tweets and had a better performance in age detection task. Furthermore, we fine-tuned Multi-dialect-Arabic-BERT on 2 M Arabic tweets and distilled it to 6 layers in order to have a better throughput and speed at inference.

For the age detection task of Arabic twitter users, we split ExaAUAC to train and test data as mentioned before in Sect. 3.1, and fine-tuned the BERT model on ExaAUAC train data by a supervised method on 6 classes for 10 epochs. For the training process we employ batch size of 16, Adam optimizer with learning rate of 5e-05 and weight decay of 0.01. Our training procedures are performed on the two 1080ti GPUs with 11 GB RAM. Calculated accuracy, precision, recall, and F1-score-measure on test data are shown in Table 7. Test data consists of accounts, labeled by taggers manually.

The age detection model with BERT on our corpus has F1-score of 0.44 according to 6 classes of labels. A comparison between model with LSTM and CNN, and BERT model on ExaAUAC, as shown in Table 8, specifies a 9 percent increase in F1-score. The advantages of using the proposed model is that we just used a maximum of 50 tweets per user and we omit using metadata so model input is lighter than LSTM+CNN model and it causes better performance in usage, has better inference speed, and additionally, is more accurate. In the following, the only Arabic model to compare with the presented model is the one from the article by Zhang et al. [22], which is based on BERT with three classes: -25, 25–34, and +35. To ensure a proper comparison, we evaluated the presented model on these three age groups and obtained an F1-score of 0.58 and an accuracy of 59%, concluding with up to a 5% increase in accuracy.

6 Conclusion

In this article, we presented a new corpus called ExaAUAC to detect the age of Arabic Twitter users. We described four different methods and resources to gather data and constructed the large-scale corpus including 181k user profiles with coverage of diverse age range in 6 class of labels consisting of −18, 18–24, 25–34, 35–49, 50–64, +65. The key advantage of presented corpus is the large size of data with balance in different classes. Since proposing a new corpus requires models for evaluation and assessment, we described two models: (1) LSTM+CNN model and (2) BERT-based model. We demonstrated how to use ExaAUAC to train an age detection model. Our results show that the accuracy of the user age detection model can be increased by using the BERT language model in combination with chunking tweets and user information. We conclude that by using only the text of tweets and user information, our best model outperforms the LSTM+CNN model by 9% and achieved F1-score of 44 on 6 age groups on 8200 test data. The advantages of the proposed model include increased accuracy and lighter input with a maximum of 50 tweets per user, resulting in better inference

speed. Also, the presented model achieved F1-score of 58 and Accuracy of 59 on 3 age groups including −25, 25–34, +35 and yields an improvement of up to 5% in accuracy compared to similar BERT-based Arabic models. The results show that it is possible to extract some information about the author only by viewing the texts written by him/her with proper accuracy. In the future, ExaAUAC will expand the data in class of +65 and also will cause increase in accuracy of age detection model by using better language model. The proposed method can also be used to provide data and detect other demographic information such as gender, education, race, etc. in Social Networks.

Acknowledgements The authors would like to appreciate Exa Company for providing the infrastructure, human resources, technical and non-technical knowledge that results in achieving this article and its goals.

Author contributions Reyhaneh Sadeghi train the models and evaluate them and wrote the experiments and result and done the experiments, and Ahmad Akbari prepared related works and Figures. Mohammad Mehdi Jaziriyan wrote introduction and tagger agreements parts of paper and did prepare paper in latex format and submitted. Abstract and conclusion are written and reviewed by all authors. All authors reviewed the final version of the manuscript.

Data availability The Corpus is available on the provided link in abstract section.

Declarations

Competing interests The authors declare no Competing interests.

References

1. Abdul-Mageed M, Elmadany A, Nagoudi E. ARBERT & MARBERT: deep bidirectional transformers for Arabic. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021; 7088-7105
2. Antoun W, Baly F, Hajj H. Arabert: transformer-based model for arabic language understanding. LREC 2020 Workshop Language Resources and Evaluation Conference, 2020; 9
3. Bradski G. The OpenCV Library. Dr Dobb's J Softw Tools. 2000;25:120. https://github.com/opencv/opencv/wiki/CiteOpenCV.
4. Chamberlain B, Humby C, Deisenroth M. Probabilistic inference of twitter users' age based on what they follow. Lecture Notes In Computer Science (including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics). 10536 LNAI 2017; 191-203
5. Culotta A, Ravi N, Cutler J. Predicting twitter user demographics using distant supervision from website traffic data. J Artif Intell Res. 2016;55:389–408.
6. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings Of The 2019 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long And Short Papers). 2019; 4171-4186, https://aclanthology.org/N19-1423.
7. Fleiss J. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76:378–82.
8. Juola P, Baayen R. A controlled-corpus experiment in authorship identification by cross-entropy. Lit Linguist Comput. 2005;20:59–67. https://doi.org/10.1093/llc/fqi024.
9. Klein A, Magge A, Gonzalez-Hernandez G. ReportAGE: automatically extracting the exact age of twitter users based on self-reports in tweets. PLoS ONE. 2022;17: e0262087. https://journals.plos.org/plosone/article/citation?id=10.1371/journal.pone.0262087.
10. Morgan-Lopez A, Kim A, Chew R, Ruddle P. Predicting age groups of Twitter users based on language and metadata features. PLoS ONE. 2017;12: e0183537. https://journals.plos.org/plosone/article/citation?id=10.1371/journal.pone.0262087.
11. Mubarak H, Hassan S, Abdelali, A. Adult content detection on arabic twitter: analysis and experiments. 2021.
12. Pandya A, Oussalah M, Monachesi P, Kostakos P, Loven L. On the Use of URLs and Hashtags in Age Prediction of Twitter Users. 2018 IEEE International Conference On Information Reuse And Integration (IRI). 2018; 62-69.
13. Pandya A, Oussalah M, Monachesi P, Kostakos P. On the use of distributed semantics of tweet metadata for user age prediction. Future Gener Comput Syst. 2020;102:437–52.
14. Pennebaker J, Mehl M, Niederhoffer K. Psychological aspects of natural language. Use: our words, our selves. Ann Rev Psychol. 2003;54:547–77.
15. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep Contextualized Word Representations. Proceedings Of The 2018 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018; 2227-2237, https://aclanthology.org/N18-1202.
16. Rangel F, Rosso P, Koppel M, Stamatatos E, Inches G. Overview of the Author Profiling Task at PAN 2013. Working Notes Papers of the CLEF 2013 Evaluation Labs. CEUR-WS. Org. 2013; 1179.
17. Rangel F, Rosso P, Chugur I, Potthast M, Trenkmann M, Stein B, Verhoeven B, Daelemans W. Overview of the 2nd Author Profiling Task at PAN 2014. Working Notes Papers of the CLEF 2014 Evaluation Labs. CEUR-WS. Org. 2014; 1180.

18. Rangel F, Rosso P, Charfi A, Zaghouani W, Ghanem B, Sanchez-Junquera J. Overview of the track on author profiling and deception detection in Arabic. Work Notes FIRE 2019 CEUR-WS Org. 2019;2517:70–83.
19. Safaya A, Abdullatif M, Yuret D. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020; 2054-2059.
20. Schler J, Koppel M, Argamon S, Pennebaker J. Effects of age and gender on blogging. AAAI Spring Symposium: Computational Approaches To Analyzing Weblogs. 2006.
21. Talafha B, Ali M, Za'ter M, Seelawi H, Tuffaha I, Samir M, Farhan W, Al-Natsheh H. Multi-dialect Arabic BERT for Country-level Dialect Identification. Proceedings of the Fifth Arabic Natural Language Processing Workshop. 2020; pp. 111-118, https://aclanthology.org/2020.wanlp-1.10.
22. Zhang C, Abdul-Mageed M. BERT-based Arabic social media author profiling. CEUR Workshop Proc. 2019;2517:84–91.