BERT-Based Arabic Social Media Author Profiling

Chiyu Zhang and Muhammad Abdul-Mageed

Natural Language Processing Lab
The University of British Columbia
chiyuzh@mail.ubc.ca, muhammad.mageeed@ubc.ca

Abstract. We report our models for detecting age, language variety, and gender from social media data in the context of the Arabic author profiling and deception detection shared task (APDA) [32]. We build simple models based on pre-trained bidirectional encoders from transformers (BERT). We first fine-tune the pre-trained BERT model on each of the three datasets with shared task released data. Then we augment shared task data with in-house data for gender and dialect, showing the utility of augmenting training data. Our best models on the shared task test data are acquired with a majority voting of various BERT models trained under different data conditions. We acquire 54.72% accuracy for age, 93.75% for dialect, 81.67% for gender, and 40.97% joint accuracy across the three tasks.[1]

Keywords: author profiling identification, BERT, Arabic, social media

1 Introduction

The proliferation of social media has made it possible to collect user data in unprecedented ways. These data can come in the form of usage and behavior (e.g., who likes what on Facebook), network (e.g., who follows a given user on Instagram), and content (e.g., what people post to Twitter). Availability of such data have made it possible to make discoveries about individuals and communities, mobilizing social and psychological research and employing natural language processing methods. In this work, we focus on predicting social media user age, dialect, and gender based on posted language. More specifically, we use the total of 100 tweets from each manually-labeled user to predict each of these attributes. Our dataset comes from the Arabic author profiling and deception detection shared task (APDA) [32]. We focus on building simple models using pre-trained bidirectional encoders from transformers (BERT) [12] under various data conditions. Our results show (1) the utility of augmenting training data, and (2) the benefit of using majority votes from our simple classifiers.

In the rest of the paper, we introduce the dataset, followed by our experimental conditions and results. We then provide a literature review and conclude.

2 Data

For the purpose of our experiments, we use data released by the APDA shared task organizers. The dataset is divided into train and test by organizers. The TRAIN set is distributed with labels for the three tasks of age, dialect, and gender. Following the standard shared tasks set up, the test set is distributed without labels and participants were expected to submit their predictions on test. The shared task predictions are expected by organizers at the level of users. The distribution has 100 tweets for each user, and so each tweet is distributed with a corresponding user id. As such, in total, the distributed training data has 2,250 users, contributing a total of 225,000 tweets. The official task test set contains 720,00 tweets posted by 720 users. For our experiments, we split the training data released by organizers into 90% TRAIN set (202,500 tweets from 2,025 users) and 10% DEV set (22,500 tweets from 225 users). The *age* task labels come from the tagset {*under-25*, *between-25 and 34*, *above-35*}. For dialects, the data are labeled with 15 classes, from the set {*Algeria*, *Egypt*, *Iraq*, *Kuwait*, *Lebanon-Syria*, *Lybia*, *Morocco*, *Oman*, *Palestine-Jordan*, *Qatar*, *Saudi Arabia*, *Sudan*, *Tunisia*, *UAE*, *Yemen*}. The *gender* task involves binary labels from the set {*male*, *female*}.

3 Experiments

As explained earlier, the shared task is set up at the user level where the age, dialect, and gender of each *user* are the required predictions. In our experiments, we first model the task at the *tweet* level and then port these predictions at the user level. For our core modelling, we fine-tune BERT on the shared task data. We also introduce an additional in-house dataset labeled with dialect and gender tags to the task as we will explain below. As a baseline, we use a small gated recurrent units (GRU) model. We now introduce our tweet-level models.

3.1 Tweet-Level Models

Baseline GRU. Our baseline is a GRU network for each of the three tasks. We use the same network architecture across the 3 tasks. For each network, the network contains a layer unidirectional GRU, with 500 units and an output linear layer. The network is trained end-to-end. Our input embedding layer is initialized with a standard normal distribution, with $\mu = 0$, and $\sigma = 1$, i.e., $W \sim N(0, 1)$. We use a maximum sequence length of 50 tokens, and choose an arbitrary vocabulary size of 100,000 types, where we use the 100,000 most frequent words in TRAIN. To avoid over-fitting, we use dropout [43] with a rate of 0.5 on the hidden layer. For the training, we use the Adam [22] optimizer with a fixed learning rate of $1e-3$. We employ batch training with a batch size of 32 for this model. We train the network for 15 epochs and save the model at the end of each epoch, choosing the model that performs highest accuracy on DEV as our best model. We present our best result on DEV in Table 1. We report all

our results using accuracy. Our best model obtains 42.48% for age, 37.50% for dialect, and 57.81% for gender. All models obtain best results with 2 epochs.

BERT. For each task, we fine-tune on the BERT-Base Muultilingual Cased model relesed by the authors [12] [2]. The model was pre-trained on Wikipedia of 104 languages (including Arabic) with 12 layer, 768 hidden units each, 12 attention heads, and has 110M parameters in entire model. The vocabulary of the model is 119,547 shared WordPices. We fine-tune the model with maximum sequence length of 50 tokens and a batch size of 32. We set the learning rate to $2e-5$ and train for 15 epochs. We use the same network architecture and parameters across the 3 tasks. As Table 1 shows, comparing with GRU, BERT is 3.16% better for age, 4.85% better for dialect, and 2.45% higher for gender.

Data Augmentation. To further improve the performance of our models, we introduce in-house labeled data that we use to fine-tune BERT. For the gender classification task, we manually label an in-house dataset of 1,100 users with gender tags, including 550 *female* users, 550 *male* users. We obtain 162,829 tweets by crawling the 1,100 users' timelines. We combine this new gender dataset with the gender TRAIN data (from shared task) to obtain an extended dataset, to which we refer as `EXTENDED_Gender`. For the dialect identification task, we randomly sample 20,000 tweets for each class from an in-house dataset gold labeled with the same 15 classes as the shared task. In this way, we obtain 298,929 tweets (*Sudan* only has 18,929 tweets). We combine this new dialect data with the shared task dialect TRAIN data to form `EXTENDED_Dialect`. For both the dialect and gender tasks, we fine-tune BERT on `EXTENDED_Dialect` and `EXTENDED_Gender` independently and report performance on DEV. We refer to this iteration of experiments as BERT_EXT. As Table 1 shows, BERT_EXT is 2.18% better than BERT for dialect and 0.75% better than BERT for gender. [3]

Table 1. Tweet level results on DEV

	Age	**Dialect**	**Gender**
GRU	42.48	37.50	57.81
BERT	45.64	42.35	60.26
BERT_EXT	-	44.53	61.01

3.2 User-Level Models

Our afore-mentioned models identify user's profiling on the tweet-level, rather than directly detecting the labels of a user. Hence, we follow the work of Zhang

[2] https://github.com/google-research/bert/blob/master/multilingual.md

[3] We note that it was not possible for us to use external age-labeled data and hence we do not report on the age task with this data augmentation setting.

& Abdul-Mageed [47] to identify user-level labels. For each of the three tasks, we use tweet-level predicted labels (and associated softmax values) as a proxy for user-level labels. For each predicted label, we use the softmax value as a threshold for including only highest confidently predicted tweets. Since in some cases softmax values can be low, we try all values between 0.00 and 0.99 to take a softmax-based majority class as the user-level predicted label, fine-tuning on our DEV set. Using this method, we acquire the following results at the user level: `BERT` models obtain an accuracy of 55.56% for age, 96.00% for dialect, and 80.00% for gender. `BERT_EXT` models achieve 95.56% accuracy for dialect and 84.00% accuracy for gender.

3.3 APDA@FIRE2019 submission

First submission. For the shared task submission, we use the predictions of BERT_EXT as out first submission for gender and dialect, but only BERT for age (since we have no BERT_EXT models for age, as explained earlier). In each case, we acquire results at tweet-level first, then port the labels at the user-level as explained in the previous section. For our second and third submitted models, we also follow this method of going from tweet to user level. **Second submission.** We combine our DEV data with our EXTENDED_Dialect and EXTENDED_Gender data, for dialect and gender respectively, and train our second submssions for the two tasks. For age second submsision, we concatenate DEV data to TRAIN and fine-tune the BERT model. We refer to the settings for our second submission models collectively as BERT_EXT+DEV.

Third submission. Finally, for our third submission, we use a majority vote of (1) first submission, (2) second submission, and (3) predictions from our user-level BERT model. These majority class models (i.e., our third submission) achieve best results on the official test data. We acquire 54.72% accuracy for age, 81.67% accuracy for gender, 93.75% accuracy for dialect, and 40.97% joint accuracy.

Table 2. Results of our submissions on official test data (user level)

	Exp. Condition	**Age**	**Dialect**	**Gender**	**Joint**
Submission 1	**BERT_EXT**	54.72	93.33	77.08	38.75
Submission 2	**BERT_EXT+DEV**	54.72	92.64	81.67	40.97
Submission 3	**MAJ_CLASS**	54.72	93.75	81.67	40.97

4 Related Works

Arabic. *Arabic* is a term that refers to a collection of languages, varieties, and dialects. The standard variety, Modern Standard Arabic (MSA), is the one usually used in formal communication and educational settings. Arabic also has a

wide range of under-studied varieties and dialects that classically used to be categorized in a coarse-grained fashion (e.g., Levantine, North African) [17, 46, 1, 13, 2]. More recent treatments focus on fine-grained categorizations such as country and city levels [25, 39, 3, 40, 31, 47]. Differences between varieties of Arabic happen at various linguistic levels, including including phonological, morphological, lexical, and syntactic [19, 6, 27, 1].

Social Media Author Profiling. *Author profiling* is the term usually used to refer to detecting a host of attributes of (often social media) users. This include identifying attributes such as age, gender, educational level, economic class, stance or ideology [10, 30], personality [42, 23, 7, 18], moral traitstraits [21, 28], and other socilogical and psychological constructs. Author profiling based on text [15, 5] is rooted in computational stylometry [16, 44] and has traces in the early work of Holmes [20]. The task of author profiling has also been approached from network perspective where cues based on friending, following, mentioning, and commenting have been leveraged for identifying author attributes [24] for author profiling. In addition, the PAN author profiling shared task [34, 33, 37, 36, 35] was established to advance related work. More information about PAN can be found in [29].

Age and Gender. A number of studies have been conducted on English-based age and gender detection, including [38, 14, 8, 45, 11]. Many of these works use feature engineering such as text n-gram and topic models [42, 41]. In these works, age is either cast as a multi-class classification task with, e.g., labels from the set {*10-19, 20-29, 30-39*} or as a regression task [26]. Other works model age with both classification and regression combined [9]. With rare exceptions [35, 4], we do not know of work on Arabic targeting age and gender.

5 Conclusion

In this work, we described our submitted models to the Arabic author profiling and deception detection shared task (APDA) [32]. We focused on detecting age, dialect, and gender using BERT models under various data conditions, showing the utility of additional, in-house data on the task. We also showed that a majority vote of our models trained under different conditions outperforms single models on the official evaluation. In the future, we will investigate automatically extending training data for these tasks as well as better representation learning methods.

6 Acknowledgement

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Social Sciences Research Council of Canada (SSHRC), and Compute Canada (www.computecanada.ca).

References

1. Abdul-Mageed, M.: Subjectivity and sentiment analysis of Arabic as a morophologically-rich language. Ph.D. thesis, Indiana University (2015)
2. Abdul-Mageed, M.: Modeling arabic subjectivity and sentiment in lexical space. Information Processing & Management (2017)
3. Abdul-Mageed, M., Alhuzali, H., Elaraby, M.: You tweet what you speak: A city-level dataset of arabic dialects. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018) (2018)
4. Alrifai, K., Rebdawi, G., Ghneim, N.: Arabic tweeps gender and dialect prediction. In: CLEF (Working Notes) (2017)
5. Argamon, S.E.: Register in computational language research. Register Studies **1**(1), 100–135 (2019)
6. Bassiouney, R.: Arabic sociolinguistics. Edinburgh University Press (2009)
7. Bleidorn, W., Hopwood, C.J.: Using machine learning to advance personality assessment and theory. Personality and Social Psychology Review p. 1088868318772990 (2018)
8. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the conference on empirical methods in natural language processing. pp. 1301–1309. Association for Computational Linguistics (2011)
9. Chen, J., Cheng, L., Yang, X., Liang, J., Quan, B., Li, S.: Joint learning with both classification and regression models for age prediction. In: Journal of Physics: Conference Series. vol. 1168, p. 032016. IOP Publishing (2019)
10. Colleoni, E., Rozza, A., Arvidsson, A.: Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. Journal of communication **64**(2), 317–332 (2014)
11. Daneshvar, S., Inkpen, D.: Gender identification in twitter using n-grams and lsa. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018) (2018)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Elaraby, M., Abdul-Mageed, M.: Deep models for arabic dialect identification on benchmarked data. In: Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018). pp. 263–274 (2018)
14. Flekova, L., Preoţiuc-Pietro, D., Ungar, L.: Exploring stylistic variation with age and income on twitter. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 313–319 (2016)
15. Gamon, M.: Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: Proceedings of the 20th international conference on Computational Linguistics. p. 611. Association for Computational Linguistics (2004)
16. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers age and gender. In: Third international AAAI conference on weblogs and social media (2009)
17. Habash, N.Y.: Introduction to arabic natural language processing. Synthesis Lectures on Human Language Technologies **3**(1), 1–187 (2010)
18. Hinds, J., Joinson, A.: Human and computer personality prediction from digital footprints. Current Directions in Psychological Science p. 0963721419827849 (2019)

19. Holes, C.: Modern Arabic: Structures, functions, and varieties. Georgetown University Press (2004)
20. Holmes, D.I.: The evolution of stylometry in humanities scholarship. Literary and linguistic computing **13**(3), 111–117 (1998)
21. Johnson, K., Goldwasser, D.: Classification of moral foundations in microblog political discourse. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 720–730 (2018)
22. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Matz, S.C., Kosinski, M., Nave, G., Stillwell, D.J.: Psychological targeting as an effective approach to digital mass persuasion. Proceedings of the national academy of sciences **114**(48), 12714–12719 (2017)
24. Mitrou, L., Kandias, M., Stavrou, V., Gritzalis, D.: Social media profiling: A panopticon or omniopticon tool? In: Proc. of the 6th Conference of the Surveillance Studies Network. Barcelona, Spain (2014)
25. Mubarak, H., Darwish, K.: Using twitter to collect a multi-dialectal corpus of arabic. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). pp. 1–7 (2014)
26. Nguyen, D., Smith, N.A., Rosé, C.P.: Author age prediction from text using linear regression. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 115–123. Association for Computational Linguistics (2011)
27. Palva, H.: Dialects: classification. Encyclopedia of Arabic Language and Linguistics **1**, 604–613 (2006)
28. Pang, D., Eichstaedt, J.C., Buffone, A., Slaff, B., Ruch, W., Ungar, L.H.: The language of character strengths: Predicting morally valued traits on social media. Journal of personality (2019)
29. Potthast, M., Rosso, P., Stamatatos, E., Stein, B.: A decade of shared tasks in digital text forensics at pan. In: European Conference on Information Retrieval. pp. 291–300. Springer (2019)
30. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L.: Beyond binary labels: political ideology prediction of twitter users. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 729–740 (2017)
31. Qwaider, C., Saad, M., Chatzikyriakidis, S., Dobnik, S.: Shami: A corpus of levantine arabic dialects. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018) (2018)
32. Rangel, F., Rosso, P., Charfi, A., Zaghouani, W., Ghanem, B., Snchez-Junquera, J.: Overview of the track on author profiling and deception detection in arabic. In: Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS.org, Kolkata, India, December 12-15 (2019)
33. 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. In: CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014. pp. 1–30 (2014)
34. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. pp. 352–365. CELCT (2013)

35. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working Notes Papers of the CLEF (2017)
36. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al. pp. 750–784 (2016)
37. Rangel Pardo, F.M., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: CLEF 2015 Evaluation Labs and Workshop Working Notes Papers. pp. 1–8 (2015)
38. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proceedings of the 2nd international workshop on Search and mining user-generated contents. pp. 37–44. ACM (2010)
39. Sadat, F., Kazemi, F., Farzindar, A.: Automatic identification of arabic language varieties and dialects in social media. Proceedings of SocialNLP p. 22 (2014)
40. Salameh, M., Bouamor, H.: Fine-grained arabic dialect identification. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1332–1344 (2018)
41. Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., Schwartz, H.A.: Developing age and gender predictive lexica over social media. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1146–1151 (2014)
42. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one **8**(9), e73791 (2013)
43. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
44. Verhoeven, B.: Two authors walk into a bar: studies in author profiling. Ph.D. thesis, University of Antwerp (2018)
45. Verhoeven, B., Daelemans, W., Plank, B.: Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In: Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari, Nicoletta [edit.]; et al. pp. 1–6 (2016)
46. Versteegh, K.: The arabic language. Edinburgh University Press (2014)
47. Zhang, C., Abdul-Mageed, M.: No army, no navy: Bert semi-supervised learning of arabic dialects. In: Proceedings of the Fourth Arabic Natural Language Processing Workshop. pp. 279–284 (2019)