DBMS-KU Approach for Author Profiling and Deception Detection in Arabic

Al Hafiz Akbar Maulana Siagian[1,3][0000−0003−0311−3134] and Masayoshi Aritsugi[2][0000−0003−0861−849X]

[1] Computer Science and Electrical Engineering, Graduate School of Science and Technology, Kumamoto University, Japan
[2] Big Data Science and Technology, Faculty of Advanced Science and Technology, Kumamoto University, Japan
[3] Indonesian Institute of Sciences, Indonesia
alha002@dbms.cs.kumamoto-u.ac.jp
aritsugi@cs.kumamoto-u.ac.jp

Abstract. This paper presents DBMS-KU team approach for task 1, i.e., author profiling in Arabic tweets, and task 2, viz., deception detection in Arabic texts, of Author Profiling and Deception Detection in Arabic (APDA). Our approach utilizes word n-grams, character n-grams, word and character n-grams combinations, and function words as features for classifiers to deal with these two tasks. We then examine several term frequency thresholds to attributes of the features. Our obtained results indicated that our approach could work well in both tasks of this APDA.

Keywords: Author profiling · Deception detection · Character n-grams · Word and character n-grams combinations · Function words.

1 Introduction

This paper elucidates the participation of DBMS-KU team in Author Profiling and Deception Detection in Arabic (APDA) of Forum for Information Retrieval and Evaluation (FIRE) 2019. In this APDA, we participate in task 1, i.e., author profiling (gender, age, and variety) in Arabic tweets, and task 2, viz., deception detection (lie or truth) in Arabic texts (twitter and news) [2, 13–20, 25, 26]. To do this, we utilize word n-grams, character n-grams, word and character n-grams combinations, and function words as features for classifiers. Previous studies have shown that using those features for classifiers could contribute well in author profiling and deception detection tasks [1–4, 7–11, 13, 20, 22–24]. To improve our classification performance, we perform particular preprocessing techniques to the dataset and apply term frequency thresholds to attributes of our utilized features in this work. Our obtained results could show encouraging evaluation scores in both tasks of the APDA.

2 Dataset

The APDA organizer provided dataset that consisted of training and testing sets for each task 1 and task 2 [2, 13–20, 25, 26]. The training and testing sets of task 1 consisted of 2250 and 720 tweets, respectively. In task 2, the training sets consisted of 532 instances for twitter and 1443 instances for news, while the testing ones had 241 instances for twitter and 370 instances for news.

3 Method

This section explains the employed features, term frequency thresholds, preprocessing, and classifiers in our approach to deal with task 1 and task 2 of the APDA. To examine the performance of our approach, we conduct 10 fold cross-validation in our experiments. Accuracy and F1-score are used to measure the performance of our approach in task 1 and task 2, respectively. This measurement is the same as the evaluation metric used by the APDA organizer [16].

3.1 Features

Word N-grams We use the word n-grams, namely, unigrams (uni), unigrams + bigrams (2grams), and unigrams + bigrams + trigrams (3grams), as our features to deal with both tasks in this work. Our word n-grams features are segregated by space(s) and excluding numbers.

Character N-grams In this work, we examine character n-grams features where n is between 3 (3chars) and 15 (15chars). The character n-grams features are constructed from all existing characters in the dataset.

Word and Character N-grams Combinations To use word and character n-grams combinations as features, we follow the method of [22, 24]. For simplicity, we consider combining word n-grams, i.e., uni, 2grams, and 3grams, with only the best character ones as our combinations features in this work, e.g., uni+3chars, 2grams+3chars, and 3grams+3chars.

Function Words To deal with this APDA, we utilize a set of 248 Arabic function words (AFW) from [12] as features. Following the work of [6, 23, 24], we employ the AFW features as combinations with our other utilized features in this work, e.g., uni+AFW, 3chars+AFW, and uni+3chars+AFW.

3.2 Term Frequency Thresholds

We examine several term frequency thresholds to attributes of our utilized features. This examination is to optimize the use of features [7, 13] when used by classifiers in this work. The examined value of the term frequency thresholds (TH) of the features in this work is from 1 to 10, e.g., uni_TH2, 3char_TH2, and uni_TH2+3chars_TH2.

3.3 Preprocessing

To improve the classification performance [7, 8, 13], we perform certain preprocessing techniques to the dataset of task 1 and task 2. In task 1, we consider three preprocessing options, namely, removing all mentioned @users (No@user), removing all mentioned @users and URLs (No@user-url), and using Arabic alphabets (Arabic-only). Meanwhile, we use all contents in the dataset and replace all numbers by zero [8], e.g., 2020 replaced by 0000, (To-zero) as our preprocessing choice in task 2. The different preprocessing considerations between task 1 and task 2 are not only because of the different purpose of classification tasks but also due to the different characteristic of each dataset in task 1 and task 2.

3.4 Classifiers

We utilize Support Vector Machine (SVM), Naive Bayes (NB), and Multinomial Naive Bayes (MNB) classifiers in this work. To do this, we use the WEKA [5] implementation of those classifiers.

4 Experimental Settings and Results

We show our experimental settings and obtained results in our submissions for classifying testing sets of task 1 in Tables 1 and 2, respectively. In this task 1, using DBMS-KU.2 settings could perform a better classification for gender and age, while utilizing DBMS-KU.3 ones could perform the best classification for variety. These good results might be due to the used preprocessing techniques, i.e., No@user for gender, No@user-url for age, and Arabic-only for variety, to our utilized features in this task 1. This conjecture might be corroborated by our finding that the obtained variety classification results of using Arabic-only preprocessing options, i.e., DBMS-KU.1 and DBMS-KU.3, were better than those

Table 1. The used experimental settings in our submissions for classifying the testing sets in task 1. The SVM classifier was used for all our submissions in the task 1.

Submissions	Settings	Gender	Age	Variety
DBMS-KU.1	Features	10chars_TH3	8chars_TH2	11chars_TH6
	Preprocessing	Arabic-only	Arabic-only	Arabic-only
DBMS-KU.2	Features	7chars_TH5	7chars_TH5	9chars_TH7
	Preprocessing	No@user	No@user-url	No@user-url
DBMS-KU.3	Features	uni_TH1+10chars_TH3	uni_TH1+8chars_TH2	uni_TH1+11chars_TH6
	Preprocessing	Arabic-only	Arabic-only	Arabic-only

Table 2. The obtained results of our submissions in task 1. Results in bold indicate the best accuracy among our three submissions.

Submissions	Gender	Age	Variety	Joint
DBMS-KU.1	0.7778	0.5792	0.9736	0.4347
DBMS-KU.2	**0.7944**	**0.5861**	0.9722	**0.4556**
DBMS-KU.3	0.7833	0.5819	**0.9778**	0.4444

Table 3. The used experimental settings in our submissions for classifying the testing sets in task 2. The MNB classifier was used for all our submissions in the task 2.

Submissions	Settings	Twitter	News
DBMS-KU.1	Features	11chars_TH2	11chars_TH2
	Preprocessing	To-zero	To-zero
DBMS-KU.2	Features	uni_TH1+11chars_TH2	uni_TH1+11chars_TH2
	Preprocessing	To-zero	To-zero
DBMS-KU.3	Features	11chars_TH2+AFW_TH1	11chars_TH2+AFW_TH1
	Preprocessing	To-zero	To-zero

Table 4. The obtained results of our submissions in task 2. Results in bold indicate the best F1-score among our three submissions.

Submissions	Twitter	News	Average
DBMS-KU.1	0.7877	0.7188	0.7533
DBMS-KU.2	**0.8125**	**0.7352**	**0.7739**
DBMS-KU.3	0.7877	0.7188	0.7533

of using No@user-url ones, viz. DBMS-KU.2 (Table 2). Nevertheless, it would be valuable to analyze further this supposition in the future.

Next, we display the used experimental settings and obtained results in our submissions for classifying testing sets of task 2 in Tables 3 and 4, respectively. According to the results in Table 4, our DBMS-KU.2 settings could perform better than other ones, i.e., DBMS-KU.1 and DBMS-KU.3, for detecting deception in both twitter and news of this task 2. This DBMS-KU.2 achievement might be due to the ability of word and character n-grams combinations features which was better to capture deception attributes than that of character n-grams, i.e., DBMS-KU.1, and character n-grams combined with Arabic function words (AFW) ones, viz., DBMS-KU.3. However, it should be worthwhile to investigate further of this presumption in the future.

5 Conclusion

This paper has described the participation of DBMS-KU team in task 1 and task 2 of the APDA of FIRE 2019. Our obtained results indicated that our approach could be useful and give promising evaluation scores in both tasks of this APDA. As future work, examining more various preprocessing techniques to the dataset, such as in [8], and using other types of character n-grams features, e.g., character n-grams in token [1,2] and typed character n-grams [7,8,20,21], must be valuable to improve the performance of our approach.

Acknowledgment A.H.A.M. Siagian thanks RISET-Pro (Research and Innovation in Science and Technology Project) KEMENRISTEKDIKTI (Ministry of Research, Technology and Higher Education of the Republic of Indonesia) scholarship for his study.

References

1. Cagnina, L., Rosso, P.: Classification of deceptive opinions using a low dimensionality representation. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 58–66. WASSA@EMNLP 2015, Association for Computational Linguistics, Stroudsburg, PA, USA (2015). https://doi.org/10.18653/v1/W15-2909
2. Cagnina, L., Rosso, P.: Detecting deceptive opinions: Intra and cross-domain classification using an efficient representation. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **25**(Supplement-2), 151–174 (2017). https://doi.org/10.1142/S0218488517400165
3. Fusilier, D.H., Montes-y-Gómez, M., Rosso, P., Cabrera, R.G.: Detecting positive and negative deceptive opinions using pu-learning. Information Processing and Management **51**(4), 433–443 (2015). https://doi.org/10.1016/j.ipm.2014.11.001
4. Fusilier, D.H., Montes-y-Gómez, M., Rosso, P., Cabrera, R.G.: Detection of opinion spam with character n-grams. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 285–294. CICLing 2015, Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-18117-2_21
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009). https://doi.org/10.1145/1656274.1656278
6. Markov, I., Chen, L., Strapparava, C., Sidorov, G.: CIC-FBK approach to native language identification. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 374–381. BEA 2017, Association for Computational Linguistics, Stroudsburg, PA, USA (2017). https://doi.org/10.18653/v1/W17-5042
7. Markov, I., Gómez-Adorno, H., Sidorov, G.: Language- and subtask-dependent feature selection and classifier parameter tuning for author profiling. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum. CLEF 2017, http://ceur-ws.org/Vol-1866/paper_96.pdf
8. Markov, I., Stamatatos, E., Sidorov, G.: Improving cross-topic authorship attribution: The role of pre-processing. In: Computational Linguistics and Intelligent Text Processing. pp. 289–302. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-77116-8_21
9. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: What yelp fake review filter might be doing? In: International AAAI Conference on Web and Social Media. ICWSM-2013 (2013), https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6006
10. Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 497–501. NAACL-HLT 2013, Association for Computational Linguistics, Stroudsburg, PA, USA (2013), https://www.cs.cornell.edu/home/cardie/papers/NAACL13-Negative.pdf
11. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 309–319. ACL-HLT 2011, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2002472.2002512

12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011), http://dl.acm.org/citation.cfm?id=1953048.2078195
13. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: Proc. of the 17th Int. Conf. on Intelligent Text Processing and Computational Linguistics. pp. 156–169. CICLing 2016, Springer-Verlag (2018). https://doi.org/10.1007/978-3-319-75487-1_13
14. Rangel, F., Rosso, P.: On the impact of emotions on author profiling. Information Processing and Management **52**(1), 73–92 (2016). https://doi.org/https://doi.org/10.1016/j.ipm.2015.06.003
15. Rangel, F., Rosso, P., Charfi, A., Zaghouani, W.: Detecting deceptive tweets in arabic for cyber-security. In: Proceedings of the 17th IEEE International Conference on Intelligence and Security Informatics. ISI, IEEE, New York, NY, USA (2019)
16. Rangel, F., Rosso, P., Charfi, A., Zaghouani, W., Ghanem, B., Sánchez-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. CEUR-WS.org, Kolkata, India, December 12-15 (2019)
17. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the author profiling task at PAN 2014. In: Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 Labs and Workshops, Notebook Papers. pp. 898–927 (2014), http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-RangelEt2014.pdf
18. 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. In: Cappellato L., Ferro N., Goeuriot L, Mandl T. (Eds.) CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (2017), http://ceur-ws.org/Vol-1866/invited_paper_11.pdf
19. Rosso, P., Rangel, F., Hernández-Farías, I., Cagnina, L., Zaghouani, W., Charfi, A.: A survey on author profiling, deception, and irony detection for the arabic language. Language and Linguistics Compass **12**(4), e12275 (2018). https://doi.org/10.1111/lnc3.12275
20. Sánchez-Junquera, J., Villaseñor-Pineda, L., Montes-y-Gómez, M., Rosso, P.: Character n-grams for detecting deceptive controversial opinions. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - Proc. of the 9th Int. Conf. of the CLEF Association. pp. 135–140. CLEF 2018, Springer-Verlag (2018). https://doi.org/10.1007/978-3-319-98932-7_13
21. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 93–102. NAACL-HLT 2015, Association for Computational Linguistics, Denver, Colorado (2015). https://doi.org/10.3115/v1/N15-1010
22. Siagian, A.H.A.M., Aritsugi, M.: Combining word and character n-grams for detecting deceptive opinions. In: 2017 IEEE 41st Annual Computer Software and Applications Conference. pp. 828–833. COMPSAC, IEEE Computer Society, Washington, DC, USA (2017). https://doi.org/10.1109/COMPSAC.2017.90

23. Siagian, A.H.A.M., Aritsugi, M.: Exploiting function words feature in classifying deceptive and truthful reviews. In: 2018 Thirteenth International Conference on Digital Information Management. pp. 51–56. ICDIM, IEEE Computer Society, Washington, DC, USA (2018). https://doi.org/10.1109/ICDIM.2018.8846971
24. Siagian, A.H.A.M., Aritsugi, M.: Robustness of word and character n-grams combinations in detecting deceptive and truthful opinions. J. Data and Information Quality (In press). https://doi.org/10.1145/3349536
25. Zaghouani, W., Charfi, A.: Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. LREC, European Languages Resources Association (ELRA), Paris, France (2018), https://www.aclweb.org/anthology/L18-1111
26. Zaghouani, W., Charfi, A.: Guidelines and annotation framework for arabic author profiling. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. LREC, European Language Resources Association (ELRA), Paris, France (2018), http://lrec-conf.org/workshops/lrec2018/W30/pdf/5_W30.pdf