Fake or Not: Distinguishing Between Bots, Males and Females

Notebook for PAN at CLEF 2019

Matej Martinc[1,2], Blaž Škrlj[1,2], and Senja Pollak[1,3]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[3] Usher Institute of Population Health Sciences, Medical School, University of Edinburgh, UK
matej.martinc@ijs.si, blaz.skrlj@ijs.si, senja.pollak@ijs.si

Abstract For the PAN 2019 Author profiling task, we present a two step author profiling system which in the first step distinguishes between bots and humans, and in the second step determines the gender of the human authors. The system relies on a Logistic Regression classifier and employs a number of different word and character n-gram features and a simple type-to-token-ratio feature, which proved useful for the bot prediction task. Experiments show that on the provided datasets of tweets, distinguishing between bots and humans is an easier task than determining the gender of the human authors. The proposed approach was 16^{th} in the global ranking of PAN 2019 Author profiling shared task.

CEUR-WS.org/Vol-2380/paper_204.pdf

1 Introduction

Social media enables members to interact and share content in an online environment but has recently seen a rise in automated social accounts linked to spamming, fake news dissemination and even manipulation of public opinion. This has had a negative effect on the level of the online discourse and also threatens services such as advertising and search for reliable content [3]. To counteract this tendency, social media companies and the research community have proposed several approaches to identify these bots automatically. This detection relies on differences in content produced by humans and bots and also on differences in an online behaviour.

Once a social media user is successfully identified as human, another field of research, generally known as author profiling (AP), deals with learning about the demographics and psychological characteristics of a person based on the text she or he produced. This type of research has already shown a potential for applications in marketing, social and psychological research, security, and medical diagnosis. The most commonly addressed task in AP is the prediction of an author's gender, which has been the main focus of a series of scientific events and shared tasks on digital text forensic called PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse)[4] since

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

[4] http://pan.webis.de/

2011, when the first PAN event took place. The first AP shared task was organized in 2013 [19].

In this paper, we describe our approach to the PAN 2019 AP shared task [18] which deals with the construction of a two step prediction model. In the first step, the system distinguishes between bots and humans and in the second step it determines the gender of human Twitter users. The rest of the paper is structured as follows: in Section 2 the findings from related work are presented. Section 3 describes the corpus and how it was preprocessed. In Section 4 we present the methodology, Section 5 presents the results, while in Section 6 we display the results of the conducted ablation study. In Section 7, we conclude the paper and present ideas for a future work.

2 Related Work

A very successful strategy for detecting bots on Twitter was proposed by [9] and is based on the deployment of honeypots for harvesting deceptive spam profiles on social media. Harvested spammers were then analyzed and findings were used in the implementation of classifiers capable of detecting new bot spammers. For classification, they used text features such as n-grams and also meta statistical features, such as the ratio between the number of URLs in the 20 most recently posted tweets and the number of tweets, and the ratio between the number of unique URLs in the 20 most recently posted tweets and the number of tweets. They report the F1-score of 88.80% achieved with the Weka Decorate meta-learner. A more recent classification approach which relied on statistical meta features (age of the account, number of tweets, followers-to-friends ratio, retweets per tweet...) was proposed by [6]. They achieved an accuracy of 86.44% in the 5-fold cross validation setting with a Random Forest classifier.

Another interesting approach was proposed by [5] who among other features (e.g., average number of hashtags and repeated tweets, latent Dirichlet allocation identified topics, graph-theoretic statistics...) also leveraged sentiment-related factors for bot identification. They used a Gradient boosting classifier and also employed statistical features derived from text, such as average number of hashtags, average number of user mentions, links and emoticons.

Traditional classifiers with extensive feature engineering seem to be pervasive in the literature about distinguishing between bots and humans but there was also some attempts to tackle the task with neural networks. [3] proposed a behavior enhanced deep model (BeDM) that regards user content as temporal text data instead of plain text and fuses content information and behavior information using a deep learning method. They report an F1-score of 87.32% on a Twitter dataset.

Gender prediction became a mainstream research topic with the work by Koppel et al. [7]. Based on experiments on a subset of the British National Corpus, they found that women have a more relational writing style (e.g. using more pronouns) and men have a more informational writing style (e.g. using more determiners). Later gender prediction research remained focused on English, yet the attention quickly shifted to social media applications [2,23,15] and other languages. The most relevant findings for the gender classification task at hand comes from PAN shared tasks in 2016 and 2017 [21,20], where one of the goals was to predict gender of the user on English and

Table 1. PAN 2019 training set statistics

Language	Bots	Male humans	Female humans	All authors	All tweets
English	2,060	1,030	1,030	4,120	412,000
Spanish	1,500	750	750	3,000	300,000

Spanish tweet datasets. In PAN 2016, the best score was achieved by [13], who used word unigrams, word bigrams and character tetragrams features. They used Logistic Regression classifier for learning. A somewhat similar Support vector machine (SVM) based system with simpler features (word unigrams, bigrams and character three- to five-grams) was used by the winners of the PAN 2017 competition [1]. Second ranked team in the PAN 2017 competition [11] also used a combination of word and character n-grams [11], as well as POS n-grams, sentiment from emojis and character flooding as features in the Logistic Regression classifier.

3 Dataset Description and Preprocessing

PAN 2019 training set consists of tweets in English and Spanish languages grouped by tweet authors (100 tweets per author) with gender and type labels (Table 1). Gender and type categories are balanced in both languages. We used this training set in our experiments for feature engineering, parameter tuning and training of the classification models.

First, all tweets belonging to the same author are concatenated and used as one document in further processing. After that, three distinct dataset transformations were employed on the documents and all these three levels of preprocessing were used in the feature engineering step:

- *Cleaned level*: replacing all hashtags, mentions and URLs with specific placeholders #HASHTAG, @MENTION, HTTPURL, respectively.
- *No punctuation level*: removing punctuation from the cleaned level;
- *No stopwords level*: stopwords are removed from the no punctuation level.

4 Feature Construction and Classification Model

Our feature construction and classification approach can be considered a simplification of the approach we used in the PAN 2017 AP shared task [11], since the winners of the PAN 2017 competition [1] conducted experiments which suggest that adding too sophisticated features negatively affects the performance of the gender classification model. For this reason, our model mostly relies on different types of n-grams and the hypothesis was, that the simplification of the model would also improve the performance of the bot classification model.

4.1 Features

The following n-gram features were used in our final model:

- *word unigrams*: calculated on lower-cased no stopwords level, TF-IDF weighting (parameters: minimum document frequency = 10, maximum document frequency = 80%);
- *word bigrams*: calculated on lower-cased no punctuation level, TF-IDF weighting (parameters: minimum document frequency = 20, maximum document frequency = 50%);
- *word bound character tetragrams*: calculated on lower-cased cleaned level, TF-IDF weighting (parameters: minimum document frequency = 4, maximum document frequency = 80%);
- *suffix character tetragrams* (the last four letters of every word that is at least four characters long [22]): calculated on lower-cased Tweets-cleaned, TF-IDF weighting (parameters: minimum document frequency = 10%, maximum document frequency = 80%).

The only somewhat more sophisticated feature used in the experiments was calculated on the cleaned level and was inserted in order to improve the performance of the bot classification model:

- *Type-to-token ratio*: calculated by dividing the number of distinct words in the document by the number of all words in the document. The intuition behind this feature is that bots tend to have a higher word repetition frequency and limited vocabulary, therefore low type-to-token ratio could be a good indication that text was produced by a non-human.

All features were normalized with the MinMaxScaler from the Scikit-learn library [14]. For example, a vector x was rescaled as:

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)}; \tag{1}$$

yielding feature in range [0,1] (if feature values are all positive).

4.2 Classification Model

Several classifiers from Scikit-learn and libSVM were tested:

- Linear SVM [4]
- SVM with RBF kernel [4]
- Logistic Regression [14]
- Random Forest [14]
- Gradient boosting [14]

We performed an extensive grid search to find the best hyper-parameter configuration for all tested classifiers. Best results were obtained with the Logistic Regression with C=1e2 and fit_intercept=False parameters. The Scikit-learn FeatureUnion[5] class also allows to define weights for different types of features we used, which influence the penalties given to specific features during the training process. The weights were adjusted with the help of the following procedure already described in [11]:

1. Initialize all feature weights to 1.0.
2. Iterate the list of features. For every feature repeat adding or subtracting 0.1 to the weight until the accuracy on the validation set is improving. When the best weight is found, move to the next feature on the list.
3. Repeat step 2 until the accuracy cannot be improved anymore.

The weights in our final Logistic Regression model were the following:

- word unigrams and word bound character tetragrams: 0.8
- suffix character tetragrams: 0.4
- type-to-token ratio: 0.3
- word bigrams: 0.1

This weight configuration proved optimal for both classification tasks and both languages and is almost identical to the configuration used in [11].

5 Experiments and Results

English and Spanish tweet datasets were split into train (containing 2,880 authors for English and 2,080 authors for Spanish) and validation (containing 1,240 authors for English and 920 authors for Spanish) sets according to the recommendation of the PAN organizers to avoid overfitting. In the training and validation experiments, gender and bot classification are considered as separate problems, while the predictions on the official test sets were generated in a sequential order, by first determining if an author is either a human or a bot and then conducting gender classification for authors identified as humans. Results of the experiments on the unofficial validation sets and official test sets in terms of accuracy are presented in Table 2. Both classes are balanced, so for bot and gender classification the majority classifier's accuracy is 0.50. On the unofficial validation sets, distinguishing between bots and humans is an easier task for the classifier, achieving 90.16% accuracy on English and 88.04% accuracy on Spanish. Accuracies for gender classification are lower with the classifier achieving 79.52% accuracy on English and 66.96% accuracy on Spanish. This difference in accuracy could also be partially contributed to smaller training set sizes for gender classification. The Spanish gender classification results are also much lower than previous results with a very similar classifier achieved in the scope of the PAN 2017 gender profiling task [11].

On the official test sets, the accuracies of English and Spanish bot classification are lower (89.39% and 87.44% respectively), which might suggest some overfitting. On the other hand, gender classification results are better on the official test sets for both

[5] http://scikit-learn.org/stable/modules/generated/sklearn.pipeline.FeatureUnion.html

languages. While on the English official gender classification test set the accuracy is marginally better, the difference on Spanish is almost 9 percentage points. All in all, we ranked 16^{th} in the official TIRA [16] evaluation, beating the LDSE [17] and word embeddings baselines but falling behind the character and word n-gram baselines.

6 Ablation Study

In order to evaluate the contribution of type-to-token and n-gram features in both classification tasks, an ablation study was conducted. Table 3 presents results for three feature configurations. While using only the type-to-token ratio feature for classification produces classification accuracies very similar to the majority classifier (see column *No n-grams*), combining this feature with n-gram features on average improves bot classification accuracy by 0.35 percentage point. On the other hand, type-to-token ratio feature negatively affects gender classification accuracy, reducing it on average by 0.34 percentage point.

The largest gains in accuracy, when the type-to-token ratio feature is used, are achieved on the English bot classification task (gain of 0.81 percentage point). On the other hand, on the Spanish bot classification task the type-to-token ratio feature marginally reduces the accuracy (reduction of 0.11 percentage point) of the classifier. When it comes to gender classification, the results of the ablation study show that the type-to-token ratio feature has a marginally positive effect on the Spanish dataset (gain of 0.21 percentage point) but also reduces the accuracy of the English gender classification by about 0.5 percentage point.

7 Conclusion and Future Work

In this paper we have presented our approach to the PAN 2019 AP task, which deals with distinguishing between humans and bots and with determining the gender of the human authors. First we presented findings from the related work that were considered during the planning phase of our research and influenced this research the most. After that, we described the datasets used in our experiments, the preprocessing and feature engineering techniques used, and the classification algorithms employed in our experiments. Finally, we presented the experiments together with results on the unofficial validation sets.

According to our experiments, distinguishing between bots and humans is a somewhat easier task than distinguishing between male and female humans. We also used

Table 2. Accuracy results on the unofficial validation set and the official test set

	Unofficial		Official	
	Bot	Gender	Bot	Gender
English	0.9016	0.7952	0.8939	0.7989
Spanish	0.8804	0.6696	0.8744	0.7572

exactly the same approach for both classification tasks, even though some related work suggested different sets of features for these two tasks. Different types of word and character n-grams proved as the most useful features in both tasks. There is however a difference in effect of the type-to-token ratio feature when it comes to both tasks. While this feature negatively affects the accuracy of the gender classifier, it does improve the accuracy of the bot classifier by 0.35 percentage point.

Another interesting observation is that even though we conducted an extensive grid search to find the best classifier with the best configuration of hyper-parameters, the final choice is identical to the choice of a classifier and hyper-parameters used in our previous study of gender classification [11], despite the additional problem of bot classification. In addition, very similar setting was also selected as best for our approach in the PAN 2019 Celebrity profiling task [12].

We believe an unexploited opportunity is the body of semantic background knowledge, such as for example the word taxonomies. Approaches such as SRNA [24] could be used to investigate, whether such knowledge contributes to learning for the task at hand.

Another line of future work will deal with the evaluation of the model on additional datasets from other social media platforms besides Twitter in order to test how well our model generalizes across different social media content. For gender identification, online workflows have been proposed [10] in the ClowdFlows environment [8] and we plan to expand the set of workflows to also cover bot identification.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency core research programme Knowledge Technologies (P2-0103). The work of the second author was funded by the Slovenian Research Agency through a young researcher grant. This paper is also supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153 - project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

1. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-gram: New groningen author-profiling model. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. (2017)

Table 3. Results of the ablation study on the unofficial validation set

	All features		No type-to-token ratio		No n-grams	
	Bot	Gender	Bot	Gender	Bot	Gender
English	0.9016	0.7952	0.8935	0.8000	0.5040	0.4984
Spanish	0.8804	0.6696	0.8815	0.6717	0.5021	0.5000
Average	**0.8910**	0.7324	0.8875	**0.7358**	0.5031	0.4992

2. 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)
3. Cai, C., Li, L., Zengi, D.: Behavior enhanced deep bot detection in social media. In: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI). pp. 128–130. IEEE (2017)
4. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2(3), 27 (2011)
5. Dickerson, J.P., Kagan, V., Subrahmanian, V.: Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In: Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 620–627. IEEE Press (2014)
6. Gilani, Z., Kochmar, E., Crowcroft, J.: Classification of twitter accounts into automated agents and human users. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 489–496. ACM (2017)
7. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
8. Kranjc, J., Podpečan, V., Lavrač, N.: ClowdFlows: A cloud based scientific workflow platform. In: Flach, P.A., Bie, T.D., Cristianini, N. (eds.) Proc. of ECML/PKDD (2). LNCS, vol. 7524, pp. 816–819. Springer (2012)
9. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 435–442. ACM (2010)
10. Martinc, M., Pollak, S.: Reusable workflows for gender prediction. In: chair), N.C.C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Paris, France (may 2018)
11. Martinc, M., Škrjanec, I., Zupan, K., Pollak, S.: Pan 2017: Author profiling-gender and language variety prediction. CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers (2017)
12. Martinc, M., Škrlj, B., Pollak, S.: Who is hot and who is not? profiling celebs on twitter. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019) (2019)
13. Modaresi, P., Liebeck, M., Conrad, S.: Exploring the effects of cross-genre machine learning for author profiling in PAN 2016. CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers (2016)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 12, 2825–2830 (2011)
15. Plank, B., Hovy, D.: Personality traits on twitter -or- how to get 1,500 personality tests in a week. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA). Lisbon, Portugal (2015)
16. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF. Springer (2019)
17. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 156–169. Springer (2016)

18. Rangel, F., Rosso, P.: Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)
19. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. Notebook Papers of CLEF pp. 23–26 (2013)
20. 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.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2017)
21. 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: CLEF 2016 Working Notes. CEUR-WS.org (2016)
22. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015. pp. 93–102 (2015), http://aclweb.org/anthology/N/N15/N15-1010.pdf
23. 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) (2013)
24. Škrlj, B., Kralj, J., Lavrač, N., Pollak, S.: Towards robust text classification with semantics-aware recurrent neural architecture. Machine Learning and Knowledge Extraction 1(2), 575–589 (2019), http://www.mdpi.com/2504-4990/1/2/34