Bot Or Not: A Two-Level Approach In Author Profiling

Notebook for PAN at CLEF 2019

Flóra Bolonyai, Jakab Buda, Eszter Katona

Eötvös Loránd University, Budapest
f.bolonyai@gmail.com
bakajb@gmail.com
katona.eszter.rita@gmail.com

Abstract. In this notebook, we summarize our work process of preparing a software for the PAN 2019 Bots and Gender Profiling task. We propose a Machine Learning approach to determine whether an unknown Twitter user is a bot or a human, and if the latter, their gender. We use logistic regressions to identify whether the author is a bot or a human and we use neural networks to attribute their gender. We were able to achieve an accuracy of 91%, 83% for bot/human and 75%, 69% for gender in English and Spanish respectively.

1 Introduction

Author Profiling is a fast-breaking scientific application in line with the expanding usage of the "Big Data" paradigm. Originally, public texts were produced only by professional authors (such as journalists, scientific / literary authors). However, as the internet spread, the amount of content produced by everyday users skyrocketed. Website texts, blogs, comments and social media posts now account for a significant part of digitally produced contents. Although these texts are unstructured, they contain data about people's opinions, preferences, attitudes, and even actions in an enormous amount compared to pre-internet times.

As the online community is growing, we face increasing threats from the anonymity of online life. If we think of politics, bots can influence the outcome of elections, or they can create and share fake news. Also, for example, identifying sexual predators can be an important task, thus, identifying potential anomalies between self-declared and true data has an increasingly important role.

The aim of the PAN 2019 Bots and Gender Profiling task [6, 18, 20] was to investigate whether the author of a given Twitter feed is a bot or a human, and in case of human, identify their gender. The training and test sets of the task consisted of English and Spanish Twitter feeds.

We followed a mixed approach and created models on two levels, for both Spanish and English tweets. To identify bots, we used two logistic regressions; one on the level of tweets and another one on the level of accounts. To identify the gender of the author, we first fitted neural networks that predicted the gender of the author of each tweet and then aggregated the 100 tweet level results to get an account level prediction for each author. Our final software for distinguishing between bots and humans performed well on the test set, but our model was less accurate for the task of discriminating between male and female Twitter users.

In Section 2 we present the related work on author profiling. In Section 3 we describe our approach in detail, including the extracted features and the fitted models. Section 3 consists of two subsections, one for the bot-or-human distinction and one for the gender attribution. In Section 4 we present our results. In Section 5 we discuss some potential future work and in Section 6 we conclude our notebook.

2 Related works

There are multiple approaches to collect information about the latent characteristics of the author of a text. There is no one best technique, and the literature shows high variance in the performance of different methods on different corpora. It is possible to roughly define two groups of variables that can be used to collect information about the authors of texts: most such variables are either dictionary or style-based.

A well performing model in gender differentiation based on n-grams was built by Boulis and Ostendorf [4], who also emphasized that tf-idf representation was not necessarily a useful measure to find differences between the texts of male and female authors, as the most used words can be very different for the two groups and therefore inverse weighting can be less effective.

Garera and Yarowsky [7] found that removing stop-words and lemmatizing were not useful when trying to differentiate between texts written by men and women as the distribution of stop-words and certain grammatical forms differ in the case of female and male authors.

Peersman, Daelemans and Varenbergh [16] compared the performance of SVM models based on character and word n-grams to predict the age and gender of social media post authors. Their results show that token-based variables are more informative than character-based ones.

Schler, Koppel, Argamon and Pennebaker [21] analyzed blog posts to gain information about the relationship between style and content-based variables and age and gender. They found that women use more pronouns and words that express emotions, agreements and disagreements, while men use more articles, prepositions and links. In their case, style-based variables proved to be more informative than content-based ones.

Goswami, Sarkar and Rustagi [8] looked for stylometric differences by age and gender by including slang words and average sentence length as new explanatory variables. Their results show that the frequency of certain slang words is very different by both age and gender but there is no significant difference regarding the length of sentences among the groups.

Word embeddings is another approach that does not belong to either of the categories above. Embeddings capture the essence of words well, and therefore, by combining these representations with neural networks, it is possible to gain knowledge about the latent characteristics of a text, among others information about its author. In the 2018 PAN competition, multiple well performing models used neural networks based on word embeddings to classify texts by the gender of their authors. However, these models did not clearly prove to be superior to traditional machine learning ones regarding gender classification [19].

Overall, there is no consensus about the types of variables and models that work best in identifying latent characteristics of authors, and therefore our approach was to gain as much information as possible from the train corpora.

3 Our approach

For the two tasks of the competition, i.e. bots and gender profiling, we trained two substantially different models. To differentiate between humans and bots, we used a system of two logistic regressions based on features extracted from the texts, whereas to classify authors by gender, we implemented a recurrent neural network based on word embeddings. In the following sections, we provide a detailed overview of our methods for both tasks. Our codes are available on GitHub[1].

3.1 Identifying bots

Features

As our classifier system consists of two logistic regressions, one that predicts per tweet and one that predicts per author, we created variables on two levels. On the one hand, we extracted features on the level of tweets, and on the other, we also created some aggregate features on the level of authors. The features were slightly different on the two levels. For example, we investigated on the tweet-level if there was another user tagged in the tweet and counted on the account-level how many different people a user tagged. For both Spanish and English tweets we extracted the same features.[2]

It should be noted that we had no internet connection during testing the software, so we could not include some of the planned information (such as expanding and examining the shared links).

To extract features, we primarily utilized online available Python packages[3] and in some cases regex.

We distinguish between 3 types of features. For some of our features, we had to use predefined dictionaries, so we call those dictionary-based features. Another group

[1] https://github.com/pan-webis-de/bolonyai19

[2] For efficient data handling we used *numpy* [22] and *pandas* [14] packages

[3] We used libraries from the following packages: *spanish_sentiment_analysis* [9], *emoji* [11], *spacy* [10], *lexical_diversity* [12], *NLTK* [3], *textblob* [13]

of features consists of those that describe the tweets grammatically and structurally, which we call style-based features. Finally, we differentiate a third group of variables which describe some meta information that could be extracted from the tweets. The extracted features are summarized in the following tables.

3.1.1. Feature extraction on tweet-level

Table 1: Dictionary based features

Feature	***Package***	***Info***
Emojis	emoji [11]	We checked how many different emojis were used in the texts and we also created a variable for the proportion of emojis in the tweets.
Proportion of stopwords	stopwords from NLTK [3]	We used the built-in stopword lists for both languages from the NLTK package.
Sentiment score	textblob [13] spanish-sentiment-analysis [9]	We used the polarity information for the English tweets based on the textblob package and the sentiment score based on spanish-sentiment-analysis for the Spanish tweets.
Number of misspelled words	brown from NLTK [3] cess_esp from NLTK [3]	Our assumption is that humans misspell words but bots do not. Misspelled words can be used as an indicator for using short forms of expressions, too.

Table 2: Style based features

Feature	***Package***	***Info***
Lexical diversity	Lex_div from lexical_diversity [12]	There are various methods to measure lexical diversity (e.g. simple ttr, log ttr); we used root ttr in our analysis.
POS-features	Spacy [10]	We identified the word class of each word and created features measuring the: • proportion of nouns • proportion of verbs • proportion of adjectives in the tweets.
Text characteristics	regex [2]	• Number and proportion of apostrophes • Number and proportion of uppercase letters • Number and proportion of numbers • Number and proportion of points • Number and proportion of commas

		• Number and proportion of punctuation marks • Number and length of words • Number of letters • Length of sentences • Number of character flooding usage

Table 3: Meta features

Feature	***Package***	***Info***
Retweet	regex [2]	Dummy for retweet
Links	regex [2]	Number of shared links
Calls	regex [2]	Number of mentions

3.1.2. Feature extraction on author-level

Table 4: Author-level features

Feature	***Package***	***Info***
Retweet	regex [2]	• Number of retweets • Number of different users retweeted
Calls	regex [2]	• Number of tags • Number of different users tagged

Logistic regression models

To differentiate between humans and bots, we fitted two logistic regressions[4] for each language on the provided training set. First, we fitted a logistic regression using a total of 30 features extracted from individual tweets. This model predicted separately for each tweet whether its author was a human or a bot. In our second logistic regression, we used two types of explanatory variables: some of them were collected from the original texts (e.g. the number of different usernames that occurred in the tweets of an author), while other features came from the results of our first logistic regression. The latter group consisted of the minimum, maximum, median, mean, standard deviation and rounded mean of the hundred predictions for each author. The structure of our system is illustrated by Figure 1.

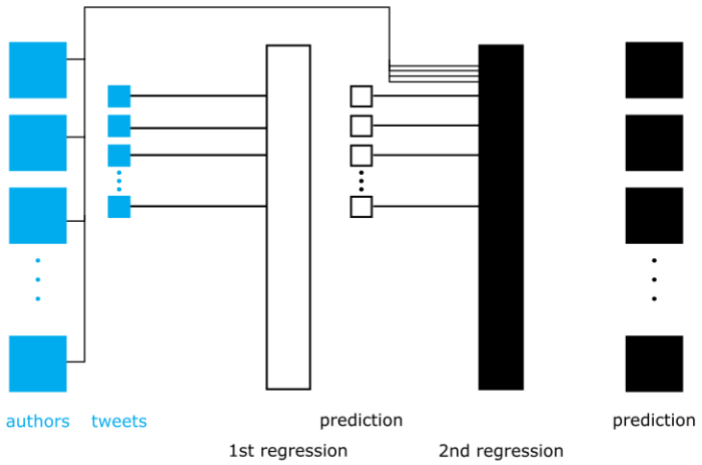
To avoid overfitting, we tuned the hyperparameters of both logistic regression classifiers based on their performance on the development set. We applied grid search using some sensible parameters (i.e. C = {10-5, 10-4, 10-3, 10-2, 10-1, 100, 101, 102,103,104,105}; Intercept = {True, False}). Our final sets of hyperparameters of the models are summarized in Table 5.

[4] We used the built-in logistic regression from *scikit-learn* [15].

Table 5: Final set of hyperparameters

Model	*Regularization strength*	*Intercept*
English tweet-based	C = 1	True
English aggregate	C = 1	True
Spanish tweet-based	C = 0.001	False
Spanish aggregate	C = 0.001	True

Figure 1: Our system of logistic regressions for differentiating bots and humans



3.1 Gender attribution

On the gender attribution task, the method used for bot identification did not yield satisfactory results. Based on related works we trained a recurrent neural network[5] on the word embedding representation of the texts. Although logistic regression based on text features did not work well for gender attribution, we kept the system of first training a model for the tweets individually than aggregating the results of this model for each author for the final classification. We used the 25 dimensional pretrained GloVe vectors [17] trained on tweets as representation for both English and Spanish texts. The 25d GloVe are space and time efficient and they contain a large number of Spanish words, despite the fact that the embedding was primarily trained on English tweets. When we experimented with a higher dimensional embedding space, it did not perform significantly better on the development set. Before transforming the texts to the vector space, we applied minimal preprocessing. We replaced all links with the

[5] We used *tensorflow* and *keras* to fit our neural networks [1, 5].

"https" string, and all mentions with the "@" character, and finally separated all non-alphanumeric characters from the words to form a separate word.

The word counts of tweets vary a lot (in the English training set the mean is 18.5 and the maximum is 97), but equal input length is required for computational efficiency. Thus, we had to pad or truncate the tweets to the same length. Because of the great variance in the word counts, padding all tweets to the length of the longest tweet would yield more padding tokens than actual words in the case of most tweets, hence rendering the training slow and inefficient. We chose 38 tokens as common length for training the neural network as 90 percent of the tweets are shorter, and the longer ones are generally tweets with many tags, which, with the embedding used, do not contain much information. We padded the end of the shorter tweets with padding tokens (0 vectors) and truncated the longer ones. Based on experimental training sessions, truncation of the beginning of the tweets gave the best results (this is probably due to the fact that longer tweets are long because they have a lot of mentions at the beginning, which do not attribute to the character count, and we used the embedding of the "@" symbol for all mentions), so we kept only the end of the longer tweets.

During the training of the RNN on the full training set, despite the various levels and types of regularizations tried, we observed heavy overfitting. This can be probably attributed to the fact that there are relatively few authors, each with many tweets in the training set, and authors have a more distinct tweeting pattern than genders. As a result, the network can learn to identify each author and attribute a gender to the authors more easily than learn the distinction between the genders, but this cannot be generalized to new authors. To avoid this possibility, during the first part of the training of the RNN we used only 1/10 of the training set, randomly selecting 10 tweets from each author. After achieving convergence on this training set, we continued the training of the RNN on the full set for a few epochs.

After some experimental training with different RNN architectures, our best performing RNN on the English dev set was a unidirectional RNN with GRU units (with recurrent dropout value of 0.35) followed by a dropout layer ($p = 0.5$) and a sigmoid unit. On the Spanish texts we used a slightly different architecture, a bidirectional RNN. On the English set, after a total of 110 epochs, the performance of the model converged. The Spanish model converged after a total of 140 epochs.

Following the tweet level prediction of the RNN, for the English texts we did not found a better performing aggregation method than computing the rounded mean of the tweet-level predicted probabilities and interpreting this as the final predicted probability for each author. For the Spanish tweets, we trained a similar logistic regression as for the bot prediction. For input variables we used the mean, the deviation, the minimum and the maximum of the tweet-level predictions for each author.

4 Results

In this section, we provide a detailed description of the performance of our classifiers. In all cases, i.e. for bots and gender profiling and for English and Spanish

language, we used the default development set defined by the hosts of the competition to tune the hyperparameters of the models before submitting our software on TIRA [18] and testing it on the real test set. Therefore, we report two sets of results for each model: performance on the pre-defined dev set, and performance on the actual test set. For both the development and test sets, we report the accuracy score as a performance metric. As we used a two-level approach to train our classifiers, we also report about the performance of our tweet-based classifiers on the development set.

Table 6: Development set performance: Identifying bots

Model	***Accuracy***
English tweet-based	0.79
English aggregate	0.90
Spanish tweet-based	0.75
Spanish aggregate	0.85

Table 7: Development set performance: Gender attribution

Model	***Accuracy***
English tweet-based	0.59
English aggregate	0.80
Spanish tweet-based	0.58
Spanish aggregate	0.65

Table 8: Test set performance

Task	***Language***	***Accuracy***
Bots vs. humans	English	0.9136
	Spanish	0.8389
Female vs. male	English	0.7572
	Spanish	0.6956

It is clear that adding the second layer of regressions accounts for significant improvements, particularly when differentiating between bots and human authors. However, we did not experiment with pure aggregate models, so we do not know how our two-level approach would perform against classifiers that use only author-level features.

5 Future works

Although our logistic regression gave encouraging results in identifying bots and humans, it did have a shortcoming, i.e. extracting features from the texts was rather slow. This time we did not experiment with feature selection, but it is likely that our features are not all significant predictors to differentiate bots from humans.

As our models for English tweets generally outperformed the ones for Spanish tweets, it is likely that some of the packages we used for feature extraction are more reliable for English texts than for Spanish ones. For example, in the case of POS tagging, the function we used for English texts was based on a corpus from the web in general, while the one we used for Spanish texts was trained merely on Spanish news.

To achieve better performance with the RNN (aside from using word vectors trained explicitly on the language used), one possible solution could be to construct a deeper network with two or three layers or use the sequence returned by the RNN as the input data for another model. Although it could increase the risk of overfitting, this could be compensated by changing the random subset of the training set multiple times during the initial training. Using a higher dimensional word representation and training more epochs could also yield better accuracy, but at a great computational cost.

6 Conclusion

In this notebook, we summarized our work process of preparing a software for the PAN 2019 Bots and Gender Profiling task [6, 18, 20]. Overall, we followed different approaches for the two tasks: to differentiate between bots and humans, we used logistic regressions with mostly text based explanatory variables, and to differentiate between female and male authors, we trained recurrent neural networks based on word embeddings. In both cases, we built classifiers on two levels. First, we fitted models to predict a response for individual tweets. Second, we created an aggregate classifier that gave us a prediction for each author. In the case of bots vs. humans, we used logistics regressions to get our final predictions. Besides the descriptive statistics of the tweet-level predictions, we also included some author-level features extracted from the texts as explanatory variables. To predict the gender of the author, we used different approaches for the English and Spanish texts. For English tweets, we simply took the rounded average of predictions of all tweets belonging to an author. For Spanish tweets, we again opted for a logistic regression, using descriptive statistics of the tweet-level predictions as input variables.

Our results show that our classifiers for English tweets tend to outperform our classifiers for Spanish tweets. Additionally, we achieved a higher accuracy in identifying humans and bots than in identifying the gender of the authors.

7 References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems, (2015).

2. Barnett, M: regex. https://pypi.org/project/regex/ (2019)

3. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009)

4. Boulis, C., Ostendorf, M.: A Quantitative Analysis of Lexical Differences Between Genders in Telephone Conversations. in: Proceedings of the 43rd Annual Meeting of ACL, pp. 435-442. (2005)

5. Chollet, F. et al.: Keras. https://keras.io (2015)

6. Daelemans, W., Kestemont, M., Manjavancas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Tschuggnall, M., Wiegmann, M., Zangerle, E.: Overview of PAN 2019: Author Profiling, Celebrity Profiling, Cross-domain Authorship Attribution and Style Change Detection. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D., Heinatz, G., Cappellato, L., Ferro, N. (eds.) Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Springer (2019)

7. Garera, N, Yarowsky, D.: Modeling latent biographic attributes in conversational genres. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Vol 2, pp. 710-718. (2009)

8. Goswami, S., Sarkar, S., Rustagi, M.:Stylometric Analysis of Bloggers' Age and Gender. Pattern Recognition and Machine Intelligence, Third International Conference, PReMI, New Delhi, India, December 16-20, 2009 Proceedings, pp. 205-212. (2009)

9. Hofman, E: spanish-sentiment-analysis 1.0.0. https://pypi.org/project/spanish-sentiment-analysis/ (2018)

10. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear. (2017)

11. Kim, T., Wurster, K.: emoji. https://pypi.org/project/emoji/ (2019)

12. Kyle, K.: lexical-diversity. https://pypi.org/project/lexical-diversity/ (2018)

13. Loria, S.: textblob Documentation, Release 0.15.2 (2018)

14. McKinney, W.: Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, pp. 51-56. (2010)

15. 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, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, pp. 2825-2830. (2011).

16. Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting age and gender in online social networks. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents, pp. 37-44. (2006)

17. Pennington, J., Socher, R., Manning, C. D.: GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543. (2014)

18. 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)

19. Rangel, F., Rosso, P., Franco, M. A Low Dimensionality Representation for Language Variety Identification. In: Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'16), Springer-Verlag, LNCS (9624), pp. 156-169. (2018)

20. Rangel, F., Rosso, P. Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In: Cappellato L., Ferro N., Müller H, Losada D. (Eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org. (2019)

21. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In AAAI spring symposium: Computational approaches to analyzing weblogs, vol. 6, pp. 199-205. (2006)

22. van der Walt, S., Colbert, S. Ch., Varoquaux, G.: The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering, 13, pp. 22-30. (2011)