Supervised Classification of Twitter Accounts Based on Textual Content of Tweets

Notebook for PAN at CLEF 2019

Fredrik Johansson

Swedish Defence Research Agency (FOI)
fredrik.johansson@foi.se

Abstract In our implemented system submitted to the bots and gender profiling task of PAN 2019, we use a two-step binary classification approach in which we classify accounts as being bot or not based on a combination of term occurrences and aggregated statistics fed to a random forest classifier. Accounts classified as human are further distinguished as male or female through a logistic regression classifier taking data-driven function words as input. We obtain highly competetive bot and gender classification accuracies on English (0.96 and 0.84, resepectively) while performing worse on Spanish (0.88 and 0.73, respectively).

1 Introduction

In the modern society a large number of communication and interactions take place on various social media platforms such as Twitter and Facebook. Social media analytics is being used for everything from finding out what people think about a specific brand or product [16] to estimating citizens' reactions to crisis events [2]. Profiling of authors (or their corresponding user accounts) is an important part of the analytical process, since the presence of e.g., Twitter bots otherwise can give a skewed view of things like the true public opinion on a product or a political party. Likewise, it can be benificial to be able to estimate whether a social media account is likely to belong to a man or a woman, since gender is often an important variable for product marketers, political campaign staff, or crisis managers to keep track of. In the PAN 2019 bots and gender profiling task [10], the challenge is to given a Twitter feed, determine whether its author is a bot or a human, and in case of a human, to identify the gender (male or female) of the author. Since social media is highly multilingual, there is both an English and a Spanish subset of the task. To further increase the complexity of the task, only the textual content of the tweets is available, i.e., no metadata or other social network information are available during neither training nor testing.

Our implemented solution is intended to be as generic as possible so that it is useful also for other languages than English and Spanish, as well as being useful for other types of Twitter accounts than the specific distribution of accounts present in this challenge. For this reason we have attempted to only include features which are likely to

generalize, rather than to only aim for as high predictive performance as possible on the PAN dataset. The rest of this paper is structured as follows. We start by giving a brief overview of related work in Section 2. In Section 3 we present the overall design choices of the implemented solution, such as how we have designed the classification problem to solve. Next, we are in Section 4 presenting the features and classifiers used in our implemented system. The obtained experimental results are described in Section 5 and we present overall conclusions in Section 6.

2 Related Work

Twitter is a social media platform on which bots have been present for a long time. There are several types of bots, ranging from spam bots to bots spreading malware and automated accounts used for various kinds of information operations in political and military conflicts [13]. An exhaustive overview of previous work on detecting bots on Twitter is outside the scope of this paper, but most of them rely on machine learning classifiers applied to metadata-based features such as account age, the used client, the ratio between followers and friends, the ratio between tweets and retweets, and average number of tweets per hour or day. One early example of such an approach in which the authors classify a Twitter account as being either human, bot, or cyborg is the work by Chu et al. [4]. In 2015, DARPA conducted a Twitter Bot Detection Challenge, in which participating teams were supposed to develop and test the effectiveness of their bot detection methods [12]. The participating teams used metadata-based features as those already discussed, but also network-based features such as degree centrality, number of accounts classified as bots in the cluster a user account belongs to, as well as content-based features [12]. A last example of previous work on bot classification is [14]. They use a total of 1150 features, including features related to user metadata, friends, network, and timing. They also make use of content-based features such as frequency of POS tags and features related to sentiment. However, they restrict their work to English-speaking Twitter users.

Some previous work has also been devoted to gender classification of Twitter users. One well-known example is [3], in which the authors use the textual content as well as user metadata consisting of full name, user name, and the description from the user profile. The authors make use of word- and character-level n-grams. There have also been previous work on classifying Twitter accounts as male or female on PAN. However, this challenge has involved posted tweets as well as user images.

3 Overall Design Choices

A first design choice to make is whether to track the classification problem as a multi-class problem with the three possible classes *male*, *female*, and *bot*, or to treat it as a two-stage binary classification problem with classes *human*, *bot* and *male*, *female*, respectively. We have experimented with both versions, but have in the end decided to model it as a two-stage binary classification problem. The main reason for this is that we in our applied research work have several practical scenarios related to detection of information operations in which we have a need for classifying Twitter accounts as

being bot or not, and a few security-related applications in which we need to separate between male and female owners of Twitter accounts. However, we see few real-world examples in which we would like to be able to do both kinds of classification in the same scenario. Moreover, the task has been presented as a two-stage classification problem in the task description at the PAN web page. This being said, it is not obvious that this problem formulation will yield better results than a multi-class definition of the problem would result in.

In many classification problems the instances are classified independently of each other, but here the individual instances correspond to single tweets while the task is to classify the accounts or authors creating these tweets. Hence, there are two options:

1. To classify tweet by tweet and to somehow weigh these individual classifications into an aggregate classification of the Twitter account as such (e.g., by using majority voting).
2. To first calculate aggregate statistics from all tweets by a specific account and to use these as features on which classifiers can be applied directly.

We have in this work used the latter approach as we think there are many signs of Twitter bots which are not shown on the level of individual tweets, but which become more clear on an aggregate level. An example of this is how similar all tweets from a single Twitter account are. If all tweets are highly similar, this may be a good indicator of this account being a bot, but tweet similarity is not an available feature if classifying the tweets independently of each other. The specific features which have been used in our implementation are described more thoroughly in Section 4.

4 Implemented Features and Classifiers

Our implemented features and classifiers used for distinguishing between bots and humans are first described in Section 4.1. Next, we describe the features and classifiers used for distinguishing between male and female authors in Section 4.2.

4.1 Bot Classification

When designing our text-based bot vs. human features, we have thought about which metadata we previously have found useful for classifying Twitter accounts as bot or human in [13] and reasoned about how such features can be extracted or derived from the textual content alone when no metadata is available. Some of the first identified features on tweet level are tweet length (number of characters) and number of capital letters, as these are thought to be potential indicators for e.g., very basic spam bots. We then calculate aggregate statistics for max, min, avg, and std for an account based on the counts from the tweet level, since the classification is supposed to take place on account level rather than individual tweet level, as discussed previously. In the same manner, we calculate statistics based on serveral other counts:

- Number of URLs (number of words starting with "http")
- Number of mentions (number of words starting with "@")

- Fraction of retweets (number of tokens starting "RT")
- Number of lowercase letters

For all these features we can treat the tweets from a specific account independently from each other. However, we also think that bot behavior can be captured by measuring the similarity among subsequent tweets. For this reason we use the ordering in which the tweets from the same account appear in the dataset (assuming this corresponds to some time ordering) and calculate the edit distance (or more specifically the Damerau-Levenshtein distance [5]) between all consecutive tweets. These distanes are then aggregated into max, min, avg, and std features in the same way as the less complex features.

Finally, an important clue to whether a Twitter account is automated or operated by a human author is the actual wordings of the tweets. For example, tweets containing the word "VIAGRA" or "SALE" are probably more likely to have been created by a spam bot than a human. However, there is a great risk of overfitting to a particular type of dataset (reducing the possibility for real-world usage) unless care is taken to regularize or in other ways constrain the classifier's reliance on specific words. We have chosen to simply concatenate all tweets belonging to the same Twitter account and then apply tf-idf to the most common unigrams and bigrams. We have experimented with various choices of the number of features to extract from the tf-idf, but settled for 800, which is quite restrictive. Moreover, we have found term presence to work at least equally good as term frequencies, which further reduce the complexity of the extracted features. A higher number of extracted features could potentially have given better results on the unseen test set, but we decided to set it low in order to mainly keep words likely to generalize well also to other datasets. More modern natural language processing techniques such as word embeddings [7] or more sophisticated architectures such as Long Short-Term Memories (LSTMs) [6] or Transformers [15] would likely have given better results than a simple tf-idf, however, this have not been experimented with due to the rather long time it takes to run even simple classifiers on the TIRA [9] virtual matchines (on which the evaluation of the submitted solutions have to take place).

We have now presented all the used features for bot vs. human classification, but there are also various choices to be made on how to combine the features into actual classifications. One obvious way would be to combine all discussed features into a single feature vector onto which a single classifier is applied. The feature vectors from the tf-idf are, however, quite sparse compared to the aggregated statistical features and we have found that we get better performance if we first train a separate classifier on the tf-idf features only and then add the output from this first classifier as an extra feature to the rest of the statistical features. We have experimented with both the crisp class predictions as well as probability distributions over the different class labels and found that using the probability distributions as an extra feature to work slightly better on our hold-out test data. We have also experimented with various classifiers. In our final implementation of the submitted system we have used a logistic regression classifier (with regularization set to $C = 1.0$) for the tf-idf features and a random forest classifier (with 500 estimators and a minimum of one sample per leaf) which have been applied to the combination of the class probability output from the logistic regression classifier and the aggregated statistical features.

4.2 Gender Classification

For the Twitter accounts that are classified as $human$ in the first step, we are in the next step applying a gender classifier intended to distinguish between $male$ and $female$ authors. For this problem we have experimented with the same features as described above. However, in the end we used only tf-idf, with the 300 most common lowercased word unigrams in the training set, as this performed better than if these features were combined with any of the other features describe above. Upon inspection it can be seen that these unigrams are basically corresponding to (data-driven) function words in the English and Spanish languages, respectively. Function words have previously been found to work well for gender classification, see e.g., [11]. Other features that have been shown to work well for gender classification of e.g., blogs are part-of-speech (POS) tags [1]. We have experimented with POS tags obtained using SpaCy[1], however, the POS tagging takes some extra time and did not seem to provide any extra predictive power, so in the end we only used the tf-idf features described above for the final submission. These features are used as input to a logistic regression classifier (with regularization set to $C = 1.0$).

5 Experimental Results

All the features and classifiers described in the last section have been implemented using scikit-learn [8]. In order to get an idea of what kind of features and classifiers that work well for the PAN 2019 bots and gender profiling task, we initially set aside a randomly selected subset of 10 percent of the English and Spanish Twitter accounts in the available dataset, respectively. This gave very high initial accuracies and F1-scores also for very basic algorithms. However, when discovering and utilizing the train/dev split recommended by the PAN organizers on their web page, more reasonable results were obtained. The reason for this difference is unknown to the author, but we assume it is a consequence of how the datasets have been collected or constructed. The numbers reported in this section have all been computed on the dev part of the dataset recommended by the PAN organizers, unless it is explicitly mentioned that the numbers have been achieved on the final submission through TIRA [9], i.e., pan19-author-profiling-test-dataset1-2019-03-20 (here after referred to as testset1) and pan19-author-profiling-test-dataset2-2019-04-29 (here after referred to as testset2).

When training the presented system on the train part and evaluating on the dev part of the train/dev dataset obtained as part of the PAN task, the results summarized in Table 1 were obtained.

When submitting the best performing version of the system to TIRA, the accuracies presented in Table 2 were obtained for the different languages and the different classification problems. Based on these results we can say that the classification of accounts as being bot or human is, as expected, an easier task than to discriminate between male and female authors. Moreover, it is also shown that the implemented system performs

[1] https://github.com/explosion/spaCy

Table 1. Accuracies obtained on the bots and gender profiling tasks on the local evaluation.

Classification task	**Language**	**Tf-idf features**	**Tf-idf + statistical features**
Bots profiling	en	0.914	**0.948**
Bots profiling	es	0.865	**0.892**
Gender profiling	en	**0.752**	0.713
Gender profiling	es	**0.648**	0.617

Table 2. Accuracies obtained on the bots and gender profiling tasks on the TIRA test datasets.

Classification task	**Language**	**testset1**	**testset2**
Bots profiling	en	0.958	0.960
Bots profiling	es	0.806	0.882
Gender profiling	en	0.792	0.838
Gender profiling	es	0.606	0.728

significantly better on English than on Spanish for both the bots and gender prediction tasks, despite that we have not used any resources or features that are specifically designed for English.

If we compare the obtained accuracies to the other participating teams, it seems like the accuracy of our submitted bot classifier is performing better than all other teams on English, while it is only among the 15 best performing submissions on Spanish. On gender profiling it is among the top-5 submissions on English, while it is only among the 20 best performing submissions on Spanish. This clearly shows that our submitted system works well on English, while it underperforms on Spanish (and have plenty of room for improvement). When the obtained accuracy scores per language are averaged we obtain an average score of 0.852, yielding a final ranking of place 9 out of 55 participating teams.

6 Conclusions

Our submission to the PAN 2019 bots and gender profiling task relies on a two-step binary classification in which we first predict whether Twitter accounts are bots or not (i.e., human), and next predict whether the owners of the accounts classified as human are male or female. For both classifications we have made use of standard tf-idf features and mildly regularized logistic regression classifiers, but for the bot or not classification we have combined the output from the logistic regression classifier with many other statistical features derived from the textual content of the tweets, including Damerau-Levenshtein distance-based metrics and the number of retweets. These combined features have been input to a random forest classifier, which produced the best performance in our conducted experiments. In our performed local experiments we obtained accuracies of approximately 0.95 (English) and 0.89 (Spanish) on the bot classification task,

and approximately 0.75 (English) and 0.65 (Spanish) on the gender classification task. When submitting the same models to TIRA and evaluating them on previously unseen test data we received results that were inline with these numbers as well.

Since we have used quite constrained features and have based them on reasoning about what kind of features that are likely to be useful for bot and gender classification, we expect the implemented system to be useful also in real-world settings and not only within the PAN competition. The basic nature of the implemented features and classifiers leave plently of room for improvement by using more state-of-the-art embeddings or deep learning architectures.

Acknowledgments

This work was supported by the R&D programme of the Swedish Armed Forces.

References

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday 12(9) (2007)
2. Brynielsson, J., Johansson, F., Jonsson, C., Westling, A.: Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. Security Informatics 3(1), 7 (Aug 2014), https://doi.org/10.1186/s13388-014-0007-3
3. 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. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2145432.2145568
4. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Trans. Dependable Secur. Comput. 9(6), 811–824 (Nov 2012), http://dx.doi.org/10.1109/TDSC.2012.75
5. Damerau, F.J.: A technique for computer detection and correction of spelling errors. Commun. ACM 7(3), 171–176 (Mar 1964), http://doi.acm.org/10.1145/363958.363994
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (Nov 1997), http://dx.doi.org/10.1162/neco.1997.9.8.1735
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 3111–3119. NIPS'13, Curran Associates Inc., USA (2013), http://dl.acm.org/citation.cfm?id=2999792.2999959
8. 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, 2825–2830 (2011)
9. 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)
10. 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)

11. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. In: Computational Approaches to Analyzing Weblogs - Papers from the AAAI Spring Symposium, Technical Report. vol. SS-06-03, pp. 191–197 (8 2006)
12. Subrahmanian, V.S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., Menczer, F.: The darpa twitter bot challenge. Computer 49(6), 38–46 (Jun 2016), https://doi.org/10.1109/MC.2016.183
13. Teljstedt, C., Rosell, M., Johansson, F.: A semi-automatic approach for labeling large amounts of automated and non-automated social media user accounts. In: 2015 Second European Network Intelligence Conference. pp. 155–159 (2015)
14. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions:detection, estimation, and characterization. In: Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017) (2017)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017), https://arxiv.org/pdf/1706.03762.pdf
16. Zhang, L., Liu, B.: Sentiment Analysis and Opinion Mining, pp. 1152–1161. Springer US, Boston, MA (2017)