Bots and Gender Profiling using a Multi-layer Architecture

Notebook for PAN at CLEF 2019

Régis Goubin*, Dorian Lefeuvre*, Alaa Alhamzeh*,**, Jelena Mitrović**, and Előd Egyed-Zsigmond*, Leopold Ghemmogne Fossi*

* Université de Lyon - INSA Lyon - LIRIS UMR5205,
**Universität Passau
regis.goubin@insa-lyon.fr, dorian.lefeuvre@insa-lyon.fr, alaa.alhamzeh@insa-lyon.fr,
jelena.mitrovic@uni-passau.de,
elod.egyed-zsigmond@insa-lyon.fr,leopold.ghemmogne-fossi@insa-lyon.fr

Abstract In this paper, we introduce the architecture used for our PAN@CLEF-2019 author profiling participation. In this task, we had to predict if the author of 100 tweets was a bot, a female human, or a male human user. This task is proposed from a multilingual perspective, for English and Spanish. We handled this task in two steps, using different feature extraction techniques and machine learning algorithms. In the first step, we used random forest classifier with different features in order to predict if the users were bots or humans. In the second step, we recovered all the users predicted as humans. We then used a 2-layers architecture to predict the gender of the users detected as humans.

1 Introduction

Nowadays, the need for author profiling is growing as people share more and more content on the internet, especially on social networks. The profiling task is useful for several domains, such as security and forensics, marketing, target advertising, politics, etc. A lot of information can be recovered via the content shared by the users such as their gender, age, affiliation, etc.

Author profiling from tweets has been introduced by the PAN annual challenge since 2013 [20,18,16,22,21,19]. However, until now, the prediction was based only on human tweets, with an aim to predict some of their characteristics (age, gender, language variety, etc.), while this year, bots appear on the scene. Another difference to the previous challenges is the absence of the images attached to the tweets. The goal of the 2019 PAN author profiling shared task is to investigate whether the author of a Twitter feed is a bot or a human. Furthermore, in case of a human user, the goal is to predict the gender of the author, in two different languages: English and Spanish. In this paper, we will describe our approach to achieve that goal.

The reminder of the paper is structured as follows. Section 2 introduces state of the art corresponding to author profiling. In section 3, we present consecutively the overall architecture of our approach, the methods of bot detection, and gender prediction. In section 4, we present the results. Finally, in section 5, we come to a conclusion and discuss future steps.

2 Related Work

2.1 Bot Detection

A social media bot is a piece of software used to automatically generate messages, reply to messages, or to share particular hashtags, act as a follower of users, and as a fake account to gain followers itself. Varol et al. [25] have estimated that 9-15% of Twitter accounts may be bots. Twitter bots are a well-known example of how fake social media accounts can create convincing online personas capable of influencing real people on cultural and political topics. The first signs of this appeared during the 2016 U.S. election where the Russian government apparently leveraged bots to spread divisive messaging. Other governments, enterprises and groups use this technique as well. Although social bots are made to appear as if they were human accounts, it is still possible to identify them as bots based on their profile i.e. the user name, profile photo, time of posting and other meta-data. This information has been used efficiently to identify bots during the 2017 French elections [8]. However, the challenge this year is to detect bots only from textual data and not from the entirety of meta-data.

A.H. Wang [26] proposed to detect bots based on the number of friends, the number of followers, and the follower ratio of a user alongside the number of tweets containing HTTP links and the number of replies/mentions from the last 20 tweets of a user.

Ferrara et al. [9] used numerous features grouped in five different classes to detect bots. Varol et al. [25] expended the available features and grouped them in six different classes in the BotOrNot (now called Botometer) framework. The number of features available also increased in 2018 [28]. Although not all features identified by Varol et al. are relevant to the PAN challenge, sentiment and content related features can be used to detect bots with good accuracy. To the best of our knowledge, there are no previous works studying bot detection based on textual information only In another paper, Ferrara also showed that similar accuracy can be achieved when only user meta-data are used [8].

2.2 Gender Prediction

The relationship between personal traits and the use of language has been widely studied by Pennebaker[14] in the psycholinguistic research field. He showed how the usage of language varies depending on personal traits. For example, he found out that, at least in English, women use negations or talk in first person more than men, because they are more "self-conscious", whereas men use more prepositions in order to describe their environment, i.e. they speak about "concrete things" more. These finding are the basis of LIWC (Linguistic Inquiry and Word Count) [24] that is one of the most often used tools in author profiling. Pioneering research in gender profiling such as Argamon et al. [1], Burger et al. [4] and Schler et al. [23] focused mainly on formal texts and blogs, reporting accuracies in range of 75%-80% in most cases, as mentioned in the last PAN overview [19]. However, most recent investigations focus on social media such as Facebook and Twitter, where we have to think about handling short phrases, with many potential typos, less formal and more spontaneous language.

Most recent studies use features such as sequence of words and characters (unigram, bigrams and n-gram), and submit them to an SVM classifier, reporting an average accuracy of 75%-82% according to the last PAN edition [19].

In the 2017 edition of this challenge, the winners, Basile et al. [2], reached an accuracy of 0,8253 on gender, all languages (Arabic, English and Spanish) combined. They used a single SVM classifier and character 3- to 5- grams and word 1- to 2- grams with Tf-idf where tf is replaced by (1 + log(tf)) as features.

In 2017, another team of the PAN authors profiling task, Kheng and al. [12], had a different approach. After several experiments, they removed stop-words in English and Arabic, used a Tf-idf model based on 1- and 2-gram model and trained these features on a Naive Bayes Classifier.

Daneshvar et al. [6] used only text during the 2018 edition of the PAN author profiling task. They reached the first place on text classification and an overall second place with a general accuracy of 0,8170. In 2018, Ciccone et al. [5] based their research on the previous work of Kheng et al. [12] in order to improve it and construct an efficient text classifier. They performed approximately the same preprocessing steps, and also used the n-grams and Tf-idf model. Furthermore, they took into consideration the experiments done by Kheng in his master thesis [11] (he was able to improve his results and finally got an f-score of approximately 0,8 in 10-fold cross-validation) and obtained an overall accuracy of 0,7981 on texts, while Kheng et al. obtained an accuracy of 0,7002.

We based our work on the methods of Ciccone et al. [5]. We reproduced and improved their text model by integrating their set of features in a more complex architecture.

3 Our Approach

The classification part of our architecture consists of two steps. The first step classifies each user as a human or as a bot. After that, the users classified as humans are reclassified according to their gender. Figure 1 shows the overall architecture.

3.1 Bot Detection

For bot detection, our approach followed a classical pipeline. First of all, in the preprocessing phase, we choose for both languages to remove repeating characters occuring more than three times in order to recover the real word. The text is also tokenized with NLTK's TweetTokenizer [3].

After pre-processing the tweets, we choose to use the features extracted in the work of Varol et al. [25] in addition to other features based on our personal observations of the dataset.

As we mentioned in the last section, the work of Varol et al. extracts numerous groups of features. We used only the ones based on the tweet text. These features are: word entropy, the ratio of tweets that contain emojis, and Part-of-Speech (POS) distribution. We used these features for the early bird submission.

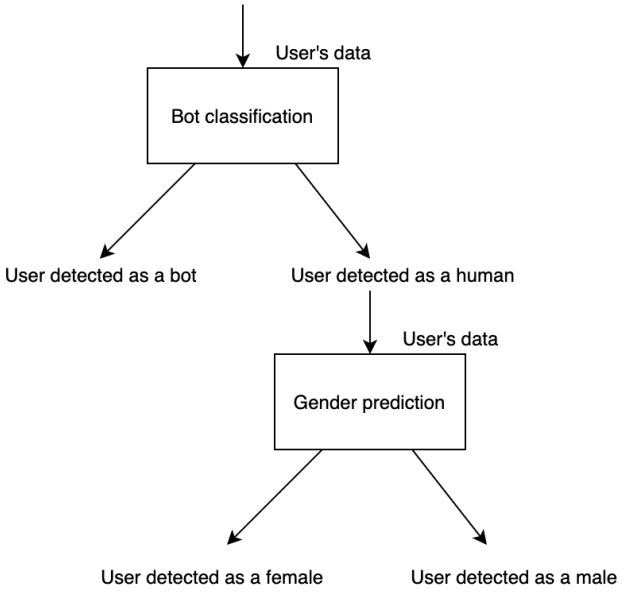

Figure 1: Schema of the overall architecture

Emojis are an important part of communication in Twitter, and they are used by bots and humans alike. However, they are not always used in the same way. For example, some bots only use emojis to communicate. Based on this observation, we choose to consider the ratio of tweets that contain emojis as a feature for a user, but we also choose to consider the number of emojis used in a tweet.

We were able to test our system during the early bird evaluation phase of this challenge, the results were quite good for English but unfortunately, for Spanish, we observed a gap between the model efficiency on the training dataset compared to the test dataset used by the early bird evaluation. That is why we thought about adding new features to the Spanish bot detection subtask.

To improve the result in the early bird, we have analyzed a sample of the dataset. We then identify some potential features to extract:

- Average number of words by tweet
- Average of the tweets' lengths for one user
- The standard deviation of the number of words
- Number of URLs
- Number of Hashtags
- Number of user mentions
- Percentage of uppercase letters
- Number of emojis

- Number of first person used
- Number of pronouns used
- Number of negations used

After running a feature selector (Shapley value based one) on these features, we choose to retain for bot classification for both languages:

- Average number of words per tweet
- Number of URLs
- Number of Hashtags
- Number of emojis
- The standard deviation of the number of words

In order to check the usefulness of these features, we tried to classify the dataset using these features alone. We achieved an accuracy of around 80% on both languages which supports our choice to use them afterward.

We choose to perform sentiment analysis on the tweets. For this purpose, we added more pre-processing steps for Spanish texts. We choose also to use DAL (Dictionary of Affect in Language), which is an instrument designed to measure the emotional meaning of words and texts. It does this by comparing individual words to a word list of 8742 words which have been rated by people for their activation, evaluation, and imagery. This concept was introduced by Whissell et al. [27] in 1986 for the English language. Later on, in 2013, a Spanish Dictionary for Affect in Language was produced by MDA Ríos [7]. Since we use DAL for Spanish tweets, we must be positive that we can find most of the words inside tweets using DAL. Moreover, when writing text, people tend to follow grammar rules. Therefore, we also apply stemming on Spanish words using NLTK's Snowball Stemmer [3]. This DAL includes the pleasantness, activation and imagery of around 2500 Spanish words [10] . For each tweet of an author, we rate his pleasantness, activation and imagery according to this DAL. The use of these features did not seem to improve the accuracy on the training data by much (around 0.2%). Therefore, we did not extract these features for the English set.

In order to train a classifier to identify bots, we perform 10-fold cross-validation on the classifier. After some testing on different classifiers we choose to use Random Forest Classifier from which we witness the best overall performances. We implement this classifier using the Sklearn Random Forest with 100 estimators.

On the final submission we use the following features for bot classification:

- Average number of words by tweet
- Number of URLs
- Number of Hashtags
- Number of Emojis
- The standard deviation of the number of words
- Word entropy
- Emojis ratio in the tweets
- Part-of-Speech (POS) tagging distribution
- Sentiment analysis (For Spanish)

3.2 Gender Prediction

The gender prediction is based on a 2-layer architecture. We distinguish between two types of classifiers:

- Low classifiers, which transform features into prediction
- Meta classifier, which takes as input the prediction of the low classifier and makes its own prediction

We use different classifiers provided by the sklearn library [13]. Figure 2 shows the architecture of this part. Each component of this architecture is described in the subsection below.

Low Classifiers The Tf-idf classifier is based on the work of Ciccone et al. [5]. The first task is the pre-processing. Twitter's data has to be cleaned, as users often make typos, use a lot of emojis, repeat characters to express happiness or anger, use upper cases in a different way than in classic texts. Thus, we proceed with different cleaning steps to remove the noise caused by the particularities of tweets and thus, create more useful features.

First of all, we remove repeating characters occurring more than three times. For instance, we transform "I'm really happyyyyyyyyyy" into "I'm really happy". The purpose was to recover the real word, since we did not want to create a vocabulary with the same word written in different way.

We remove the punctuation, URLs and user mentions. We also remove stop words but keep some of them with a transformation. Thus, we transform any used pronoun except the first person into '<pronoun>', while the first person pronouns into '<first_person>' and the negation sign into '<negation>'. This last pre-processing step, in which the focus is on the fact that both genders do not write in the same manner, is based on the work of Pennebaker and associates [14].

Finally, for English texts, we remove the plural endings, such as the final -s from nouns, taking into account all other possibilities and exceptions in plural formation in the English language. We then tokenize the text using the TweetTokeniser of NLTK [3]. As features, we kept the Tf-idf model, based on 1- and 2- word n-grams. We tried different classifiers and datasets to set up the final classifier. We tried to train an SVM and a Bagging Classifier with both 2019 and 2018 training datasets.

We realized that the two datasets are very different. It seems that the 2019 training dataset is really specific and works only on itself. On the other hand, the 2018 dataset seemed to be very complete and the classifier trained on it had good results on both datasets. To perform our experiments, we ran our classifier on the 2018 test dataset and 2019 dev split provided by PAN organizers. As we did not know if the test dataset was going to be closer to the 2018 or the 2019 training dataset, we trained an SVM on both datasets. We use a LinearSVC (from sklearn library) with a maximum of 500 iterations. As a consequence, our Tf-idf model combined the completeness of the 2018 training dataset and the particularities of the 2019 one.

Bagging classifier is, in some cases, useful in order to avoid over-fitting or to improve the accuracy. We choose to implement an architecture of 10 SVMs, with the same

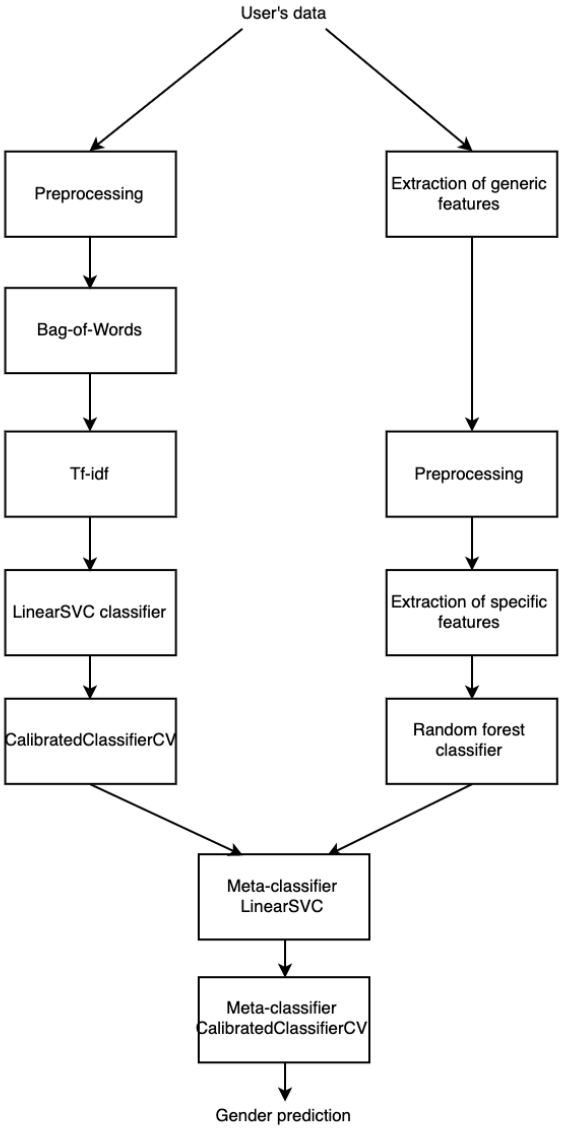

Figure 2: Schema of the gender prediction architecture

parameters we used so far. As bagging didn't provide us with significant improvement, and, in several experiments, it decreased the results, we chose to keep only SVM in the architecture.

Potential Features		Used Features	
		English	**Spanish**
Generic	Average number of words by tweet	✓	✓
	Average of the tweet lengths	✓	✓
	Number of URLs	✓	✓
	Number of Hashtags	✓	✓
	Number of user mentions	✓	✓
	Number of emojis	✓	✓
	Emoji ratio in the tweets		
	Word entropy	✓	✓
	Percentage of uppercase letters		
	The standard deviation of the number of words		✓
Specific	Number of first person used	✓	
	Number of pronouns used	✓	
	Number of negations used	✓	✓

Table 1: List of potential and used features in both languages

We expend the architecture defined last year by Ciccone et al. Indeed, we considered other features to predict the user's gender.

The specific (language dependant) and generic features (not related to the language) classifiers recover different features from the text in order to improve the prediction provided by the Tf-idf classifier. We did not make any preprocessing in order to recover generic features, as it would have modified the features. However, we made one preprocessing step to recover specific features: we removed the punctuation marks. It prevents some typos like "It is our.s" where "our.s" would have been an undetected pronoun. Obviously, we did not remove apostrophes as we wanted to detect all the negation such as "don't" or "wouldn't". We ran a feature selector (Shapley value based one) on these features to only keep the most useful and accurate ones. In the table 1, we sum up the features we extracted and the selected features per language.

We trained this classifier on 70% of the dataset. On the last 30%, it reached an accuracy of 0,7032 in Spanish and 0,7081 in English. We tried different classifiers on these features. The best performance was reached by a Random Forest classifier. We use the Sklearn implementation with 100 estimators. For the other parameters, we kept the default values defined by Sklearn.

Meta classifier The meta classifier is a simple SVM classifier which takes a few features as input. The Spanish meta classifier uses the prediction of the Tf-idf classifier and the generic and specific feature classifiers, while the English meta classifier uses the prediction of the Tf-idf classifier and the generic and specific features.

The meta classifiers are trained with 30% of the dataset. To assess this relevance, we train it with 10-fold cross-validation. As the training sets were small (620 human users for the English, 460 human users for Spanish), we essentially focused on the average

English		Spanish	
Bot	Gender	Bot	Gender
0.9356	0.8295	0.7444	0.6667

Table 2: Accuracy obtained on the early bird submission

English		Spanish	
Bot	Gender	Bot	Gender
0.9034	0.8333	0.8678	0.7917

Table 3: Accuracy obtained on the final submission

accuracy of the 10-fold. Then, we trained both classifiers with all the humans contained in dev splits.

We used a meta classifier for two reasons. We needed an "arbitrator" in order to determine a final prediction according to both low classifier predictions. Besides, it improves the accuracy obtained by the low classifiers. Indeed, the meta classifier tries to trust the low classifier according to their relevance and gives some importance to each classifier to take advantage of both ones.

4 Results

In this section, we compare the results we obtained against our early bird results.

Bot Improvement We have added features to our English bot classifier used for the early bird submission system. These features have slightly improved our results on the training data but do not seem to improve our results on the final submission and might have worsened the results. On the other hand, the modifications on Spanish have greatly increased our results. The early bird and the final submission datasets are not the same so we cannot draw conclusions before further testing.

Gender Improvement For the gender, we cannot consider the absolute accuracy, as it is totally dependant on the bot detection accuracy. As a consequence, we chose to determine our progress by considering the percentage of accuracy lost, using the following formula:

$$relative\_gender\_accuracy = \frac{gender\_accuracy}{type\_accuracy}$$

For both submissions, the relative gender accuracy is shown in table 4. In this table, it can be seen that we improved the relative accuracy. Obviously, the differences between bot detection accuracy and gender accuracy are caused by bad classification of human gender. Thus, the results can be influenced by a huge proportion of bots in

	English	Spanish
Early bird submission	0.8866	0.8960
Final submission	0.9224	0.9123

Table 4: Relative gender accuracy per submission

gender accuracy. Nevertheless, as the improvement is over 3.5% in English and over 1.5% in Spanish, we can say with confidence that our modification had a good impact on gender prediction.

Consequently, we suppose the feature selector allowed better results. This supposition means some extracted features confused our classifier and could be considered as noise. Furthermore, it seems that the training with both 2018 and 2019 datasets was a good choice.

It is difficult to give an overall conclusion to the gender results. Indeed, the accuracy obtained on human detection is completely hidden. The comparison with the state-of-the-art [6] is not possible.

5 Conclusion and Future Work

In this paper, we introduced our architecture for the author profiling shared task, proposed by the PAN @CLEF challenge.

For bot detection we based our approach on the work of Varol et al. [25]. As we only have tweet text as data, we were only able to extract some of the features proposed and thus, we identified other potential features and we chose to add some of these features to our classifier.

For the gender prediction, we implemented a 2-layers architecture. We based our work on the solution of Ciccone et al.[5]. Nevertheless, we improved this solution by integrating it into a more complex architecture. We took advantages of features not used by the classifier based on Tf-idf.

We also presented our results and the improvement we obtained in both early bird and final submissions we have made.

As future work, we would like to improve our results and be close (or better than) the state-of-the-art of the challenge. Notably, we plan to integrate other features in the gender part, such as features based on pre-trained word embedding models.

References

1. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. TEXT 23, 321–346 (2003)
2. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-GrAM: New Groningen Author-profiling Model. arXiv:1707.03764 [cs] (Jul 2017), http://arxiv.org/abs/1707.03764, arXiv: 1707.03764
3. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media, Inc., 1st edn. (2009)

4. 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
5. Ciccone, G., Sultan, A., Laporte, L., Granitzer, M.: Stacked Gender Prediction from Tweet Texts and Images p. 11 (2018)
6. Daneshvar, S., Inkpen, D.: Gender Identification in Twitter using N-grams and LSA: Notebook for PAN at CLEF 2018. In: CLEF (2018)
7. Dell' Amerlina Ríos, M., Gravano, A.: Spanish DAL: A Spanish Dictionary of Affect in Language. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 21–28. Association for Computational Linguistics, Atlanta, Georgia (Jun 2013), https://www.aclweb.org/anthology/W13-1604
8. Ferrara, E.: Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. First Monday 22(8) (Jul 2017), http://arxiv.org/abs/1707.00086, arXiv: 1707.00086
9. Ferrara, E., Varol, O., Menczer, F., Flammini, A.: Detection of Promoted Social Media Campaigns p. 4 (2016)
10. Gravano, A., Rıos, M.G.D.: Spanish DAL: A Spanish Dictionary of Affect in Language p. 24
11. Kheng, G.: Author Profiling : author "gender" and "variety of language" retrieval in tweets. Master's thesis, University of Passau and INSA de LYON (Sep 2017)
12. Kheng, G., Laporte, L., Granitzer, M.: INSA LYON and UNI PASSAU's participation at PAN@CLEF'17: Author Proling task p. 11 (2017)
13. 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)
14. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology 54(1), 547–577 (2003), https://doi.org/10.1146/annurev.psych.54.101601.145041, pMID: 12185209
15. 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)
16. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Proling Task at PAN 2015 p. 40
17. 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)
18. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd Author Proling Task at PAN 2014 p. 30
19. Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Proling Task at PAN 2018: Multimodal Gender Identication in Twitter p. 38 (2018)
20. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Proling Task at PAN 2013 p. 13
21. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Proling Task at PAN 2017: Gender and Language Variety Identication in Twitter p. 26 (2017)
22. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th Author Proling Task at PAN 2016: Cross-Genre Evaluations p. 35
23. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging SS-06-03, 191–197 (8 2006)

24. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of Language and Social Psychology 29(1), 24–54 (2010), https://doi.org/10.1177/0261927X09351676
25. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online Human-Bot Interactions: Detection, Estimation, and Characterization. arXiv:1703.03107 [cs] (Mar 2017), http://arxiv.org/abs/1703.03107, arXiv: 1703.03107
26. Wang, A.H.: Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach. In: Foresti, S., Jajodia, S. (eds.) Data and Applications Security and Privacy XXIV. pp. 335–342. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2010)
27. Whissell, C.M.: Chapter 5 - THE DICTIONARY OF AFFECT IN LANGUAGE. In: Plutchik, R., Kellerman, H. (eds.) The Measurement of Emotions, pp. 113–131. Academic Press (Jan 1989), http://www.sciencedirect.com/science/article/pii/B9780125587044500116
28. Yang, K.C., Varol, O., Davis, C.A., Ferrara, E., Flammini, A., Menczer, F.: Arming the public with AI to counter social bots. arXiv:1901.00912 [cs] (Jan 2019), http://arxiv.org/abs/1901.00912, arXiv: 1901.00912