Bots and Gender Profiling of Tweets using Word and Character N-Grams

Notebook for PAN at CLEF 2019[1]

Yaakov HaCohen-Kerner, Natan Manor, Michael Goldmeier

Dept. of Computer Science, Jerusalem College of Technology – Lev Academic Center
21 Havaad Haleumi St., P.O.B. 16031, 9116001 Jerusalem, Israel

kerner@jct.ac.il, natanmanor@gmail.com, mmgoldmeier@gmail.com

Abstract. Author profiling deals with the identification of various details about the author of the text (e.g., age and gender). In this paper, we describe the participation of our team (hacohenkerner19) in the PAN 2019 shared task on Bots and Gender Profiling in two languages: English and Spanish. Given a Twitter feed, we should determine whether its author is a bot or a human. In the case of human, we should identify her/his gender. In this paper, we describe our pre-processing methods, feature sets, five applied machine learning methods, and accuracy results. The best accuracy result for the English dataset (84.8%) was obtained by LinearSVC using 2,000 word unigrams. The same result (84.8%) was also obtained by LR by using four preprocessing methods, 2,000 word unigrams, and 1,000 word bigrams with maximal skips of 2 words. The best accuracy result (75,54%) for the Spanish dataset was achieved using LinearSVC with only the HTML tag removal preprocessing method and a combination of 1,000 word unigrams, 1,000 word bigrams, and 3,000 character trigrams.

Keywords: Bot Profiling, Gender Profiling, Character N-grams, Word N-grams, Supervised Machine Learning.

1 Introduction

Gender profiling deals with the analysis of a given text while inferring the gender of the author of the given text. This task is of growing importance all over the world. Important and interesting applications can use gender detection in various domains such as business intelligence, forensics, and psychology. A linguistic analysis of a given text can help to identify certain characteristics of the author. For example, companies would like to know, based on the analysis of online product reviews, the gender of people who like or dislike each of their products.

Another interesting task with increasing importance is to identify whether its author is a bot or a human. This task is important because companies, organizations, and individuals might use bots in various social media platforms in order to influence users with commercial, political or ideological purposes. For instance, bots could artificially increase the popularity of a product by promoting it and/or writing positive ratings, as well as underestimate the importance of competitive products by writing negative ratings and comments. There is a greater danger when the use of the bot is political or ideological. Moreover, bots are also used for fake news spreading, which can severely harm organizations or individuals. Therefore, the automatic distinction between people and bots is of high importance from the point of view of marketing, forensics, and security.

In this paper, we describe the participation of our team hacohenkerner19 in the PAN 2019 shared task on bots and gender profiling. More specifically, the shared task is as follows. Given a Twitter feed, determine whether its author is a bot or a human. In the case of human, identify her/his gender. The addressed languages are English and Spanish.

In our research, we consider the application of several supervised machine learning (ML) methods, various types of feature sets and various numbers of features from each feature set.

The rest of the paper is organized as follows. Section 2 provides background and presents some related work on text classification in general and author profiling in particular. Section 3 introduces the feature sets that we have implemented and used in this study. Section 4 presents the experimental setup, the experimental results for the datasets written in English and Spanish and their analysis. Finally, Section 5 summarizes and suggests ideas for future research.

2 Related Work

This section presents a general background and presents several previous studies related to text classification in general and author profiling in particular.

2.1 Text classification

Text classification (TC) is the supervised learning task of assigning natural language text documents to one or more predefined categories [Meretakis and Wuthrich, 1999]. There are two main types of TC: topic-based classification and style-based classification.

These two classification types often require different types of feature sets to achieve the best performance. Topic-based classification is typically performed using word uni-grams and/or n-grams (n > 2) [Argamon et al., 2007; HaCohen-Kerner et al., 2008A]. Style-based classification is typically performed using linguistic features such as quantitative features, orthographic features, part of speech (POS) tags, function words, and vocabulary richness features [HaCohen-Kerner et al., 2010A; HaCohen-Kerner et al., 2010B].

2.2 Author profiling

The author profiling task is of growing importance during recent years. Various author profiling applications are found in business intelligence, forensics, psychology, and security systems. The general aim of an author profiling task is to determine various |demographic information about the text's author(s), e.g., age, cultural background, gender, native language and/or dialect, and various personality traits. In this paper, we will limit ourselves to bots and gender profiling because the PAN 2019 shared task is on these tasks.

The gender classification is a relatively simple (binary classification) and probably the most frequent profiling task. However, such classification can be effective only if the writing style between genders does differ [Eckert et al., 2013] and if such stylistic differences can be detected [Jockers and Witten, 2010].

In contrast to most other demographic traits, the link between gender and word use has been extensively studied [Pennebaker et al., 2003]. Differences in women's and men's language have received relatively high attention within the scientific community as well as in the popular media. However, early studies on gender classification, mainly on formal texts and blogs, reported on accuracies around 75%-80% in most cases [Schler et al., 2006; Holmes and Meyerhoff, 2008; Burger et al., 2011].

Recently, various bot profiling tasks have been published. Oentaryo et al. [2016] presented a new categorization of bots and developed a systematic bot profiling framework with a rich set of features and classification methods. They carried out extensive empirical studies to analyze on broadcast, consumption and spam bots, as well as how they compare with regular human accounts. They discovered that the diversities of timing patterns for posting activities (i.e., tweet, retweet, mention, and hashtag, and URL) constitute the key features to effectively identify the behavioral traits of different bot types. Stella et al. [2018] analyzed nearly 4 millions Twitter posts. They showed that bots act from peripheral areas of the social system to target influential humans, with violent contents, increasing their exposure to negative and inflammatory narratives and exacerbating social conflict online.

Rangel et al. [2015] presented in their overview paper the framework and the results for the Author Profiling task at PAN 2015, which dealt with the identification of age, gender, and personality traits of Twitter users. In comparison to previous years of PAN [Rangel et al., 2013; Rangel et al., 2014] the PAN-15 systems achieved significantly higher accuracy values for gender identification. This may suggest that irrespective the shorter length of individual tweets and their informality, the number of tweets per author is sufficient to profile age and gender with high accuracy. Regarding the features, it was not clear which ones (style-based or content-based) were the most important ones, because of the high number of different ones used and combined by the teams.

A similar phenomenon occurred in the gender classification tasks in [Rangel et al., 2016A]. It was difficult to highlight the contribution of any particular feature since the teams used many of them. Second order representations based on relationships among terms, documents, profiles, and sub-profiles were used by teams that achieved first positions in some of the tasks. Likewise, the distributed representations achieved the first position in gender identification on the Dutch final evaluation.

The best resulting approaches that took part in the gender classification tasks in PAN 2017 [Rangel et al., 2017] took advantage from combinations of n-grams, other content-based features, and style-based features. The best final average gender ranking (for English, Portuguese, and Spanish) shows that the best overall result (82.53%) has been obtained by Basile et al. [2017], who used the scikit-learn[2] LinearSVM implementation trained with combinations of character 3- to 5-grams and word 1- to 2-grams with TF-IDF weighting with sublinear term frequency scaling. New research about celebrity profiling by Wiegmann et al. [2019] will appear in ACL 2019.

3 Features

In this section, we present the various types of features that we applied for the profiling task. Our features include various n-grams sets, where each one of them is defined by the following template: *number_k-n_type* where number is the number of the features in the set, k is the size of the wanted skip (0 – no skip, 1 – skip of one unit, 2 – skip of 2 units, …) only for text inside word boundaries, n is code of the grams (1 for unigrams 2 for bigrams, 3 for trigrams, …), and type is W for words or C for characters. All values are represented by TF-IDF values. The specific various n-grams sets that were applied will be presented later in the framework of the experiments.

4 Experimental Setup and Results

In this section, we present presents the experimental setup, the experimental results for the datasets written in English and Spanish and their analysis.

4.1 Experimental setup

The PAN CLEF 2019 [Daelemans et al., 2019] launched an evaluation campaign. The algorithms of the teams that have participated in this campaign have been evaluated using the TIRA platform Potthast et al. [2019]. The algorithms and the results of the participated teams in this Bots and Gender Profiling tournament have been overviewed in Rangel and Rosso [2019]. The Low Dimensionality Statistical Embedding (LDSE) method was presented in Rangel et al. [2016B]. The results of this method set a pretty high bar, which was ahead of most of the teams participating in the competition

General approach: Our approach to authorship profiling is to apply supervised ML methods to TC as was suggested by Sebastiani [2002]. The process is as follows. First, given a corpus of training documents, where each document is a Twitter feed with 100 tweets, which is labeled as either ‘male’ or ‘female’ or ‘bot’, we processed each document to produce values for different combinations of word and character n-grams. Second, we applied several popular ML methods on the generated combinations of features. Third, we tried additional combinations of features. Finally, the best models for the training set were tested on the test set.

[2] http://scikit-learn.org/stable/index.html

Preprocessing: There is a widespread variety of text preprocessing types such as: conversion of uppercase letters into lowercase letters, HTML object removal, stopword removal, punctuation mark removal, reduction of different sets of emoticon labels to a reduced set of wildcard characters, replacement of HTTP links to wildcard characters, word stemming, word lemmatization, correction of common misspelled words, and reduction of replicated characters. Not all of them are considered as effective by all TC researchers. Many systems use only a small number of simple preprocessing types (e.g., conversion of all the uppercase letters into lowercase letters and/or stopword removal).

In our classification experiments, we applied the following text preprocessing types for the English language: L – converting uppercase letters into lowercase letters, P – punctuation mark removal, M – word lemmatization, S – stopword removal, H – HTML tags removal, R- reduction of Replicated characters, C – Error Correction. The application of the S preprocessing type deletes all instances of 423 stopwords for English text (421 stopwords from Fox [1989] plus the letters "x" and "z" that are not found in Fox [1989], yet are included in many other stopword lists). For the Spanish language, we applied the following text preprocessing types: L – converting uppercase letters into lowercase letters, P – punctuation mark removal, S – stopword removal (321 stopwords for Spanish[3]), H – HTML tags removal, and U – URL tags removal.

ML methods: We applied five ML methods: MLP– Multilayer Perceptron[4], LinearSVC – SVM with a linear kernel[5], LR - Logistic regression[6], RF - Random Forest[7], and MNB - Multinomial Naive Bayes[8].

A brief description of these ML methods is as follows. MLP is a feedforward neural network ML method [Jain et al., 1996] where artificial neural network (ANN) can be viewed as a weighted directed graph in which nodes are artificial neurons and directed edges (with weights) are connections from the outputs of neurons to the inputs of neurons. Support vector machine (SVM, also called support vector network) [Cortes and Vapnik, 1995] is a model that assigns examples to one of two categories, making it a non-probabilistic binary linear classifier. LinearSVC is SVM with a linear kernel, which is recommended for TC because most of TC problems are linearly separable [Joachims, 1998] and training an SVM with a linear kernel is faster compared to other kernels. Logistic regression (LR) is a variant of a statistical model that tries to predict the outcome of a categorical dependent variable (i.e., a class label) [Cessie and Van Houwelingen, 1992; Hosmer et al., 2013]. Random Forest (RF) is an ensemble learning method for classification and regression [Breiman, 2001]. RF operates by constructing a multitude of decision trees at training time and outputting classification for the case at hand. RF combines the "bagging" idea presented by Breiman [1994] and random selection of features introduced by Ho [1995] to construct a forest of decision trees. MNB is a Multinomial Naive Bayes Classifier that belongs to the family of Naive Bayes classifiers, which are classifiers based on applying Bayes' probabilistic theorem with independence assumptions between the features. The MNB classifier assumes that its

[3] http://snowball.tartarus.org/algorithms/spanish/stop.txt

[4] http://scikit-learn.org/stable/modules/neural_networks_supervised.html

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

[6] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[7] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[8] https://scikit-learn.org/stable/modules/naive_bayes.html

features (usually words) are chosen independently from multinomial distribution (McCallum and Nigam, 1998). A multinomial distribution is useful to model feature vectors where each value represents, for example, the number of occurrences of a term or its relative frequency.

Tools and information sources: We used the following tools:

- scikit-learn - a library for ML methods
- NLTK[9]- a library that produces the various n-gram features and provides a corpus of synonyms
- Numpy[10] - a library that performs fast mathematical processing
- Autocorrect[11]- a library that automatically corrects spelling errors.
- Emoji – a library that provides an option of translating emojis to words that express the emoji

4.2 Experimental results and their analysis

In Tables 1-4, we present our experimental results carried out on the training set supplied by the organizers of the competition. These results were obtained on the official split of the training set (2/3 for training and 1/3 for test).

The baseline accuracy results of the TC experiments that were performed for the English and the Spanish datasets are shown in Tables 1 and 2. The best results for each of the two best ML methods in both tables are bolded.

Table 1. Baseline accuracy results for the English dataset.

Features	MLP	LinearSVC	LR	RF	MNB
1000 Word Unigrams	0.783	0.841	0.825	0.750	0.691
2000 Word Unigrams	0.794	**0.848**	**0.834**	0.766	0.706
3000 Word Unigrams	0.790	0.846	0.827	0.760	0.716
4000 Word Unigrams	0.798	0.839	0.823	0.745	0.734
5000 Word Unigrams	0.796	0.837	0.816	0.749	0.735

Table 2**.** Baseline accuracy results for the Spanish dataset.

Features	MLP	LinearSVC	LR	RF	MNB
1000 Word Unigrams	0.7109	0.7413	0.7457	0.6957	0.6489
2000 Word Unigrams	0.6913	0.7293	0.7446	0.7089	0.6652
3000 Word Unigrams	0.7098	0.7349	0.7467	0.6685	0.6587
4000 Word Unigrams	0.7272	**0.7435**	**0.7500**	67.93	66.20
5000 Word Unigrams	0.7250	0.7413	0.7478	0.6533	0.6630

[9] https://www.nltk.org/

[10] http://www.numpy.org/

[11] https://github.com/phatpiglet/autocorrect

Tables 1 and 2 show various baseline results for the English and Spanish datasets, respectively. The best accuracy result for the English dataset 84.8% was obtained by the SVC method using 2,000 word unigrams. The second best ML method was LR with an accuracy of 83.4% using 2,000 word unigrams. The best accuracy result for the Spanish dataset 75% was obtained by LR using 4,000 word unigrams. The second best ML method was SVC with an accuracy of 74.35% using 4,000 word unigrams.

The best accuracy result for the English dataset in Table 3 (84.8%) was obtained by LR by using four preprocessing methods (L – converting uppercase letters into lowercase letters, M – word lemmatization, P – punctuation mark removal, and R-reduction of Replicated characters), 2,000 word unigrams, and 1,000 word bigrams with maximal skip of 2 words. Unfourtantly, although we have done dozens and maybe hundreds of different experiments so far we did not succeed to improve the best baseline result (also 84.8%) that was obtained by LinearSVC using 2,000 word unigrams.

The best accuracy result for the Spanish dataset in Table 4 (75.54%) was obtained by LinearSVC with only the HTML tag removal preprocessing method and by the combination of 1,000 word unigrams, 1,000 word bigrams, and 3,000 character trigrams. This result shows 1,000 word unigrams, a slight improvement of 0.54% comparing to the best baseline accuracy result in Table 2 (75%).

Table 3. Best accuracy results for the English dataset.

Features	Best ML	Preprocessing	Accuracy
2000 W Unigrams, 1000 W Bi-gram with skips of 2	LR	LMPR	**0.848**
2000 W Unigrams, 1000 W Bi-grams with skips of 2	SVC	LMPR	0.845
1000 W Unigrams, 1000 C Unigrams, 3000 C Tri-grams	SVC	H	0.842
1000 W Unigrams, 1000 C Unigrams, 3000 C Tri-grams	LR	H	0.830
1000 W Unigrams, 1000 C Bi-grams	SVC	H	0.823

Table 4. Best accuracy results for the Spanish dataset.

Features	Best ML	Preprocessing	Accuracy
1000 W Unigrams, 1000 C Bi-grams, 3000 C trigrams	SVC	H	**0.7554**
3000 W Unigrams, 3000 C bigrams	SVC	LHS	0.7522
1000 W Unigrams, 3000 C trigrams, 1000 C Bi-grams	SVC	L	0.7511
5000 W Unigrams	LR	NONE	0.7472
2000 W Unigrams, 1000 W unigrams with skips of 2	LR	NONE	0.7452

Table 5. Official accuracy results obtained for the test set on TIRA.

Baseline/Our Team	Bots vs. Human		Gender		Average
	English	Spanish	English	Spanish	
Random baseline	0.4905	0.4861	0.3716	0.3700	0.4296
Majority baseline	0.5	0.5	0.5	0.5	0.5
Our team (hacohenkerner19)	0.4163	0.4744	0.7489	0.7378	0.5944
LDSE baseline	0.9054	0.8372	0.7800	0.6900	0.8032

In Table 5, we present part of the official accuracy results obtained for the test set on TIRA. The results of our model (hacohenkerner19) were better than the results obtained by the Random and Majority baseline methods. Our results were comparable to the LDSE baseline results for the gender task for both languages. However, our results were significantly lower than those of the LDSE baseline results for the Bots vs. Human task for both languages. Possible explanations for this phenomenon could be (1) our model only classifies each profile to one of three types: male or female or bot; We did not classify whether it is a bot or a person, and only if it a person to classify whether it is a male or a female and/or (2) our models presented in Tables 3 and 4 are overfitting.

5 Summary and Future Work

In this paper, we describe the participation of our team (hacohenkerner19) in the PAN 2019 shared task on bots and gender profiling of tweets in English and Spanish. We tried various pre-processing types, a wide variety of feature sets, and five ML methods.

Regarding the experimental results carried out on the training set, the best accuracy result for the English dataset (84.8%) was obtained by LinearSVC using 2,000 word unigrams. The same result (84.8%) was also obtained by LR by using four preprocessing methods, 2,000 word unigrams, and 1,000 word bigrams with a maximal skip of 2 words. The best accuracy result (75,54%) for the Spanish dataset was achieved using LinearSVC with only the HTML tag removal preprocessing method and a combination of 1,000 word unigrams, 1,000 word bigrams, and 3,000 character trigrams.

Regarding the official accuracy results obtained for the test set on TIRA, the results of our model were better than the results obtained by the Random and Majority baseline methods. However, our results were significantly lower than those of the LDSE baseline results for the Bots vs. Human task for both languages. Possible explanations could be (1) our model only classifies each profile to one of three types: male or female or bot; We did not classify whether it is a bot or a person, and only if it a person to classify whether it is a male or a female and/or (2) our models presented in Tables 3 and 4 are overfitting.

Future research proposals include (1) applying additional combinations of feature sets; (2) tuning each model separately; (3) applying various deep neural models; and

(4) building and applying model(s) that will use also keyphrases [HaCohen-Kerner et al., 2007], expansions of abbreviations [HaCohen-Kerner et al., 2008B], and summaries [HaCohen-Kerner et al., 2003] that can be extracted from the tweet profiles.

Acknowledgments. This work was partially funded by the Jerusalem College of Technology (Lev Academic Center) and we gratefully acknowledge its support.

References

1. Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., Levitan, S.: Stylistic text classification using functional lexical features: Research articles. Journal of the American Society for Information Science and Technology. 58, 6, 802–822 (2007).
2. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-GrAM: New Groningen Author-profiling Model. arXiv preprint arXiv:1707.03764 (2017).
3. Breiman, L.: Bagging predictors. Univ. California Technical Report No. 421. (1994).
4. Breiman, L.: Random forests. Machine learning, 45(1), 5-32 (2001).
5. 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).
6. Cessie, S. Le, Van Houwelingen, J. C.: Ridge estimators in logistic regression, Applied statistics, pp. 191-201 (1992).
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning, 20(3), 273-297 (1995).
8. 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).
9. Eckert, P., McConnell-Ginet, S.: Language and gender. Cambridge University Press (2013).
10. Fox, C.: A stop list for general text. In Acm Sigir forum (Vol. 24, No. 1-2, pp. 19-21). ACM (1989).
11. HaCohen-Kerner, Y., Malin, E., Chasson, I.: Summarization of Jewish Law Articles in Hebrew. In CAINE, pp. 172-177 (2003).
12. HaCohen-Kerner, Y., Stern, I., Korkus, D., Fredj, E.: Automatic machine learning of keyphrase extraction from short HTML documents written in Hebrew. Cybernetics and Systems: An International Journal, 38(1), 1-21 (2007).
13. HaCohen-Kerner, Y., Mughaz, D., Beck, H., Yehudai, E.: Words as classifiers of documents according to their historical period and the ethnic origin of their authors. Cybernetics and Systems: An International Journal, 39(3), 213-228 (2008A).
14. HaCohen-Kerner, Y., Kass, A., Peretz, A.: Combined one sense disambiguation of abbreviations. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers (pp. 61-64). Association for Computational Linguistics (2008B).

15. HaCohen-Kerner, Y., Beck, H., Yehudai, E., Rosenstein, M., Mughaz, D.: Cuisine: Classification using stylistic feature sets &/or name-based feature sets. Journal of the American Society for Information Science and Technology 61 (8), 1644–57 (2010A).
16. HaCohen-Kerner, Y., Beck, H., Yehudai, E., Mughaz, D.: Stylistic feature sets as classifiers of documents according to their historical period and ethnic origin. Applied Artificial Intelligence 24 (9), 847–62 (2010B).
17. Ho, T. K.: Random Decision Forests. Proceedings of the 3rd Int. Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. 278–282 (1995).
18. Holmes, J., Meyerhoff, M. (Eds.).: The handbook of language and gender (Vol. 25). John Wiley & Sons (2008).
19. Hosmer Jr, D. W., Lemeshow, S., Sturdivant, R. X.: Applied logistic regression (Vol. 398). John Wiley & Sons (2013).
20. Jain, A. K., Mao, J., Mohiuddin, K. M.: Artificial neural networks: A tutorial. Computer, 29(3), 31-44 (1996).
21. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer, Berlin, Heidelberg (1998).
22. Jockers, M. L., Witten, D. M.: A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing, 25(2), 215-223 (2010).
23. Meretakis, D., Wuthrich, B.: Extending naive Bayes classifiers using long itemsets, Proc. of the 5th ACM-SIGKDD Int. Conf. Knowledge Discovery, Data Mining (KDD'99), San Diego, USA, 165-174 (1999).
24. Oentaryo, R. J., Murdopo, A., Prasetyo, P. K., Lim, E. P.: On profiling bots in social media. In International Conference on Social Informatics (pp. 92-109). Springer, Cham (2016).
25. 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).
26. 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)
27. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In CLEF Conference on Multilingual and Multimodal Information Access Evaluation, CELCT, pp. 352-365 (2013).
28. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., ... Daelemans, W.: Overview of the 2nd author profiling task at pan 2014. In CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014, pp. 1-30 (2014).
29. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In CLEF p. 2015 (2015).
30. 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 Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al., pp. 750-784 (2016A).
31. 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 (CICLing'16), Springer, Cham, pp. 156-169 (2016B).

32. 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. Working Notes Papers of the CLEF (2017).
33. Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018).
34. 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)
35. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47 (2002).
36. 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).
37. Stella, M., Ferrara, E., De Domenico, M.: Bots increase exposure to negative and inflammatory content in online social systems. Proceedings of the National Academy of Sciences, vol. 115 (49), pp. 12435-12440 (2018).
38. Wiegmann, M., Stein, B., Potthast, M.: Celebrity Profiling. To appear in 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), July 2019. Association for Computational Linguistics (2019).