Bot and Gender Detection of Twitter Accounts Using Distortion and LSA

Notebook for PAN at CLEF 2019

Andrea Bacciu, Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, Valerio Neri, and Julinda Stefa

Affiliation
Department of Computer Science, Sapienza University of Rome, Italy
{lamorgia, mei, nemmi, stefa}@di.uniroma1.it
bacciu.1747105@studenti.uniroma1.it, neri.1754516@studenti.uniroma1.it

Abstract In this work, we present our approach for the Author Profiling task of PAN 2019. The task is divided into two sub-problems, bot, and gender detection, for two different languages: English and Spanish. For each instance of the problem and each language, we address the problem differently. We use an ensemble architecture to solve the Bot Detection for accounts that write in English and a single SVM for those who write in Spanish. For the Gender detection we use a single SVM architecture for both the languages, but we pre-process the tweets in a different way. Our final models achieve accuracy over the 90% in the bot detection task, while for the gender detection, of 84.17% and 77.61% respectively for the English and Spanish languages.

1 Introduction

The ability to profile the author of a message automatically with the advent of social networks and online platforms is more than ever a crucial issue. As an example, be able to profile users that write on a specific topic can provide useful insight to address advertisement campaigns. In forensic investigations, detailed profiling of an anonymous user can speed up the investigative process. Moreover, be able to profile an author also include the ability to understand if the author of a set of messages is a human or a bot. On social networks, especially on Twitter, the presence of bots is heavy. Usually, bots have the task to drum up the attention of human users on a specific event but, unfortunately, they are often used to spread misleading information and fake news.

In this paper, we present our approach for the Author Profiling task of PAN 2019, that focuses on the bot and gender detection of Twitter accounts. We start introducing the problem, the dataset provided by PAN, and the evaluation framework used in this task. In section 3, we describe the features selected to build our model for the bot detection, the architecture, how single features perform on the task, and the final evaluation. In section 4, we address the problem of gender detection, also here we describe all the

phases we faced to build the final model. Finally, in section 5, we make a brief survey of the works that inspired our methodology, and section 6 concludes the paper with the final considerations about the work and the possibility to improve it.

2 Problem

In this section, we describe the problem addressed in this work. More in details, we start defining the Author Profiling competition, then we report about the data, and the evaluation framework used by the organizers of the task to evaluate the classifiers of the participants.

2.1 Pan Task

PAN is a series of scientific events and shared tasks on digital text forensics and stylometry. Currently, we are at the 19th [11] edition of the event, and the 7th on the author profiling task [34]. In each edition, the organizers propose several challenges related to the forensic analysis, provide the guidelines and the dataset to accomplish the tasks. Participants team can concur for one or all of them. At the end of the challenges, all the participants are invited to submit a notebook where they explain the methodology and the idea developed for the solution of the task.

2.2 Author Profiling Task

The Author Profiling Task 2019 [34] aims to identify the nature of Twitter accounts, detecting if the writer is a Bot or a Human being and in the last case, the gender of the account owner. The task is proposed in two different languages: English and Spanish. Participants can address just one problem, Bot or Gender detection, just one language or the complete challenge.SV

2.3 The Dataset

The dataset is made by two sets of Twitter accounts, one for the English language, and the other for the Spanish. Each account in the dataset is a collection of 100 tweets. Both the datasets are already split into Train and Dev partitions. Moreover, the complete dataset contains also two Test partitions, but they are not public, both the dimension and the number of instances for each class are unknown to the participants. Test sets can be used only in the remote evaluation phases. Instead, the known part of the dataset has the ground truth. Here, each account is described by two columns, where the first one defines the nature of the account and the second the gender.

To maintain a realistic scenario, no cleaning operation on the tweets was performed by the organizers. In this way, participants can be aware if a message is a tweet or a retweet, on the other side there is no guarantee that all the tweets of the same users are in the same language. All the links in the dataset appear, as usual on Twitter, in the short format (https://t.co/id). Hence, since during the evaluation phase is not possible

to use the internet connection, neither the link can be resolved nor is possible to extract information about their contexts such as the hostname, or the page content.

The English part of the dataset is made globally by 4120 accounts, of which 2880 belong to the Train set and 1240 to the Dev. Instead, the Spanish one has 3000 users, with 2080 accounts in the Train partition and the remaining users for the Dev set. In all the partitions users are evenly divided between bots and humans, and humans in their turn are divided in half between men and women. As we can see in Table 1, the length of the tweets can vary a lot. We have tweets made by only one character and tweets longer than 900.

Table 1. Datasets composition

Partition	users	max char	min char
En Train	2880	933	1
En Dev	1240	646	1
Es Train	2080	932	1
Es Dev	920	876	1

Evaluation Framework As we previously said, PAN does not disclose to the participants the Test dataset, in order to simulate a real case scenario. So, it is possible to evaluate the performance of the developed solution on the test set only using (TIRA). TIRA [17, 32] — Testbed for Information Retrieval Algorithms, is a platform that focuses on hosting shared tasks and facilitates the submission of software. Pan provides to all the participants a Virtual Private Server on Tira, where participants can deploy their model and run it on preset datasets. For the author profiling task of 2019, two Test sets were released on TIRA. The first Test was used during the pre-evaluation phase, at this stage it was possible to execute multiple runs on the Test set and to request the evaluation for all of them to the moderators. Instead, for the second Test set, the final one, it was possible to request the evaluation for just one run. Performances of runs are computed with the accuracy metric. Finally, it is important to note that the evaluation of the gender is based on the output of the bot detection task, this means that only the accounts classified as human by the bot detector are used to compute the accuracy of the gender classifier.

3 Bot Detection

Our classifiers to address the bot detection problem rely on the meta-estimator AdaBoost for the English language and on a single SVM architecture for the Spanish. Following we describe step by step how we built our models.

3.1 Features

In this section, we describe the features that we use for the bot detection classifier. For each account we compute the following features:

- **Emojis**: The average number of emojis used in each tweet.
- **Web link**: The average number of links shared in each tweet.
- **Hashtag**: The average number of hashtags used.
- **Len of Tweets**: The average length of the tweets.
- **Len of ReTweets**: The average length of the retweets.
- **Semicolons**: The average number of semicolons used.
- **Cosine Similarity score**: For all the tweets that belong to the same user, we weight each word with the Tf-Idf. Then, for each pair of tweets, we compute the cosine-similarity. Finally, we get as feature the average of the scores. The idea is that the cosine similarity of bots is higher than that of humans. Since the messages that belong to the same Bot, tend to be very similar among themselves.
- **Sentiment Analysis**: We perform the sentiment analysis to each tweet, then we use as features the average neutral sentiment score and the average on compound score. To perform the sentiment analysis, we use the Vader Sentiment Analysis tool [19]. It provides for the analyzed sentence a positive, a negative, a neutral and a compound score where the compound score is a single uni-dimensional measure that sums up the sentiment of the sentences.
- **Text Distortion**: We use the function of [39] to distort the text. The distortion technique is a pre-processing method that consists in masking some part of the text before the feature extraction. The distortion is used to emphasize the use of special characters and punctuation. This function transforms every ASCII characters into * and leaves unchanged no-ASCII characters. An example of distorted text is shown in Tab. 2. So, we first concatenate all the tweets belonging to the same user. Then, we replace all the emoticons with the tag ::EMOJI::, in this way we normalize all the emoticons, and at the same time after the text distortion we can still have a recognizable pattern (::*****::). Finally, we apply the distortion and we extract from the text the first 1000 char-grams by frequency of length from 2 up to 8, that appear at least 5 times among all the tweets of the account. The selected char-grams are then weighted with Tf-Idf in which the term frequency is logarithmically scaled.

In Tab 3 are reported the final dimension of each feature, and the total dimension of our features set.

Table 2. Text distortion

Original Text	**Distorted Text**
RT @BIBLE_Retweet: Great men are not always wise - Job 32:9	** @*****_*******: ***** *** *** *** ****** **** - *** **:*
I don't know. Just making conversation with you, Morty. What do you think, I-I-I... know everything about everything?	* ***'* ****. **** ****** ********** **** ***, *****. **** ** *** *****, *-*-*... **** ********** ***** **********?

Table 3. Dimension of features

Feature type	En Dimension	Es Dimension
Emojis	1	1
Web link	1	1
Hashtag	1	—
Len of Tweets	1	1
Len of ReTweets	1	1
Semicolons	1	1
Cosine Similarity Score	1	1
SA: Neutral	1	1
SA: Compound	1	1
Distortion Text tf-idf	1000	1000
Total	1009	1008

3.2 Features Reduction

Once extracted all the features, we use the PCA [18, 28] —Principal Component Analysis, to reduce the dimensionality of the features space with a minimum loss of information. In particular, in our algorithm we use the PCA implementation of sklearn. For the English bot classifier, we reduce the dimension of the features from 1009 to 56, while for the Spanish bot classifier from 1008 to 46. In both the cases we set the *withen* parameter of PCA equals to True.

3.3 Bot Classifiers

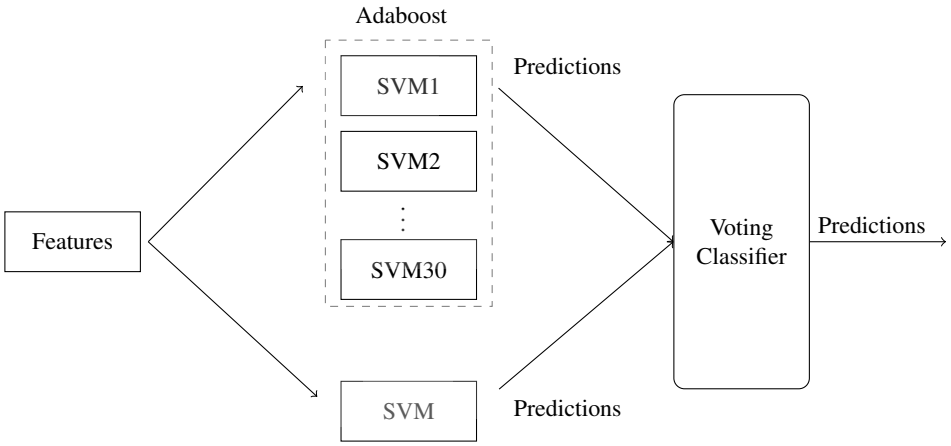
We built two different classifiers one to detect the English Bot and one for the Spanish one. In Fig. 1 is shown the architecture for the English Bot classifier. As we can see, the classifier is made by two layers. In the first layer, we have a single SVM [9] with an RBF kernel, and an AdaBoost instance. AdaBoost is a meta-estimator [15], that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. We use AdaBoost with 30 estimators and SVM as a base estimator, with a learning rate of 65% and the SAMME.R algorithm. In the second layer, we have a Soft-Voting Classifier that ensemble the predictions of the previous layer. To ensemble the predictions we sum up for each class the probability of being the right one as predicted by our classifiers, then we pick as final prediction, the class with the highest value.

For the Spanish bot classification, we use just a single SVM, since in this case, after some experiments, we found that it performs better than the ensemble architecture used for the English bot detection. The SVM on our experiment has Radial Basis Function kernel, with the hyper-parameters left to the default values.

3.4 Evaluation

To evaluate our model, accordingly with the metric evaluation of the task, we use the accuracy. As first experiment, we evaluate the features we extracted one by one, with

Figure 1. English bot classifier architecture



the same architecture described above. In Tab 4 are shown the performances. As we can see, all the features have similar performances for both the languages. Of particular relevance is the Distortion feature that alone achieves an accuracy of 90.48% for the English language, and the ratio of web links shared with more than 79%. The compound feature has a strange behavior, in fact on the English language it is very close to the random guess, while for the bot detection in Spanish it is around 68%.

In Tab. 5 are shown the results with our final model on the three evaluation datasets provided by PAN. As we can see, our model performs better for the English language than for the Spanish.

4 Gender Profiling

In this section, we describe our approach to gender detection. For this task, we use a single SVM architecture for both the languages.

4.1 Pre-Processing

Before extracting the features from the text, we pre-process the tweets in a different way depending on the language taken into account. For both the languages, we process all the tweets belonging to the same account as a single document. **English**: for the English accounts, we start our pre-processing phase tokenizing the text. To accomplish

Table 4. Features performance

Feature type	En Score	Es Score
Emojis	71.05%	70.65%
Web link	79.19%	87.93%
Hashtag	76.53%	—
Len of Tweets	59.59%	53.36%
Len of ReTweets	86.12%	81.41%
Semicolons	64.27%	55.21%
Cosine Similarity Score	60.00%	64.67%
SA: Neutral	61.53%	66.19%
SA: Compound	51.37%	68.36%
Distortion Text tf-idf	90.48%	81.73%
all without PCA	77.90%	76.73%

Table 5. Bot detection evaluation results

Test	English	Spanish
Dev	95.56%	92.06%
Test1	93.18%	88.89%
Test2	94.32%	90.78%

this operation, we use the TweetTokenizer of the NLTK python package [4]. We choose this tokenizer since it was designed exactly to web content. It can handle out of the box ASCII emoticons, and it replaces character sequences of length greater than 3 with sequences of length 3. After this step, we stem the tokenized text. We use the Snowball Stemming [31], from the NLTK library. Stemming is a technique of Natural Language Processing that reduces an inflected word to its base form. For example, given this list of words ['denied', 'died', 'agreed', 'owned'], the Snowball Stemmer produces as output the following words ['deni', 'die', 'agre', 'own'].

Spanish: Also, for the Spanish tweets, we start our pre-processing with the tokenization of the text. After this stage, we lemmatize the tweets instead of stemming them as done for the English. Lemmatizing like Stemming generate the root form of the inflected word, but while the stemmer operates on a single word without knowledge of the context, a lemmatizer does a full morphological analysis to identify the lemma for each word accurately. For instance, if we consider the word "meeting" that can be either the base form of a noun or a form of a verb ("to meet"), the lemmatizer tries to understand from the context of the sentence the right form and lemmatize the word accordingly. To lemmatize the tweets, we use the es_core_news_sm library of the SpaCy [1] framework.

[1] `https://spacy.io`

4.2 Features Extraction and Features Reduction

After the pre-processing phase, we move to the features extraction. Here, we use the same kind of feature, Word n-gram, but with different settings for English and Spanish. For the English gender detection, we select the 90000 most frequent words n-grams of length from 1 up to 5, that have a document frequency higher than 4. Then, we weight the n-grams with the Tf-Idf, in which the term frequency is logarithmically scaled. In the Spanish case, we use words n-grams of length from 1 up to 8, that appear in each document at least 7 times, then we weight them with the Tf-Idf. In this last case, we select the 50,000 most frequents n-grams. Once we have extracted all the features, we apply the Latent Semantic Analysis to reduce their dimensionality, the effectiveness of low dimensionality representation in the Author Profiling task is shown in [33]. Latent Semantic Analysis (LSA) [14] is a statistical approach to extracting relations among words by means of their contexts of use in documents. It makes no use of natural language processing techniques for analyzing morphological, syntactic, or semantic relations, nor does it use humanly constructed resources. LSA starting from the Tf-Idf , applies a reduced-rank singular value decomposition (SVD) on the Tf-Idf matrix to reduce the number of rows while preserving the similarity structure among columns. After this step, we have that each account for both the languages is represented by 11 features.

4.3 Classifier and Evaluation

In our experiments, we try different classifiers: Logistic Regression, Random Forest, and SVM. To evaluate our classifiers, we measure their performances using only the humans in the Dev set. In this way, we have not the bias of the bot miss-classified by the Bot detector. Unfortunately, we cannot repeat the same experiment on the Test sets, since they are not public. In Table 6 are shown the performances of the aforementioned classifiers. As we can see, the best result is achieved for both the languages with a single SVM with a radial base function kernel. We tune the hyper-parameter of the classifiers through grid search. We find that the best configuration for our models is to use for the English model a penalty coefficient of 8192, while of 4096 for the Spanish model, we left the remaining hyper-parameters to the default values. In Tab. 7 are shown the final results of our classifiers, in this last case, we made a full evaluation accordingly with the PAN guidelines, so we evaluate all the users that the bot detector classified as Human. Investigating on the miss-classified English speaking accounts of the Dev set, we note that most of them are humans that we erroneous classify as bot. So, we look inside their messages, and we found that most of them have strange posting behavior. As an example, they share almost the same messages multiple times, or the majority of their tweets are bible verses.

5 Related Work

Bot Detection. Since 2006, the year Twitter was released to the public, it catches the attention of the research community both from a sociological and technical point of view.

Table 6. Evaluation results of gender detection on humans of Dev set

Test	English	Spanish
Logistic Regression	75.16%	64.43%
Random Forest	72.58%	63.91%
SVM	85.48%	71.30%

Table 7. Final results of gender detection

Test	English	Spanish
Dev	80.42%	63.68%
Test1	80.68%	69.44%
Test2	84.17%	77.61%

Among these works, some of them focused on detecting the nature of the accounts, such as fake followers [10], bots, spammers, advertising account [40], with different approaches. Morstater et al. [26] clustered the approaches by the kind of features used in the detection outlining three different categories. The first one is characterized by the detection algorithm that exploits the content of the messages [45]. The second to those exploit the information contained in the profile description or account details like the mail address used for the sign-in or the device used for tweeting [41, 42]. Lastly, the one that takes into account the network structure and the connection of the accounts under investigation [24]. Chu et al. [8] were the first to address the problem of the bot detection on Twitter. They focused on the classification of Twitter accounts into three categories: Human, Bot and Cyborg — hybrid accounts managed by a human assisted by software and vice versa. They noticed that human have complex timing behaviors in terms of posting while bots and cyborgs posting at regular times. Moreover, they found that Bots posts contents are very often spam messages and that Bots share more links than human. In their classification, they achieve an average accuracy of 96% with a Bayesian classifier on a balanced dataset made by 6,000 samples. Chavoshi et al. [6] showed that it is possible to detect automated account exploiting only the time series of their posting, achieving with their methodology an accuracy of 94%. Varol et al. [44] estimated that between 9% and 15% of active Twitter accounts are bots. For the estimation they used a Random Forest classifier and 1150 different kinds of features belonging to the three categories described above, their classifier can separate the classes with an Area Under the Curve of 0.94. Finally, Koulompis et al. [21] were the first to introduce the sentiment analysis in the task of bot detection on Twitter. In [3, 13] in conjunction with other kinds of features, they exploited the sentiment analysis to analyze the presence of bots in the tweets related to the Indian and U.S elections respectively.

Gender identification. The problem of automatic profiling the author of a text is a crucial problem in many application scenarios, such as forensic, security, and commercial settings [2]. The goal of the task is to infer as much as possible information about an unknown author. The profile traits most studied in the literature are: Age, gender,

personality traits, native language. Following we focus on the works that address the gender identification problem. The linguistic community was the first researchers to face the problem of gender identification [22, 23, 43]. Pennebaker et al. [30] address the problem from a psychological point of view, they conclude that the stylistic differences between women and men are consistent with a sociological framework of gender differences in access to power. In the field of the automatic gender identification several approaches were proposed and on different datasets. Cheng et al. [7] explore the problem on the Reuters and Enron corpus using three different classifiers SVM, Decision tree, and logistic regression. In [1, 5, 12, 25] the authors deal with different twitter datasets built by themselves. Also, the participant of the PAN author profiling task of the years 2015 [38],2017 [36] and 2018 [36] address the problem on a twitter dataset provided by the task organizers. Finally, datasets built from the heterogeneous social network was used in [2, 35, 37] From the point of view of the features, we can divide the features used in these works into three macro-categories: Syntactic features, that capture the writing style of the authors. As example in [7, 27] the authors found that women and men have different habits of using punctuation; for example, women tend to use more question marks with respect to the men. Function words are the words that have an ambiguous meaning and express grammatical relationships among other words within a sentence. Instead, in [16, 20, 29] the authors noticed that women tend to use 'I', 'me', and 'my' more frequently than men, or that the women use intensive adverbs and positive adjectives more than men. Finally, we have the char and the word n-grams highly adopted in every stylometric task.

6 Conclusion and future work

In this work, we presented our approach to the Author profiling task of PAN2019. We develop 4 different classifiers, one for each problem subset. For the Bot detection of English written messages, we used an ensemble architecture where the AdaBoost outputs are ensembled by a soft-voting classifier. For each one of the other problems, we use a single fine-tuned SVM. We achieve excellent performances, especially in the bot detection task, where we record a score of about 95% on the English Dev and the English final Test set. Regarding gender detection, our model achieves an accuracy of 85.48% on English Dev set and of 71.30% on the Spanish one, when there is no bias introduced by the Bot detection. If we consider the full pipeline of the task, that means to evaluate the gender on all the accounts classified as human by the bot classifiers, our model achieves an accuracy of 80.42% and 63.68% respectively for the English and Spanish datasets. Globally our models perform better on the English accounts than the Spanish ones, so we believe that more work is needed to fill this gap. Finally, we believe that the performances of the classifiers can even be better if the model can take into account at least the hostname of the link found in the dataset. Moreover, in a real case scenario, also the profile description of the account, the username, and the profile image can help to boost the performance as shown in the previous editions.

Acknowledgment

This work was supported in part by the MIUR under grant “Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University.

References

1. Alowibdi, J.S., Buy, U.A., Yu, P.: Empirical evaluation of profile characteristics for gender classification on twitter. In: 2013 12th International Conference on Machine Learning and Applications. vol. 1, pp. 365–369. IEEE (2013)
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM 52(2), 119–123 (2009)
3. Badawy, A., Ferrara, E., Lerman, K.: Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 258–265. IEEE (2018)
4. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O’Reilly Media, Inc." (2009)
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. Chavoshi, N., Hamooni, H., Mueen, A.: Debot: Twitter bot detection via warped correlation. In: ICDM. pp. 817–822 (2016)
7. Cheng, N., Chandramouli, R., Subbalakshmi, K.: Author gender identification from text. Digital Investigation 8(1), 78–88 (2011)
8. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Transactions on Dependable and Secure Computing 9(6), 811–824 (2012)
9. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
10. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: Fame for sale: Efficient detection of fake twitter followers. Decision Support Systems 80, 56–71 (2015)
11. 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 (Sep 2019)
12. Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., Hu, W.: Gender identification on twitter using the modified balanced winnow. Communications and network 4(3), 189–195 (2012)
13. Dickerson, J.P., Kagan, V., Subrahmanian, V.: Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In: Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 620–627. IEEE Press (2014)
14. Dumais, S.T.: Latent semantic analysis. Annual review of information science and technology 38(1), 188–230 (2004)
15. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55(1), 119–139 (1997)

16. Gleser, G.C., Gottschalk, L.A., John, W.: The relationship of sex and intelligence to choice of words: A normative study of verbal behavior. Journal of Clinical Psychology 15(2), 182–191 (1959)
17. Gollub, T., Stein, B., Burrows, S., Hoppe, D.: Tira: Configuring, executing, and disseminating information retrieval experiments. In: 2012 23rd International Workshop on Database and Expert Systems Applications. pp. 151–155. IEEE (2012)
18. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of educational psychology 24(6), 417 (1933)
19. Hutto, C.J., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media (2014)
20. Jaffe, J.M., Lee, Y., Huang, L., Oshagan, H.: Gender, pseudonyms, and cmc: Masking identities and baring souls. In: 45th Annual Conference of the International Communication Association, Albuquerque, New Mexico (1995)
21. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: Fifth International AAAI conference on weblogs and social media (2011)
22. Labov, W.: The intersection of sex and social class in the course of linguistic change. Language variation and change 2(2), 205–254 (1990)
23. Lakoff, R.: Language and woman's place. Language in society 2(1), 45–79 (1973)
24. Lee, K., Eoff, B.D., Caverlee, J.: Seven months with the devils: A long-term study of content polluters on twitter. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
25. Liu, W., Ruths, D.: WhatâĂŹs in a name? using first names as features for gender inference in twitter. In: 2013 AAAI Spring Symposium Series (2013)
26. Morstatter, F., Wu, L., Nazer, T.H., Carley, K.M., Liu, H.: A new approach to bot detection: striking the balance between precision and recall. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 533–540. IEEE (2016)
27. Mulac, A.: The gender-linked language effect: Do language differences really make a difference? Lawrence Erlbaum Associates Publishers (2006)
28. Pearson, K.: Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2(11), 559–572 (1901)
29. Pennebaker, J.W., King, L.A.: Linguistic styles: Language use as an individual difference. Journal of personality and social psychology 77(6), 1296 (1999)
30. 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)
31. Porter, M.F.: Snowball: A language for stemming algorithms (2001)
32. 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)
33. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 156–169. Springer (2016)
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. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, 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)

36. 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. Working Notes Papers of the CLEF (2018)
37. 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 (2016)
38. Rangel Pardo, F.M., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: CLEF 2015 Evaluation Labs and Workshop Working Notes Papers. pp. 1–8 (2015)
39. Stamatatos, E.: Authorship attribution using text distortion. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 1138–1149 (2017)
40. Stringhini, G., Egele, M., Kruegel, C., Vigna, G.: Poultry markets: on the underground economy of twitter followers. ACM SIGCOMM Computer Communication Review 42(4), 527–532 (2012)
41. Thomas, K., Grier, C., Paxson, V.: Adapting social spam infrastructure for political censorship. In: Presented as part of the 5th {USENIX} Workshop on Large-Scale Exploits and Emergent Threats (2012)
42. Thonnard, O., Dacier, M.: A strategic analysis of spam botnets operations. In: Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference. pp. 162–171. ACM (2011)
43. Trudgill, P.: Sex, covert prestige and linguistic change in the urban british english of norwich. Language in society 1(2), 179–195 (1972)
44. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: Detection, estimation, and characterization. In: Eleventh international AAAI conference on web and social media (2017)
45. Wang, A.H.: Detecting spam bots in online social networking sites: a machine learning approach. In: IFIP Annual Conference on Data and Applications Security and Privacy. pp. 335–342. Springer (2010)