Celebrity Profiling on Twitter using Sociolinguistic Features

Notebook for PAN at CLEF 2019

Luis Gabriel Moreno-Sandoval[1,3], Edwin Puertas[2,1,3], Flor Miriam Plaza-del-Arco[4], Alexandra Pomares-Quimbaya[1,3], Jorge Andres Alvarado-Valencia[1,3], and L. Alfonso Ureña-López[4]

[1] Pontificia Universidad Javeriana, Bogotá, Colombia
{edwin.puertas,jorge.alavarado,morenoluis,pomares}@javeriana.edu.co
[2] Universidad Tecnológica de Bolívar, Cartagena, Colombia
epuerta@utb.edu.co
[3] Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA)
[4] Universidad de Jaén, Jaén, Andalucía, Spain.
{fmplaza, laurena}@ujaen.es

Abstract Social networks have been a revolutionary scenario for celebrities because they allow them to reach a wider audience with much higher frequency than using traditional means. These platforms enable them to improve or sometimes deteriorate, their careers through the construction of closer relationships with their fans and the acquisition of new ones. Indeed, networks have promoted the emergence of a new type of celebrities that exists only in the digital world. Being able to characterize the celebrities that are more active on social networks, such as Twitter, gives an enormous opportunity to identify what is their real level of fame, what is their relevance for an age group, or a specific gender or occupation. These facts may enrich decision making, especially in advertising and marketing. To achieve this aim, this paper presents a novel strategy for the characterization of celebrities profile on Twitter based on the generation of socio-linguistic features from their posts that serve as input to a set of classifiers. Specifically, we produced four classifiers that describe the level of fame, the gender, the birth date, and the possible occupation of a celebrity. We obtained the training and test data sets as part of our participation at PAN 2019 at CLEF. Results of each classifier are reported including the analysis of which features are more relevant, which classification techniques were more useful and which were the final precision and recall results.

Keywords: celebrity profiling, socio-linguistic feature, user profiling, computational linguistic, natural language processing, author profiling, twitter

1 Introduction

Author profiling is a sub task of authorship analysis whose objective is the analysis of shared content in order to predict different characteristics of authors such as gender, age, personality or native language [15].

Knowing the profile of an author could be of vital importance in multiple areas. In marketing to understand what types of people like or dislike some products and analyzing their online reviews [14]. In safety to identify psychological traits that allow to detect profiles with abnormal behaviors that may cause harm to other users [6] or to discover fake profiles (one person can have multiple profiles for fraudulent and other misdeeds)[5].

With the increasing usage of social media and the rapid expansion of user generated content, author profiling task has gained a lot of interest in the last years.This task is a research topic in the natural language processing community on which various shared tasks have been generated recently. Perhaps one of the best-known shared tasks is the one organized at PAN [18] at the Conference and Labs of the Evaluation Forum (CLEF)[5] since 2013. Specifically, the focus has been on gender and age identification.

Social media have meant a real revolution for famous people, that is, celebrities, such as artists, sportsmen, among others, who take advantage of Facebook, Twitter or other platforms to get closer to their fans, and in turn, get a new way to earn income [10]. Studying the profile of each celebrity allows us to extract certain characteristics, such as the vocabulary they use to refer to their profession, the way of writing, the way of communicating with their fans, and their possible age or profession.

In this paper, we describe our submission as part of our participation at PAN 2019 [4] at CLEF. In particular, we have participated in the celebrity profiling task [20]. It is the first year that this task is organized and it consists of determining the degree of fame, occupation, age, and gender of a celebrity, given his/her social media feed. Our main contribution is to generate and analyze specific features from celebrity of digital social networks and incorporate them into different machine learning classifiers.

The rest of the paper is structured as follows. In Section 2, we introduce the relate work. In Section 3, we explain the data set used in our strategy for celebrity characterization. Section 4 presents the details of the proposed strategy. In Section 5 and 6, we discuss the analysis and evaluation results. We conclude in Section 7 with remarks and future work.

2 Related work

Early research on the profile of authors focused mainly on formal texts and blogs. However, today's researchers focus primarily on social media platforms such as Twitter of Facebook, where the language is less formal and users post messages continuously [15]. The contribution of the different researchers who used the PAN datasets is remarkable.

Most of the strategies presented at PAN used combinations of features based on styles such as frequency of punctuation marks, capitalization, together with part-of-speech tags and content-based features such as bag of words, dictionary-based words,

[5] http://clef-initiative.eu

topic-based words, entropy-based words or term frequency inverse document frequency (TF-IDF) [17]. In the Author Profiling Task at PAN 2017 and PAN 2018, more participants employed deep learning techniques, which perform automatic feature selection. However, in the gender and language variety subtasks, the best performances belonged to a logistic regression classifier with combinations of character, word, and POS n-grams, emojis, sentiments, character flooding, an SVM trained with combinations of character and TF-IDF n-grams [16]. Basile et al. [2] used word unigram and character n-grams. They extracted character three to five grams and word unigrams to bigrams with TF-IDF weighting. Authors in [9] combined POS (Part Of Speech) tags n-grams with syntactic dependencies to model the use of amplifiers, verbal constructions, pronouns, subjects and objects, types of adverbials, as well as the use of interjections and profanity. The authors in [15] used the counts of stopwords, punctuation marks, emoticons, and slang words.

Copland et al. [3] showed that through the study of the "personal pronoun", specially the use of "me" and "us" is possible to identify important sociolinguistic variables. These variables can be associated with the social status of a person and the school from which the person comes (e.g. Christian, Lutheran). Hence some of the approaches we applied to identify features of celebrities take into account the use of personal pronouns in the texts.

Regarding the machine learning approaches, the most commonly used classifiers have been Logistic Regression [12,8], Support Vector Machines [19,1], Multilayer Perceptron [8] and distance-based methods.

3 Data Description

The training dataset of the celebrity profiling task at PAN 2019 [20] consists of English tweets with the following features: degree of fame, occupation, age, and gender that includes 48335 user profiles with 2181 tweets avg. per user. Table 1 provides details about some attributes of the dataset. The task is to predict four traits of a celebrity from their social media communication.

4 System Description

In this section, we describe the predictive model used in our submission. The model used for the task of profiling celebrities at PAN 2019 was designed to identify four types of classes: profession, gender, fame and year of birth. In accordance with the characteristics of the data set and the goals of the task we defined four hypotheses, which are described in detail in Table 2.

In addition, for each of the hypotheses, two types of strategies were used. The first strategy is related to the vocabulary associated with the words of all tweets. For the other strategy, tweets statistics were generated by user profiles to determine the global use of words, hashtags, mentions, URLs, and emoji. Taking into account the above-mentioned assumptions and strategies.

On the basis of the proposed hypotheses and strategies, the "Training System" was designed. Figure 1 shows the proposed system to predict celebrities, which consists of

Table 1. Characteristics of dataset

Label	#Tweets	Value
Degree of fame	7116	Superstar
	25230	Star
	1490	Rising
Gender	24221	Male
	9583	Female
	32	Undefined
Occupation	13481	Sport
	9899	Performer
	5475	Creator
	2835	Politics
	818	Science
	525	Professional
	768	Manager
	35	Religious

the following stages: preprocessing, standardization and transformation, extraction of characteristics, configuration and classifiers, and testing.

4.1 Preprocessing

In the preprocessing stage, we use the concatenated vocabulary of each user's tweets, in order to have only one document per user profile. In addition, the re-labeling of the hashtags is applied, which was done with the word "label_hashtag", the mentions word with the word "label_mention", the URLs with the word "label_url", and the emojis by UTF-8 were replaced with the word "label_emoji". Finally, globally re-tagged words are searched and counted.

4.2 Normalization and Transformation

The next stage is associated with normalization and transformation process. The normalization process is related to the balance of the classes and the generation of random samples for the training and testing process. With respect to the transformation process, the vector representation of words is performed and the use of the features for each user profile is calculated. This process can be configured in such a way that the vectorial representation of the words can be done with "N-gram" and the global features related to the tweets of the user profiles can also be parameterized.

4.3 Feature Extraction

The features based on the use of words, hashtags, mentions, URLs, and emojis that are calculated for each one of the tweets by profile in the celebrity system are Table 3.

Table 2. Description of hypothesis – H0

Class	Description - H0
Profession	The profession is mainly associated with the use of "specialized" vocabulary. Therefore, the classification process must be based on the vocabulary collected by each profession.
Gender	In gender, we want to establish features for the use of emojis, hashtags, mentions, RT and URLs. For this, it is expected that the features associated with the words added to those found in the user profiles will improve the classifications.
Fame	Fame is perhaps the most important label in establishing features such as the use of emojis, hashtags, mentions, RT and URL. In addition, it is verified if the message is written in first, second or third person. With the above, it is expected that the features associated with the words added to those found in the usage profiles will improve the classifications.
Birth years	This label is perhaps the most difficult to classify because the wide range of years from 1940 to 2011. For this reason, groups were established in order to generate features of use of emojis, hashtags, mentions, RT and URLs. Also, it was contemplated if the message was written in first, second and third person. With the above, it is expected that the features associated with the words added to those found in the usage profiles will improve the classifications.

These metrics allow us to see the distribution of each feature in the profile, and for some of them kurtosis and asymmetry are calculated.

It measures have them as a complement of the averaged data in the associated topics of the size of each word or the number of words per tweet. The main idea is to be able to have a real form of these two measures given that the average may not show the complete information. The rest of the average features regarding the characteristics of a social network such as hashtags, mentions, emojis, URLs, and retweets are also used in the profile.

The lexical diversity was represented using the feature Text-Type Ratio (TTR) [11]. This measure allows us to see what is the use of vocabulary concerning all the words included in the texts, which we think is very useful for detecting bots or specific kind of people. Finally, the use of the first, second, or third person, singular or plural could also show us social characteristics.

4.4 Settings and classifiers

At the configuration stage, the system will adjust machine hardware parameters such as processors and threads. In addition, different scenarios can be configured for the use of the classifiers. Finally, the system may be adjusted to store the best performing vector words and qualifiers. It should be noted that during the execution of the system, the data set was divided into 60 % for training and 40 % for tests for all our experiments.

On the other hand, based on the previous tasks carried out in the PAN, several classifiers were examined, such as Naive Bayes (NB), Gaussian Naive Bayes (GNB), Naive Bayes Complement (CNB), Logistic Regression (LR), and Random Forests (RF).

Figure 1. System Training.

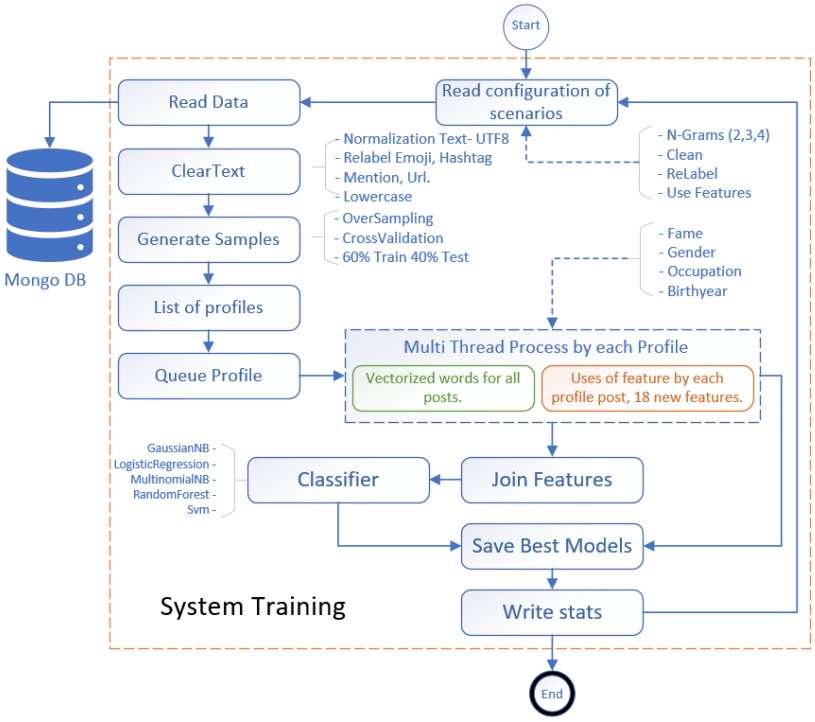


Table 4 shows the most successful classification models after running 552 scenarios with and without the user feature. In addition, the number of runs per model is shown for birthyear, fame, gender, and occupation classes.

4.5 Test

In the test stage, a software component that performs the following activities was developed. First, the test data sets are read. The tweets are processed by each user. Afterwards, the features of the use are calculated. As shown in Table 4, different models were created looking for the best classifications. Subsequently, vector representation is made. The best classifiers for Fame, Birthyear, Occupation, and Gender classes are then calculated. Finally, the best predictors are exported. Figure 2 shows the "System Test" used by our models.

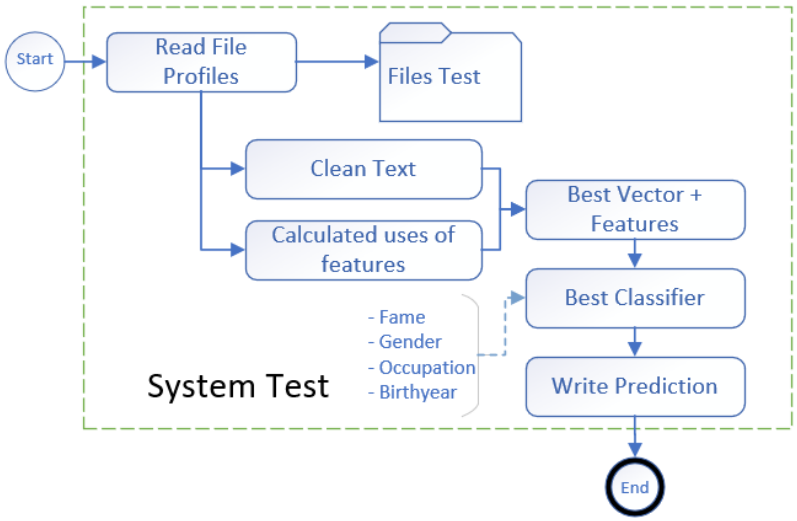
5 Experiments and Analysis of Results

During the pre-evaluation phase, we carried out different experiments, and the best ones were taken into account for the evaluation phase. The system has been evaluated using the usual competition metrics, including Accuracy (Acc), Precision (P), Recall (R) and

Table 3. Features Description

#	Feature	Description
1	stats_avg_word	Average word size per tweet
2	stats_kur_word	Kurtosis of the variable stats_avg_word
3	stats_label_emoji	Amount of emojis per tweet for the profile
4	stats_label_hashtag	Number of hastags per tweet for the profile
5	stats_label_mention	Number of mentions per tweet for the profile
6	stats_label_url	Number of urls per tweet for the profile
7	stats_label_retweets	Number of retweets per tweet for the profile
8	stats_lexical_diversity	Lexicon diversity for all tweets by profile
9	stats_label_word	Number of words per tweet for the profile
10	kurtosis_avg_word	Kurtosis of the variable stats_kur_word
11	kurtosis_label_word	Kurtosis of the variable stats_label_word
12	skew_avg_word	Statistical asymmetry of the variable stats_avg_word
13	skew_label_word	Statistical asymmetry of the variable stats_label_word
14	stats_person_1_sing	Number of tweets used by the first person of the singular
15	stats_person_2_sing	Number of tweets used by the second person singular
16	stats_person_3_sing	Number of tweets used by the third person singular
17	stats_person_1_plu	Number of tweets used by the first and second person of the plural
18	stats_person_3_plu	Number of tweets used by the third person plural

Figure 2. System Training.



F1-score (F1). The best systems in the pre-evaluation phase will be explained in detail in the following sections.

As can be seen in Table 5, the summary shows the performance of each label calculated for the challenge. For each label, it is observed the best classification model, the accuracy obtained with it and the features that worked best for the classification. The

Table 4. Classification models with better results

Models	Birthyear	Fame	Gender	Occupation	Total
ComplementNB	-	-	-	6	6
GaussianNB	5	5	6	3	19
GaussianNB + userfeatures	5	5	6	3	19
LogisticRegression	5	5	6	9	25
LogisticRegression + userFeatures	5	5	6	3	19
MultinomialNB	-	-	-	6	6
RandomForest	5	5	6	9	25
RandomForest + userFeatures	5	5	6	3	19
Total	30	30	36	42	138

Table 5. Summary of results in celebrity profiling for class

Class	Acc	Model
Fame	0.65	Logistic Regression
Gender	0.88	Logistic Regression
Occupation	0.567	Multinomial NB
Birthyear	0.387	Logistic Regression

classifiers that obtained the best performance were Logistic Regresion and Multinomial Naive Bayes. Finally, it describes the pre-processing performed, whether the dataset has been cleaned or not, whether the 18 characteristics have been used in the classification and the minimum word frequency in the vector words.

It should be noted that the system presented was trained and tested with the celebrity dataset provided by the official site of PAN 2019 [20]. Also, the presentations were made on the TIRA [13] platform in which we configure a virtual server with ten processors; we set up the environment to perform the tests [7].

5.1 Fame classification

The variable fame is perhaps the most important for the competition. As it was raised in the hypothesis the 18 proposed features has an impact on its classification . The results obtained for this variable are as follows. For the model that was evaluated with the 18 proposed features it has an accuracy of 0.65. While the model that was evaluated only with the traditional bag-of-words has an accuracy of 0.51. The results show that the features used in our model describe this variable more accurately.

As can be seen in Table 6, this variable achieved the best performance with the logistic regression classifier where the proposed features were used. Moreover, we performed the following steps: pre-processing, standardization, cleaning, and re-labeling. Finally, only words with frequencies higher than three were taken into account in the vocabulary counting matrix.

Table 6. Fame classification

Class	Precision	Recall	F1-Score	Support
0 - rising	0.69	0.71	0.7	551
1 - star	0.56	0.54	0.55	784
2 - superstar	0.7	0.71	0.71	820
Micro avg	0.65	0.65	0.65	2155
Macro avg	0.65	0.65	0.65	2155

Table 7. Gender classification

Class	Precision	Recall	F1-Score	Support
0 - female	0.87	0.89	0.88	790
1 - male	0.89	0.88	0.88	813
2 - nonbinary	0.36	0.4	0.38	10
Micro avg	0.88	0.88	0.88	1613
Macro avg	0.71	0.72	0.71	1613

5.2 Gender classification

The gender variable has an additional variation because it includes an extra non-binary variable. The inclusion of the non-binary variable presented us with a significant challenge because there was a significant imbalance with this new value. Based on the above, it was hypothesized that the proposed use of the characteristics would have an impact on the classification. Subsequently, it was corroborated that the addition of the non-binary viable has a significant effect on the model, given that it describes it extensively with an accuracy of 0.36. And with the addition of 18 characteristics to the model, it resulted in an accuracy of 0.88.

As can be seen in Table 7, this variable achieved the best performance with the logistic regression classifiers where the features were used. Moreover, we performed the following steps: pre-processing, normalization, cleaning, and re-labeling. Finally, only words with frequencies greater than 9 are taken into account in the vocabulary counting matrix.

5.3 Birth year classification

as it was proposed in the hypothesis, the birth year model have a better accuracy when using the new features proposed. The result of the words vector model, adding the features increases accuracy to 0.37. The addition of the 18 features in the model gave a gain of 0.29 in the accuracy.

For the birth year classification we discretized the variable using a a window size m based on the birth year. The value increases linearly from about 2 years for 2012 to about 9 years for 1940.

As can be seen in Table 8, this variable achieved the best performance with the logistic regression classifier where the features were used. Moreover, we performed the

following steps: pre-processing, normalization, cleaning, and re-labeling. Finally, in the vocabulary counting matrix, only words with frequencies higher than six are taken into account, and the significant imbalance of this variable also led to an oversampling process.

Table 8. Birth year classification

Class	Precision	Recall	F1-Score	Support
0 - [1940, 1950]	0.19	0.16	0.17	182
1 - [1950, 1960]	0.38	0.22	0.28	401
2 - [1960, 1970]	0.29	0.43	0.35	385
3 - [1970, 1980]	0.28	0.25	0.26	390
4 - [1980, 1990]	0.42	0.41	0.41	412
5 - [1990, 2000]	0.66	0.66	0.66	404
6 - [2000, 2012]	0.25	0.37	0.3	163
Micro avg	0.37	0.37	0.37	2338
Macro avg	0.35	0.36	0.35	2338

In this variable, we have evaluated the accuracy with the data initially delivered by the competition, and we did not use any additional changes for the final evaluation.

5.4 Occupation classification

The occupation variable, as we said in the initial hypothesis, is initially based on specialized vocabulary. However, we did not use new approaches such as embeddings, ontologies or other technologies. The results showed that using user profile identification in the occupation variable calculation, did not significantly affect it. After applying the vocabulary and the 18 characteristics to the models used, an accuracy of 0.57 was obtained.

As can be seen in Table 9, this variable achieved the best performance with the Multinomial Naive Bayes classifier where only the vocabulary was used without any cleaning. Finally, only words with frequencies higher than three are taken into account in the vocabulary counting matrix.

6 Result Test

As shown in Table 10, the models were tested using the training dataset, the test1 dataset and the test2 dataset. In the ranking of the task, we occupied the second position.

7 Conclusions and Future Work

The task of celebrities CLEF-PAN 2019 generated several challenges that are worth highlighting. First, we have four classes to calculate a celebrity, but the number of

Table 9. Occupation classification

Class	Precision	Recall	F1-Score	Support
0 - creator	0.47	0.42	0.44	402
1 - manager	0.58	0.18	0.28	288
2 - performer	0.53	0.79	0.64	395
3 - politics	0.66	0.82	0.73	391
4 - professional	0.31	0.13	0.18	0.191
5 - religious	0.25	0.14	0.18	14
6 - science	0.49	0.43	0.46	298
7 - sports	0.69	0.88	0.77	405
Micro avg	0.57	0.57	0.57	2384
Macro avg	0.5	0.47	0.46	2384

Table 10. Summary of results in the Datasets

Class	Dataset Training	Dataset Test1	Dataset Test2
Fame	0.82	0.56	0.51
Gender	0.64	0.64	0.56
Occupation	0.54	0.46	0.41
Birthyear	0.56	0.51	0.51
C-Rank	0.63	0.54	0.49

values that had each of them was a problem. The most critical was the birth year class in which its dimensionality was reduced, creating groups of profiles every ten years for better accuracy in the classification.

On the other hand, the training dataset of celebrities had an evident imbalance in some of the classes. For example: the birth year was imbalanced, gender has only 32 samples of the type non-binary, and finally occupation values like religion had only 35 samples. Some of these challenges were solved with strategies of balancing examples by performing oversampling.

The volume of data was another important challenge, it was necessary to process more than 53 million tweets associated with the profiles analyzed. To deal with that, we work on a cluster of 10 servers.

The novelty in the analysis presented in this paper is to analyze specific features of digital social networks for each profile. The use of sociolinguistic features in the user profile has shown many quirks in topics social, cultural, and of gender. These characteristics describe the sociolect of celebrities linked in this study; we also find it is essential to understand if the text was written in the first, second or third person, and the lexical diversity that each profiles had.

As future work, we plan to analyze the models with real samples with a similar or greater volume of messages. Finally, we want to review the posts and context data to have models that respond socially to variables that represent real phenomena in the network.

Acknoledgments

We thank the Center for Excellence and Appropriation in Big Data and Data Analytics (CAOBA), Pontificia Universidad Javeriana, and the Ministry of Information Technologies and Telecommunications of the Republic of Colombia (MinTIC). The models and results presented in this challenge contribute to the construction of the research capabilities of CAOBA. Also, Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) and LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government. Finally, the author Edwin Puertas gives thank Universidad Tecnológica de Bolívar. Needless to say, we thank the organizing committee of PAN, especially Paolo Rosso, Francisco Rangel, Matti Wiegmann and Martin Potthast for their encouragement and kind support.

References

1. Aragón, M.E., López-Monroy, A.P.: A straightforward multimodal approach for author profiling. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018) (2018)
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. Copland, F., Shaw, S., Snell, J.: Linguistic ethnography: interdisciplinary explorations. Springer (2016)
4. 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)
5. Fatima, M., Hasan, K., Anwar, S., Nawab, R.M.A.: Multilingual author profiling on facebook. Information Processing & Management 53(4), 886–904 (2017)
6. Ferrari, A., Consoli, A.: Building accurate hav exploiting user profiling and sentiment analysis. arXiv preprint arXiv:1609.07302 (2016)
7. 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)
8. HaCohen-Kerner, Y., Yigal, Y., Elyashiv Shayovitz, D.M., Breckon, T.: Author profiling: Gender prediction from tweets and images (2018)
9. Karlgren, J., Esposito, L., Gratton, C., Kanerva, P.: Authorship profiling without using topical information: Notebook for pan at clef 2018. In: 19th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2018, Avignon, France, 10 September 2018 through 14 September 2018. vol. 2125. CEUR-WS (2018)
10. Khamis, S., Ang, L., Welling, R.: Self-branding,'micro-celebrity'and the rise of social media influencers. Celebrity Studies 8(2), 191–208 (2017)
11. McCarthy, P.M., Jarvis, S.: vocd: A theoretical and empirical evaluation. Language Testing 24(4), 459–488 (2007), https://doi.org/10.1177/0265532207080767
12. Nieuwenhuis, M., Wilkens, J.: Twitter text and image gender classification with a logistic regression n-gram model. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018) (2018)

13. 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)
14. Rangel, F., Rosso, P.: Use of language and author profiling: Identification of gender and age. Natural Language Processing and Cognitive Science 177 (2013)
15. 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)
16. 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)
17. Rosso, P., Rangel, F., Farías, I.H., Cagnina, L., Zaghouani, W., Charfi, A.: A survey on author profiling, deception, and irony detection for the arabic language. Language and Linguistics Compass 12(4), e12275 (2018)
18. Stamatatos, E., Rangel, F., Tschuggnall, M., Stein, B., Kestemont, M., Rosso, P., Potthast, M.: Overview of pan 2018: Author identification, author profiling, and author obfuscation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Association, CLEF 2018. Avignon, France, September 10-14/Bellot, Patrice [edit.]; et al. pp. 267–285 (2018)
19. Tellez, E.S., Miranda-Jiménez, S., Moctezuma, D., Graff, M., Salgado, V., Ortiz-Bejar, J.: Gender identification through multi-modal tweet analysis using microtc and bag of visual words. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018) (2018)
20. Wiegmann, M., Stein, B., Potthast, M.: Overview of the Celebrity Profiling Task at PAN 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)