Computational Social Networks

RESEARCH Open Access

Celebrity profiling through linguistic analysis of digital social networks

Luis G. Moreno-Sandoval[1,2*], Alexandra Pomares-Quimbaya[1,2] and Jorge A. Alvarado-Valencia[1,3]

*Correspondence:
morenoluis@javeriana.edu.co
[2] Department of System Engineering, Pontificia Universidad Javeriana, Bogota, Colombia
Full list of author information is available at the end of the article

Abstract

Digital social networks have become an essential source of information because celebrities use them to share their opinions, ideas, thoughts, and feelings. This makes digital social networks one of the preferred means for celebrities to promote themselves and attract new followers. This paper proposes a model of feature selection for the classification of celebrities profiles based on their use of a digital social network Twitter. The model includes the analysis of lexical, syntactic, symbolic, participation, and complementary information features of the posts of celebrities to estimate, based on these, their demographic and influence characteristics. The classification with these new features has an F1-score of 0.65 in Fame, 0.88 in Gender, 0.37 in Birth year, and 0.57 in Occupation. With these new features, the average accuracy improve up to 0.14 more. As a result, extracted features from linguistic cues improved the performance of predictive models of Fame and Gender and facilitate explanations of the model results. Particularly, the use of the third person singular was highly predictive in the model of Fame.

Keywords: Celebrity profile, Natural Language Processing, Author profile, Demographic features, Influential feature

Introduction

Digital social networks (DSN) have become popular as a means of spreading information and connecting people with like-minded ones [1]. The capacity to spread opinions shows a general phenomenon with relevant implications in the context of social influence [2].

Public accessibility of DSN along with the ability to share and exchange opinions, thoughts, and feelings, among others, allow people to connect not only with friends and family, but also with any celebrity on the network [3]. This ability has been evident in the growth of DSN communication [4]. However, the success of such communication attempts depends on the level of trust that members have with each other [1, 5], considering that opinions have helped to influence the feelings and emotions of the public [6].

For this reason, the interest of researching on micro blog communities with services such as Twitter is growing exponentially [7] due to the massive production of written information that each user generates. This information includes personal data such as name, photograph, location, etc.; quantitative data such as the number of followers and

people they follow; and also their timeline, which is the chronology of their messages both public and private. Likewise, a user can follow another one by accepting to receive the messages that the other user posts [8].

On the other hand, language variation is permanent and evident in the new ways of writing in DSN. Such variation is not necessarily random, but highly related to social factors [9]. In fact, the linguistic ethnography holds that:

> *"to a considerable degree, language and the social world are mutually shaping, and that close analysis of situated language use can provide both fundamental and distinctive insight into the mechanisms and dynamics of social and cultural production in everyday activity" [10]*

When people share their opinions in DSN (such as Twitter), they might also be revealing demographic, social and/or psychosocial information about themselves. For example, Schwartz and colleagues [8] indicate that his research has been driven to an integral exploration of the language that differentiates people, giving a new perspective to psychosocial processes that yield results on how to identify the words most commonly used by people with self-esteem issues or how possessive words may vary from men and women to refer to their sentimental companions.

Rangel and colleagues [11] point out that due to the huge amount of information available on social networking platforms, it is possible to obtain information about different attributes such as gender, age, personality, native language, or political orientation from the analysis of an author's profile.

Considering that celebrities use DSN frequently to communicate and connect with their followers [12]; and understanding that user's behavioral profile is reflected in the message according to their writing patterns [13], it is essential to detect whether a user is a celebrity is essential in order to determine the influence they may have on other users of social networks [14] and to know what would be the impact of a comment made by this user. This provides information to measure the influence of celebrities on their followers by means of the corpus of their texts.

Our motivation to write this paper is to explore the predictive and explanatory capacities of linguistic features on demographics and influence variables of celebrities using DSN. In fact, the research's main objective is understand how these linguistic features, which are found in the texts that celebrities publish on DSN, generate new information that allows to classify celebrities according to their demographics and influence variables. Moreover, these new variables derived from the texts, can indicate the use of language which shows specific sociolects and idiolects useful to analyze the celebrity's profile, and increase the accuracy level in the classification models.

Recognizing this opportunity, this article formally addresses the study of linguistic analysis observed from celebrities using DSN and proposes a model with 18 features that can quantify the outcome of five types of analysis: lexical, syntactic, symbolic, participation, and complementary information. From the lexical analysis, the average use of words and lexical diversity are analyzed. The syntactic analysis studies the personal pronouns most commonly used by celebrities. The symbolic analysis studies how symbolic contents such as emojis and hashtags are used; the participation analysis quantifies the features of participation in the network (mentions and retweets). Finally, the

complementary information analyzes the reference that the celebrity makes to other media (URLS).

The difference between this paper and the one presented [12] at the Conference and Labs of the Evaluation Forum (CLEF) is that this paper proposes a new model of characteristic selection and explains how this model helps to increase the accuracy value. At the Plagiarism analysis, Authorship identification, and Near-duplicate detection (PAN) at CLEF they only presented several classification models and showed the accuracy obtained with different principles but there is nothing associated with characteristic selection.

This study presents eight sections. The second section presents a summary of the background, showing the authors who worked in areas related to the analysis of DSN, identification of profiles in texts, detection of demographic and social variables in texts, and influence of celebrities. The third section presents the methodology with the necessary steps to determine the features of the digital identity describing celebrities' characteristics. The fourth section illustrates the data preparation, which is the corpus description, exclusion of redundant measures, and the methodology application. The fifth section shows the results of the constructed explanatory models with the significance from each one of the features found in the digital identity. The sixth section shows the results of making the celebrity classification model validation and prediction ability with the features selected to quantify the improvement of the accuracy. Finally, in the seventh and eighth section, the conclusions and future works are presented.

Background

Relevant background on celebrity detection has three elements: first, a basic background in social networks; second, a review of the works related to author profiling, including Machine Learning classification models tested and demographics and social variables that have been found as valuable in the task of Author profiling; finally, a review of works that address particularly the study and prediction of celebrities influence.

Social networks

According to Aggarwal [15] a social network is defined as:

> *"a network of interactions or relationships, where the nodes consist of participants and the edges consist of relationships or interactions between these participants."*

Social network analysis (SNA), therefore, seeks to discover different types of patterns in the relationship of the different nodes found inside the network [16], allowing them to describe these communities. Thanks to the Internet, there is an interactive dialogue platform of digital relationships, emulating physical interactions [17, 18], which makes possible to keep the different participants of the network in contact [17, 18] creating not only new forms of sharing information, but also new forms of communication, which, a possible effect would be a transformation of personal opinion or decision due the influence from the new contacts [19]. Therefore, nowadays SNA are of great interest to determine how languages can be used to describe communities and their collective subjectivity from sociolects.

With the vast exchange of information over the Internet, users in social networks are leaving a digital trail; for example, every day, Facebook members post 3.2 billion likes and comments, and 340 million tweets are sent out on Twitter [19]. This trail contains associated information given in texts, images, URLs, or audios, thus, generates a social structure programmed by each user in their own network based on the connections with other users [20]. Therefore, the availability of large amounts of data on the web has given a new motivation to use of statistical and computational tools in the area of Social Network Analysis (SNA) because of their growing popularity [15], combined with Natural Language Processing (NLP). Fan et al. [21] apply that combination for reducing the harmful effects caused by the spread of rumor in a social network through independent cascade (IC) model and the linear threshold (LT) model.

Consequently, the work oriented to computational linguistics has focused on the analysis of the corpus found in conversations shared in social networks to analyze opinions, feelings, emotions and in general, the expression of private status on certain individuals [2, 22, 23].

Author profiling

Author Profiling has been approached from different aspects that converge searching how to describe or profile an author. One of these aspects has studied the problem from a computational point of view, giving all the relevance to classification. Other aspects are from the sociolinguistic point of view, where language is understood as a process of social construction that develops along the time and describes dialects, sociolects, or chronolects associated to the authors.

Therefore, some examples of the aspects mentioned are the methodologies of the first, third, and fifth place of celebrity profiling in the PAN at CLEF event. First, Radivchev and colleagues [24] vectorized with a Term Frequency-Inverse Document Frequency (TF-IDF) the users' tweets taking into account the top 10,000 features from word bigrams to use a combination of logistic regression and Support Vector Machines (SVM). In contrast, Martinc and colleagues [25] selected a Logistic regression classifier with word unigram and character tetragram features where the Logistic regression classifier and its hyper-parameters were chosen with a grid search. Finally, Petrik and Chuda [26] extracted the text features with TF-IDF using bigrams and trigrams to capture word relationships, then, they combined it with Random forest with 200 decision trees as a classification model.

Profile classification

Theoretical and empirical studies have demonstrated a strong relationship between social factors and linguistic attitudes, since language is perceived as a social activity that reflects and influences social reality [11, 27].

In fact, for Rangel and colleagues [11], the analysis of shared contents aims to:

> *" predict different attributes of the authors, such as gender, age, personality, native language, or political orientation. Therefore, social networks are playing a vital role in identifying what people think because they can reinforce political ideas or even influence the way of thinking."*

The relationship between personality traits and the usage of language has been widely studied by psycholinguistics, analyzing the use of language and how it varies depending on personal characteristics. Initial researches on author profiles focused mainly on formal texts and blogs. However, at present time, researchers mainly focus on digital social networks, where language is more spontaneous and less formal [11].

Then, there are connections which are not captured with traditional analysis because a common feature of social media communication is that this is delivered through short messages. These messages do not often use standard language variations [28], and the data itself drives an integral exploration of the language that differentiates people, finding connections that cannot be captured with traditional analysis such as word categorization of vocabulary [8].

Consequently, social activities represent a great challenge for the selection and identification of the user profile, which is caused mainly by the diversity of texts and complex social structures [11, 29, 30].

Demographic and social variables

Jadhav and Mhetre [20] and Simaki and colleagues [27] indicate a connection between social networks and personal behavior on the web, identifying the relationship and influence between social factors and a person's language. In fact, Milroy and Milroy [31] point out that one of the most important contributions of Labov's (1972) "quantitative paradigm" on the study of language has been the systematical examination of the relationship between language variation and the variables of "speaker" such as age, ethnicity, gender, social network, and social class.

Due to this growing interest, the extraction of demographic information from the text has been studied, and important approximations have been made by authors like Przybyla and Teisseyre [32], who identified demographic characteristics such as education, party association, and year of birth. In contrast, Simaki and colleagues [27] used texts to determine an author's gender, from a qualitative to a quantitative analysis, or [33] exploring the differences between male and female writing in a large subset of the British National Corpus.

The authors Nguyen and colleagues [22] and Romaine [34], state that linguistic variations occur over long and non-immediate periods of time in a sociolect. This means that the corpus of each generation has its own linguistic characteristics in which people of different gender and age tend to have different linguistic features. This is strongly related to the social influence and identity they have in the usage of language [27].

As for the characterization of "occupation", authors such as Sloan and colleagues [35] used a search engine designed to identify the socioeconomic group of a tweet. The 2010 Standard Occupational Classification (SOC) system is used by U.S. federal statistical agencies to classify workers and jobs into occupational categories.

Celebrities' influence

Celebrities are some of the most common users of DSN, by promoting their careers, and obtaining followers [36]. Therefore, social networks have been a revolutionary scenario for these individuals because these platforms allow them to share any information with

their fans [12]. This demonstrates that a minority influences an exceptional number of people, becoming an important factor in the creation of public opinion [37].

In order to know the celebrity's influence on the network, it is necessary to specify who influences who. However, this evidence of influence on real-world networks is limited, and it is something that only a few studies have attempted empirically [38].

To determine this influence it is necessary to know that there are celebrities who use only one social network. For example, words like "YouTuber" referring to a person whose primary social network is YouTube, or a person who only uses this social network in search of having a high reputation.

The development of micro-celebrities is more evident on Instagram, Facebook, Twitter, and other social platforms [39], leading to find different categories of celebrities on different social networks, therefore the data base of this study shows the celebrity profiling by hierarchical levels.

It is well known identifying profiles is not easy, and although there are exciting approximations, computational linguistics requires an integrated approach providing elements to understand patterns of linguistic variation [31] related to ethnographic and social factors, presenting a model and its validation to detect celebrities from variables identified and explained in the development of this study.

When trying to identify a user as a celebrity on Twitter, authors such as Wang and Kraut [40] argued that the specific topic and its continued usage in the user's tweets affect the number of followers in two modalities: hemophilia and network externalities. However, Hutto and colleagues [41] created a theory based on forecasting models that although it included the topic of tweets, unlike Wang and Kraut, they did not find a prediction with more followers based on continues usage of a topic. Therefore, it is important to raise new proposals for the prediction of celebrities, not only for the number of followers, but also because more work is required to understand the importance of the contents published to engage an audience [42].

Meanwhile, Li and colleagues [18] indicated that to detect opinion leaders in social networks, academic studies generally consider the semantic analysis of user's comments or the emotional analysis of contents published by users based on positive or negative comments; also, by analyzing feelings to define the relevance of the connection between users and followers. However, the detection of opinion leaders with semantic analysis or analysis of emotions is not always suitable for complex social networks, so Wang [43] proposed a method of extracting community opinion leaders based on a hierarchical structure.

Deep learning applied on feature selection in social network

Neural networks have the basic idea of representing the process of pattern recognition and classification that the human brain performs [44]. Therefore, research fields have applied this basic idea to evolve the models to increase their performance in classification models. Casas [45] mentions the phenomenon of replacement in statistical and optimization models to understand Geography's travel behaviors and traffic management.

Now, popularity is a critical issue in celebrities' behavior since an increase in the degree of their fame is often the result of the implementation of marketing strategies in the networks. Thus, research with neural networks becomes more relevant when

concluding the critical factors for popularity or the key active times of popularity for making posts on social networks. Hsu et al. [46] developed research to improve the performance of classifiers in social network popularity prediction tasks; they implemented a multimodal approach by integrating the images included in the post and their social information into a Convolutional Neural Network (CNN). Huang et al. [47] performed a deep neural network model (Long short-term memory (LSTM) and CNN) with embedding in the responsible factors to improve the predictions of long-time popularity in social networks.

In turn, the social network's characteristics involve vital indicators for the promotion of popularity. Retweets or hashtags contain relevant information about the interests of the communities participating in a separate communication thread generating topics of interest for celebrities, which can influence and achieve higher popularity in the network. Zhang et al. [48] proposed a neural network to predict retweeting behavior by weighting a layer of different interests from a clustering process to identify the core tweets of the cluster. Li et al. [49] modeled a CNN and LSTM-RNNs by improving existing classifiers to make hashtag recommendations by tweet representations that included word embedding generation, sentence composition, tweet composition, and hashtag classification.

PAN at CLEF is an international initiative that has been promoting the research of its excellence network on the fields of Digital Text Forensics and Stylometry for ten years. As a result, the best research groups around the world in the fields of Natural Language Processing (NLP) and Information Retrieval (IR) meet annually to participate in the Author Profiling, Author Verification, Authorship Attribution, and Style Change Detection tasks. In the last version at 2019 with the tasks of Bots and Gender Profiling, Celebrity Profiling, and Style Change Detection, we participated in the task for Celebrity Profiling obtaining the second place.[1]

Table 1 presents proposals for profiling celebrities, including the characteristics used by authors who have worked on identifying profiles through DSN.

Celebrity feature selection methodology

To achieve the objective of the study and to be able to determine the attributes that describe the characteristics of celebrities analyzing their texts in social networks, this methodology includes 4 phases (see Fig. 1):

1. Modeling digital identity using lexical, syntactic, symbolic, participation, and complementary information features extracted from a user's publications.
2. Calculating the central tendency and dispersion measurements for each feature.
3. Reducing the dimensionality considering the calculated measures.
4. Constructing a model of significance analysis of each attribute of the digital identity over the person's characteristics in the real world.

[1] See more information in https://pan.webis.de/clef19/pan19-web/celebrity-profiling.html in the Fame category.

Table 1 Demographic and social variables for profile detection

Characteristic	Authors	Tittle	Database
Gender	Simaki et al. [50]	Evaluation and Sociolinguistic Analysis of Text Features for Gender and Age Identification.	The collection of blog posts on the website of 19,320 bloggers. These publications were extracted from blogger.com in August 2004. The corpus size is 681,288 publications containing more than 140 million words.
Gender	Argamon et al. [33]	Gender, Genre, and Writing Style in Formal Written Texts	British National Corpus
Gender	Simaki et al. [27]	Sociolinguistic Features for Author Gender Identification: from qualitative evidence to quantitative analysis.	2936 posts from blog hosting sites and blog search engines.
Age	Moreno-Sandoval et al. [77]	Age Classification from Spanish Tweets: The Variable Age Analyzed by using Linear Classifiers	Spanish Colombian Tweets pre-processing 50,819 accounts of people linked to universities and 734,037 accounts of people linked to celebrities.
Age and gender	Johannsen et al. [51]	Cross-lingual syntactic variation over age and gender	International user review websites
Age and gender	Peersman et al. [28]	Predicting Age and Gender in Online Social Networks	Chat texts from the Belgian social networking site, Netlog.
Age and gender	Rangel et al. [11]	Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter	Twitter texts and images from a corpus covering Arabic, English and Spanish were analyzed.
Age, gender, socioeconomic level	Moreno-Sandoval et al. [78]	Spanish Twitter Data Used as a Source of Information About Consumer Food Choice	1.3 million Spanish Twitter texts where 11,691 tweets mentioned food with an initial food knowledge base of 1128 words and an own generation of Knowledge Base.
Personality, gender and age	Schwartz et al. [8]	Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach	Facebook messages from 75,000 volunteers
Age, gender, occupation and fame	Radivchev et al. [24]	Celebrity Profiling using TF-IDF, Logistic Regression, and SVM–Notebook for PAN at CLEF 2019	Approximately 53 million PAN tweets at CLEF 2019
Age, gender, occupation and fame	Petrik and Chuda [26]	Twitter feeds profiling with TF-IDF–Notebook for PAN at CLEF 2019	Approximately 53 million PAN tweets at CLEF 2019
Age, Gender, Occupation and Fame	Matinc et al. [25]	Who is hot and who is not? Profiling celebs on Twitter–Notebook for PAN at CLEF 2019	Approximately 53 million PAN tweets at CLEF 2019
Gender, education, party affiliation and yearbirth	Przyby and Teisseyre [32]	Analyzing Utterances in Polish Parliament to Predict Speaker's Background	100 statements by the same author and multilevel annotations from the corpus source.
Gender and class Role	Milroy [31]	Mechanisms of change in urban dialects: the role of class, social network and gender	Previous sociolinguistic studies
Age, occupation and social class	Sloan et al. [35]	Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-data	Profile of Twitter Users in the United Kingdom (UK). Metadata data collected through the Collaborative Social Networking Observatory (COSMOS)
Occupation	Huang et al. [29]	Multi-source integration framework for user occupation inference in social media systems	Micro blog platforms: Sina Weibo.

Modeling digital identity

Among the different contents created by people in DSN using texts called "post" that form a corpus from which it is possible to extract information that may help to determine some people's characteristics such as occupation, age, gender, and degree of fame. Through text mining methods, new knowledge emerges to extract relevant information analyzing and identifying vast amounts of unstructured data through text mining methods [52]. This phase proposes a model that relates these previous characteristics to groups of features that can be found in the posts available on digital social networks.

The proposed groups of linguistic features are classified as lexical, syntactic, symbolic, participation, and complementary information type (see Fig. 2). These characteristics represent the contents of the "posts" of the digital user identity. Alternatively, these features might be associated with the digital user identity, particularly with their demography and influence features, focusing on Gender, Birth year, Occupation and Fame. The study assumes that by analyzing these attributes, characteristics of the real user identity can be obtained.

Although some features are standard in different digital social networks within the "post", there may be others that can be specific to a particular social network.

For example, there are some features shared by social networks such as Facebook, Twitter, and Instagram (see Fig. 3) which a Least Cost Influence (LCI) problem study could find a set of users with minimum cardinality to influence a certain fraction of users in multiple social networks [19]; however, these may vary in their use and application. For example, for Facebook, a "like" can be determined by different emojis that allow having a higher degree of granularity when identifying the emotion that generates to share that "post".

Therefore for this phase, common features in digital social networks are analyzed (see Fig. 4).

Lexical features seek to estimate the size of words, style, and diversity of the text. Syntactic features correspond to expressions in the use of personal pronouns; Symbolic features refers to the inclusion of semiotics from the use of symbols such as emojis or hashtags which represent an implicit content. Participatory features allow to link different participants or social dynamics that may represent a confirmation or question message or a reinforcement of a common idea. Complementary information features allow to extend or argue comments.

Calculating central tendency and dispersion measurements

To quantify each qualitative feature described in the previous phase, this study must calculate measurements allowing a statistical analysis of the usage and distribution of the variables, such as central tendency and the level of dispersion measurements.

The central tendency measurements, include the mean, the mode, and the median that present in a single value a value set represented by the center where a data set is located. Besides, it is necessary to determine the dispersion by employing statistical parameters such as the standard deviation (indicating how far data are according to the central measurements), the skewness and kurtosis identify the bias and sharpness of data distribution, respectively.

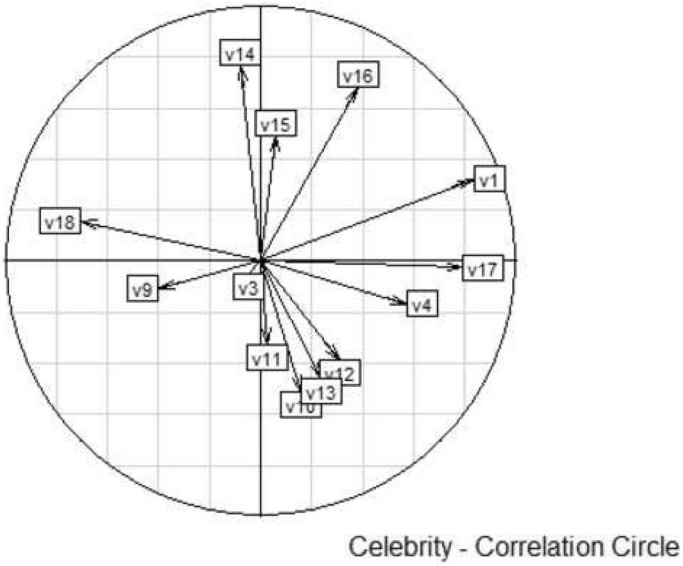

Fig. 1 Methodology

Reducing the dimensionality

Since all measurements should be presented to build a significance analysis model, it is necessary to debug those that may be redundant. To do this, it is proposed to make a correlation analysis. The purpose is to indicate if there is a relationship between two variables and what is the strength degree of such relationship using "corrplot" package.

Then, the explanatory variables are normalized using these variables, a principal component analysis (PCA) is performed (using "ade4" package).

PCA and LDA are both the earliest data representation learning algorithms. PCA is an unsupervised method [53] that converts existing high-dimensional data into a low-dimensional space. PCA retains the data's variance to the maximum extent to get the data's low dimensional representation from a global perspective [54] and preserve the global information of data in the learned feature space [53]. Besides, PCA has been widely used for dimensionality reduction [55], and authors keep holding that "although it is one of the earliest multivariate techniques, it continues to be the subject of much research, ranging from new model-based approaches to algorithmic ideas from neural networks. It is incredibly versatile, with applications in many disciplines." [56].

Nowadays, representation learning has evolved the dimensionality reduction task. Its techniques include neural network models and non-negative constraint matrix factorization models [57] that, with the use of incremental learning, automatically adjust feature selection according to what is learned whenever new examples or data sets emerge.

Those recent outputs of those unsupervised approaches through clustering techniques produce a selected subset of the features can reduce the computation cost and improve the clustering performance [58], which outstanding the performance of classification problems. On the other hand, these models can consider the information discrepancy between the original feature space and the lower-dimensional subspace, which efficiently reduces the loss of information, and the structure-preserving term is based on the low

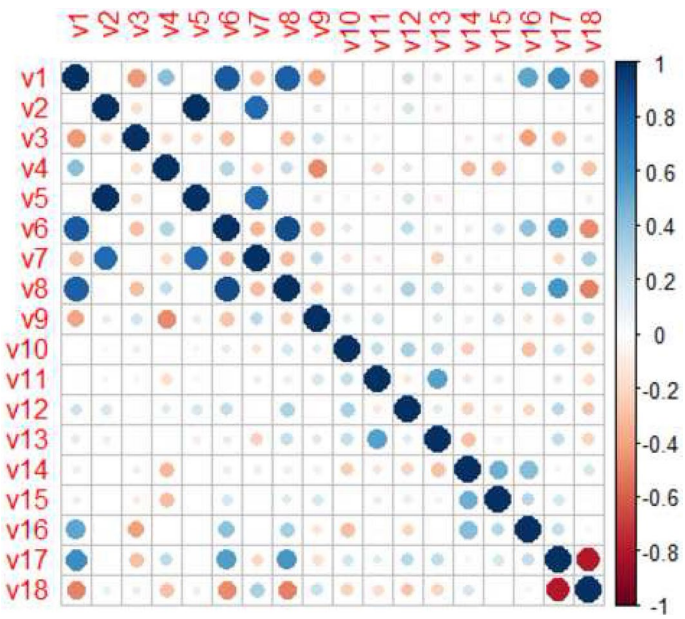

Fig. 2 Real identity model using features extracted from a user's publications

Fig. 3 Features of influence in the different social networks

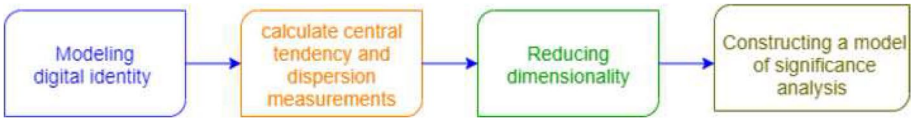

Fig. 4 Feature characterization

rank sparse graph, which acquires adequate discriminative information and avoids problems of parameters selection [54]. Hence, further analysis can then be facilitated efficiently, and achieve better performance alleviating the issue of scalability in the term of computational complexity to some extent. However, those unsupervised approaches are not standard models because how to further reduce the computational complexity while

keeping models powerful is still an issue worth studying [59]. Another example of the disadvantage of those models is probably because Learning Sparse computes feature's score independently, but it neglects the possible correlation between different features, thus failing to produce an optimal feature subset [54].

Finally, the multiple correspondence analysis (MCA) is used to reveal the relationship between the different physical characteristics of the analyzed profiles (using "FactoMineR" package).

Constructing a model of significance analysis

To evaluate whether each variable contributes significantly to the user's characteristics, it was decided to use a multinomial logit model as described by [60]. For the purpose of our analysis, Y is going to be all the characteristics of the real user and X corresponds to be the measurements obtained from the dimensionality reduction phase.

According to Peña [60], to evaluate whether each variable contributes significantly to the model, the p-values established by the Wald statistic 1 are used:

$$W_{\beta_i} = \frac{\hat{\beta}_i}{Var(\hat{\beta}_i)}, \tag{1}$$

where the test hypotheses are:

$$H_0 : \beta_i = 0,$$

$$H_a : \beta_i \neq 0.$$

The Wald test rejects the null hypothesis if the p-value is less than 5%; as a result, the coefficients are considered to be different from 0, inferring, this variable is statistically significant.

For example, Sluban et al. [61] distinguish some studies that search the outstanding features on Twitter to measure influence, which is the principal measure for celebrities. Avnit (2009) [62] shows the million follower fallacy, where an account with more retweets gets a higher level of influence than one pursue a large number of followers. Therefore, Suh et al. [63] states that URLs, hashtags, the number of followers and followees, the age of the account involve the number of retweets.

Data preparation

Description of the corpus

The study used a corpus from the PAN@CLEF2019 data sets corresponding for celebrities posts in social network Twitter based on the English language. The first data set[2] was the input for this paper, providing 68,583,577 Tweets and 31,203 profiles divided in 60/40 proportion for training and test data as a mechanism to avoid overfitting. Later, TIRA implemented a blind evaluation; it refers to "an evaluation process where the authors of a to be-evaluated piece of software cannot access the test data and hence cannot (unwittingly) optimize their algorithm against it" [64] with a second and third

[2] See Availability of data and materials section to download this data set

data sets[3] which contain approx 6,000 and 60,000 profiles, correspondingly. The software itself is packaged within a virtual machine, and new performance results[4] were achieved. Table 2 presents the performance results of the data sets mentioned above.

Four types of analysis were performed, which consisted on identifying occupation, gender, fame, and birth year as shown in Tables 3 and 4. However, each of these categorical variables were extracted from the information provided by the different profiles in the social network database. Specifically for the variable "Fame", a celebrity means a person who has verified his Twitter account and is notable according to Wikipedia's notoriety criteria, definition granted by the PAN @ CLEF 2019.

Application of central tendency and dispersion measurements to selected features

Central tendency and dispersion measurements were applied, however, for this study, mean, skewness, and kurtosis were analyzed as the measures that contribute the most to the analysis (see Fig. 5). The standard deviation was not selected because it was too high compared to the average due to the large amount of atypical data. Similarly, mode and median were not considered as they did not provide relevant information.

Dimensionality reduction

After selecting the measures to be used for the analysis, it is proposed to build a model with 18 variables corresponding to the analysis of lexical features (V1 to V8); syntactic features (V14 to V18); symbolic features (V9 and V10), participation features (V11 and V13) and complementary information features (V12). For the variables associated with the lexical features shown in Table 5, it is calculated, for example, for each t_i (being t_i each of the user's tweets) the average number of characters per word in the profile.

For the variable V14 associated with the corpus syntactic analysis, the following calculation is made: for each t_i the number of times the post is written in the first person singular is calculated. The average according to the total tweets of the profile is also calculated. Similarly to the calculation of variable V14, all the other variables described in Table 6 are calculated.

To analyze the variables related to the symbolic, participation, and complementary information features, in each t_i it is calculated the average number of times that each emoji, hashtag, URL, mention or retweet is used. Each of these language elements in the social network was taken as a variable, as shown in Tables 7, 8, and 9. In particular, separating the URL features into an individual group of feature means recognizing that a tweet is more informative when accompanied by URLs [65] considering the limit of the tweet. Therefore, the scope of the information acquired with a URL is beyond knowing about a clickable link (hashtag) that facilitates an easy search of tweets that with same hashtag [63].

There are highly correlated variables (see Fig. 6), such as variables v2 "Kurtosis character" with V5 "Kurtosis avg character". Equivalently, variable V2 correlates positively with variable V7 "Skew avg character". Similarly, variable V1 "Avg character" is

[3] You can require access to download these data sets in TIRA (Evaluation as a service who their main task is improving the replicability of shared tasks in computer science)https://www.tira.io/task/celebrity-profiling/dataset/pan19-celebrity-profiling-test-dataset2-2019-05-02/.

[4] The evaluation blind results made are expressed in accuracy measure and you can search them inhttps://pan.webis.de/clef19/pan19-web/celebrity-profiling.html.

Table 2 Performance results of data sets

	Dataset 1 **Training** **F1-score**	**Dataset 2** **Test** **F1-score**	**Dataset 3** **Test** **F1-score**
Fame	0.8258	0.5628	0.5176
Gender	0.6475	0.6442	0.5606
Birth year	0.5689	0.5176	0.5156
Occupation	0.5456	0.469	0.4183

Table 3 Characteristics database

Characteristic	**Profiles**	**Value**
Fame	6160	Superstar
	23658	Star
	1384	Rising
Gender	22283	Male
	8887	Female
	32	Undefined

positively related to variable v6 "Kurtosis label word". In contrast, there is a negative relationship between variable V7 and V3: "Lexical diversity".

For the construction of the model, only variables V1, V3, and V4 will be taken, because they reflect the same information as variables V2, V5, V6, V7, and V8. Therefore, they will not be included in the model since doing so generates collinearity problems and are not explanatory from a sociolinguistic perspective.

There is also a strong correlation between the variables V16 "person 3 singular", V17 "person 1 plural" and V18 "person 3 plural" (see Fig. 6) , which corresponds to syntactic analysis of the corpus and measures the use of different pronouns. Although these variables are highly related, they are going to use because they represent sociolinguistic and idiolect variables which become in explanatory variables of interest in the study as they help in the task of characterizing the celebrities.

After this, the predictor variables (those corresponding to Tables 5, 6, 7, 8, and 9) were normalized. With these variables, a principal component analysis (PCA) is performed (see Fig. 7) .

There is no relevant relationship between the variables (see Fig. 7). The only thing shown is that the variable V15 "person 2 singular" is in opposite relationship to the variable V11 "label mention" and V3 "diversity lexical", which means that the use of second singular person relates negatively with the mentions used in the tweets and the lexical diversity.

Table 4 Continued from previous page

Characteristic	Profiles	Value
Occupation	12586	Sport
	9195	Performer
	5110	Creator
	2353	Politics
	731	Science
	497	Professional
	698	Manager
	32	Religious
Birth year	1230	1940
	2763	1950
	4487	1960
	6575	1970
	9660	1980
	6097	1990
	389	2000
	1	2010

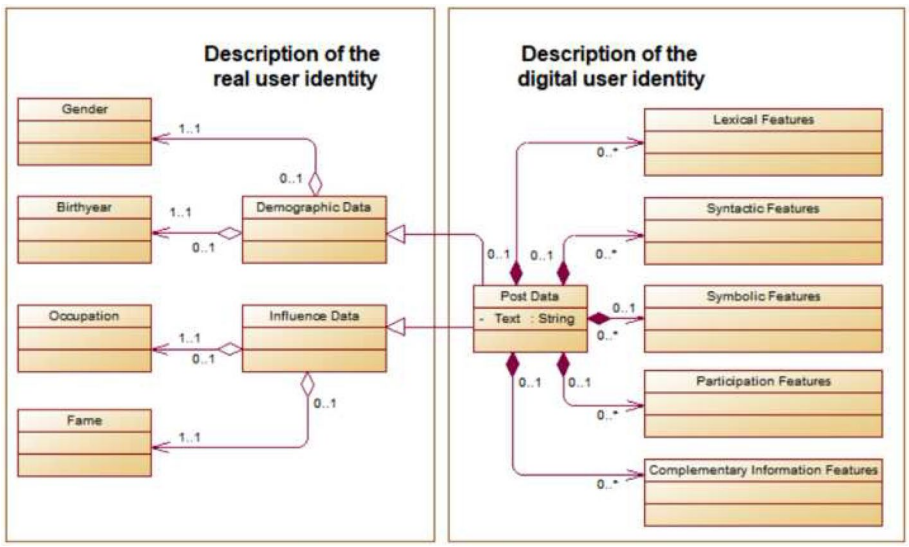

Fig. 5 Analytical measures

Although the birth year variable is discrete, it was grouped by decades. This grouping method changed the Birth year variable to be categorically treated as the Fame, Occupation, and Gender variables.

For the characteristics of the people, a multiple correspondence analysis (MCA) was used (see Fig. 8), which made it possible to reveal the relationships between these celebrities' profiles.

Table 5 Feature group of lexical analysis

Label	Name	Description
V1	Avg_character	Average number of characters per word in profile
V2	Kurtosis_character	Avg_character variable kurtosis
V3	Lexical_diversity	Lexical diversity of all profile tweets
V4	Label_word	Average number of words in the profile divided by the number of tweets
V5	Kurtosis_avg_character	Kurtosis_character variable kurtosis
V6	Kurtosis_label_word	Label_word variable kurtosis
V7	Skew_avg_character	Statistical skew of the Avg_character variable
V8	Skew_label_word	Statistical skew of the label_word variable

Table 6 Feature group related to syntactic analysis

Label	Name	Description
V14	Person_1_singular	Average number of tweets using singular first-person pronoun
V15	Person_2_singular	Average number of tweets using singular second-person pronoun
V16	Person_3_singular	Average number of tweets using singular third-person pronoun
V17	Person_1_plural	Average number of tweets with plural first and second-person pronouns
V18	Person_3_plural	Average number of tweets with plural third-person pronoun

Table 7 Group feature related to symbolic analysis

Label	Name	Description
V9	Label_emoji	Average of emojis used in a tweet for the profile
V10	Label_hashtag	Average of hashtags used in a tweet for the profile

Table 8 Group feature related to participation analysis

Label	Name	Description
V11	Label_mention	Average of mentions used in a tweet for the profile
V13	Label_retweets	Average of retweets used in a tweet for the profile

Results of celebrity feature selection

The significance models along with the p-values, described for each Wald test, are shown in Tables 10, 14, 17, 18, 22, and 23 for each of the variables previously selected as a result of multivariate analysis.

Table 9 Group feature related to the analysis of complementary information

Label	Name	Description
V12	Label_url	Average of URLs used in a tweet for the profile

Fig. 6 Corplot explanatory variables represent the correlation degree on the entire matrix of correlations between the analyzed features shown by the intensity of the colors. Each cell contains the result of measuring the correlations between a pair of features within a color scale with extreme values of 1 or − 1. For this purpose, it was used the R version 3.6.1. Thus, a high correlation between a pair of features is when it has a solid color tone close to the values of 1 or − 1 on the color scale shown on the right side. You can find more information in Availability of data and materials section

Fame

Table 10 shows the degree of "Fame" and its relationship with the different groups of lexical, syntactic, participation, and complementary information features.

Table 10 presents the coefficients of the model posed in equation 2. Therefore, the expression[5] of the model for the first category (star) of fame is:

$$\begin{aligned} logit[p(Star = 1)] = {} & 0.99 + 0.01V1 - 2.49V3 + 0.4V4 \\ & + 0.48V9 + 0.53V10 \\ & + 0.82V11 + 0.56V12 - 1.41V13 \\ & + 1.07V14 + 1.1V15 + 3.75V16 \\ & + 3.23V17 + 0.22V18. \end{aligned} \tag{2}$$

Tables 11 and 12 summarize the results using as referent group the category "Rising"[6] that can be seen in Table 10. Thus in Table 10, since the estimators are relative to the

[5] When the remaining category estimators are used, it produces a similar equation for the analyzed characteristic. In this case, the resulting equation would model the Superstar category.

[6] It should be noted that the variables of the characteristics analyzed are categorical, i.e., the objective is to quantify a qualitative variable by expressing it in a binary form 1 or 0. In other words, the multinomial logit model predicts n-1 categories of the variables when it is processed by any statistical program to avoid problems of collinearity and its interpretation is made by comparing the levels of the variable that were not omitted, without falling into the error of explaining this omitted variable.

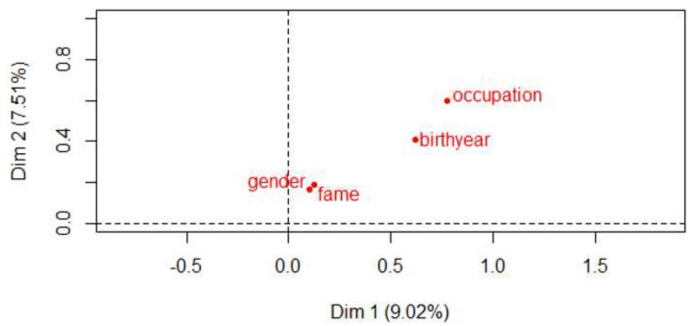

Fig. 7 PCA correlation circle also known as variable correlation plots. This figure was made in R version 3.6.1. The arrows grouped together indicate positively correlated variables; whereas, arrows in opposed quadrants are negatively correlated variables. You can find more information in Availability of data and materials section

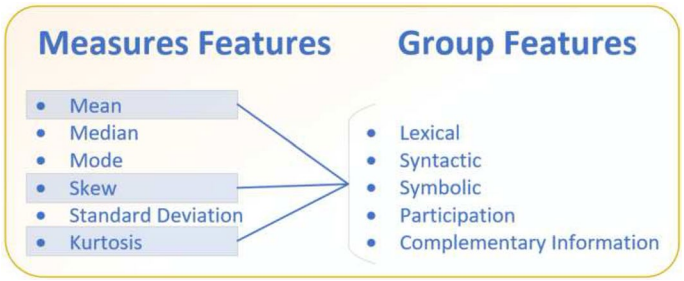

Fig. 8 Multiple correspondence analysis MCA made in R version 3.6.1. The closeness between the characteristics indicates the relationship degree. As a result, a relationship between the characteristics of "gender" and "fame" is evident, while "occupation" and "birth year" do not have a marked relationship. You can find more information in Availability of data and materials section

referent group, for a unit change in the feature, the logit of outcome relative to the Rising group celebrities is expected to change by its respective estimator given the other features in the model are held constant. For example, if a celebrity were to increase his use of lexical diversity by one point, the multinomial log-odds for Star celebrity relative to a Rising celebrity would be expected to decrease by 2.49 units while holding all other features in the model constant.

Celebrities use all the syntactic features, complementary information, participation, and symbols along with their different categories. We can suppose that common helpful in both categories indicates a feature group is more helpful like Table 13 shows.

Gender

Table 14 presents the coefficients of the model posed in Eq. 3. Therefore, the expression[7]

[7] When the remaining category estimators are used, it produces a similar equation for the analyzed characteristic. In this case, the resulting equation would model the Nonbinary category. of the model for the first category (Male) of gender is:

Table 10 Model of the person's characteristic Fame

Feature groups	Variable	Star		Superstar	
	Predictor	Estimator	P-value	Estimator	P-value
Lexical	(Intercept)	0.99	0.028	1.33	0.008
	V1	0.01	0.22	0.03	0.055
	V3	− 2.49	0.003	− 12.91	<0.001
	V4	0.4	<0.001	− 1.15	<0.001
Syntactic	V14	1.07	<0.001	2.03	<0.001
	V15	1.1	<0.001	2.14	<0.001
	V16	3.75	<0.001	5.8	<0.001
	V17	3.23	<0.001	4.45	<0.001
	V18	0.22	<0.001	0.21	<0.001
Symbolic	V9	0.48	<0.001	0.91	<0.001
	V10	0.53	<0.001	0.92	<0.001
Participation	V11	0.82	<0.001	0.67	<0.001
	V13	− 1.41	<0.001	− 1.94	<0.001
Complementary information	V12	0.56	<0.001	2.51	<0.001
Residual deviance: 41544.39					
Akaike information criterion (AIC): 41604.39					

$$\begin{aligned} logit[p(Male = 1)] = {} & 3.41 + 0.02V1 + 26.42V3 - 0.29V4 \\ & - 0.77V9 - 0.24V10 \\ & - 0.09V11 - 1.02V12 - 0.46V13 \\ & - 3.03V14 - 2.45V15 + 2.25V16 \\ & - 1.32V17 + 0.13V18. \end{aligned} \tag{3}$$

Table 15 shows the inference of the results using as reference the category "Female"[8], which can be seen in Table 14. Thus, if a celebrity were to increase his use of lexical diversity by one point, the multinomial log-odds for Male celebrity relative to Female celebrity would be expected to increase by 26.42 units while holding all other features in the model constant.[9]

In summary, the celebrity gender in all its categories highlights the use of lexical diversity; they post in singular first person and employ social network features such as hashtag and retweet. However, there are not a common helpful, which indicates the absence of a more useful feature group like Table 16 shows.

[8] It should be noted that the variables of the characteristics analyzed are categorical, i.e., the objective is to quantify a qualitative variable by expressing it in a binary form 1 or 0. In other words, the multinomial logit model predicts n-1 categories of the variables when it is processed by any statistical program to avoid problems of collinearity and its interpretation is made by comparing the levels of the variable that were not omitted, without falling into the error of explaining this omitted variable.

[9] This same analysis is replicable for the remaining features of male celebrities as long as their P-value is below 0.05.

Table 11 Results for the person's characteristic *Fame*

Category	Description
Star	Coefficients of the model are not very high in the majority except for the use of the pronoun of third-person singular and first-person plural. These variables help the explanation in a significant way, the use of the pronouns of the first and third plural persons are typical. There is a negative relationship between lexical diversity and average use of retweets. The average use of characters does not contribute significantly to the explanation of the category.

Table 12 Results for the person's characteristic *Fame*

Category	Description
Superstar	Lexical diversity, retweets, and average word usage have a negative relationship. In contrast, the highest positive coefficients are found in the first and third plural personal pronoun. However, if Wald's test for the coefficients is reviewed, it shows that all these variables confirm that there is a significant contribution of these variables, i.e., the probability that the coefficients of the model are zero is very low.

Table 13 Helpful features group for Fame

Feature group	Star Helpful	Superstar Helpful
Lexical	x	x
Syntactic	o	o
Symbolical	o	o
Participation	o	o
Complementary information	o	o

Occupation

Tables 17 and 18 describe the characteristic occupation. It is evident that the highest level of significance of the variables is in the political category, while it is not as strong in the religious and managerial categories.

Table 17 presents the coefficients of the model posed in Eq. 4. Therefore, the expression[10] of the model for the first category (Manager) of occupation is:

$$\begin{aligned} logit[p(Manager = 1)] = & -4.7 + 0.02V1 + 1.85V3 + 0.88V4 \\ & + 0.22V9 + 0.63V10 \\ & + 0.12V11 - 0.91V12 - 0.16V13 \\ & - 0.9V14 + 0.88V15 - 3.89V16 \\ & + 4.43V17 - 0.12V18 \end{aligned} \tag{4}$$

The description in the behavior of each one of the categories of the characteristic Occupation using as reference the category "Creator"[11] category are shown in Tables 19 and

[10] When the remaining category estimators are used, it produces a similar equation for the analyzed characteristic. In this case, the resulting equations would model the n categories - manager category.

[11] It should be noted that the variables of the characteristics analyzed are categorical, i.e., the objective is to quantify a qualitative variable by expressing it in a binary form 1 or 0. In other words, the multinomial logit model predicts n-1 categories of the variables when it is processed by any statistical program to avoid problems of collinearity and its interpretation is made by comparing the levels of the variable that were not omitted, without falling into the error of explaining this omitted variable.

Table 14 Model of the characteristic of the person Gender

Feature groups	Variable	Male		Nonbinary	
	Predictor	Estimator	P-value	Estimator	P-value
Lexical	(Intercept)	3.41	<0.001	− 1.97	0.562
	V1	0.02	<0.001	− 0.02	0.83
	V3	26.42	<0.001	− 483.78	<0.001
	V4	− 0.29	<0.001	− 0.75	0.402
Syntactic	V14	− 3.03	<0.001	3.27	0.002
	V15	− 2.45	<0.001	1.77	0.187
	V16	2.25	<0.001	− 5.03	0.112
	V17	− 1.32	<0.001	0.21	0.935
	V18	0.13	<0.001	− 0.13	0.415
Symbolic	V9	− 0.77	<0.001	− 0.48	0.363
	V10	− 0.24	<0.001	− 1.6	0.055
Participation	V11	− 0.09	0.016	− 0.72	0.209
	V13	− 0.46	<0.001	2.87	0.017
Complementary information	V12	− 1.02	<0.001	0.41	0.639
Residual deviance: 37703.85					
AIC: 37763.85					

Table 15 Results for the characteristic of *Gender*

Category	Description
Male	Most of the variables have a negative relationship in their coefficients. However, the average use of characters, lexical diversity and the average use of the third person both in the singular and in the plural show a positive effect indicating the use of these characters.
Nonbinary	The use of retweets and first-person singular are significant variables. On the contrary, hashtags and lexical diversity are significant but not used.

20. Hence, if a celebrity were to increase his words using singular first-person pronoun by one point, the multinomial log-odds for Manager celebrity relative to Creator celebrity would be expected to decrease by 0.9 units while holding all other features in the model constant[12].

A common feature across the various categories of celebrity occupations is the use of the third person singular. Within the group represented by Table 17, the relevance on

[12] This same analysis is replicable for the remaining features of male celebrities as long as their P-value is below 0.05.

Table 16 Helpful features group for Gender

Feature group	Male Helpful	Nonbinary Helpful
Lexical	o	x
Syntactic	o	x
Symbolical	x	x
Participation	o	x
Complementary information	o	x

Table 17 Model of the person's characteristic Occupation

Group Features	Variable	Manager		Performer		Politics		Professional	
	Predictor	Estimator	P-value	Estimator	P-value	Estimator	P-value	Estimator	P-value
Lexical	(Intercept)	− 4.7	<0.001	6.68	<0.001	− 12.13	<0.001	− 7.87	<0.001
	V1	0.02	0.298	− 0.11	<0.001	0.17	<0.001	0.06	<0.001
	V3	1.85	0.369	− 7.95	<0.001	− 9.86	<0.001	− 1.49	0.662
	V4	0.88	<0.001	− 1.25	<0.001	2.93	<0.001	1.21	<0.001
Syntactic	V14	− 0.9	0.005	1.3	<0.001	− 0.5	0.038	− 0.58	0.113
	V15	0.88	0.013	2.22	<0.001	− 1.92	<0.001	0.69	0.093
	V16	− 3.89	<0.001	− 2.58	<0.001	− 7.42	<0.001	− 3.89	<0.001
	V17	4.43	<0.001	− 0.64	0.111	5.07	<0.001	4.8	<0.001
	V18	− 0.12	0.004	− 0.1	<0.001	− 0.23	<0.001	− 0.05	0.243
Symbolic	V9	0.22	0.193	0.66	<0.001	− 0.92	<0.001	0.29	0.157
	V10	0.63	<0.001	0.55	<0.001	1.25	<0.001	0.75	<0.001
Participation	V11	0.12	0.265	0.03	0.544	− 0.24	0.006	0.22	0.07
	V13	− 0.16	0.546	− 0.45	<0.001	1.5	<0.001	− 0.37	0.226
Complementary information	V12	− 0.91	<0.001	− 0.15	0.091	− 2.69	<0.001	− 0.36	0.117
Residual deviance: 100703.9									
AIC: 100913.9									

the number of words in the post and the use of hashtag are standard features; on the other hand, the use of mentions, the plural first person, and the care on the number of characters in the tweet are common features in Table 18. However, there is no common helpful indicating the absence of a more useful feature group like Table 21 shows.

Birth year

Tables 22 and 23 show the behavior of the model according to the variables' significance level of the decade to which the celebrity belongs.

Table 22 presents the coefficients of the model posed in equation 5. Therefore, the expression[13] of the model for the first category (1950) of Birth year is:

[13] When the remaining category estimators are used, it produces a similar equation for the analyzed characteristic. In this case, the resulting equations would model the n decades - 1940 decade.

Table 18 Model of the person's characteristic Occupation

Group features	Variable	Religious		Science		Sports	
	Predictor	Estimator	P-value	Estimator	P-value	Estimator	P-value
Lexical	(Intercept)	− 3.7	0.254	− 10.76	<0.001	5.8	<0.001
	V1	0.1	0.026	0.12	<0.001	− 0.11	<0.001
	V3	− 11.86	<0.001	− 7.56	0.156	− 1.15	0.211
	V4	− 0.03	0.965	2.02	<0.001	0.08	0.352
Syntactic	V14	− 1.82	0.157	− 0.37	0.253	− 1	<0.001
	V15	− 0.32	0.772	− 1.16	0.01	− 1.43	<0.001
	V16	− 4.74	0.007	− 4.79	<0.001	− 3.25	<0.001
	V17	6.74	<0.001	4.57	<0.001	− 1.27	0.002
	V18	− 0.01	0.947	− 0.01	0.76	-0.14	<0.001
Symbolic	V9	0.51	0.51	− 0.73	0.007	1.2	<0.001
	V10	0.46	0.424	0.38	0.006	1.41	<0.001
Participation	V11	− 1.87	0.005	0.34	0.001	− 0.46	<0.001
	V13	0.11	0.941	− 0.28	0.262	0.85	<0.001
Complementary information	V12	− 1.35	0.1	− 0.98	<0.001	− 4.47	<0.001
Residual deviance: 100703.9							
AIC: 100913.9							

Table 19 Results for the person's characteristic *Occupation*

Category	Description
Manager	The average number of words used varies. The same as the lexical diversity, the use of emojis is little, as the use of mentions and the singular first-person pronoun.
Performer	No recurrent use of mentions and URL within their corpus, and no use of singular first-person pronoun.

$$
\begin{aligned}
logit[p(1950 = 1)] =& 1.05 - 0.01V1 - 2.85V3 - 0.09V4 \\
&+ -0.32V9 - 0.02V10 \\
&+ 0.27V11 + 0.2V12 + 0.01V13 \\
&- 0.06V14 - 0.07V15 + 0.02V16 \\
&+ 0.59V17 - 0.04V18
\end{aligned} \tag{5}
$$

The description in the behavior of the characteristic Birth year using as reference the category "1940"[14] Decades are shown in Tables 24 and 25. Consequently, if a celebrity were to increase mentions in his posts by one point, the multinomial log-odds for born celebrities in 1950 relative to born celebrities in 1940 would be expected to increase by 0.27 units while holding all other features in the model constant.[15]

[14] It should be noted that the variables of the characteristics analyzed are categorical; i.e., the aim is to quantify a qualitative variable by expressing it in a binary form 1 or 0. In other words, the multinomial logit model predicts n-1 categories of the variables when it is processed by any statistical program is to avoid problems of collinearity and its interpretation is made by comparing the levels of the variable that were not omitted, without falling into the error of explaining this omitted variable.

[15] This same analysis is replicable for the remaining features of born celebrities in 1950 as long as their P-value is below 0.05.

Table 20 Results for the person's characteristic *Occupation*

Category	Description
Politician	All the measured variables enrich their corpus, i.e., they have a vast lexicon, and a high average of words. They use elements of the social network and all the pronouns in their corpus.
Professional	They have a very varied lexical diversity. They do not use many elements of the social network, only hashtag, and the use of pronouns is limited to the singular third-person and plural first-person pronouns.
Religious	They have a high average word usage (long tweets) with great lexical diversity. They only use the mentions of the elements from the social network, and for syntactic analysis, it is evident the use of plural first person and singular third-person pronouns.
Science	They have high lexical diversity. They do not use retweets, nor singular first-person pronoun.
Sport	They do not use any syntactic attribute, URL, or mention. They use symbolic features and retweets instead.

Table 21 Helpful features group for Occupation

Feature group	Manager Helpful	Performer Helpful	Politics Helpful	Professional Helpful	Religious Helpful	Science Helpful	Sport Helpful
Lexical	x	o	o	x	x	x	x
Syntactic	x	x	o	x	x	x	o
Symbolical	x	o	o	x	x	o	o
Participation	x	x	o	x	x	x	o
Complementary Information	o	x	o	x	x	o	o

Celebrities of different ages do not present a common feature of the groups analyzed for this paper, i.e., there is no common helpful, which indicates the absence of a more useful feature group like Table 26 shows.

However, in Table 22, a common feature of the use of mentions can be seen for the decades analyzed; on the other hand, the common feature in Table 23 is the use of the first person plural. Besides, complementary information since 1980 decade at 2000 decade seems to be a feature group more helpful.

Validation of the celebrity classification model with selected features

Classifiers models with the selected features were created using the PAN CLEF 2019 celebrity analysis data set. These models were divided into a training subset with 60% of the samples, and a test subset with 40% of the samples, with these subsets we developed a performance training and testing for each one of the models.

Different classification models were programmed for texts with a scikit-learn library [66] such as multinomial Naive Bayes (NB), Gaussian Naive Bayes (GNB), Naive Bayes Complement (NBC), Logistic Regression (LR), and Random Forest (RF) called from now on classical classifiers, and Deep Neural Networks (DNN). The model with the best performance on each variable: gender, birth year, occupation, or fame, was selected and replicated for the famous actor data set. Table 27 describes the configured parameters of the best-performing classifiers. Each classifier model was trained with a group of terms associated with each of the celebrity profiles.

Table 22 Model of the characteristic of the person Birth year

Feature groups	Variable	1950		1960		1970		1980	
	Predictor	Estimator	P-value	Estimator	P-value	Estimator	P-value	Estimator	P-value
Lexical	(Intercept)	1.05	0.099	3.22	<0.001	5.61	<0.001	6.94	<0.001
	V1	− 0.01	0.561	− 0.02	0.043	− 0.04	<0.001	− 0.13	<0.001
	V3	− 2.85	0.166	− 1.34	0.411	− 2.69	0.094	− 2.26	0.153
	V4	− 0.09	0.517	− 0.64	<0.001	− 1.1	<0.001	− 0.74	<0.001
Syntactic	V14	− 0.06	0.812	0.2	0.423	1.53	<0.001	2.53	<0.001
	V15	− 0.07	0.822	0.06	0.839	− 0.15	0.59	− 0.25	0.385
	V16	0.02	0.966	− 0.02	0.96	− 0.35	0.363	− 2.56	<0.001
	V17	0.59	0.123	0.21	0.569	− 0.86	0.031	− 2.31	<0.001
	V18	− 0.04	0.125	0	0.924	− 0.02	0.384	− 0.01	0.781
Symbolic	V9	− 0.32	0.156	0.48	0.014	1.26	<0.001	1.95	<0.001
	V10	− 0.02	0.818	0.05	0.615	0	0.993	0.41	<0.001
Participation	V11	0.27	0.01	0.54	<0.001	0.54	<0.001	0.22	0.019
	V13	0.01	0.954	− 0.02	0.929	0.07	0.725	0.35	0.102
Complementary information	V12	0.2	0.2	0.21	0.158	− 0.11	0.427	− 1.91	<0.001
Residual deviance: 92807.86									
AIC: 92989.86									

Table 23 Model of the characteristic of the person Birth year

Feature groups	Variable	1990		2000		2010	
	Predictor	Estimator	P-value	Estimator	P-value	Estimator	P-value
Lexical	(Intercept)	7.99	<0.001	0.33	0.755	− 4.91	<0.001
	V1	− 0.34	<0.001	− 0.09	0.001	− 0.1	0.704
	V3	− 4.56	0.008	− 9.64	<0.001	0.31	0.256
	V4	− 0.03	0.838	− 0.31	0.213	− 0.96	0.46
Syntactic	V14	3	<0.001	1.6	<0.001	3.06	0.258
	V15	0.57	0.068	1.33	0.003	0.86	0.736
	V16	− 3.72	<0.001	− 0.74	0.389	− 0.58	0.398
	V17	− 5.31	<0.001	− 6.99	<0.001	0.92	0.024
	V18	− 0.08	0.005	− 0.03	0.567	0.02	0.951
Symbolic	V9	2.61	<0.001	1.96	<0.001	1	0.724
	V10	0.68	<0.001	− 0.37	0.075	0.87	0.04
Participation	V11	− 0.96	<0.001	0.42	0.015	0.31	0.861
	V13	4.13	<0.001	1.12	0.005	− 0.41	0.392
Complementary information	V12	− 4.04	<0.001	0.57	0.037	1.82	0.43
Residual deviance: 92807.86							
AIC: 92989.86							

Table 24 Results for the person's characteristic *Birth year*

Category	Description
The 1950s	They repeatedly use mentions. They do not use the rest of the features of the social network. They do not have a marked lexical diversity. The average corpus size is variable.
The 1960s	Their corpus has great use of words and characters. The features of the social network are emojis and mentions. Their lexical diversity is very variable, and there is no predominance of any personal pronoun.
The 1970s	They have similar behavior to the corpus of the previous decade. However, in this decade, there is evidence of the use of first-personal pronouns in both singular and plural.

Table 25 Results for the person's characteristic *Birth year*

Category	Description
The 1980s	A total generational change is evident in the 5 groups of features analyzed in the model, which means they repeatedly use almost all the features, except for the retweets. The lexical diversity is not as marked as in previous generations and they do not use personal pronouns in the singular second-person nor plural third person.
The 1990s	This generation uses texts with all the elements of the network, the non-use of singular second-person pronoun distinguishes syntactic analysis.
The 2000s	The corpus of this decade use hashtags to a lesser extent but it is more frequent the use of other features on the social network. Lexical diversity is more extensive than others, and the corpus are of a short length. They do not use the singular and plural third-person pronouns.
The 2010s	These profiles have very unequal corpus between the different profiles. Hashtags are the most common and highly used features and it is evident the use of the plural first-person pronoun.

Table 26 Helpful features group Birth year

Feature group	1950 Helpful	1960 Helpful	1970 Helpful	1980 Helpful	1990 Helpful	2000 Helpful	2010 Helpful
Lexical	x	x	x	x	x	x	x
Syntactic	x	x	x	x	x	x	x
Symbolical	x	x	x	o	o	x	x
Participation	x	x	x	x	o	o	x
Complementary information	x	x	x	o	o	o	x

A cleaning processing and homogenizing of the text in UTF8 encoding was performed. Additional characteristics were included and selected to finally obtain a group of features that comprised the new attributes in the classifier. In addition, an over-sampling technique [67] was applied with the idea to balance the data set between classes with small samples and the other classes.

This allowed the models to improve the average precision up to 0.14 higher than a classifier model that only uses the group of words[16] from each celebrity. In addition, a data set of famous actors was analyzed to see if the results of the models are similar. Consequently, created a data set using the A (The elite of acting circles), and B (quite famous, but not super famous as an A-list) lists published in IMDb website[17]. The celebrity

[16] Improvement performance can deduce from the public result from blind evaluation with the PAN@CLEF2019 dataset located in Table 2 in Fame F1-score for Dataset 3 Test.

[17] *You can find the A list in*https://www.imdb.com/list/ls008173417/, *B list in*https://www.imdb.com/list/ls024783564/ *and C list in*https://www.imdb.com/list/ls064876597/

Table 27 Description of the parameters used in the best classifier

Logistic regression		Multinomial naive Bayes	
Parameter	**Value**	**Parameter**	**Value**
Penalty to minimize cost function (penalty)	l2	Additive smoothing (alpha)	Laplace
Dual formulation (dual)	Primal formulation	Learn class prior probabilities (fit_prior)	True
Tolerance level for stopping criteria (tol)	1.00E-04		
Relative strength of regularization (c)	1		
Calculate the intercept (fit_intercept)	True		
Intercept scaling (intercept_scaling)	1		
Pseudo-random number generator (random_state)	0		
Solver algorithm (solver)	L-BFGS		
Number of maximum iteration(max_iter)	100		
Approach for handling multiple classes (multi_class)	multinomial		

* Note the parameter label in parentheses is the code needed to define these parameters in Python [66]

listings were manually extracted from IMDb, but IMDb experts had already defined the ratings creating three different fame level lists. The famous actor dataset involves a label annotation of the data where the manual label quality is controlled by brings the same number of actors (100 actors) for each fame category. Therefore, the A-list published on IMDb becomes the simile of the celebrities' dataset's superstar category. Similarly, the B list is the counterpart of the category of stars in the celebrities' dataset. IMDb is the world's most popular and authoritative source for movie, TV, and celebrity content. Tables 28 at 31 show the model results for each one of the variables with the test data sets.

The fame variable with a multinomial logistic regression classifier obtained a final average F1-score of 0.65 for PAN at CLEF dataset and the 0.44 F1-score for famous actors considering the list come from different IMDb reviewers , as shown in Table 28.

The gender variable with a multinomial logistic regression classifier obtained a final average F1-score of 0.88 for PAN at CLEF data set and 0.89 F1-score for famous actors, which outstanding in the previous result, as shown in Table 29.

The birth year variable with a multinomial logistic regression classifier obtained a final average F1-score of 0.37 for PAN at CLEF and 0.25 F1-score for famous actors, which is slightly lower, as shown in Table 30.

The occupation variable obtained a final average F1-score of 0.57 with a multinomial naive Bayes classifier, as shown in Table 31.

As shown in Table 32, the classifier maintains similar results in all classes except occupation; in terms of fame, the paper data set has a better result than the famous actors' data set; in the gender and birth year variables, the results are similar in both data sets.

Deep learning

A deep learning model was trained to compare the performance of this model with the model proposed in the paper and the baseline ones. Specifically, two models were

Table 28 Fame classification using multinomial logistic regression classifier

Data set	Class	Precision	Recall	F1-score	Support
PAN at CLEF2019	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
Famous actors	0 - rising	0.00	0.00	0.00	0
	1 - star	0.50	0.25	0.34	51
	2 - superstar	0.46	0.68	0.55	47
	Micro avg	0.46	0.46	0.46	98
	Macro avg	0.32	0.31	0.30	98
	Weighted avg	0.48	0.46	0.44	98

Table 29 Gender classification using multinomial logistic regression classifier

Dataset	Class	Precision	Recall	F1-score	Support
PAN at CLEF2019	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
Famous actors	0 - female	0.92	0.95	0.93	92
	1 - male	0.90	0.83	0.86	46
	2 - nonbinary	0.00	0.00	0.00	0
	Micro avg	0.89	0.89	0.89	98
	Macro avg	0.60	0.59	0.59	98
	Weighted avg	0.90	0.89	0.89	98

generated under two approaches: the first one adds and analyzes the lexical features; the second one adds the new features proposed in this paper.

Table 33 describes the parameters of two Deep Learning models with neural networks using the Keras library with the Tensorflow framework in Python to compare the performance of these models with the one proposed by the paper and baseline models. The model uses a densely connected regular layer Dense as a sequential model API, an activation function for the hidden layers Rectified Linear Unit (ReLU), a Softmax function on the output layers, a Sparse categorical cross-entropy loss function, and an Adam optimizer to do the weighting calculations performed by this optimization method in order to reduce the error on the target output.

Baselines

The baseline models were generated for the PAN at CLEF2019 contest. The details on how they obtained the lexical features are not explicitly published. Instead, they describe the following:

Table 30 Birth year classification using multinomial logistic regression classifier

Dataset	Class	Precision	Recall	F1-score	Support
PAN at CLEF2019	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
Famous actors	0 - [1930–1940]	0.00	0.00	0.00	4
	1 - [1940–1950]	0.04	0.33	0.07	3
	2 - [1950–1960]	0.35	0.43	0.39	14
	3 - [1960–1970]	0.17	0.05	0.07	21
	4 - [1970–1980]	0.29	0.27	0.28	26
	5 - [1980–1990]	0.38	0.31	0.34	16
	6 - [1990–2000]	0.44	0.29	0.35	14
	7 - [2000–2012]	0.00	0.00	0.00	0
	Micro avg	0.24	0.24	0.24	98
	Macro avg	0.21	0.21	0.19	98
	Weighted avg	0.29	0.24	0.25	98

Table 31 Occupation classification using multinomial naive Bayes classifier

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 - politician	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

> *"baseline-uniform randomly draws from a uniform distribution of all classes and reflects the data-agnostic lower bound, baseline-rand randomly selects a class according to the prior likelihood of appearance in the test dataset, and baseline-mv always predicts the majority class of the test dataset." [68]*

The models implemented in this research differ on the treatment over lexical features. Paper data set and Famous Actors' model use n-gram to configure the vectorial representation of words with a minimum frequency of 9 for gender, 6 for the birth year of birth, 3 for occupation, and none for fame. A differential treatment

Table 32 Performance F1-score results of data sets

	Classical classifiers	
	Paper data set	**Famous Actors**
Fame	0.650	0.440
Gender	0.880	0.890
Birth year	0.370	0.250
Occupation	0.570	0.900

Table 33 Deep Neural Network configuration

Parameter	**Value**
Model	Keras sequential - Tensorflow
Loss	sparse_categorical_crossentropy
Optimizer	Adam
Epochs	5
Batch size	10
Activation input layer	ReLu
Activation output layer	Softmax
Dense input layer	10
Dense output layer	3, 8, 70
Features (proposed features)	1365
Features (lexical features)	1000

performed a pre-processing of texts to replace hashtags, mentions, URLs, and emojis with special tokens and apply a lemmatization method on the 1.000 most frequent words penalized with the TF-IDF process.

Table 34 reports the results obtained using different classification techniques taking as input lexical features and the proposed features. Results presented in columns 1, 2, and 3 are baseline models using lexical features; the fourth column contains the best performance on each class using the paper dataset, classical classifiers and the proposed features. Finally, the fifth and sixth columns report the results using a deep neural network with lexical features only and the features proposed in this paper (i.e., syntactic, symbolic, participation and complementary information features), respectively.

Neural network with lexical features has three out of four best performances; the gender and fame are the classes with the best F1-score, 0.88 and 0.75, respectively. In contrast, year of birth is the class with the lowest F1-score overall models with 0.04. The best score for this class was obtained using the classical machine learning model (0.37).

Neural network with proposed features shows the gender as the the best-ranked class with an F1-score of 0.80, followed by the fame with 0.74. On the other hand, the occupation has an F1-score of 0.39, reflecting the large gap with fame, i.e., its F1-score is much lower as the number of values increases. Gender and fame are the best ones with the best F1-score.

Table 34 F1-score obtained with the models using the paper data set

	Baseline rand	Baseline uniform	Baseline mv	Classical classifiers	Neural network with lexical features	Neural network with proposed features
Fame	0.341	0.099	0.285	0.650	0.750	0.744
Gender	0.344	0.266	0.278	0.880	0.887	0.804
Birth year	0.123	0.117	0.071	0.370	0.047	N/A
Occupation	0.125	0.152	0.121	0.570	0.684	0.398

In general, the performance of the classical classifiers is considerably lower than the obtained with the neural network model, except on the gender class, where is very similar. The neural networks models that include the proposed features have a decrease on the F1-score, especially on the Occupation class. Finally, the classical classifiers that used the proposed features have one out of four of the best performances in the analyzed classes, making the classical models the best option to classify a class with a large number of possible values. It is very clear, that deep learning models offer new options to explore the problem of celebrity classification.

Conclusions

After analyzing a rigorous selection of features, a measurement group applied to the feature group was achieved, determining the significance of each of them in order to identify a person's characteristics such as fame, gender, occupation, and birth year. This paper presents a new approach that addresses the characterization of profiles using relations of lexical, syntactic, symbolic, complementary, and information variables achieving a better approach in order to identify significant features that help in the identification of celebrity profiles.

The use of these new features improves the initial classifications made with only the words for the characteristics of Gender, Birth year, Occupation, and Fame. These new features, derived from texts in DSN achieving an increase on the average F1-score up to 0.14, this occurs by including the set of features in the classification task of the characteristic. We used the proposed features in several classifiers. We found that they represent a greater contribution when using classical classifiers (e.g., logistic regression, multinomial naive Bayes). On the contrary, they decreased the performance on deep learning models

As a result, the best-performing models are "Gender" and "Fame" with a residual deviation: 37703.85 and 41544.39 with an AIC: 37763.8 and 41604.39. In contrast, the worst-performing models are "Age" and "Occupation" with a residual deviation of 92807.86 and 100703.9 both with an AIC: 92989.86 and 100913.9.

It is evident from the model that regardless of fame, occupation, and gender, celebrities write recurrently in the third person singular. However, new generations (those born in the last five decades) use more the first person plural.

The occupation that uses all groups of lexical and syntactic features is politician. This is prone to happen due to the occupation's nature in which being at the forefront of language and trends is vital to hold good.

The most recurrent gender characteristics are the use of the first-person singular in both men and women, but this is not evident in the nonbinary users.

It is also important to note that the most commonly used feature from social network along the decades are Mentions followed by Emojis.

Discussion and future works

In the analysis of user profiles, and even more deeply in the analysis of celebrity profiles, multiple documents analyze a user's comments to determine the attributes of demographic, sociological or psychographic variables [33] [11] [69].

However, some models only use lexical characteristics that concentrate efforts on preprocessing, and other studies look for other types of obtained variables from the text, as is the case with sociolinguistic studies. In this type of study, we can observe the analysis of variables in the use of some words that denote the social use of "sociolect" or "idiolect" languages [70].

Therefore, it is essential to have a group of documents that can describe a user's style and not to have only one document per profile. These group of documents can be seen in studies such as Copland and colleagues [71] that analyzed the use of the first person singular or the first person plural in a group of students at a school, and according to the results.

It was inferred that this use denotes social status according to the use of possessives pronouns ("my school", "our school"), identifying qualities of the property to identify a higher socioeconomic level. So, the challenge of DSN is broad, as they are designed to have personal interactions, where people can share different types of information with different features, for example, text messages.

On Twitter, text messages are associated with the length of characters, also with the use of symbols, emojis, and expressions such as hashtags that can indicate semiotics. Texts are also used to make comments to other users to themselves by creating mentions within the network and finally referring to external sources of information contained in the URLs that can guide or give context to the messages. These messages imply a different measurements than the use of lexical or syntactic characteristics.

By studying these and other additional characteristics, it is possible to improve the precision of the classification processes on demographic, sociological, psychographic, and behavioral variables of users in a social network. However, based on the specific analysis of celebrities, interesting information can be obtained due to the large number of followers they have, and this type of analysis is essential for celebrities due to the influential power they can have on their followers [72].

As future work, the collection and analysis of other language-related elements, such as sociolects and idiolects is recommended. This collection will enable a higher and more accurate profile of social network users, as it will be possible to analyze the digital user's text, in particular celebrity texts, in a more granular way. Also, the use of synonyms and antonyms, or more than one language and other typical elements of DSN may indicate higher measurements of ranking performance.

On the other hand, it is also possible to use non-linguistic elements that social networks have as data that are not necessarily linguistic, such as the number of followers

or the reciprocity of links in publications. The type of data from social network analysis, with the social graph of these characteristics, which comes from extracts in networks, can could help to predict demographic or influencing variables of digital users. It is also possible to explore new strategies in classification models and deep learning techniques that explore other types of architectures such as LMTS or CNN with different configurations to enhance the ranking of celebrities on social networks.

Finally, identifying profiles is not yet an easy task. The proposal to build new models of celebrity profiles from their texts is an interesting approach and, specially, to have new types of features that allow to increasing the accuracy in this type of natural language processing task. The social phenomenon in which users use a language to express their private states with other digital users has meeting points in language and social fields, and it is possible to generate other phenomena such as homophilia or find new patterns of relationship.

Abbreviations

MCA: Multiple correspondence analysis; PCA: Principal component analysis; SOC: Standard occupational classification; LCI: Least cost influence; DSN: Digital social networks; NLP: Natural Language Processing; SVM: Support vector machines; TF-IDF: Term Frequency-Inverse Document Frequency; CLEF: Conference and labs of the evaluation forum; PAN: Plagiarism analysis, authorship identification, and near-duplicate detection; SNA: Social network analysis; AIC: Akaike information criterion; NB: Naive Bayes; GNB: Gaussian Naive Bayes; NBC: Naive Bayes Complement; LR: Logistic regression; RF: Random forest; IC: Independent cascade; LT: Linear threshold; CNN: Convolutional neural networks; LSTM: Long short-term memory.

Acknowledgements

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). Models and results presented in this challenge contribute to the construction of the research capabilities of CAOBA.

Author Contributions

These authors contributed equally to this work.

Funding

This research received no external funding.

Availability of data and materials

Database can be found in https://zenodo.org/record/3530253#.Xm7LSKgzblV where an access request must be done to download the data and calculate the measures described in Tables 3, 4, 5, 6, 7, 8, 9. The software used for the statistical study was RStudio to plot Figs. 6 at 8 and export the results shown in Tables 10 at 23. Figure 6 used "corrplot" package [73] to plot the values resulting from the correlation measurement with a Spearman method. Then, Fig. 7 used the "ade4" package [74] and "s.corcircle" command plotting the "dudi.pca" command to perform a principal component analysis. Finally, Fig. 8 used a "FactoMineR" package [75] through the MCA command for categorical variables In contrast, the model for classification shown in Tables 28, 29, 30, 31 was made with Python 3.7 with the scikit-learn library. [66]. The parameters θ_y for Multinomial Naive Bayes classifier is estimated by a smoothed version of maximum likelihood [66]. $\hat{\theta_{yi}} = \frac{N_{yi}+\alpha}{N_y+\alpha n}$ The smoothing priors $\alpha \geq 0$ accounts for features not present in the learning samples and prevents zero probabilities in further computations. Setting $\alpha = 1$ is called Laplace smoothing, while $\alpha < 1$ is called Lidstone smoothing. Finally, the working notes of Clef 2019 [76] contain more related papers about the celebrity profiling task celebrated over PAN Lab on Digital Text Forensics and Stylometry challenge.

Declarations

Competing interests

The authors declare no conflict of interest. The author declares no competing interests.

Author details

[1]Center of Excellence and Appropriation in Big Data Analytics (CAOBA), Bogota, Colombia. [2]Department of System Engineering, Pontificia Universidad Javeriana, Bogota, Colombia. [3]Department of Industrial Engineering, Pontificia Universidad Javeriana, Bogota, Colombia.

Received: 25 April 2020 Accepted: 26 July 2021
Published online: 26 August 2021

References

1. Sherchan, W., Nepal, S., Paris, C.: A survey of trust in social networks. ACM Comput. Surv. **45**(4), 47–14733 (2013). https://doi.org/10.1145/2501654.2501661
2. Cercel, D.-C., Trausan-Matu, S.: Opinion propagation in online social networks: a survey. ACM International Conference Proceeding Series (2014). https://doi.org/10.1145/2611040.2611088
3. Allor, M.: Relocating the site of the audience. Crit. Stud. Mass Commun. **5**(3), 217–233 (1988). https://doi.org/10.1080/15295038809366704
4. Reynolds, W.N., Salter, W.J., Farber, R.M., Corley, C., Dowling, C.P., Beeman, W.O., Smith-Lovin, L., Choi, J.N.: Sociolect-based community detection. In: 2013 IEEE International Conference on Intelligence and Security Informatics, pp. 221-226 (2013). https://doi.org/10.1109/ISI.2013.6578823
5. Golbeck, J.: Trust and nuanced profile similarity in online social networks. ACM Trans. Web **3**(4), 12–11233 (2009). https://doi.org/10.1145/1594173.1594174
6. Mansouri, F., Abdelalim, S., Ikram, E.A.: A modeling framework for the moroccan sociolect recognition used on the social media. In: Proceedings of the 2Nd International Conference on Big Data, Cloud and Applications. BDCA'17, pp. 34–1345. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3090354.3090389
7. Zanzotto, F.M., Pennacchiotti, M., Tsioutsiouliklis, K.: Linguistic redundancy in twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '11, pp. 659–669. Association for Computational Linguistics, Stroudsburg, PA, USA (2011). http://dl.acm.org/citation.cfm?id=2145432.2145509
8. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P., Ungar, L.H.: Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS ONE **8**(9), 73791 (2013). https://doi.org/10.1371/journal.pone.0073791
9. Yang, Y., Eisenstein, J.: Putting things in context: community-specific embedding projections for sentiment analysis (2015)
10. Rampton, B., Tusting, K., Maybin, J., Barwell, R.D.: UK linguistic ethnography: a discussion paper coordinating committee UK linguistic ethnography forum **1**, (2004)
11. Rangel, F.M., Rosso, P., Montes-yGómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. In: Notes Papers of the CLEF (2018)
12. Moreno-Sandoval, L.G., Puertas, E.A., Plaza-del-Arco, F.M., Pomares-Quimbaya, A., Alvarado-Valencia, J.A., Alfonso, L., Ureña-López: Celebrity profiling on twitter using sociolinguistic features notebook for pan at clef 2019. (2019)
13. Phad, P.V., Chavan, M.K.: Detecting compromised high-profile accounts on social networks. In: 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–4 (2018). https://doi.org/10.1109/ICCCNT.2018.8493851
14. Singh, M., Bansal, D., Sofat, S.: Who is who on twitter–spammer, fake or compromised account? A tool to reveal true identity in real-time. Cybern. Syst. **49**(1), 1–25 (2018). https://doi.org/10.1080/01969722.2017.1412866
15. Aggarwal, C.C.. In: Aggarwal, C.C. (ed.): An Introduction to Social Network Data Analytics, pp. 1–15. Springer, Boston, MA (2011). https://doi.org/10.1007/978-1-4419-8462-3_1
16. Scott, J.: Social network analysis: developments, advances, and prospects. Soc. Netw. Anal. Min. **1**(1), 21–26 (2011). https://doi.org/10.1007/s13278-010-0012-6
17. Vatrapu, R., Mukkamala, R.R., Hussain, A., Flesch, B.: Social set analysis: a set theoretical approach to big data analytics. IEEE Access **4**, 1–1 (2016). https://doi.org/10.1109/ACCESS.2016.2559584
18. Li, C., Bai, J., Zhang, L., Tang, H., Luo, Y.: Opinion community detection and opinion leader detection based on text information and network topology in cloud environment. Inf. Sci. **504**, 61–83 (2019). https://doi.org/10.1016/j.ins.2019.06.060
19. Zhang, H., Nguyen, D., Zhang, H., Thai, M.: Least cost influence maximization across multiple social networks. IEEE/ACM Trans. Netw. **24**, 1–11 (2015). https://doi.org/10.1109/TNET.2015.2394793
20. Jadhav, K.U., Mhetre, N.A.: Mass users behaviour prediction in social media: a survey. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **5**, 3286–3288 (2014)
21. Fan, L., Wu, W., Zhai, X., Xing, K., Lee, W., Du, D.-Z.: Maximizing rumor containment in social networks with constrained time. Soc. Netw. Anal. Min. (2014). https://doi.org/10.1007/s13278-014-0214-4
22. Nguyen, D., Doğruöz, A.S., Rosé, C.P., de Jong, F.: Computational sociolinguistics: a survey. Comput. Linguist. **42**(3), 537–593 (2016). https://doi.org/10.1162/COLI_a_00258
23. Tsytsarau, M., Palpanas, T.: Survey on mining subjective data on the web. Data Min. Knowl. Discov. **24**(3), 478–514 (2012). https://doi.org/10.1007/s10618-011-0238-6
24. Radivchev, V., Nikolov, A., Lambova, A.: Celebrity profiling using tf-idf, logistic regression, and svm—notebook for pan at clef 2019. In: Cappellato, L., Ferro, N., Losada, D.E., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers, vol. 2380. CEUR-WS.org, Switzerland (2019). http://ceur-ws.org/Vol-2380/
25. Martinc, M., Škrlj, B., Pollak, S.: Who is hot and who is not? Profiling celebs on Twitter—notebook for PAN at CLEF 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers, vol. 2380. CEUR-WS.org, Switzerland (2019). http://ceur-ws.org/Vol-2380/
26. Petrik, J., Chuda, D.: Twitter feeds profiling with TF-IDF—notebook for PAN at CLEF 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers, vol. 2380. CEUR-WS.org, Switzerland (2019). http://ceur-ws.org/Vol-2380/
27. Simaki, V., Aravantinou, C., Mporas, I., Kondyli, M., Megalooikonomou, V.: Sociolinguistic features for author gender identification: from qualitative evidence to quantitative analysis. J. Quant. Linguist. **24**(1), 65–84 (2017). https://doi.org/10.1080/09296174.2016.1226430
28. Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting age and gender in online social networks. In: Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents. SMUC '11, pp. 37–44. , ACM, New York, NY, USA (2011). https://doi.org/10.1145/2065023.2065035
29. Huang, Y., Yu, L., Wang, X., Cui, B.: A multi-source integration framework for user occupation inference in social media systems. World Wide Web **18**(5), 1247–1267 (2015). https://doi.org/10.1007/s11280-014-0300-6

30. Sánchez-Rebollo, C., Puente, C., Palacios, R., Piriz, C., Fuentes, J.P., Jarauta, J.: Detection of jihadism in social networks using big data techniques supported by graphs and fuzzy clustering. Complexity **2019**, 1–13 (2019). https://doi.org/10.1155/2019/1238780
31. Milroy, J., Milroy, L.: Mechanisms of change in urban dialects: the role of class, social network and gender. Int. J. Appl. Linguist. **3**(1), 57–77 (1993). https://doi.org/10.1111/j.1473-4192.1993.tb00043.x
32. Przybyła, P., Teisseyre, P.: Analysing utterances in polish parliament to predict speaker's background. J. Quant. Linguist. **21**(4), 350–376 (2014)
33. Argamon, S., Fine, J., Rachel Shimoni, A.: Gender, genre, and writing style in formal written texts. Text (2003). https://doi.org/10.1515/text.2003.014
34. Romaine, S.: Language and Social Class, pp. 281–287. (2015). https://doi.org/10.1016/B978-0-08-097086-8.53015-3
35. Sloan, L., Morgan, J., Burnap, P., Williams, M.: Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. PLOS ONE **10**(3), 1–20 (2015). https://doi.org/10.1371/journal.pone.0115545
36. Wiegmann, M., Stein, B., Potthast, M.: Celebrity profiling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2611–2618. Association for Computational Linguistics, Florence, Italy (2019). https://www.aclweb.org/anthology/P19-1249
37. Watts, D., Dodds, P.: Influentials, networks, and public opinion formation. J. Consum. Res. **34**, 441–458 (2007). https://doi.org/10.1086/518527
38. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Trans. Web (2007). https://doi.org/10.1145/1232722.1232727
39. Djafarova, E., Trofimenko, O.: 'instafamous'—credibility and self-presentation of micro-celebrities on social media. Inf. Commun. Soc. **22**(10), 1432–1446 (2019)
40. Wang, Y.-C., Kraut, R.E.: Twitter and the development of an audience: those who stay on topic thrive! In: CHI (2012)
41. Hutto, C.J., Yardi, S., Gilbert, E.: In: A longitudinal study of follow predictors on twitter, In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '13, pp. 821–830. , ACM, New York, NY, USA (2013). https://doi.org/10.1145/2470654.2470771
42. Chang, S., Kumar, V., Gilbert, E., Terveen, L.: Specialization, homophily, and gender in a social curation site: Findings From Pinterest, pp. 674–686 (2014). https://doi.org/10.1145/2531602.2531660
43. Wang, Chun: Ya Jun Du, Ming Wei Tang: Opinion leader mining algorithm in microblog platform based on topic similarity. In: 2016 2nd IEEE International Conference on Computer and Communications (ICCC), pp. 160-165 (2016). https://doi.org/10.1109/CompComm.2016.7924685
44. Kiang, M.Y.: Neural networks. In: Bidgoli, H. (ed.) Encyclopedia of Information Systems, pp. 303–315. Elsevier, New York (2003). https://doi.org/10.1016/B0-12-227240-4/00121-0 . https://www.sciencedirect.com/science/article/pii/B978008044910400482X
45. Casas, I.: Neural networks. In: Kitchin, R., Thrift, N. (eds.) International Encyclopedia of Human Geography, pp. 419–422. Elsevier, Oxford (2009). https://doi.org/10.1016/B978-008044910-4.00482-X . www.sciencedirect.com/science/article/pii/B978008044910400482X
46. Hsu, C.-C., Lee, Y.-C., Lu, P.-E., Lu, S.-S., Lai, H.-T., Huang, C.-C., Wang, C., Lin, Y.-J., Su, W.-T.: Social media prediction based on residual learning and random forest, In: Proceedings of the 25th ACM International Conference on Multimedia. MM '17, pp. 1865-1870. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3123266.3127894
47. Huang, J., Tang, Y., Hu, Y., Li, J., Hu, C.: Predicting the active period of popularity evolution: a case study on twitter hashtags. Inf. Sci. **512**, 315–326 (2020). https://doi.org/10.1016/j.ins.2019.04.028
48. Zhang, Q., Gong, Y., Wu, J., Huang, H., Huang, X.: In: Retweet prediction with attention-based deep neural network. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM '16, pp. 75-84. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2983323.2983809
49. Li, J., Xu, H., He, X., Deng, J., Sun, X.: Tweet modeling with lstm recurrent neural networks for hashtag recommendation, pp. 1570–1577 (2016). https://doi.org/10.1109/IJCNN.2016.7727385
50. Simaki, V., Mporas, I., Megalooikonomou, V.: Evaluation and sociolinguistic analysis of text features for gender and age identification. Am. J. Eng. Appl. Sci. **9**, 868–876 (2016). https://doi.org/10.3844/ajeassp.2016.868.876
51. Johannsen, A., Hovy, D., Søgaard, A.: Cross-lingual syntactic variation over age and gender. (2015). https://doi.org/10.18653/v1/K15-1011
52. Namugera, F., Wesonga, R., Jehopio, P.: Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda. Comput. Soc. Netw. (2019). https://doi.org/10.1186/s40649-019-0063-4
53. Zhong, G., Wang, L.-N., Dong, J.: An overview on data representation learning: from traditional feature learning to recent deep learning. J. Financ. Data Sci. (2016). https://doi.org/10.1016/j.jfds.2017.05.001
54. Wan, Y., Chen, X., Zhang, J.: Global and intrinsic geometric structure embedding for unsupervised feature selection. Expert Syst. Appl. (2017). https://doi.org/10.1016/j.eswa.2017.10.008
55. Sirovich, L., Kirby, M.: Low-dimensional procedure for the characterization of human faces. J. Opt. Soc. Am. A Opt Image Sci. **4**, 519–24 (1987). https://doi.org/10.1364/JOSAA.4.000519
56. Jolliffe, I.. In: Lovric, M. (ed.) Principal Component Analysis, pp. 1094–1096. Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-04898-2_455
57. Peng, H., Bao, M., Li, J., Bhuiyan, M., Liu, Y., He, Y., Yang, E.: Incremental term representation learning for social network analysis. Future Gener. Comput. Syst. **86**, 1503–1512 (2018). https://doi.org/10.1016/j.future.2017.05.020
58. Wang, S., Tang, J., Liu, H.: Embedded unsupervised feature selection. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI'15, pp. 470-476. AAAI Press. (2015)
59. Zhang, B., Xiang, J., Wang, X.: Network representation learning with ensemble methods. Neurocomputing **380**, 141–149 (2020). https://doi.org/10.1016/j.neucom.2019.10.098
60. Peña, D.: Análisis de Datos Multivariantes. S.A. MCGRAW-HILL / INTERAMERICANA DE ESPAÑA, España (2002)

61. Sluban, B., Smailović, J., Battiston, S., Mozetič, I.: Sentiment leaning of influential communities in social networks. Comput. Soc. Netw. (2015). https://doi.org/10.1186/s40649-015-0016-5
62. Avnit, A.: The million followers fallacy. Pravda Media Group (2009)
63. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. In: 2010 IEEE Second International Conference on Social Computing, pp. 177-184 (2010)
64. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture, pp. 123–160 (2019). https://doi.org/10.1007/978-3-030-22948-1_5
65. Yazdanfar, N., Thomo, A.: Link recommender: Collaborative-filtering for recommending urls to twitter users. Procedia Computer Science 19, 412–419 (2013). https://doi.org/10.1016/j.procs.2013.06.056. The 4th International Conference on Ambient Systems, Networks and Technologies (ANT 2013), the 3rd International Conference on Sustainable Energy Information Technology (SEIT-2013)
66. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
67. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. **18**(17), 1–5 (2017)
68. 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, vol. 2380. CEUR-WS.org, Switzerland (2019). http://ceur-ws.org/Vol-2380/
69. Lim, K.H., Datta, A.: Finding twitter communities with common interests using following links of celebrities. (2012). https://doi.org/10.1145/2310057.2310064
70. Stoop, W., Van den Bosch, A.: Using idiolects and sociolects to improve word prediction, pp. 318–327 (2014). https://doi.org/10.3115/v1/E14-1034
71. Copland, F., Shaw, S., Snell, J.: Linguistic Ethnography: Interdisciplinary Explorations. Springer, London (2016)
72. Choi, C.J., Berger, R.: Ethics of celebrities and their increasing influence in 21st century society. J. Bus. Ethics **91**(3), 313–318 (2010). https://doi.org/10.1007/s10551-009-0090-4
73. Friendly, M.: Corrgrams: exploratory displays for correlation matrices. Am. Stat. **56**, 316–324 (2002)
74. Chessel, D., Dufour, A.-B., Thioulouse, J.: The ade4 package - I: one-table methods. R News **4**(1), 5–10 (2004)
75. Lê, S., Josse, J., Husson, F.: FactoMineR: an R package for multivariate analysis. J. Stat. Softw. Artic. **25**(1), 1–18 (2008). https://doi.org/10.18637/jss.v025.i01
76. Cappellato, L., Ferro, N., Losada, D.E., Müller, H. (eds.): CLEF 2019 Labs and Workshops, Notebook Papers, vol. 2380. CEUR-WS.org, Switzerland (2019)
77. Moreno-Sandoval, L.G., Mendoza-Molina, J.F., Puertas-Del Castillo, E.A., Duque-Marín, A., Pomares-Quimbaya, A., Alvarado-Valencia, J.A.: Age classification from Spanish tweets - the variable age analyzed by using linear classifiers. In: Hammoudi, S., Smialek, M., Camp, O., Filipe, J. (eds.) Proceedings of the 20th International Conference on Enterprise Information Systems (ICEIS 2018), pp. 275–281 (2018). https://doi.org/10.5220/0006811102750281
78. Moreno-Sandoval, L.G., Sanchéz-Barriga, C., Espíndola-Buitrago, K., Pomares-Quimbaya, A., Garcia, G.C.: Spanish Twitter data used as a source of information about consumer food choice. In: Holzinger, A., Kieseberg, P., Tjoa, A., Weippl, E. (eds.) Machine Learning and Knowledge Extraction. International Cross-Domain Conference for Machine Learning and Knowledge Extraction. CD-MAKE 2018. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99740-7_9

Publisher's Note