PAN @ CLEF 2019 - CELEBRITY PROFILING

CELEBRITY PROFILING ON TWITTER USING SOCIOLINGUISTIC FEATURES

Luis Gabriel Moreno Sandoval, Edwin Puertas, Flor Miriam Plaza del Arco, Alexandra Pomares Quimbaya, Jorge Andrés Alvarado Valencia, and L. Alfonso Ureña López
Pontificia Universidad Javeriana, Bogotá, Colombia
{edwin.puertas,jorge.alavarado,morenoluis,pomares}@javeriana.edu.co
Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA)
Universidad de Jaén, Jaén, Andalucía, Spain.
{fmplaza, laurena}@ujaen.es

Summary

Social networks have been a revolutionary scenario for celebrities because they allow them to reach a wider audience with much higher frequency than using traditional means. These platforms enable them to improve or sometimes deteriorate, their careers through the construction of closer relationships with their fans and the acquisition of new ones. Indeed, networks have promoted the emergence of a new type of celebrities that exists only in the digital world.

Being able to characterize the celebrities that are more active on social networks, such as Twitter, gives an enormous opportunity to identify what is their real level of fame, what is their relevance for an age group, or a specific gender or occupation. These facts may enrich decision making, especially in advertising and marketing.

Hypothesis - H0

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

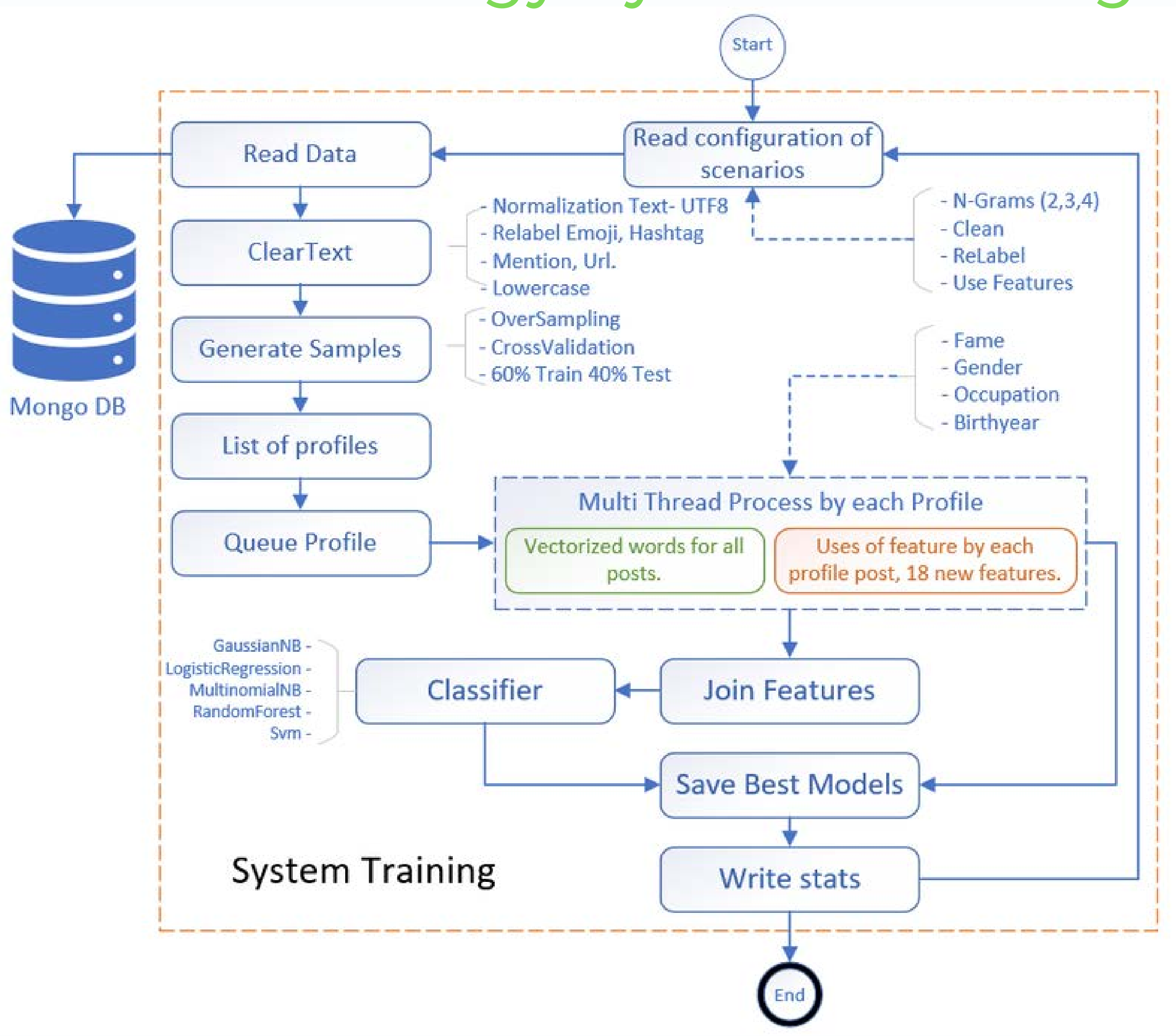
Features for Classification Model

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

+ Words Vector from Tweets

1) Network Features,
2) Lexical Features,
3) Sociolinguistic features
4) Text Tweet

Methodology System Training



Results

The novelty in the analysis presented in this paper is to analyze specific featuresof digital social networks for each profile. The use of sociolinguistic features in theuser profile has shown many quirks in topics social, cultural, and of gender. Thesecharacteristics describe the sociolect of celebrities linked in this study; we also find itis essential to understand if the text was written in the first, second or third person, andthe lexical diversity that each profiles had.As future work, we plan to analyze the models with real samples with a similaror greater volume of messages. Finally, we want to review the posts and context datato have models that respond socially to variables that represent real phenomena in thenetwork.

Result Classifier Accuracy

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

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

ACKNOLEDGMENTS

Center for Excellence and Appropriation in Big Data and Data Analytics (CAOBA), Pontificia Universidad Javeriana, and the Ministry of Information Technologies and Telecommunications of the Republic of Colombia (MinTIC). The models and results presented in this challenge contribute to the construction of the research capabilities of CAOBA. Also, Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) and LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government.