SINAI at PoliticEs 2022: Exploring Relative Frequency of Words in Stylometrics for Profile Discovery

César Espin-Riofrio[1,*], Jenny Ortiz-Zambrano[1,*] and Arturo Montejo-Ráez[2,*]

[1]University of Guayaquil, Delta Av. s/n, Guayaquil, 090510, Ecuador
[2]University of Jaén, Las Lagunillas s/n, Jaén, 23071, Spain

Abstract

In this article we summarise our participation in the PoliticEs task within the IberLEF evaluation forum in its 2022 edition. This task is entitled Spanish Author Profiling for Political Ideology. We proposed a Voting Classifier model that leverages the use of several classical classifiers using as features the combination of stylometry measures with embeddings obtained from a Spanish RoBERTa model for text representation. Our final work achieved an F1 score of 0.785 for Gender prediction, 0.753 for Profession, 0.784 for Ideology_Binary and 0.561 for Ideology_Multiclass, with a final macro average for F1 of 0.721. These results indicate that the combination of stylometric features can be useful in the determination of user profiles.

Keywords

Stylometry, Author profiling, Transformer model, Voting classifier, Ensemble learning

1. Introduction

Social media has undergone a great evolution, becoming an essential factor in the communication of today's society [1], giving rise to new sources of data. Many organisations use this data as a tool to analyse some of their events or members [2].

The Natural Language Processing (NLP) task known as Author Profiling aims to identify the personal traits of an author from his or her writings. The traits, such as gender, age, linguistic variety or personality, are of great interest for fields such as forensics, security and also marketing [3]. Traditional research has been carried out mainly in English [4].

Political ideology is a psychographic trait that can be used to understand individual and social behaviour, including moral and ethical values, as well as inherent attitudes, judgements, biases and prejudices [5]. The relationship between personality traits and political ideology was shown in [6]. The aim of the IberLEF 2022 task - PoliticES Spanish Author Profiling for Political Ideologies [7], is to extract author-specific information, such as political ideology, gender and profession, from a set of tweets of a given user. Political ideology has been posed both as a binary problem (left-right) and as a multi-class problem.

IberLEF 2022, September 2022, A Coruña, Spain.
*These authors contributed equally.
cesar.espinr@ug.edu.ec (C. Espin-Riofrio); jenny.ortizz@ug.edu.ec (J. Ortiz-Zambrano); amontejo@ujaen.es (A. Montejo-Ráez)
0000-0001-8864-756X (C. Espin-Riofrio); 0000-0001-6708-4470 (J. Ortiz-Zambrano); 0000-0002-8643-2714 (A. Montejo-Ráez)

Our system, to establish the authors' profiles according to the task, uses stylometric measures such as phraseological variables, usage of frequent words for Spanish, or embeddings generated by a BERT-type neural network. These features feed a voting system comprising several classical classifiers, facing this task as a machine learning problem.

In Section 2 we present our method and the different models and strategies tested, together with the dataset used. In Section 3 we show our results and finally, in Section 4, we present the conclusions of our research and directions for future work.

2. Method

This section presents the dataset and method used in the experimentation. We create vectors by extracting features from the tweets of the various authors using general and political lexical usage analysis, Transformer model for text representation through embeddings, and phraseological analysis, and then train a set of machine learning classification models. The system has been developed using Python [8] and libraries such as Scikit-learn [9] and NLTK [10].

2.1. Training data

The organizers provided an extension of the Policorpus 2020 dataset [11] proposed in the task, which used UMUCorpusClassifier [12] to collect tweets during 2020 and 2021 from several Twitter accounts of politicians and journalists in Spain. The accounts were selected mainly from: (1) members of the Spanish government, (2) members of the Spanish Congress and Senate, (3) mayors of some important Spanish cities, (4) presidents of autonomous communities, (5) former politicians, and (6) collaborators affiliated to political parties. It includes journalists from different Spanish media, such as ABC, El País, El Diario, El Mundo or La Razón, among others.

Each author was tagged with their gender, profession, and their political spectrum on two axes: binary (left, right) and multiclass (left, moderate_left, moderate_right, right). The dataset consists of about 400 different users with at least 120 tweets. For the shared task, independent training and test sets were released.

2.2. Analysis of general and political lexical usage

For this analysis, we take as reference the first most 500 frequently used words listed in the "Corpus de Referencia del Español Actual (CREA)" [13], which has been established in previous experiments as an influential resource in determining an author's writing style.

Likewise, we obtained a list of 300 political words used on Twitter through the corpus compiled in [14], consisting of Spanish tweets from fifteen Twitter accounts of politicians from the five main parties (PSOE, PP, Cs, UP and VOX) covering the Spanish election campaign on November 10, 2019 (10N Spanish Elections).

With CREA's 500 frequently used words and the 300 words of political use on Twitter from the 2019 Spanish elections, we created feature vectors for each author (just computing the frequency of the words present in the tweet that are contained in these two lexicons). We believe it is an interesting approach to get closer to determining the writing style for the proposed task.

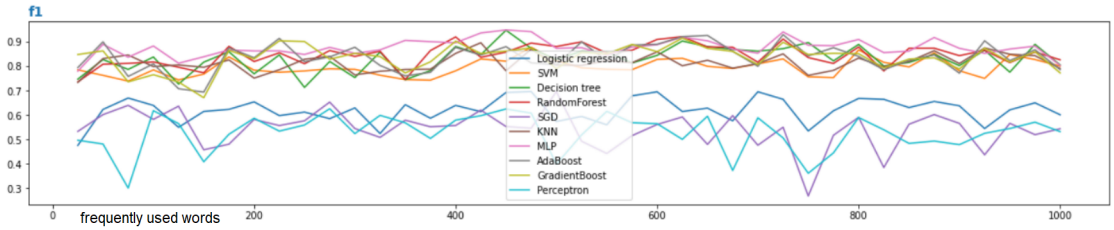

Figure 1: F-score of Classifiers in previous experimentation.

2.3. Basic statistical characteristics

From the set of tweets of each author, we obtained the following characteristics:

- Type-token ratio: this is the frequency of each grammatical category in the text, i.e. the *part-of-speech*.
- Average word length.
- Mean sentence length.
- Standard deviation of sentence length.
- Average length of tweets.

2.4. Transformer model for text vectorisation

Transformer models [15] are being widely used with good results in many Natural Language Processing tasks [16]. The Transformer architecture is especially conducive to pre-training on large text corpora, allowing for higher accuracy in tasks such as text classification [17].

We have used a Spanish RoBERTa model (MarIa) [18], a pre-trained by masking on a large Spanish corpus [19]. It is based on the RoBERTa model [20] and has been pre-trained using the largest Spanish corpus known to date, with a total of 570 GB of cleaned and processed text for this work, collected from the web crawls carried out by the National Library of Spain [21] from 2009 to 2019. To generate the embeddings of a text, the encodings for the [CLS] token were taken.

2.5. Classifiers

All the above features (RoBERTa embeddings, statistical features and frequencies for words belonging to the two former lexicons) are concatenated and scaled, at feature level, to the interval [0,1], so that we have a single vector that feeds the following supervised learning classification methods: Logistic Regression (LR), Random Forest (RF), Decision Trees (DT), Multi-layer Perceptron (MLP) and Gradient Tree Boosting (GTB), provided by the Scikit-learn library [22]. We performed previous experimentations on a wide set of classification algorithms to obtain the best performing ones according to the F-score (see Figure 1), hence the choice of the mentioned classifiers.

These classifiers are configured using default parameters set by the Sckit-learn library, and are combined using a voting system [23]. In our case, we opt for a majority vote (hard vote) for selecting the final predicted class.

3. Preliminary experimentation

We conducted a series of preliminary experiments on the provided training data aimed at evaluating the goodness of fit of our system proposal. The various experiments were carried out by increasing the number of frequently used words in the Spanish language, and Table 1 shows the different results. Our system will be based on the combination of all the evaluated classifiers, by means of a voting system as described lately. The highest averaged F1-score over the considered classifiers was obtained with 500 words. Therefore, we used this number of words as frequently used words from the CREA list.

Table 1
F-score, best performance methods.

Words	LR	DT	RF	MLP	GTB
50	0.6994	0.6965	0.7444	0.8294	0.8429
100	0.6139	0.6382	0.6993	0.7379	0.7123
150	0.6144	0.7547	0.8240	0.7872	0.8366
200	0.5943	0.8352	0.8077	0.8918	0.8425
250	0.6166	0.7884	0.7998	0.8629	0.8108
300	0.5862	0.7781	0.8195	0.8623	0.8529
350	0.6555	0.8524	0.8687	0.8774	0.7809
400	0.6490	0.7998	0.9014	**0.9385**	0.8538
450	0.5844	0.8425	0.8514	0.8819	0.8460
500	**0.7029**	0.8271	0.8931	0.8972	**0.9074**
550	0.6563	0.8857	0.8652	0.8619	0.8829
600	0.5918	0.8583	0.8570	0.8879	0.8848
650	0.6250	0.8473	0.8756	0.8578	0.8575
700	0.5968	0.8603	0.8439	0.9042	0.8433
750	0.6340	**0.9181**	0.8608	0.8575	0.8542
800	0.6054	0.8862	0.8592	0.8826	0.8824
850	0.5694	0.8570	**0.9156**	0.8422	0.8740
900	0.6721	0.7798	0.8270	0.8845	0.7652
950	0.6510	0.7577	0.8157	0.8892	0.7894
1000	0.6407	0.7971	0.8600	0.8633	0.7626

Using cross-validation and the F-score measure, Table 2 shows the results obtained for each classifier on each label of the task, including the overall evaluation of the classifiers.

Table 2
F-score. Evaluating estimator performance.

Classifier	Gender	Profession	Ideology_Binary	Ideology_Multiclass
LR	**0.67**	**0.92**	0.86	0.78
RF	0.65	0.86	0.80	0.60
DT	0.56	0.82	0.74	0.52
MLP	0.64	0.91	**0.89**	**0.79**
GTB	0.62	0.83	0.85	0.62
Ensemble	0.67	0.90	0.87	0.74

LR and MLP turned out to be the best performing classifiers. For the label Gender and Profession, the best performance was observed with LR, for Ideology_Binary and Ideology_Multiclass with MLP. On the other hand, DT was the worst performer in the evaluation. The training process lasted 18 minutes and 36 seconds in the Google Colaboratory runtime environment.

Table 3 shows our participation with SINAI team compared to the official results in the Iberlef 2022 - PoliticES task [7] provided by the organization. It refers to the first and last results as well as the calculated average of all scores. We obtained the 16th place among 20 participating teams with an Average Macro F1 of 0.721471, being below the average participation score of 0.757993. Our best result was with the Gender label, being in fifth place above the average with an F-score of 0.785714. As can be seen in the table, we are not very far from the average values of the labels, but we are far from the best systems that achieved very high and difficult to beat yields.

Table 3
Comparison of SINAI team with final results, Iberlef 2022 - PoliticES

Team Name	Average Macro F1	F1 Gender	F1 Profession	F1 Ideology Binary	F1 Ideology Multiclass
LosCalis	0.902262 (1)	0.902868 (1)	0.944327 (1)	0.961623 (1)	0.800229 (4)
NLP-CIMAT-GTO	0.890961 (2)	0.784836 (6)	0.921250 (3)	0.961482 (2)	0.896275 (1)
Alejandro Mosquera	0.889182 (3)	0.826714 (3)	0.933452 (2)	0.951519 (3)	0.845044 (3)
…					
(Average scores)	0.757993	0.742648	0.787287	0.842803	0.659235
SINAI	**0.721471 (16)**	**0.785714 (5)**	**0.753945 (15)**	**0.784689 (15)**	**0.561536 (16)**
…					
UC3M-DEEPLNLP-1	0.586437 (19)	0.648920 (17)	0.403409 (19)	0.746377 (17)	0.547044 (17)
UMUTeam	0.511228 (20)	0.576211 (19)	0.432432 (18)	0.595665 (19)	0.440603 (19)

4. Conclusions and future work

In this paper, we summarize the method and classifiers presented to solve the task proposed in Iberlef 2022 - PoliticES. We use lexical usage analysis by means of general and political frequent use words, the RoBERTa-base-bne Transformer model for text vectorization, complemented with phraseological features. Several classical machine learning algorithms were tested, with the Multilayer Perceptron as the best performing algorithm in preliminary experiments. The selected algorithms were combined in a voting system.

As for the officially obtained results, we have achieved an acceptable participation, with the best results in the determination of Gender and Ideology_Binary followed by Profession and Ideology_Multiclass. It is worth highlighting the importance of the use of frequent words in the analysis of lexical features, including also frequent words of a political nature. In relation to the training phase, the use of Voting Classifier was important to enhance the overall performance of the classifiers used. In addition, the use of RoBERTa-base-bne for text vectorization enhanced feature extraction, as it is a pre-trained Transformer model with high performance for the Spanish language.

Finally, comparing the results obtained, it can be seen that the method presented here can give even better results in combination with other techniques with Spanish lexical analysis using frequently used words. It is our intention to further explore the combination of specific

features with deep neural networks and to tune these networks to improve their performance in author profiling tasks.

We plan to continue our research by exploring how the different features contribute to the performance. Ablation tests will be carried out to this end. Also, fine-tuned RoBERTa models can enhance the system. It is our purpose to continue the investigation on the relevance of frequently used words as a main indicator of authorship. How this frequent words are selected could depend on the context of the texts, so domain oriented lists of words could be needed.

Acknowledgments

This work has been partially supported by Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, and LIVING-LANG project (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033.

References

[1] W. are social, Digital Report 2022: El informe sobre las tendencias digitales, redes sociales y mobile. - We Are Social Spain, 2022. URL: https://wearesocial.com/es/blog/2022/01/digital-report-2022-el-informe-sobre-las-tendencias-digitales-redes-sociales-y-mobile/.

[2] F. Rangel, G. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling hate speech spreaders on twitter task at pan 2021, in: CLEF, 2021.

[3] M. Potthast, P. Rosso, E. Stamatatos, B. Stein, A decade of shared tasks in digital text forensics at pan, in: European Conference on Information Retrieval, Springer, 2019, pp. 291–300.

[4] J. Savoy, Machine learning methods for stylometry, Springer, 2020.

[5] B. Verhulst, L. J. Eaves, P. K. Hatemi, Correlation not causation: The relationship between personality traits and political ideologies, American journal of political science 56 (2012) 34–51.

[6] M. Fatke, Personality traits and political ideology: A first global assessment, Political Psychology 38 (2017) 881–899.

[7] J. A. García-Díaz, S. M. Jiménez-Zafra, M. T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticES 2022: Spanish Author Profiling for Political Ideology, Procesamiento del Lenguaje Natural 69 (2022).

[8] S. Raschka, Python machine learning, Packt publishing ltd, 2015.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.

[10] E. Loper, S. Bird, Nltk: The natural language toolkit, arXiv preprint cs/0205028 (2002).

[11] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74.

[12] J. A. García-Díaz, Á. Almela, G. Alcaraz-Mármol, R. Valencia-García, Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks, Procesamiento del Lenguaje Natural 65 (2020) 139–142.
[13] [Online], 2022, CREA | Real Academia Española, URL: https://www.rae.es/banco-de-datos/crea.
[14] J. Sánchez-Junquera, S. P. Ponzetto, P. Rosso, A twitter political corpus of the 2019 10n spanish election, in: International Conference on Text, Speech, and Dialogue, Springer, 2020, pp. 41–49.
[15] [Online], 2022, Transformers, URL: https://huggingface.co/docs/transformers/index.
[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.
[17] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, X. Le Q, generalized autoregressive pretraining for language understanding. arxiv 2019; 1906.08237, 1906.
[18] [Online], 2022, [2107.07253] MarIA: Spanish Language Models, URL: arxiv.org/abs/2107.07253.
[19] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.
[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[21] [Online], 2022, Biblioteca Nacional de España, URL: http://www.bne.es/es/Inicio/index.html.
[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[23] D. Ruta, B. Gabrys, Classifier selection for majority voting, Information fusion 6 (2005) 63–81.