LosCalis at PoliticEs 2022: Political Author Profiling using BETO and MarIA

Sergio Santamaria **Carrasco**[1], Roberto Cuervo **Rosillo**[1]

[1]*Universidad Carlos III de Madrid, Computer Science Department, Av de la Universidad, 30, 28911, Leganés, Madrid, Spain*

Abstract

The specific identification of author characteristics from content, also called author profiling, is gaining ground, as it can play an important role for a wide range of applications, such as automatic tailoring of customer service communication. Political ideology is a psychographic trait strongly correlated with individual and social behavior and its analysis can contribute to uncovering various personal characteristics. This, coupled with the widespread use of social networks, which have become a tool for political expression, makes the development of systems capable of automatically extracting this information with the aim of identifying and reaching different groups a matter of great interest. In this paper, we describe a deep learning architecture for the identification of gender, profession and political ideology in social media. The architecture is based on pre-trained Spanish BERT and RoBERTa. The proposed system participated in PoliticEs and obtained a micro-F1 of 90.28%

Keywords

Author profiling, NLP, Language model, Classificattion, BERT, RoBERTa, Deep Learning

1. Introduction

With the exponential growth in the use of social networks such as Twitter, researchers have been able to study human behavior on an unprecedented scale. Different approaches suggest that through the analysis of the language used, numerous attributes such as gender, age or psychological characteristics of users can be extracted [1]. Political ideology is a psychographic trait that can be used to fully understand individual and social behaviours, including moral and ethical values as well as inherent attitudes, appraisals, biases, and prejudices [2]. Since political affinity has a great impact on our society and it influences the decisions taken during day-to-day life, the identification of this affinity through **Natural Language Processing (NLP)** techniques can be a key factor in the study of human behavior as well as in areas such as marketing, where target groups identification is crucial.

Automatic extraction of Political ideology information from texts can be addressed as author profiling problem. This problem represents one of the three major fields in Automatic Authorship Identification together with Authorship attibution and Authorship identification.

This paper describes the team LosCalis participation in PoliticEs 2022 challenge [3], which focuses on the gender, profession, and political spectrum identification from a binary and

multi-class perspective using Twitter Spanish data.

Proposed system core is based on transformers [4] and it combines the spanish pre-trained BERT [5] model known as **BETO** [6] as well as the spanish pre-trained RoBERTa [7] model known as **MarIA** [8]. It employs both architectures for document level characteristics extraction together with a MLP for labels decoding. The obtained results by the system proposed reach a 0.9028 F1 score. System results shows the good performance of the approaches based on deep learning.

The rest of the paper is organized as follows. Section 2 describes the datasets provided by the PoliticEs task, as well as an extension of this developed by our team. In Section 3, the architecture of our system is described. Section 4 presents the results obtained for our system. Section 5 draws some preliminary conclusions from our analysis.

2. Dataset

2.1. PoliticEs 2022 corpus

The dataset provided [9] contains tweets from different users selected along Spanish goverment members, Spanish Congress and Senate members, mayors of some important cities in Spain, some autonomous communities presidents, former politicians, collaborators affiliated with political parties and finally, different Spanish news media journalists. Each tweet was labelled with his author's gender (male and female), profession (politician and journalist) and political spectrum on two axes: binary (left and right) and multiclass (left, moderate left, moderate right and right).

The corpus was divided in two parts: training and test. Table 1 describes the distribution of tweets and users in each subset.

Table 1
PoliticEs 2022 corpus statistics

	Training set	Test set
No. of users	313	72
No. of tweets per user	120	120
Total No. of tweets	37,560	8,640

Figure 1 ilustrates the distribution of different labels. While dataset is balanced for most categories, it is striking how this is not the case for the profession category, where the number of politicians is much higher than the number of journalists.

Finally, as part of the data exploration analysis, Figure 2 shows the most relevant words for each of the political spectrums. It is interesting to note that while left-wing politicians refer to the political right and its representatives, right-wing politicians refer to the political left and its representatives.
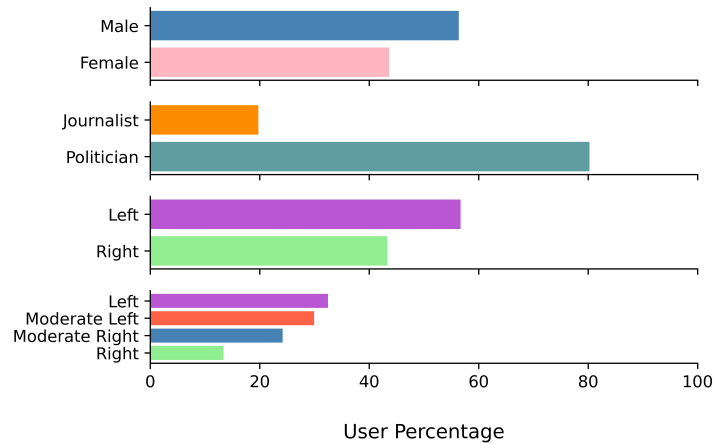
Figure 1: Distribution per user of the different categories, gender, profession, binary political ideology and multiclass political ideology, in the training subset.

Figure 2: Most relevant word clouds by political spectrum: left (top left), moderate left (top right), moderate right (bottom left) and right (bottom right).

2.2. Extending the dataset

During our participation in PoliticEs 2022, it was decided to extend the original dataset, collecting tweets from Spanish politicians and journalists posted between January 2021 and February 2022. In order to ensure that our dataset was similar to the one offered by the organization, the same preprocessing steps were followed:

1. Twitter accounts are encoded as @user.
2. URLs are removed.
3. Mentions of political parties are replaced by the token [POLITICAL_PARTY].

As a result, 361,646 different tweets were obtained from a total of 430 unique users. Due to the anonymization of users in the original dataset, it cannot be assured that there are no repeat users. However, since the publication date of the tweets in the dataset provided by the organization is 2020 and that of the tweets collected by our team is 2021 and 2022, it is guaranteed that the new data were not present in the original corpus.

An important difference from the original data is that the number of tweets per user is not fixed and varies from user to user. Despite the fact that the number of posts per user could have been limited, it was considered that our system would benefit from each example extracted.

Figure 3, which shows the distribution of users for each of the different categories, illustrates that the distribution is not far from that of the original dataset. Meanwhile, Figure 3 shows the same distribution but this time by tweets.
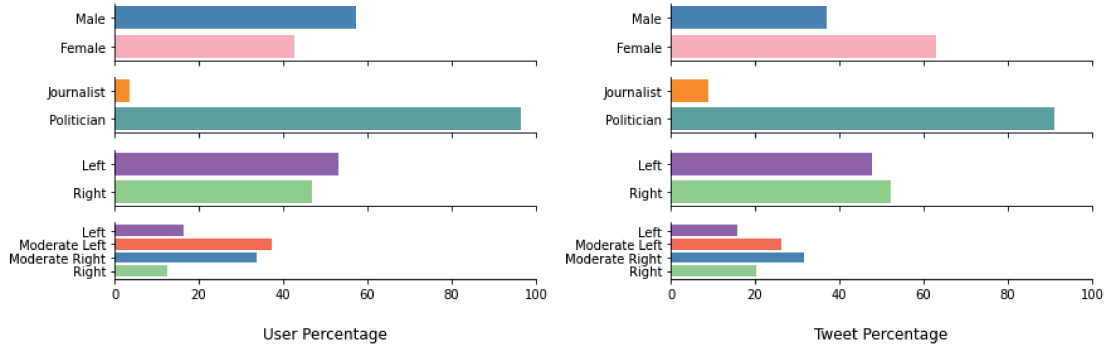

Figure 3: User (left) and Tweet (right) distribution of the different categories, gender, profession, binary political ideology and multiclass political ideology, in the dataset extension.

3. Methods and system description

Proposed system for the author profiling task is based on the fine-tuning of the combination of two pre-trained transformer-based models. The first one, BETO is the first Spanish BERT-based and has demonstrated its performance in a large multitude of different NLP tasks [10, 11]. The second, MarIA, is a recently presented RoBERTa-based model that has been pretrained using a massive corpus of 570GB of clean and deduplicated texts with 135 billion words extracted from the Spanish Web Archive crawled by the National Library of Spain between 2009 and 2019 [8]. Its choice was motivated by his outperform of the existing Spanish models across a variety of NLU tasks. Designed architecture follows the scheme shown in Figure 4.

User tweets are transformed into tokens that feed BETO and MarIA. From the classification token known as [CLS], each model generates a 768-dimensional vector representing the meaning of the entire sentence. Both vectors are concatenated and the result serves as input to four separate classification blocks, one for each classification task (gender, profession, binary ideology
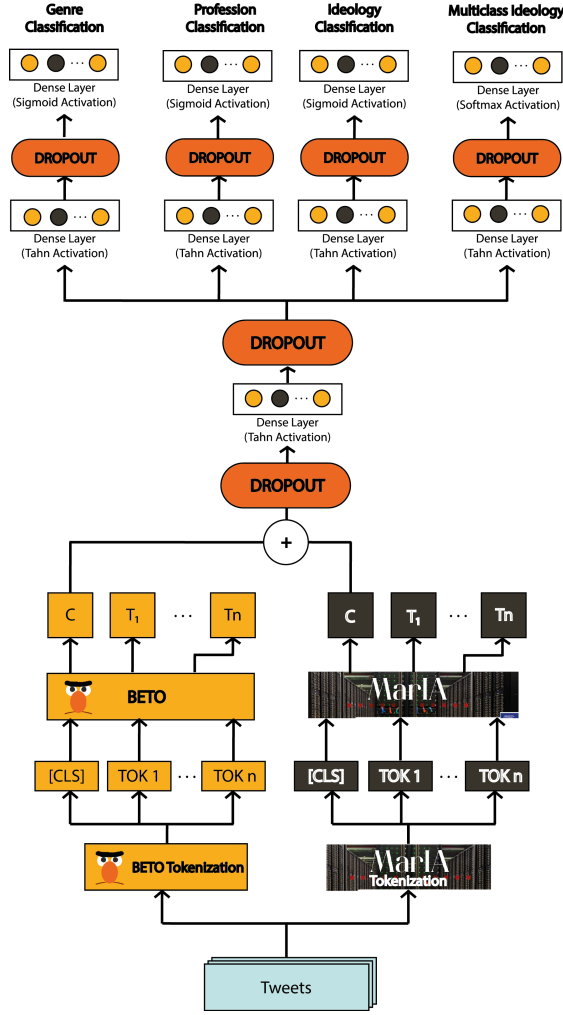
Figure 4: Architecture of the proposed model for author profiling.

and multiclass ideology). Each classification block is formed by fully connected dense layer with 768 units and $tanh$ activation function that directly connects with linear layer with a softmax or sigmoid activation on top, which returns a probability score for each class. Dropout for regularization, with probability of 0.15 to prevent overfitting, is applied after BETO, MarIA and the fully connected dense layers.

Due to the limitations of BERT-based models, where the maximum number of tokens is set to 512, it is not possible to concatenate tweets from the same user and serve them as input to our model. Consequently, tweets from the same users are grouped into blocks of maximum 512 tokens. Each block is classified individually using the previously described architecture. A voting system is used to obtain the final classification of a user of each category, which corresponds to the predominant class, as shown in Figure 5.

The system has been developed in python 3 [12] with Keras 2.9.0 [13], Tensorflow 2.8.0 [14]
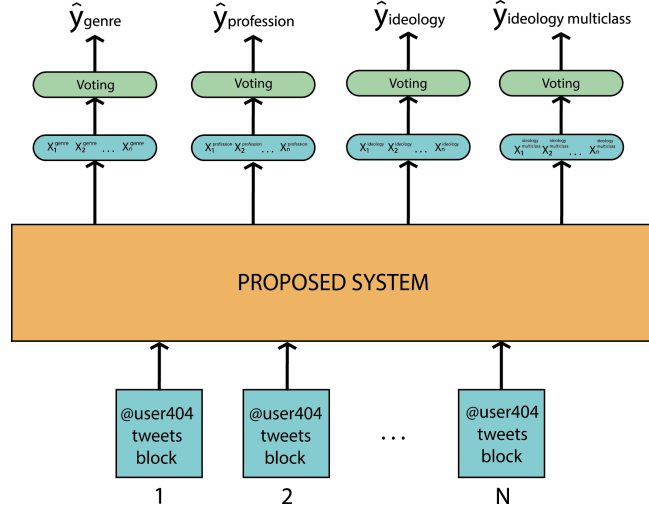
Figure 5: Voting system.

Table 2
Detailed hyper-parameter setting

Parameter	Value
Train batch size	64
Predict batch size	8
Learning rate	3e-5
Training epochs	5
Max sequence lenght	512
Dropout	0.15
Optimizer	Adam [16]
Dense layer units	768

and the transformers library provided by HuggingFace [15]. The HuggingFace transformers package is a python library providing pre-trained and configurable transformer models. It provides the BETO cased and the MarIA RoBERTa-base used in our system.

Hyperparameters used can be found in Table 2.

4. Results

In our experiments, standard measures of accuracy, recall and macro F1 score were applied to evaluate the performance of our model on each classification task. To evaluate the overall performance of the system the macro average F1 score was adopted. These metrics are also adopted as the evaluation metrics during PoliticEs task.

Throughout the experimentation, the extension of the dataset developed during the competition was used to train the system, while the dataset provided by the organization was exploited to hyperparameter fine tuning. The detailed hyper-parameter settings are illustrated in Table 3

where 'Opt.' denotes optimal.

Table 3
Detailed hyper-parameter settings in the PoliticEs task.

Parameters	Tuned range	Opt.
Train batch size	[16, 32, 64]	64
Learning rate	[1e-5, 2e-5, 3e-5, 1e-4]	3e-5
Max sequence length	[128, 256, 512]	512
Dropout	[0.10, 0.15, 0.20]	0.15
Dense layers units	[384, 516, 768, 960]	760
Epochs	[3, 4, 5, 6, 7]	5

In the course of experimentation, we were able to verify how a higher aggregation of tweets per user results in improved system performance. This was observed during the variation of the maximum sequence length hyperparameter, where increasing the maximum number of tokens increased the number of tweets per aggregation. The results obtained are shown in Table 4.

Table 4
Experimental results by varying max sequence length

Max Sequence Lenght	Macro F1-score
128	0.827612
256	0.859478
512	0.889967

Considering the optimal hyper-parameters, this configuration was used in our model to process the test set provided by PoliticEs. Our proposal, as the Table 5 shows, reaches an average macro F1-score of 0.902262. The results in the different classification tasks vary, being notably worse in the multiclass ideology classification. The reason is probably due to the fact that it is the classification task with the largest number of classes and the similarity between discourses within the political spectrum.

Table 5
Results of PoliticEs on the test set

Classification Task	Macro F1-score
Gender	0.902868
Profession	0.944327
Ideology Binary	0.961623
Ideology Multiclass	0.800229
Average Macro F1	0.902262

In addition, the errors generated by our system on the corpus set were analyzed. Confusion matrix of the gender and profession classification tasks can be seen in Figure 6, while those of binary and multiclass ideology in Figure 7. It can be observed that, while the binary ideology

classification errors are balanced, the errors in the classification of both gender and profession are unbalanced, showing a bias towards the classes with a higher presence (male and political) in the dataset used for training.
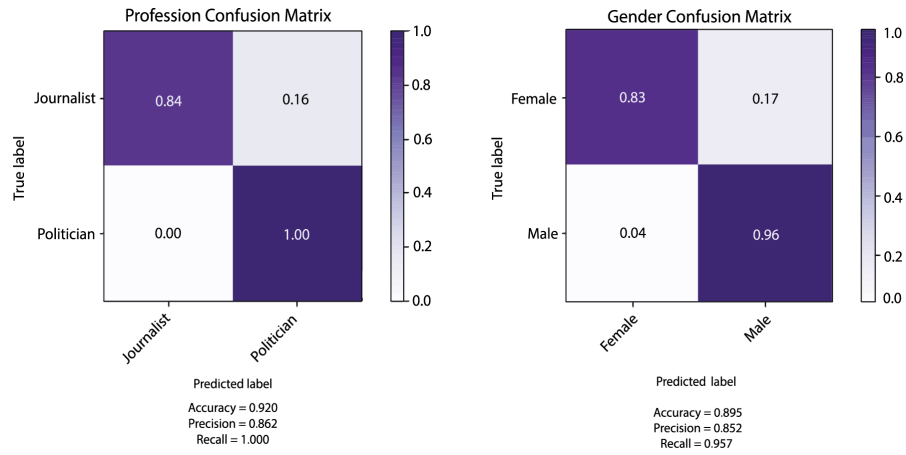

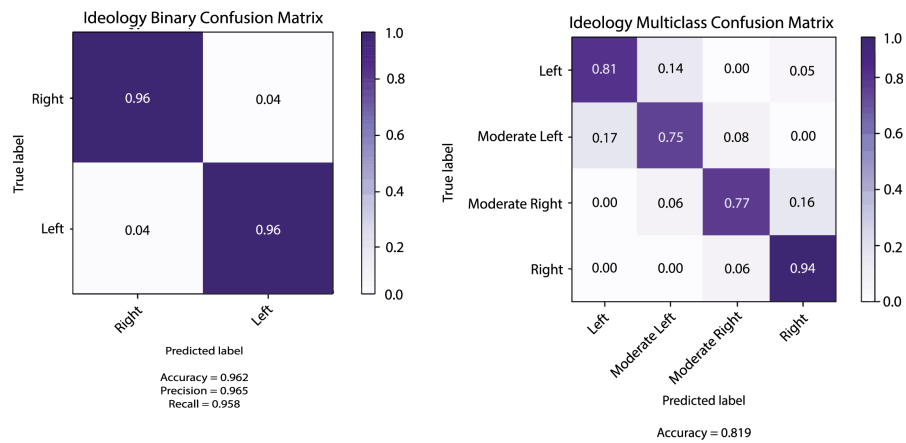**Figure 6:** Profession (left) and gender's (right) confusion matrices.

Figure 7: Binary (left) and multiclass (right) ideology's confusion matrices.

Misclassification errors in multiclass ideology generally occur between those classes that are close together on the political spectrum, such as left and moderate left or right and moderate right.

5. Conclusion

The increase of users in recent years in social networks has turned author profiling into a tool of great interest that has potential benefits in the study of human behavior and in a variety of applications such as adapting customer services or political agendas. It is important to be aware

of the potential social implications of these methods [17], as these techniques can be used for pernicious purposes. An example is the Facebook-Cambridge Analytica case [18], in which a large amount of Facebook users' data obtained without their consent was used to influence the voting intention of the social network's users during different election campaigns.

PoliticEs is the first Spanish shared task focused in extracting political ideology from a text collection with the aim of contributing to sociology research, since it is a psychographic trait strongly correlated with personality.

As a result of our participation, the original dataset provided by the organization has been extended, greatly increasing the number of manually annotated tweets. A deep learning based system is also proposed, combining two powerful pre-trained Spanish transformer-based models, BETO and MarIA, which achieves a high performance obtaining 0.902262 average macro F1-score, demonstrating the validity of the architecture for the task.

Due to the limitation of transformer-based models that are unable to process long sequences due to their self-attention operation, future work would explore Longformer [19]. The attention mechanism of this architecture replaces standard self-attention and combines a local windowed attention with a task motivated global attention, allowing longer sequences to be processed with less loss of performance. Since an increase in the maximum number of sequence length may lead to an increase in computational cost, another alternative to explore are text truncation methods, such as the one proposed in [20]. This method applies a truncation algorithm guided by tokens identified as important with known methods from the explainable artificial intelligence domain. Both alternatives would allow all tweets from the same user to be processed together, avoiding the loss of information and turning the voting system dispensable.

References

[1] F. Mairesse, M. Walker, et al., Words mark the nerds: Computational models of personality recognition through language, in: Proceedings of the Annual Meeting of the Cognitive Science Society, volume 28, 2006.

[2] 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.

[3] 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).

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[6] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[7] 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).

[8] 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.

[9] 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.

[10] Y. Fu, Z. Yang, N. Lin, L. Wang, F. Chen, Sentiment analysis for spanish tweets based on continual pre-training and data augmentation, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain, 2021.

[11] J.-C. Han, R. T.-H. Tsai, Ncu-iisr: Pre-trained language model for cantemist named entity recognition., in: IberLEF@ SEPLN, 2020, pp. 347–351.

[12] G. Van Rossum, F. L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.

[13] F. Chollet, et al., Keras, 2015. URL: https://github.com/fchollet/keras.

[14] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning, in: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.

[15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).

[16] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[17] D. Hovy, S. L. Spruit, The social impact of natural language processing, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 591–598.

[18] J. Isaak, M. J. Hanna, User data privacy: Facebook, cambridge analytica, and privacy protection, Computer 51 (2018) 56–59.

[19] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[20] K. Fiok, W. Karwowski, E. Gutierrez-Franco, M. R. Davahli, M. Wilamowski, T. Ahram, A. Al-Juaid, J. Zurada, Text guide: Improving the quality of long text classification by a text selection method based on feature importance, IEEE Access 9 (2021) 105439–105450.

A. Online Resources

The sources for the LosCalis participation are available via

- GitHub https://github.com/ssantamaria94/PoliticEs2022,