

# CIMAT\_2021 at PoliticEs 2022: Ensemble Based Classification Algorithms for Author Profiling in Spanish Language

Enrique Santibáñez-Cortés<sup>1</sup>, Azael Carrillo-Cabrera<sup>1</sup>, Yair Antonio Castillo-Castillo<sup>1</sup>, Daniela Moctezuma<sup>2</sup> and Victor Muñoz-Sánchez<sup>1</sup>

<sup>1</sup>Research Center in Mathematics (CIMAT), Monterrey, Nuevo León, Mexico

<sup>2</sup>Centro de Investigación en Ciencias de Información Geoespacial, Aguascalientes, Ags., Mexico.

## Abstract

Author profiling is a very important and useful task in the Natural Language Processing research community. Its objective is to infer some characteristics related to the author of some text, such as gender, age, and preferences, among others. In this paper, we present our solution to the Spanish Author Profiling for Political Ideology task in PoliticEs@IberLEF2022. This solution consists on specialized classification models for each subtask, specifically, we used fine-tuned BERT models for the gender and profession subtasks, XGBoost for binary ideology, and Logistic Regression for multiclass ideology. A variety of pre-processing techniques were also used to clean up the texts. With our final approach we obtained the 4th place in the PoliticEs contest.

## Keywords

Author profiling, Ensemble Learning, Deep Learning

## 1. Introduction

The current technological advance allows us to access large amounts of data generated by users on the Internet and social media platforms such as Twitter, Facebook, blogs, or public forums. Most of this data is textual, and to get some insight about the people who generate such data, is very important and useful for many real world applications. Nowadays, there are significant research efforts to tackle this problem, mainly in the field of Natural Language Processing (NLP) and Machine/Deep Learning (ML, DL).

The author profiling task is defined in [1] as "the process of identifying the data about a user interest domain". Furthermore, in [2], it is described as the extraction of demographic aspects of a person from their texts. For example, gender, age, location, occupation, socioeconomic level or native language, but also, personality traits such as extraversion or neuroticism as well as political ideology among others.

---

*IberLEF 2022, September 2022, A Coruña, Spain*

✉ [enrique.santibanez@cimat.mx](mailto:enrique.santibanez@cimat.mx) (E. Santibáñez-Cortés); [azael.carrillo@cimat.mx](mailto:azael.carrillo@cimat.mx) (A. Carrillo-Cabrera);


[yair.castillo@cimat.mx](mailto:yair.castillo@cimat.mx) (Y. A. Castillo-Castillo); [dmoctezuma@centrogeo.edu.mx](mailto:dmoctezuma@centrogeo.edu.mx) (D. Moctezuma);

[victor\\_m@cimat.mx](mailto:victor_m@cimat.mx) (V. Muñoz-Sánchez)

🌐 <https://github.com/Enriquesec> (E. Santibáñez-Cortés); <https://github.com/AzaelCC> (A. Carrillo-Cabrera)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

The type of extracted user's information can be classified as static data (e.g. native country) or dynamic data (e.g. preferences or needs) [3]. Traits such as gender, language, age, location info, bot or not, are some of the most detected in the user profiling task. The user profiling task can be tackled using different types of data, such as tweets, and behavioral usage patterns (e.g. number of clicks, typing velocity, etc.). Using Social Media or data from the Internet are resources often employed in the community. The potential applications and usage of user profiling are vast. Some examples of them are: profiling anomaly detection [4], cyber security [5], specific topic reviews [6], user's satisfaction measurement [7], and recommendation systems [8], among others.

Important advances and analysis on this topic have been done, for instance in [9] is presented a review of the user profiling task, its advances, challenges, and solutions. Here, Ifeanyi et al. describe a user profiling taxonomy where the task is grouped into several types, such as dynamic or static profiling, these could be also done by looking for the behavioral, interest, or intention traits.

The methods used to tackle this task are very diverse, for instance, supervised or non-supervised machine learning algorithms, ontology-based methods, statistical models, or even a hybrid methodology using some of the prior ones. One of the first attempts to tackle user profiling is described in [10] where a Lifestyle finder method is proposed, here the demographic information is used to generalize user-specified data related to the population. The experimental website was launched in 1996 generating a database of more than 4,000 users, at the time of the paper's publication in 1997. Since then, the methods have evolved substantially. In more current works, like [11], is proposed a dynamic user profiling on Twitter method using word embeddings to describe the user's content over time. This profiling was done based on the user's interests. A multi-source user profiling method applied to recommendation systems is proposed in [12], where the data sources include personal history, explicit preferences, and social activities like comments, and shares. As a result, is generated a continuous updating profiling improved the accuracy of the personalized recommendation system.

In this paper, the problem of Twitter user profiling is tackled, in two aspects, gender and profession identification, and also political ideology detection.

The manuscript is organized as follows, Section 2 describes the preprocessing that was performed on the data to be used. Section 3 describes the proposed methodology for both problems of gender and profession and ideology identification. Section 4 shows the obtained results and finally, Section 5 draws our main findings and conclusions.

## 2. Dataset

The dataset is conformed by 37,560 tweets, each of them is related to one person tagged with information related to gender, profession, and ideology (binary or multiclass options). The dataset contains information from 300 different users, each of them with 120 tweets. Analyzing the dataset, we observed there is no unbalanced data according to the classes distribution (*gender* and *ideology* binary), nevertheless, in classes *profession* and *ideology* multiclass are clearly unbalanced data, as in shown in Table 2 [13, 14].

Again, through the dataset review, we conclude that a pre-processing step is mandatory. For

**Table 1**  
Dataset classes distribution

| Class               | Label                 | Percentage |
|---------------------|-----------------------|------------|
| gender              | male-female           | 0.56-0.44% |
| profession          | politician-journalist | 0.80-0.20% |
| ideology binary     | left-right            | 0.56-0.43% |
| ideology multiclass | left-moderate left,   | 0.24-0.32, |
|                     | moderate right-right  | 0.29-0.13% |

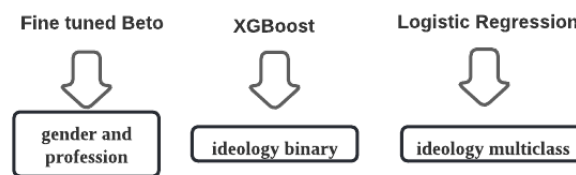
instance, in the following example of a tweet could be noted some grammatical errors, the usage of emojis, words abbreviations, and some unusual characters.

**@user 1, male, journalist, right, moderate\_right.** *check-mark-button Derrumbe de la estrategia judicial creada contra el gobierno por el 8M. check-mark-button Feijoo tritura el argumentario del [POLITICAL\_PARTY] sobre la gestión residencias . check-mark-button Unión Europea y FMI alaban al gobierno por Ingreso Mínimo Vital. Vaya fin d semana lleva ....*

The contest organizers used the special characters or tag `[POLITICAL_PARTY]`, and `@user` to refer to one political party and one Twitter user. This was done as to not have a way to get this additional information for the system proposal because this could change the main objective of the task, and the task could be somehow trivial.

### 3. Methodology

The methodology we proposed consists on the specialization of some learning algorithms to address particular subtasks related to author profiling on twitter. To this end, we perform specific fine-tuning to obtain optimal hyperparameters for the following models: BERT [15], particularly, we used the spanish pre-trained model BETO [16], logistic regression and XGBoost [17]. In Figure 1 we show a summary of the learning algorithms we used and the specific subtasks they solve.



**Figure 1:** Summary models.

In the following sections, we explain the methodology in detail.

### 3.1. Data preprocessing

On Section 2, we emphasize the need of doing some preprocessing on the dataset to deal with all the issues we observe on our data from tweets. In our case, we did some traditional pre-processing to the texts, such as lower case text transformation, punctuation cleaning, removal of unusual characters, removal of repeated words, grammatical errors correction, emojis replacement, and Spanish translation. Furthermore, lists of punctuation symbols, unusual characters, and repeated words, were done based on a sample of the dataset.

In Table 2, we show an example of the pre-processing step we perform on the corpus.

**Table 2**

Example of how we did the replacement of some words of symbols

| Original text     | Transformation result |
|-------------------|-----------------------|
| @user @user @user | @user                 |
| x                 | por                   |
| q                 | que                   |
| covid19           | coronavirus           |
| S.M.              | su majestad           |
| R. de             | Región de             |

For emoji replacement, we counted the number each emoji appearances, then we selected the top 80% of the most used emojis. Subsequently, each emoji was replaced by a textual representation of their meaning. The least frequent emojis (the remaining 20%) were replaced with the tag  $[emoji\_k]$ .

### 3.2. Models

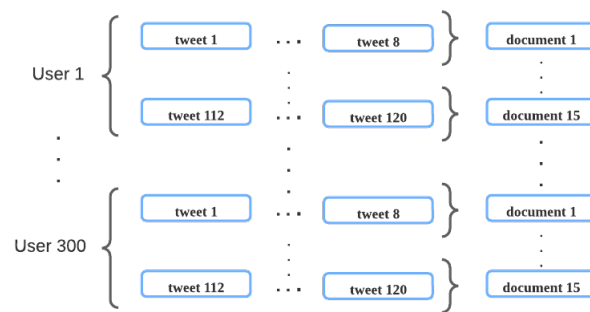
As baseline models, we used those proposed by the competition organizers, which consists on a bag of words (BOW) representation and a logistic regression model as the classifier for all prediction problems, i.e. gender, profession and ideology.

For all classification models we used, we obtained the optimal one by using a cross-validation grid-search approach defined over the main hyper-parameters for each model. For logistic regression, the fine-tuned parameters were the type of regularization ( $l1$ ,  $l2$ ) and the regularization parameter  $C \in [-3, 3]$  with stepsize of 0.1. For XGBoost, the fine-tuned parameters were the maximum depth  $max\_depth \in [1, 10]$  with stepsize 1, and the number of boosting iterations  $n\_estimators \in [50, 100, 150, 200]$ . These models gave us good results for the political ideology subtask, but that was not the case for gender and profession subtasks. For this reason, we pursued another approach.

For the gender and profession sub-tasks our proposal is based on a fine-tuned BETO model with a bagging scheme. A tweet is, by design, a very small piece of text. In fact, it is capped at 280 characters. We found that on the provided dataset the word count was around 42 words per tweet. There have been successful attempts at modeling tweets as single documents in Spanish, such as Robertuito [18], which is a RoBERTa model [19] trained on 500 million tweets. While this seemed promising at first, we found that the performance for this problem was poor. This may be due to the specific political context of the tweets from our dataset, which might not be

covered in the training corpus of Robertuito. A BETO base model has the clear advantage that it was pretrained on big corpus with a great diversity of topics and contexts, so, it can deal in a better way with other contexts. Another thing to consider is the formality of the language. BETO is pre-trained in a variety of texts, ranging from very informal, to very formal like wiki texts.

It was common for us to find tweets that did not contain enough information for us as humans to classify the user by their gender or their profession. This lead us to believe that we somehow needed to study many tweets at once in order to enrich the information about the user. Ideally, it would be great to have a user-level embedding, capable to represent and summarize the information from all 120 tweets from each user. However, it is impossible for BERT to process 120 tweets per user due to maximum sequence length constraint. To circumvent this issue, we propose to create documents composed of 12 tweets per user for the gender subtask, and 8 for the profession subtask, as is shown in Figure 2.



**Figure 2:** Transformation of tweets for the BETO model.

Pasting and dividing the dataset in this fashion, brought us important advantages. First, we were able to encode longer documents from the users, allowing to obtain useful contextual embeddings with the BETO model. Besides, even when we reduced the original 120 documents (tweets) per user to only 8 or 12, it was good enough to the fine-tuning process of BETO. Different sizes were tested to create the documents (8-12), but the ones mentioned above were the ones we obtained the best results.

For prediction, we used a Bagging scheme [20], meaning that we obtain a label prediction for each one of these 8 or 12 documents per user with the fine-tuned BETO model, then, the mean was taken and rounded to get the final prediction for each user.

### 3.3. Performance metric

To approximate the generalization of our approaches, we considered a 80% split for training and 20% for testing. Considering that there are unbalanced classes, the best metric where performance is correctly reflected is the Macro-F1. Since there are 4 classes to predict, we consider the average of the Macro-F1 metric for each of the classes to evaluate the overall performance on the task, defined in Equation (1).

$$\text{Macro-F1}_{\text{PolitiES}} = \frac{1}{4} \left( \sum_{\text{class} \in C} \text{Macro-F1}_{\text{class}} \right) \quad (1)$$

## 4. Results

The results we obtained with our proposal in the test dataset are shown in Table 4, including the baseline. Based on those results, we decided to use the fine-tuned BETO model as was described in Section 3.2 for the gender and profession subtasks, the XGBoost model for the binary ideology subtask, and the logistic regression model for the multiclass ideology classification.

**Table 3**

Results on Macro-F1 score for the baseline model, Logistic Regression, XGBoost, and BETO model evaluated in the test dataset.

| Class               | Baseline | LR    | XGB   | Beto  |
|---------------------|----------|-------|-------|-------|
| gender              | 0.576    | 0.693 | 0.793 | 0.796 |
| profession          | 0.432    | 0.881 | 0.805 | 0.860 |
| ideology binary     | 0.592    | 0.919 | 0.950 | 0.901 |
| ideology multiclass | 0.411    | 0.790 | 0.700 | -     |

With this proposal, which includes the data pre-processing we described in Section 3.1, **we secured the 4th place** on the contest Table 4. Furthermore, 2nd place on the gender and multiclass ideology subtasks was obtained.

**Table 4**

Final results

| Team Name              | Average Macro-F1 | Macro-F1 Gender | Macro-F1 Profession | Macro-F1 Ideology Binary | Macro-F1 Ideology Multiclass |
|------------------------|------------------|-----------------|---------------------|--------------------------|------------------------------|
| 1er. LosCalis          | 0.902            | 0.902           | 0.944               | 0.961                    | 0.800                        |
| 2do. NLP-CIMAT-GTO     | 0.890            | 0.784           | 0.921               | 0.961                    | 0.896                        |
| 3er. AlejandroMosquera | 0.889            | 0.826           | 0.933               | 0.951                    | 0.845                        |
| <b>4to. CIMAT_2021</b> | <b>0.879</b>     | <b>0.836</b>    | <b>0.895</b>        | <b>0.941</b>             | <b>0.845</b>                 |

## 5. Conclusions

This work proposes a solution to the PoliticEs authorship profiling problem, which consists in determining the gender, profession, and political ideology of a user from a sample of their tweets. Specifically, we used a fine tuned BETO model to predict gender and profession, an XGBoost model for binary ideology, and Logistic Regression for multiclass ideology. With our final approach we obtained the 4th place in the PoliticEs contest. It should be noted that this methodology was built from testing different approaches which include ordinal classification

methods, extract the sentiment of the tweet and use it as feature, among some other techniques that did not give better results than those presented in Section 3.

An idea to improve the previous analysis lies in rethinking the use of the BETO model for the binary ideology class, since we realized that it had very low yields compared to the rest of the models.

## References

- [1] S. Kanoje, S. Girase, D. Mukhopadhyay, User profiling trends, techniques and applications, *ArXiv abs/1503.07474* (2015).
- [2] M. Á. Á. Carmona, E. Villatoro-Tello, M. M. y Gómez, L. V. Pineda, Author profiling in social media with multimodal information, *Computación y Sistemas* 24 (2020).
- [3] A. Cufoglu, User profiling-a short review, *International Journal of Computer Applications* 108 (2014).
- [4] M. Chen, A. A. Ghorbani, et al., A survey on user profiling model for anomaly detection in cyberspace, *Journal of Cyber Security and Mobility* 8 (2019) 75–112.
- [5] R. Basti, S. Jamoussi, A. Charfi, A. B. Hamadou, Arabic twitter user profiling: Application to cyber-security., in: *WEBIST*, 2019, pp. 110–117.
- [6] E. Tutubalina, S. Nikolenko, Exploring convolutional neural networks and topic models for user profiling from drug reviews, *Multimedia Tools and Applications* 77 (2018) 4791–4809.
- [7] A. M. Flores, M. C. Pavan, I. Paraboni, User profiling and satisfaction inference in public information access services, *Journal of Intelligent Information Systems* 58 (2022) 67–89.
- [8] F. Z. Benkaddour, N. Taghezout, F. Z. Kaddour-Ahmed, I.-A. Hammadi, An adapted approach for user profiling in a recommendation system: Application to industrial diagnosis., *Int. J. Interact. Multim. Artif. Intell.* 5 (2018) 118–130.
- [9] C. I. Eke, A. A. Norman, L. Shuib, H. F. Nweke, A survey of user profiling: State-of-the-art, challenges, and solutions, *IEEE Access* 7 (2019) 144907–144924.
- [10] B. Krulwich, Lifestyle finder: Intelligent user profiling using large-scale demographic data, *AI magazine* 18 (1997) 37–37.
- [11] S. Liang, X. Zhang, Z. Ren, E. Kanoulas, Dynamic embeddings for user profiling in twitter, in: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1764–1773.
- [12] B. Veloso, F. Leal, B. Malheiro, Personalised combination of multi-source data for user profiling, in: *Proceedings of International Conference on Information Technology and Applications*, Springer, 2022, pp. 707–717.
- [13] 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).
- [14] 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.
- [15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

- [16] 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.
- [17] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794. URL: <https://doi.org/10.1145/2939672.2939785>. doi:10.1145/2939672.2939785.
- [18] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, arXiv preprint arXiv:2111.09453 (2021).
- [19] 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 abs/1907.11692 (2019).
- [20] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140. URL: <https://doi.org/10.1023/A:1018054314350>. doi:10.1023/A:1018054314350.