Dataverse at PoliticES-IberLEF2023

Pablo Emilio López-Ávila[1,*], Andrea Bethsabe García-Gutiérrez[1],
Pedro Adair Gallegos-Ávila [1], Ramón Aranda[2,3] and Miguel Ángel Álvarez-Carmona[1,3]

[1] *Centro de Investigación en Matemáticas (CIMAT), Sede Monterrey, Nuevo León, Mexico, 66629*
[2] *Centro de Investigación en Matemáticas (CIMAT), Sede Mérida, Yucatán, Mexico, 97302*
[3] *Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT), CDMX, Mexico, 03940*

Abstract

This article presents a proposal for author attribute profiling, specifically focusing on gender, profession, and political ideology (both binary and multiclass), using a BETO (transformer architecture for Spanish tasks) model. The model is applied to analyze tweets associated with Spanish authors. Originally designed and utilized in the authors' thesis project for document classification, the main objective is to assess the generalization capability of the proposed architecture across various document classification tasks, outperforming random attribute assignment.

Keywords

Author profilling, Gender, Profession, Political Ideology, Transformers, BETO, Spanish

1. Introduction

Over the past years, there has been a shift in information collection techniques, with a move away from relying solely on *telephone surveys* during election times, and a greater emphasis on utilizing social networks [1]. Social networks, especially platforms like Twitter, have become the primary spaces where individuals share their opinions, engage in debates, and passionately defend their viewpoints and ideals. In general, the task of determining important aspects of an author through his/her texts on social networks is known as author profiling [2, 3].

This article provides a detailed account of how the authors' proposal was implemented in their master's thesis project and evaluates its performance in this specific context. The classifier utilized in this project is based on the Transformers architecture, specifically a variant known as BETO [4]. Through fine-tuning [5], the model's weights are adjusted to suit the task proposed at [6, 7]. The classification layers function as linear projections, enabling the expansion of the range of considered characteristics and subsequently contracting them to the number of classes relevant to each attribute. The proposal achieved an average $F1$ score of 0.666 across all the considered attributes, with the poorest performance observed in the case of the *multiclass*

IberLEF 2023, September 2023, Jaén, Spain

pablo.lopez@cimat.mx (P. E. López-Ávila); andrea.garcia@cimat.mx (A. B. García-Gutiérrez); pedro.gallegos@cimat.mx (P. A. G.); arac@cimat.mx (R. Aranda); miguel.alvarez@cimat.mx (M. : Álvarez-Carmona)

0009-0006-8214-5693 (P. E. López-Ávila); 0009-0003-6264-3589 (A. B. García-Gutiérrez); 0009-0004-9412-3980 (P. A. G.); 0000-0001-8269-3944 (R. Aranda); 0000-0003-4421-5575 (M. : Álvarez-Carmona)

CEUR Workshop Proceedings (CEUR-WS.org)

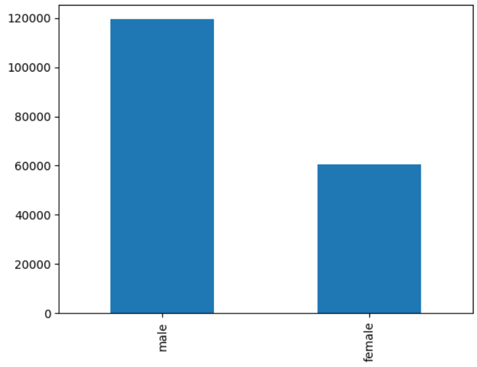

Figure 1: Gender frequency on trainin dataset

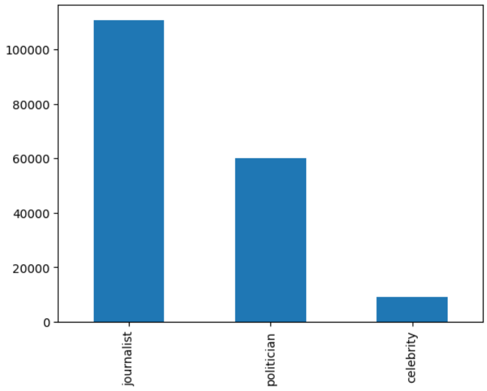

Figure 2: Profession frequency on training data set

ideology attribute. Further details regarding this will be discussed in subsequent sections. The article includes sections dedicated to Dataset Description, Methods, Results, and Conclusions.

2. Dataset description

The dataset provided in this study [8, 7] is an extension of the previous PoliticEs 2022 task. It encompasses observations gathered from Twitter profiles in Spain between 2020 and 2022. These profiles include individuals from various categories, such as politicians (government members, congress members, mayors, former politicians, and collaborators), political journalists (representing media outlets like ABC, El País, ElDiario, and El Mundo), as well as celebrities. The selection of these individuals was based on their political ideology, which could be inferred from factors such as their political party affiliation, editorial stance, or public support for a specific political party. For each user, the dataset provides the following information:

- Gender: Male or Female (Binary class)
- Profession: Politic, Journalist or Celebrity (Multiclass)
- Binary Ideology: Left or Right
- Multiclass Idelogy: Left, Moderate Left, Moderate Right or Right

2.1. Classes Distributions

The name given to these attributes is **Non ideological**, this makes reference to elements that describes the user answering the question ¿Who is posting? (e.g A female journalist)

Figure 1 shows the **Gender** on training data set seems to introduce a 67-33 ratio that's imbalanced but it's too early to conclude that **Female gender** will be problematic. Figure 2 shows information about **profession** attribute it seems to introduce a 60-30-10 ratio, in this case the **celebrity profession** might be often misclassified for any majority class. There are different approaches in order to deal with very **Imbalanced classes**, the most frequent are

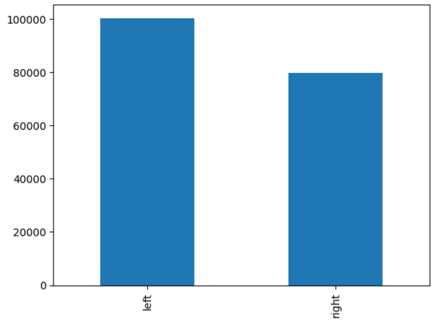

Figure 3: Binary ideology class distribution

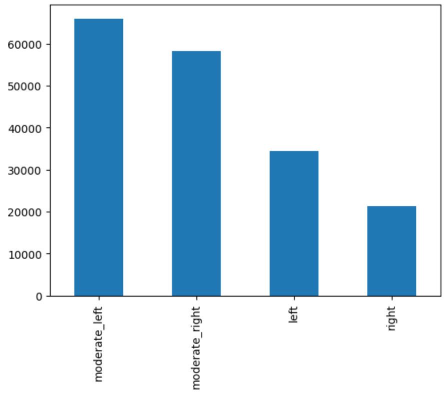

Figure 4: Multiclass ideology distribution

Subsampling and Oversampling techniques, these approaches could be a bit aggressive so we first tried fitting models using the original data set.

We called these **Political attributes** because they're related to **political ideology** for both cases, binary and multiclass.

For the **Binary** the training data set seem to introduce a $60 - 40$ ratio having the left class as the majority class (see Figure 3), in practice there's no problem dealing with these proportions at the model fitting. The **multiclass** problem introduces a $35 - 30 - 15 - 10$ ratio ((see Figure 4)), having the **moderate classes** as the majority classes and **right** as the minority, these proportions are indeed imbalanced but as the **Profession** attribute, we'll wait until face heavy misclassification problem.

2.2. Wordclouds

Most frequent words on binary problem
If a word as *Gobierno* on the following **wordclouds** is bigger than any other word as *Madrid*, we can tell that *Gobierno* has a lot more appearances than *Madrid*, We also noticed that the most frequent words on any class are the same such as **Gobierno, España, Madrid** words than will not help us to keep the classes away. A finding of interest is that the respective **Opposing ideology** appears as one of the most frequent word.

Most frequent words on Multiclass problem
We can see a similar set of most frequent words as **Gobierno, España, Madrid**, for **Moderate Left** (Figure 7) words such as **Periodistas, Mujeres, Justicia and Asturias** and finally **Moderate Right** (Figure 8) we have words such as **Impuesto, Presupuesto, Violencia, an Familia**. These key words can be taken as representation of the class, but the real challenge will be how to make distinction from **moderate classes** and **non moderate classes**

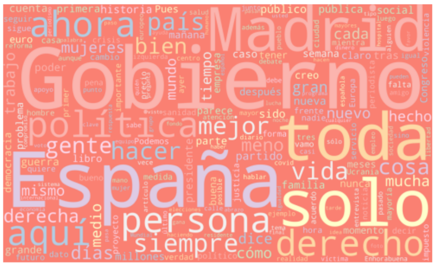

Figure 5: Left Ideology Wordcloud

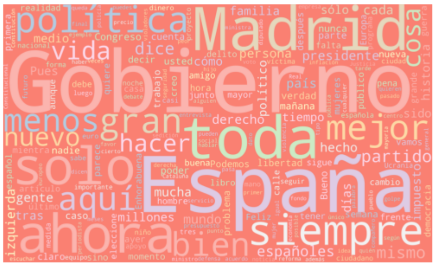

Figure 6: Right Ideology Wordcloud

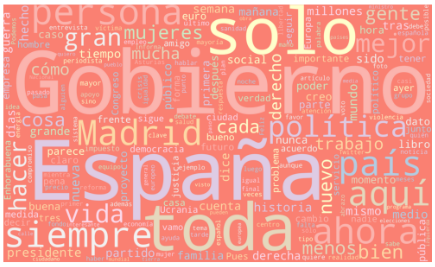

Figure 7: Moderate Left Ideology Wordcloud

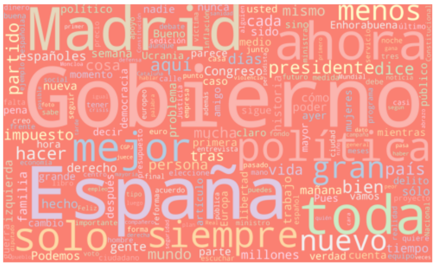

Figure 8: Moderate Right Ideology Wordcloud

3. Methods

The proposed solution starts preprocessing the content of the tweets, this includes:

- Convert all words in the tweet to lowercase.
- Remove emojis and non numerical elements.

We tried to preserve as much information from the original documents as we could, after this We proceeded to use a **Fine tuned** BETO [4], BETO is a BERT [9] model trained on the Spanish Unannotated Corpora (SUC) that has been used on different NLP tasks.
Begins taking *preprocessed tweets* to a 512 dimensional **Token representation** using **Word Piece Tokenization** this is the **BERT Tokenizer** [10], thus:

$$tweet_k = (CLS, token_1, \cdots, token_{k_{last}})$$

Evaluating all these token representation on the BERT model, produces an output of shape:

$$Num\_Sequences \times Num\_Tokens \times 768$$

In this case for Classification task, we must extract the information related to the **CLS** [9, 11] token this generates a 768 dimensional vector representation for every tweet that contains their semantic information. Finally we used the information extracted from the CLS in a Classifier based on Linear Layers (see Figure 9):

- First layer projects 768 element vectors into a 382 dimensional space

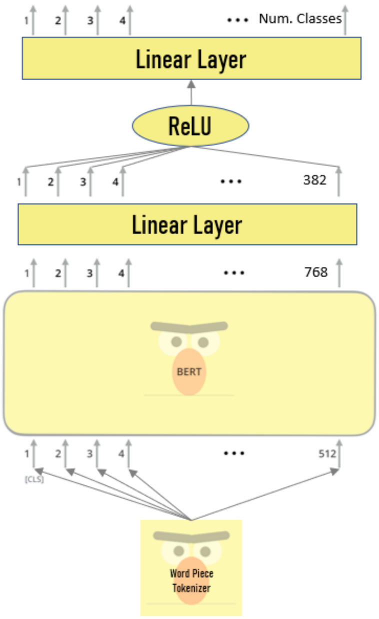

Figure 9: Proposed BETO based architecture for Political Ideology classification (Binary and Multiclass)

- A ReLU activation function between linear layers
- Last layer projects a 382 representation into a **Number of classes Space**

Training details:

- 3 epochs
- Batch size of 32 elements
- Adam Optimizer using a learning rate of $5e^{-5}$
- Unfreezed BETO weights

4. Results

In this competition [7] the proposal's performance is measured using the macro $F1$ score per trait on every class, after this an arithmetic mean of the macro F1 scores is computed, this is the final score to rank team's performance, in this time the following are our best results:

Table 1 shows the results obtanied by our methodology for the different scenarios. It is evident that classes such as Gender and Binary Ideology exhibit superior performance compared to random class assignments. Despite having three categories, the Profession class achieved remarkable results, nearly on par with binary classification tasks. However, the Multiclass Ideology category demonstrated the lowest performance, resulting in a decline in the overall average F1 score.

Table 1
Results of the proposal for the different cases.

Results	
Classification task	F1 Score
Gender	0.724442
Profession	0.731998
Ideology binary	0.784927
Ideology multiclass	0.425304
Average Score	0.666668

5. Conclusions

In conclusion, this proposal aimed to achieve high-quality BERT representations by employing different tokenization and preprocessing techniques. Overall, the approach yielded satisfactory results for most classification problems, with the exception of Multiclass Ideology. It is possible that the limited performance in this category can be attributed to the distribution of the training dataset. Insufficient examples of the targeted class may have hindered the model's ability to classify instances accurately. Additionally, it is plausible that the 'untag' test dataset contained a significant number of examples belonging to the minority classes, which could have impacted the model's performance.

To address this issue in future work, we propose employing an approach that involves generating artificial examples using class-exclusive words. This strategy aims to mitigate the imbalanced data problem and improve the performance of minority classes.

Another potential factor influencing performance is the length of the merged user tweet, which may exceed the maximum token limit allowed by BERT (512 tokens). This limitation could have affected the model's capture of all relevant information from longer tweets. In summary, while the proposal achieved satisfactory results overall, there is room for improvement, particularly in addressing imbalanced data and considering token limitations. Future research should focus on implementing strategies such as generating artificial examples and exploring techniques to handle longer tweets effectively. The preliminary analysis reveals that a majority of the merged tweets contain between 2000 and 2500 words. This poses a challenge since BERT is unable to handle such lengthy sequence lengths [9]. To address this issue in future endeavors, a potential solution could be implementing a **Vote System**, as suggested in [12]. This approach would involve classifying each individual tweet within a user's collection and assigning the most frequent (or one of the most frequent) classes found across their tweets. Furthermore, regarding the proposed **Linear classifier**, there is potential for improvement by exploring more sophisticated architectures, such as incorporating LSTM layers. Testing these more complex structures will help determine whether they outperform traditional approaches like Linear Layers. Considering alternative architectures can enhance classification performance and provide valuable insights for future investigations.

References

[1] A. Rao, N. Spasojevic, Actionable and political text classification using word embeddings and lstm, 2016. arXiv:1607.02501.

[2] M. Á. Álvarez Carmona, E. Villatoro Tello, M. Montes y Gómez, L. Vilaseñor Pineda, Author profiling in social media with multimodal information, Computación y Sistemas 24 (2020) 1289–1304.

[3] M. Á. Álvarez-Carmona, E. Villatoro-Tello, L. Villaseñor-Pineda, M. Montes-y Gómez, Classifying the social media author profile through a multimodal representation, in: Intelligent Technologies: Concepts, Applications, and Future Directions, Springer, 2022, pp. 57–81.

[4] 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.

[5] M. A. Alvarez-Carmona, R. Aranda, A. Rodriguez-Gonzalez, D. Fajardo-Delgado, M. G. A. Sanchez, H. Perez-Espinosa, J. Martinez-Miranda, R. Guerrero-Rodriguez, L. Bustio-Martinez, A. D. Pacheco, Natural language processing applied to tourism research: A systematic review and future research directions, Journal of King Saud University-Computer and Information Sciences (2022).

[6] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.

[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 at IberLEF 2023: Political ideology detection in Spanish texts, Procesamiento del Lenguaje Natural 71 (2023).

[8] 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.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[10] X. Song, A. Salcianu, Y. Song, D. Dopson, D. Zhou, Fast wordpiece tokenization, 2021. arXiv:2012.15524.

[11] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, H. Carlos, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, Journal of Information Science (2022) 01655515221100952.

[12] S. S. Carrasco, R. C. Rosillo, Loscalis at politices 2022: Political author profiling using beto and maria, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruna, Spain, 2022.