THANGCIC at PoliticEs 2022: Term-based BERT for Extracting Political Ideology from Spanish Author Profiling

Hoang Thang Ta[1,2], Abu Bakar Siddiqur Rahman[1,3], Lotfollah Najjar[3,*] and Alexander Gelbukh[1]

[1]*Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico*
[2]*Dalat University, Lam Dong, Vietnam*
[3]*College of Information Science and Technology, University of Nebraska Omaha, Omaha, Nebraska, USA*

Abstract

This paper presents our participation in the task of detecting gender, profession, and political ideology in tweets of Spanish users, in a binary and multi-class perspective. The task plays an important role in identifying political ideology of parties and politicians, especially new emerging ones. This may support relevant tasks to make predictions in the elections, or create an impact on the decision of citizens through out propagation systems. For each user, we extracted features as the most popular terms from a bunch of his/her tweets, then put them as input data for the training, which applied a transfer learning set up on pre-trained BERT models. Our quick method should be suggested as a baseline for the task with the highest F1 average macro of 72.72%. In detail, we obtained F1 Gender of 69.14%, F1 Profession of 81.47%, F1 Ideology Binary of 75.76%, and F1 Ideology Multiclass of 64.51%.

Keywords

Political Ideology, Author Profiling, BERT, Text Classification, IberLEF, SEPLN

1. Introduction

Political ideologies encapsulate a set of ethical ideas about how a country should be ruled by its government to meet specific goals. Human psychology is the main source to make decisions intentionally or unintentionally for most of the cases in our societal life. It is grown by interrelating sense from birth based on race, culture, gender, religion, nationality, specific set up rules by government to a country and geographical effects. Unspoken issues can have a significant impact to make next effective decisions about the political ideologies based on psychology in the political domain by author profiling [1]. Not only political domain but also in most online services, author or user profiling nature can be identified by the users' posts in social media such as Twitter [2, 3]. Customers' opinions and feelings for a product can be predicted

IberLEF 2022, September 2022, A Coruña, Spain.

*Corresponding author.
tahoangthang@gmail.com (H. T. Ta); abubakarsiddiqurra@unomaha.edu (A. B. S. Rahman); lnajjar@unomaha.edu (L. Najjar); gelbukh@cic.ipn.mx (A. Gelbukh)
0000-0003-0321-5106 (H. T. Ta); 0000-0002-8581-0891 (A. B. S. Rahman); 0000-0003-3960-4189 (L. Najjar); 0000-0001-7845-9039 (A. Gelbukh)

based on the psychological data from online computational advertising [4]. Online delivery of items provides a new feature based on the psychology of customer reviews. Similarly, left and right wing political ideologies can be detected from posts of the politicians and journalists on social media to reveal how people and country treated by the particular community. In which, left wing is related mostly with freedom, rights, equality of individuals with liberal beliefs, whereas right wing contains the power of the government with conservative beliefs [5]. Motivating by these, political ideology is identified by Natural Language Processing (NLP) researchers by deep learning methods from social media data.

The data can be collected from posts of journalists or politicians to analyze psychological traits for detecting political ideology. However, it is noted that if the data is collected from the citizens who posts related to political issues should be anonymized due to privacy issues and not repeating the case of collecting millions of psychological profiles from Facebook during United States (U.S.) election 2016 [6]. The citizen posts can only be used to detect political ideology to obtain information what ideology belongs to based on psychological traits by authors. Author profiling used extensively to detect fake news [7] and gender identification [8] from social media. However, there is a few works have done by author profiling to detect political ideology. García-Díaz et. al. have worked on author profiling to identify psychological traits and demographic information based on political ideology from Spanish politicians tweets [9].

In this paper, we engage the Politic 2022 challenge [10], which requires participants to identify gender, profession, and political ideology in tweets of Spanish users. We extracted the most popular terms as noun phrases ranked by their raw frequencies from tweets to use as an input for pre-trained BERT models. Gender, profession and political ideology are extracted and classified from the tweets of Spanish users by author profiling task. Politician and journalist are the profession that detect from the tweets and political ideology are in both binary and multi class classification by left wing, right wing and left, right, moderate left, moderate right respectively.

2. Related Works

Political ideology is not related by casual relationship with personality traits, rather there exists a correlation relationship between these two terms. Political attitudes can be developed eventually in a human life similar to genetic factors that develop a human personal traits later in life. Political ideology can be considered as psychographic trait that provide information to realize individual and social behavior, inherent attitudes of the people in the society with moral and ethical values, appraisals, biases, and prejudices [11]. Personal traits weigh an essential part in the study of political behavior. In another research, it showed how big five personality traits assist to understand the reflection of inner unspoken behavior in a political environment. Extraversion (energetic), agreeableness (trust and kindness), conscientiousness (thoughtfulness), emotional stability or neuroticism (sadness or threats to someone life), openness (imagination and insight with positive attitude) are the big five traits of personality where openness and conscientiousness are more likely to support liberal social policies, agreeableness associated with economic liberalism with social conservatism [12]. The relationship between personality and political ideology differ from country to country in addition with the variation of continents. However, a relationship was investigated by a correlation between big five personality traits

and political ideology by collecting data from 21 countries of all continents that covered a wide variety of behavioral nature from a variety of persons in different parts of the world. The results found from the data that conscientiousness was strongly correlated with the right wing, whereas openness to experience and agreeability were notably more correlated to the left wing [13]. Political ideology and psychological traits impact on own personal life related to health. An internet survey was experimented in the United States (U.S.) to investigate socio-political characteristics to evaluate the attitude of citizens about vaccination where the results found a correlation between the attitude and making decisions about vaccines based on political ideology [14].

Author profiling used mainly to detect age, gender, nationality, psychological behavior based on the writing styles and writing texts [15]. When a user posts for a certain period on social media, then the users post can be used for social media analysis. Rangel et. al. used the emotions of authors for author profiling by using graph analysis to identify age and gender [16]. A tweet posts by a bot or human identified by author profiling shared task on PAN 2019 [17]. Exactly 56 participants were attended for the task where the participants use logistic regression, multilayer perceptron, convolution neural network (CNN), recurrent neural network (RNN), LSTM models. TOEFL, ICLE exam dataset were collected to identify the nationality and ethnicity of the authors by author profiling tasks where traditional TF-IDF vectorizer was used to feed into machine learning classifiers [18]. Markov et al. used punctuation from the writing styles to detect the nationality of the authors where it described that native language speakers used different punctuation styles compared to non native speakers [19]. Cross domain author profiling tasks are arduous due to not having available dataset. If a model built from an origin domain with Facebook dataset and then attempt to classify text in different target domain with email dataset. In this case, cross domain author profiling identified age and gender with better accuracy by using multi-source BERT stack ensemble methods [20].

3. Task Description

The participants are required to extracting political ideology of users from a set of their gathered tweets. According to the declaration of organizers, this is the first Spanish shared task focused on the problem of author profiling in political ideology. The task is not only interesting in the way of political ideology identification, but also for gender and profession detection of users only through out their writing style of tweets.

The challenge contains 2 kinds of problems, binary problem (pib) and multi-class problem (pim). Depending on traits extracted from users' tweets, we have a list of various kinds of problems:

- `gender` (pib): detect users whether male or female.
- `profession` (pib): identify the profession (politician or journalist) of users.
- `ideology_binary` (pib): the political ideology of users is left (left-wing politics) or right (right-wing politics) in a binary perspective.
- `ideology_multiclass` (pim): the political ideology of users is left, moderate left (moderate-left politics), right, or moderate right (moderate-right politics) in a multiclass perspective.

Table 1
Several examples taken from the dataset.

Label	Gender	Profession	Ideo. Bin.	Ideo. Multi.	Tweets
@user1	male	politician	left	moderate_left	a list of tweets
@user10	male	journalist	right	moderate_right	a list of tweets
@user100	female	politician	left	moderate_left	a list of tweets
@user105	male	politician	left	moderate_left	a list of tweets
@user106	male	politician	right	moderate_right	a list of tweets
...	...	...	...	...	...

Ideo. Bin. = Ideology Binary, Ideo. Multi. = Ideology Multiclass

Table 1 shows some several examples in the training dataset which indicate values of gender, profession, ideology binary, ideology multiclass, and tweets of each author. In the datasets, the organizers separate a list of tweets into single tweets by lines with the repetition of author information.

For the evaluation, organizers declared to use Precision, Recall, and F1 values to measure the model performance of each trait. The macro-average values are used for the ranking by the arithmetic mean for all traits by teams.

4. Dataset Analysis

In this section, we present some information and several analyses of the dataset. From 2020 to 2021, Garcia et. al. had started to collect tweets of politicians and journalists related to the political topics, and used UMUCorpusClassifier to extract ideology political information [21]. Later on, they opened this challenge with the dataset as an extension of PoliCorpus 2020 [9].

In total, the training dataset contains 372 different users, account for 37560 tweets while the test set has 105 users and 12600 tweets. Figure 1 shows the distribution of gender, profession, ideology binary, ideology multiclass of users in the training set. Except trait "Profession", the distribution of other traits are roughly balanced. When concatenating tweets as long paragraphs by users, the maximum length can be up to over 7000 tokens. Not only that, there are also some paragraphs which have less than 100 tokens. Therefore, the idea of training on these long texts is impossible on BERT and even Longformer [22], due to the limitation of tokens (512 for BERT and 4096 for Longformer).

5. Methodology

We apply some preprocessing steps for tweets before grouping them by their users.

- Remove special characters, smileys, and symbols.
- Split `urls` which are sticky in text by a blank. In the experiment, there is no case for this step.

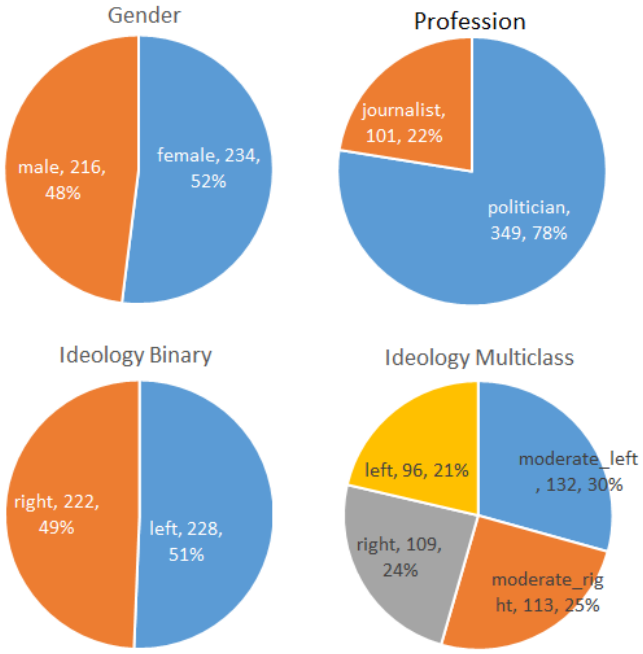

Figure 1: The distribution of gender, profession, ideology binary, and ideology multiclass by authors in the dataset.

- Normalize hashtags by removing `#`. For example, hashtags `#PGE2021`, `#CasoDina`, `#Plataforma` will be normalized as `PGE2021`, `CasoDina`, and `Plataforma`.
- Normalize `@[user]` to `@usuario`.
- Apply package `es_core_news_md` of spaCy v2.3.2 [23] to split tweets into tokens and remove redundant spaces, then combine them back as texts.

Next, for each author, we create a individual term dictionary based on his/her tweets combined as a long paragraph. Parallelly, we also create a global term dictionary for all users, which is helpful to analyze the common terms between individual users and all users. All terms in the dictionaries are lowercase for reducing the number of terms and increasing their raw frequencies. Package `es_core_news_md` is again used to extract terms as noun phrases which have a length more than 3 characters, then sort the output dictionary by frequencies.

Table 2 lists some popular terms by first 5 users in the training dataset. We can clearly see that users with different gender, profession and political ideology may have different popular terms or the orders of these terms are different. For example, the female users have a tendency of prioritizing the problems of children, feminism, or education more than male users. Meanwhile, all users also have the common terms such as `"gobierno (goverment)"`, `"madrid"`, or `"españa (spain)"` due to natural popular of these terms in the topic.

Here, we combine these terms together as texts by delimiters (spaces, commas, or [SEP] tokens) for the training process. However, the number of terms is also needed to consider. If we choose a few terms only, it is hard for the model to differentiate traits among user or capture all of their traits effectively. If we choose a large number of terms, the model can not reach to

Table 2
Top 20 popular terms of some selective users and top 20 popular terms of all users.

User information & Individual popular terms
('@user1', 'male', 'politician', 'left', 'moderate_left') ***gobierno***, madrid, españa, derecha, años, ley, congreso, virus, ayuso, movilidad salud, centros, normas, rastreadores, gestión, millones, votos, sanidad, atención, crisis
('@user10', 'male', 'journalist', 'right', 'moderate_right') ***gobierno***, ciudadanos, presidente, elecciones, sánchez, años, congreso, investidura, líder, pacto coalición, portavoz, país, mundo, rivera, puntos, senado, partido, participación, vida
('@user100', 'female', 'politician', 'left', 'moderate_left') hijos, años, suerte, españa, mujeres, centros, democracia, derecha, personas, educación ley, ayuso, madrid, justicia, titular, ***gobierno***, gente, niños, cosa, juez
('@user105', 'male', 'politician', 'left', 'moderate_left') ***gobierno***, años, ciencia, universidades, país, personas, violencia, asturias, alarma, sanidad historia, meses, planespañapuede, crisis, ley, gente, palabras, empresas, ciudadanía, portavoz
('@user106', 'male', 'politician', 'right', 'moderate_right') ***gobierno***, ciudad, años, presidente, personas, españa, placer, valencia, reto, fotos apoyo, millones, justicia, respeto, mesa, intervención, congreso, persona, ministro, euros
Global popular terms
gobierno, años, españa, personas, país, vida, ley, gente, millones, gracias presidente, libertad, apoyo, año, mujeres, crisis, congreso, medidas, mundo, sánchez

the coverage point due to the number of commons increased in all users. In the experiment, we practiced with different tests, and went to the best number of terms from 100 to 150. The removal of most popular terms by each author also helps the model learn better, but we ran of time for the challenge so not possible try this in the experiment.

6. Experiments

At first, we trained our model in tweet-level classification; however, the accuracy was about 30%-50% due to the dispersion of features that makes the model hard to identify correctly tweets to classes. The result was the same for concatenating tweets and then splitting into adjacent pieces of texts which has the number of tokens lower than the maximum value of BERT or Longformer. Finally, we thus went to the idea of extracting important terms to help the model learn better.

In the training, we did not split the dataset into training, validation, or test sets and used pretrained model `nlptown/bert-base-multilingual-uncased-sentiment`.

We also used cross-entropy loss and Adafactor optimization with `learning rate = 0.001`. Other hyperparameters are `MAX_LEN = 192`, `RANDOM_SEED = 42`, and `BATCH_SIZE = 8`. We must use a small value of `BATCH_SIZE` due to the limited resources on our server. We pick from 100 to 150 top terms for the training, so `MAX_LEN = 192` is to make sure that this length can capture all terms in the token encoding. By passing some test runs, we found that the model go best with 128 popular terms. The models were saved within 20 epochs with the highest values

Table 3
The F1 results by teams in the challenge of political ideology identification.

Team	Avg. Mac.	Gender	Profession	Ideo. Bin.	Ideo. Multi.	Rank
LosCalis	0.9022	0.9028	0.9443	0.9616	0.8002	1/20
NLP-CIMAT-GTO	0.8909	0.7848	0.9212	0.9614	0.8962	2/20
Alejandro Mosquera	0.8891	0.8267	0.9334	0.9515	0.8450	3/20
CIMAT_2021	0.8797	0.8368	0.8950	0.9416	0.8455	4/20
...	...	...	...	...	...	...
ThangCIC (ours)	**0.7272**	**0.6914**	**0.8147**	**0.7576**	**0.6451**	**14/20**
NLP_URJC	0.7219	0.6598	0.8329	0.8081	0.5867	15/20
SINAI	0.7214	0.7857	0.7539	0.7846	0.5615	16/20
UC3M-DEEPLNLP-2	0.6431	0.6938	0.4732	0.8291	0.5762	17/20
...	...	...	...	...	...	...
UMUTeam (baseline)	*0.5112*	*0.5762*	*0.4324*	*0.5956*	*0.4406*	*20/20*

Avg. Mac. = Average Macro, Ideo. Bin. = Ideology Binary, Ideo. Multi. = Ideology Multiclass

of training accuracy. There are 4 models for 4 classifiers to extract political ideology, gender, and profession of users. We train BERT models on the last hidden layer with `hidden_size=512`, and `softmax()` function is used to identify classes by the highest probability.

After submitting our test sets to organizers, we obtained the F1 average macro of 72.72%, F1 Gender of 69.14%, F1 Profession of 81.47%, F1 Ideology Binary of 75.76%, and F1 Ideology Multiclass of 64.51%. Our results and other teams are shown in Table 3. Although our method outperformed the baseline and some teams, it is still a modest result compared top teams. However, our quick method is a good way to test the task difficulty by applying transfer learning on a standard transformer model, BERT. Therefore, we suggest that our method should be counted as a baseline for the task.

7. Conclusion

In this paper, we presented our participation on extracting political ideology information from tweets. Firstly, tweets were grouped together by users, then passed through preprocessing steps, before retrieving and ranking popular terms from them for the training process. We applied a BERT pre-trained model to produce 4 models, respectively to 4 classifiers to predict political ideology information. Our result outperformed the baseline based on BoW, but it is still a big gap compared to the top teams. However, we shown that our method is a useful way to quickly check the challenge difficulty through out extracting terms in a bunch of texts and applying transfer learning from any pre-trained model, especially BERT. We suggest our method should be the baseline due to the speedy development and the popular of deep learning in recent years. In the future, we will continue to analyze meticulously the datasets to understand the correlation between traits and tweets of users. Furthermore, we will design author profiles and proper classifiers to aim to the higher performance.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20220852 and 20220859 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

[1] R. Marasco, C. Tien, Ask a Political Scientist: A Conversation with Efrén Pérez about Political Psychology and the Study of Race and Ethnic Politics, Polity 54 (2022) 181–188.

[2] D. Estival, T. Gaustad, S. B. Pham, W. Radford, B. Hutchinson, Author Profiling for English Emails 263 (2007) 272.

[3] P. Mishra, M. D. Tredici, H. Yannakoudakis, E. Shutova, Author Profiling for abuse detection. (2018) 1088–1098.

[4] S. C. M. P. S. Yun, Joseph T., J. A. . S. V. Malthouse, Edward C.and Konstan, Challenges and future directions of computational advertising measurement systems, Journal of Advertising (2020) 446–458.

[5] P. H. Hanel, N. Zarzeczna, G. Haddock, Sharing the same political ideology yet endorsing different values: Left-and right-wing political supporters are more heterogeneous than moderates, Social Psychological and Personality Science (2019) 874–882.

[6] C. Cadwalladr, E. Graham-Harrison, Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach (2018) 22.

[7] J. L. Fernández, J. Antonio López Ramírez, Approaches to the Profiling Fake News Spreaders on Twitter Task in English and Spanish (2020).

[8] P. O. Sezerer, Erhan, S. Tekir, A Turkish Dataset for Gender Identification of Twitter Users. (2019) 203–207.

[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] 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).

[11] 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.

[12] A. S. Gerber, G. A. Huber, D. Doherty, C. M. Dowling, The Big Five Personality Traits in the Political Arena (2011) 265–287.

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

[14] B. Baumgaertner, J. E. Carlisle, F. Justwan, The influence of political ideology and trust on willingness to vaccinate, PloS one 13 (2018) e0191728.
[15] F. Rlaude, R. Konow, S. Ladra, Fast compressed-based strategies for author profiling of social media texts. (2016) 1–6.
[16] F. Rangel, P. Rosso, On the impact of emotions on author profiling. (2016) 73–92.
[17] M. H. F. Siddiqui, I. Ameer, A. F. Gelbukh, G. Sidorov, Bots and Gender Profiling on Twitter., in: CLEF (Working Notes), 2019.
[18] D. N. Perera, R. Weerasinghe, R. Pushpanand, Determining the Ethno-nationality of Writers Using Written English Text. (2021).
[19] I. Markov, V. Nastase, C. Strapparava, Punctuation as native language interference (2018) 3456–3466.
[20] D. J. P. Neto, I. Paraboni, Multi-source BERT stack ensemble for cross-domain author profiling (2022).
[21] 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.
[22] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).
[23] Explosion, 2022, es_core_news_md of spaCy, URL: https://spacy.io/models/es.