Participation of ESCOM's NLP Group at PoliticES-IberLEF2023: Voting Ensemble and Basic Machine Learning Methods Applied to Political Ideology

Ana-María Acosta-Pacheco[1], Diana-Paola De-La-Cruz-Sierra[1], Hannah Gu[1], José-Manuel Suárez-Bautista[1], Jorge Hernández-Espinoza[1], Mauricio-Gael Hernández-Lom[1], Luis-Ricardo Merino-Vázquez[1] and Omar Juárez Gambino[1,*]

[1] *Instituto Politécnico Nacional (IPN) - Escuela Superior de Cómputo (ESCOM), J.D. Batiz e/ M.O. de Mendizabal s/n, Mexico City, 07738, Mexico*

Abstract

In this paper, we describe the participation of the ESCOM NLP group in the PolictES 2023 task. The competition considers four classification problems (gender, profession, binary political ideology, and multiclass political ideology). Following a supervised approach, we trained four models, one for each subtask, and used them to make predictions on the test file. Naïve Bayes, Logistic Regression, and Support Vector Machines algorithms were tunned and tested with different text representations. A voting ensemble was also used to improve the performance. Our team ranked third in the overall evaluation of the competition.

Keywords

Political Ideology, Text Classification, Machine Learning

1. Introduction

Social networks allow users to express their opinion. These opinions express users' beliefs and show psychographic traits like personality, attitude, and political ideology [1]. Political ideology refers to a citizen's political opinions and attitudes. Knowing the people's political ideologies is important because it helps their representatives interpret their demands and desires [2].

Some authors have addressed automatic political ideology identification as a Machine Learning task. In [3], a recursive neural network was used to identify the political position

IberLEF 2023, September 2023, Jaén, Spain

[*] Corresponding author.

aacostap1501@alumno.ipn.mx (A. Acosta-Pacheco); ddelacruz1800@alumno.ipn.mx (D. De-La-Cruz-Sierra); hgu1900@alumno.ipn.mx (H. Gu); jsuarez1804@alumno.ipn.mx (J. Suárez-Bautista); jhernandeze1706@alumno.ipn.mx (J. Hernández-Espinoza); mhernandezl1605@alumno.ipn.mx (M. Hernández-Lom); lmerinov1700@alumno.ipn.mx (L. Merino-Vázquez); jjuarezg@ipn.mx (O. J. Gambino)

at the sentence level. The experiments were performed on a dataset of Congressional debates annotated for ideological biased, obtaining a 70.2% of accuracy. Authors in [4] proposed using a deep neural network to detect political ideology in news articles. The articles were collected from five different online sources and tagged as conservative or liberal. The best model obtained a 90% of F1-Score.

Given the importance of this problem, IberLEF2023 [5] collocated the PolictES 2023 task [6] that aims to extract the political ideology from a given user's set of tweets. The task requires identifying gender, profession, binary ideology, and multiclass ideology from a group of tweets.

In this paper, we describe our participation in the PolictES task. The trained model obtained third place in the contest in the overall task evaluation showing a very competitive performance. The rest of this paper is organized as follows. Section 2 describes the task and the corpus. Section 3 describes the method we used. Section 4 explains the performed experiments and the obtained results. Section 5 shows our conclusions and future work.

2. Task and corpus description

PoliticES 2023 is a comprehensive task involving analyzing and classifying political attributes in social media data from various perspectives. It is divided into four subtasks, each contributing to a deeper understanding of the political landscape.

The first subtask focuses on gender classification, a binary categorization distinguishing between male and female users. This information can provide insights into gender representation and participation in political discussions on social media.

The second subtask concerns the profession of the users. This multiclass classification covers three possibilities: politicians, journalists, and celebrities. Understanding the professional background of users can shed light on the various perspectives and influences within the political discourse.

The third subtask revolves around political ideology. This binary classification aims to classify users into left or right ideological positions. This classification makes it possible to identify the predominant political leanings among social network users.

Finally, the ideology subtask extends the above classification by introducing additional categories that account for the middle ground within each ideology. This multiclass classification includes the left, moderate left, right, and moderate right distinctions. This classification allows for a nuanced analysis of political ideologies and provides a deeper understanding of the ideological spectrum present in social network discussions.

The contest corpus consists of several CSV files [1]. In the initial phase, we worked with a development file containing 14,400 examples, while in the final phase, we were provided with a larger training set of 180,000 instances and a test file with 43,760 instances. All experiments reported in section 4 used the larger training set because instances of the smaller one were already contained in the larger version. The corpus comprises tweets from various users. The instances are grouped by a tag forming a cluster of 80 tweets. Posts on social media platforms such as Twitter have an inherent word limit, necessitating considering concise expressions of political opinions. In addition, it is important to recognize that not all tweets in the corpus are directly related to politics. Therefore, during the preprocessing and training phases, special attention was paid to filter out noise and ensuring that the model focused on relevant politics.

3. Method

3.1. Preprocessing

1. Clustering. The first stage consisted of grouping the tweets according to their tag. The eighty tweets with the same tag were concatenated into a single tweet, which will be used in the following stages.
2. Tokenization. The text was divided into individual units called tokens, which could be words or other meaningful elements.
3. Lemmatization. Inflectional forms of words were reduced to their base form using spaCy library [7]. This process helped in normalizing the words and reducing variations.
4. Stop words and punctuation removal. Common words and punctuation marks that do not contribute significantly to the analysis were eliminated. This step helped improve the quality of the processed text by removing noise.

3.2. Text representation

The vector space model is utilized in text analysis to transform textual data into numerical vectors. Various representation approaches were used:

- Binarization. Indicates the presence or absence of words with a value of 1 or 0, respectively.
- Term frequency. Assigns the frequency of the words as representative values.
- TF-IDF. Determines the relevance of words within a document.

These techniques enable the development of effective Machine Learning models and support comprehensive NLP tasks[8]. To solve the tasks, we tried the three different text representations. In Section 4 we explain this in detail.

3.3. Machine Learning

We decided to address the four tasks as a text classification problem. For this problem, several Machine Learning algorithms have been used [9]. In our experiments, we used Naïve Bayes (NB), Logistic Regression (LR), and Support Vector Machines (SVM) algorithms. In addition, we use an ensemble method to improve the results. All the experiments were done using the Scikit-learn library [10]. In Section 4, we explained the model's training process, the parameter adjustment, and the obtained results.

4. Experiments and results

We experimented with different text representations, Machine Learning algorithms, and model parameters. For each subtask (i.e., gender, profession, binary political ideology, and multiclass political ideology), particular experiments were performed. The training file (180,000 tweets) was divided into a training set with 80% of the tweets (144,000) and 20% for the development set (36,000). In addition, 5-fold cross-validation was applied for model selection and parameter fitting in the training set. The experiments are explained in the following subsections.

4.1. Genre classification

For the genre subtask, the preprocessing explained in subsection 3.1 was applied. After several runs, the best text representation was the binarized version. No word occurrence limit was specified, so any term that appeared at least once was included. The three models mentioned in subsection 3.3 were tested, obtaining the best result with the Naïve Bayes algorithm by adjusting the smoothing parameter (alpha) to 2. Results obtained in the development set (20% of the training file) are shown in Table 1.

Table 1
Naïve Bayes results in the development set

Metric	Value
precision	0.76
recall	0.74
f1score	0.74

4.2. Profession classification

The profession subtask is a multiclass classification problem. Three possible classes must be predicted (i.e., politician, journalist, and celebrity). The distribution of the classes shows an imbalance, as celebrity represents only 5% of the total tweets, the politician class represents 33%, while the remaining 62% are tweets from the journalist class. To address this problem, we set a different weight for the classes according to their distribution for both Machine Learning methods. The *class_weight* parameter of Scikit-learn was set to "balanced" for adjusting weights inversely proportional to class frequencies in the input data. Classes with lower frequencies will have a higher weight, while those with higher frequencies will be assigned a lower weight.

After several experiments, the best text representation was term frequency. LR and SVM with a linear kernel obtained the best results. In Table 2, we show the results of the classifiers. As we can see, SVM retrieved more data, but LR had better precision. In order to improve these results, we applied a voting ensemble method. The result of the ensemble is shown in the last row of the table.

Table 2
Results for profession task in the development set

Algorithm	Metrics		
	precision	recall	fscore
LR	0.92	0.84	0.88
SVM	0.85	0.89	0.87
Voting ensemble	0.92	0.86	0.89

4.3. Binary political ideology classification

Political ideology reflects people's positions on the state's and society's functioning. For this subtask, the possibilities of left-wing and right-wing ideologies are considered. TF-IDF was selected for text representation. In the training process, SVM with linear kernel was the best classifier. In Table 3 we show the results for this task.

Table 3
Results for binary political ideology task in the development set

Metric	Value
precision	0.96
recall	0.94
f1score	0.95

4.4. Multiclass political ideology classification

This subtask is a fine-grained version of the previous one. The multiclass classification considers four categories: left, moderate left, right, and moderate right. The text representation selected was frequency. The Logistic Regression algorithm obtained the best results. In Table 4 we show the obtained results.

Table 4
Results for multiclass political ideology task in the development set

Metric	Value
precision	0.91
recall	0.91
f1score	0.91

4.5. Results of models in the test file

In Table 5 we show the best text representations, classifiers, and their hyperparameters set in the development set. The trained models were used in the test set provided for the contest's final phase. This test set consisted of 43,760 tweets. The same preprocessing steps explained in subsection 3.1 was applied to the test file. Each fitted model was used

Table 5
Best text representation and hyperparameters of the classifiers used in the development set

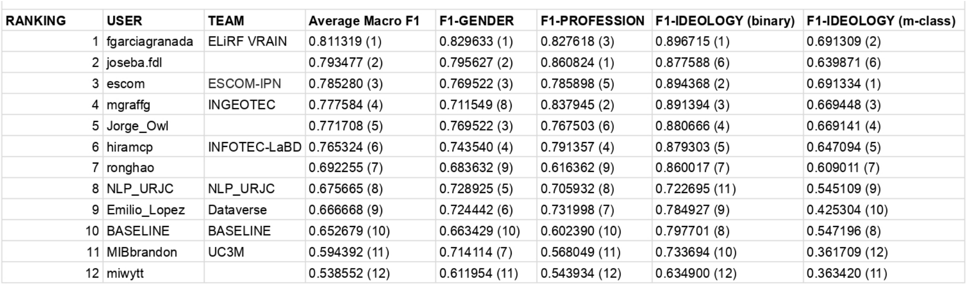
Task	Text representation	Classifier	Hyperparameters
Gender	NB	Binarized	alpha = 2
Profession	Term frequency	Voting classifier (LR + SVM)	LR: class_weight = 'balanced', C = 1.0, solver = 'lbfgs', penalty = 'l2'
			SVM: class_weight = 'balanced', C = 1.0, kernel = 'linear', max_iter = -1
Binary ideology	TF-IDF	SVM	class_weight = 'balanced', C = 1.0, kernel = 'linear', max_iter = -1
Multiclass ideology	Term frequency	LR	class_weight = None, C = 1.0, solver = 'lbfgs', penalty = 'l2'

to perform the prediction required for each subtask. Table 6 shows the performance of the models applied to the test file for each subtask, and the position achieved by our team (out of 12 teams in total) according to the results published by the contest organizers.

Table 6
Results of trained models in the test file

Subtask	f1score	Team position
Gender	0.76	3
Profession	0.78	5
Binary ideology	0.89	2
Multiclass ideology	0.69	1

Finally, we show in Figure 1 the contest participants' results. As can be seen, we obtained third place (ESCOM-IPN team), taking into account the average macro f1 score.



RANKING	USER	TEAM	Average Macro F1	F1-GENDER	F1-PROFESSION	F1-IDEOLOGY (binary)	F1-IDEOLOGY (m-class)
1	fgarciagranada	ELiRF VRAIN	0.811319 (1)	0.829633 (1)	0.827618 (3)	0.896715 (1)	0.691309 (2)
2	joseba.fdl		0.793477 (2)	0.795627 (2)	0.860824 (1)	0.877588 (6)	0.639871 (6)
3	escom	ESCOM-IPN	0.785280 (3)	0.769522 (3)	0.785898 (5)	0.894368 (2)	0.691334 (1)
4	mgraffg	INGEOTEC	0.777584 (4)	0.711549 (8)	0.837945 (2)	0.891394 (3)	0.669448 (3)
5	Jorge_Owl		0.771708 (5)	0.769522 (3)	0.767503 (6)	0.880666 (4)	0.669141 (4)
6	hiramcp	INFOTEC-LaBD	0.765324 (6)	0.743540 (4)	0.791357 (4)	0.879303 (5)	0.647094 (5)
7	ronghao		0.692255 (7)	0.683632 (9)	0.616362 (9)	0.860017 (7)	0.609011 (7)
8	NLP_URJC	NLP_URJC	0.675665 (8)	0.728925 (5)	0.705932 (8)	0.722695 (11)	0.545109 (9)
9	Emilio_Lopez	Dataverse	0.666668 (9)	0.724442 (6)	0.731998 (7)	0.784927 (9)	0.425304 (10)
10	BASELINE	BASELINE	0.652679 (10)	0.663429 (10)	0.602390 (10)	0.797701 (8)	0.547196 (8)
11	MIBbrandon	UC3M	0.594392 (11)	0.714114 (7)	0.568049 (11)	0.733694 (10)	0.361709 (12)
12	miwytt		0.538552 (12)	0.611954 (11)	0.543934 (12)	0.634900 (12)	0.363420 (11)

Figure 1: PoliticES 2023 official results

5. Conclusions and future work

Political ideology refers to a citizen's political opinions and attitudes. This task is relevant because it helps representatives interpret the people's demands and desires. In this paper, we report our participation in PoliticES 2023 task. The contest consisted of four subtasks: genre, profession, binary ideology, and multiclass ideology classification. For each subtask, different text classification and Machine Learning algorithms were tried. After several experiments, four different models were trained, one for each task. Despite using basic algorithms for classification problems (NB, LR, and SVM), the results were very competitive, obtaining third place in the overall evaluation. Even for the main subtasks of the contest, the political ideology, first place was obtained in the multiclass version and second place for the binary one.

In future work, we propose to consider more features, such as POS tags for gender classification and sentiment and emotion analysis for political ideology. We would like to explore pre-trained LLMs fine-tuned for each subtask.

Acknowledgments

This research was funded by CONACyT-SNI and Instituto Politécnico Nacional (IPN), through grants SIP-20232011 and EDI.

References

[1] 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.

[2] E. G. Carmines, N. J. D'Amico, The new look in political ideology research, Annual Review of Political Science 18 (2015) 205–216.

[3] M. Iyyer, P. Enns, J. Boyd-Graber, P. Resnik, Political ideology detection using recursive neural networks, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 1113–1122.

[4] K. M. Alzhrani, Political ideology detection of news articles using deep neural networks, Intelligent Automation & Soft Computing 33 (2022) 483–500.

[5] 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.

[6] 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).

[7] Explosio.ai, Industrial-strength natural language processing in python, 2023. URL: https://spacy.io/, accessed on june 19, 2023.

[8] D. Jurafsky, J. F. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall eBooks (2019).

[9] C.-V. García-Mendoza, O. Juárez Gambino, Web crawler and classifier for news articles, in: Advances in Computational Intelligence: 21st Mexican International Conference on Artificial Intelligence, MICAI 2022, Monterrey, Mexico, October 24–29, 2022, Proceedings, Part II, Springer, 2022, pp. 127–136.

[10] 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.