URJC-Team at PoliticEs 2022: Political Ideology Prediction using Linear Classifiers

Miguel Ángel Rodríguez-García[1,*], Soto Montalvo Herranz[1] and Raquel Martínez Unanue[3]

[1]Universidad Rey Juan Carlos, Spain
[2]Universidad Nacional de Educación a Distancia, Spain

Abstract

Different works have demonstrated the relationship between personality traits and political ideology and how they influence our daily lives. The challenge proposed in the IberLEF 2022 Task, PoliticEs, consists of extracting political ideology traits from the text by utilising Natural Language Processing (NLP) techiques. This paper describes the participation of the URJC-Team in such task. In particular, the achievement of extracting political ideology features walks through identifying the gender, the profession, and the political spectrum from a binary and multi-class perspective. In this work, we proposed two Machine Learning models to address the binary and multiclass classification problems, a Linear Support Vector Machine and Logistic Regression. The utilized dataset comprises hundreds of tweets that are cleaned and processed to generate various representations that serve as an input for the system. Between the different proposed subtasks, the proposed classification method has obtained competitive results for the binary ideology classification task, reaching 0.81. The proposal has great room for improvement, and we have planned the following steps for it.

Keywords

Political Ideology Classification, Machine Learning, Natural Language Processing, Social Media, Author Profiling,

1. Introduction

Nowadays, increasing attention exists to analyzing the link between the political preferences and behaviour of individuals [1]. Over time, several strategies have been utilized to estimate the political polarization of people, considering certain aspects such as voting behaviour, debate text, or even their speeches [2]. In fact, recent progresses in text mining have demonstrated that political ideology can be estimated from text with an elevated exactitude [3]. In this context, the arrival of Social Media and its massive popularity have provided a new incredible data source, where people publish every day their opinions, critics and comments about different topics. Such data can be utilised to successfully estimate personal information of users like

IberLEF 2022, September 2022, A Coruña, Spain.

*Corresponding author.
miguel.rodriguez@urjc.es (M. Á. Rodríguez-García); soto.montalvo@urjc.es (S. M. Herranz); raquel@lsi.uned.es (R. M. Unanue)
0000-0001-6244-6532 (M. Á. Rodríguez-García); 0000-0001-8158-7939 (S. M. Herranz); 0000-0003-1838-632X (R. M. Unanue)

personality traits, ethnicity, sexual orientation and even political orientation [4]. In particular, the challenge proposed in this shared task is devoted to extracting this political polarization from a dataset harvested from Twitter accounts of politicians and political journalists. It is required to consider the challenge as a classification problem, taking two different perspectives, binary (left-right political spectrum) and multi-class classification (includes granularity moderate). In this working note, we describe the system presented by the URJC_Team in the PoliticEs challenge [5] of IberLEF 2022. The shared task is devoted to employing NLP techniques to primary address the extracting political ideology from two different perspectives: i) binary classification, where two well-distinguished classes are utilised to label the proposed dataset, left and right; and ii) multiclass classification, where the granularity of the classification is increased with two extra classes to specify more concretely the political ideology of the extracted twitters. Furthermore, the task includes two implicit extra subtasks: gender and profession classification, defined as a binary classification problem. On the one hand, gender classification is devoted to classifying tweets by considering the male/female dimensions. On the other hand, the profession has been defined as distinguishing between tweets written by journalists or politicians. We have defined a Machine Learning (ML) based system to face the challenge. As inputs, it utilised the most distinguishable features extracted from the datasets provided by the challenge. Different tokenisers are utilised to analyse the text and create such feature distributions.

The remainder of this working note is organized as follows: Section 2 provides a review of the related work. It briefly examines the path of the strategies utilized to identify the political ideology from the initial approaches based on word collections to the current era of ML models. Section 3 specifies the distribution of the data used and the system proposed. In Section 4 the results achieved in the challenge are presented. Finally, Section 5 summarizes the conclusions obtained developing this work.

2. Related Work

Political ideology detection has started to raise interest in the last years within the NLP community [6]. The initial efforts were focused on using text representation techniques that captured the ideology variation in language. In this via, Sim et al., in [7] utilised a manually labelled corpus of various political writings whose authors are perceived as a representative of the three coarse ideologies selected: left, right or centre. The annotated corpus is utilised to automatically construct "cue lexicons", a list of terms that are strongly linked to an ideology. Following the same idea, in literature, there exist approaches based on topic model strategies. These approaches employ a statistical model to capture abstract topics from document collections. Ahmed et al. in [8] define a new topic model based on Latent Dirichlet Allocation, called mview-LDA, to construct a word distribution from a document collection. The latest approaches take advantage of the explosion of ML algorithms to address the problem. Preotiuc-Pietro et al. [9] employ a specific Linear Regression Algorithm trained with embeddings obtained from a Word2Vec model. Taking advantage of the research lines explored in those last approaches, we have utilised them as an initial hypothesis to construct the system presented for the challenge thrown.

3. Material and Methods

3.1. Data

The PoliticEs2022 dataset [10] contains tweets that have been havested from accounts of politicians and political journalists. The selection of the accounts was made by considering certain features like, political affiliation of politicians adquired from the party to which belong and political orientation of journalists guessed from their declarations on different digital channels. The tweets have been labelled considering the following topics: *label* to anonimize the user who write the tweet; *gender* to classify each tweet in terms of male/female; *profession* to disambiguate if the user is a journalist or politician; *ideology_binary* classifies the tweets in left/right ideology; and, finally, *ideology_multiclass*, which establishes a finer granularity in the ideology classification, and includes two dimension more, moderate_right and moderate_left. Table 1 shows the distribution of these labels in the training and testing datasets provided by the organizers in development and evaluation phases, respectively.

Table 1
Development and Evaluation datasets distribution.

		Development		Evaluation	
		Train	Test	Train	Test
label		5000	1000	37560	12600
gender	male	2650	800	21240	8280
	female	2350	200	16320	4320
profession	journalist	672	300	7381	3000
	politician	4328	700	30179	9600
ideology_binary	left	2750	250	21360	6840
	right	2250	750	16200	5760
ideology_multiclass	left	950	50	9120	2520
	moderate_left	1800	200	12240	4320
	moderate_right	700	250	4920	2040
	right	1550	500	11280	3720

The label row provides the total tweets that have been collected for each dataset. If we look at determining labels, it can be easily seen that there are unbalanced distributions. For instance, in the development dataset, the label "profession" has a distinguishable difference between both occupations determined. It also happens in the Evaluation dataset in the same label "profession" and the ideology multiclass, where labels "left" and "right" differ broadly.

3.2. Methods

The approaches designed have followed the same pipeline composed of three main components: i) pre-processing module, where each tweet in the dataset is analyzed for removing emojis links, hashtags, labels, interrogations and exclamations symbols. Here, we laid out the first fork, taking or not taking into account stopwords; ii) feature extractor module, where cleaned tweets are analyzed to obtain the most distinguishable features; and, finally, iii) classifier module, where the system utilizes these features to train a ML model for carrying out the classification tasks demanded.

Initially, when we started facing the challenge, we addressed the challenge from two different points of view: binary classification for all the subtasks where two classes were demanded to distinguish and multi-class classification subtask for the other remaining. However, after several tests, we decided to tackle all, considering two different supervised learning models, Linear Support-Vector Machine (SVM) and Logistic Regression (LR). Regarding the pipeline defined above, we configured two approaches with two different settings, with or without stopwords. Hence, two different pipelines dealt with the tasks demanded, four in total. In the extractor stage, two different methods were combined to create the feature distributions. We combine character and word gram models to extract sequences of 4 characters and 2 words from each tweet. Two different vectorizers were utilized, the first was trained to extract 2-word length, and the second was trained to take sequences of 4-characters length. After the training, both feature vectors were combined and utilized for training the models. After the training, both feature vectors were combined and utilized for training the models. For instance, given the following text: "Seguimos mejorando Salamanca para hacerla más accesible, cómoda y verde.", the 2-word vectorizer will extract set of words like: "accesible", "accesible cómoda" or "cómoda" and the 4-character vectorized will extract sequences like: "com", "aman" or "manc". We selected these models and extractor techniques because we found approaches in literature that they reached a good performance [11, 12, 13]. Finally, to build the models, we have utilized the scikit-learn library [14]. After testing different settings, we employed the followings: for the Linear SVM model, the regularization parameter to 1.0, the norm used in the penalization to l2 and the loss function to the square of the hinge loss; and for the LR model, we specify the inverse of regularization strength to 100, the norm of the penalty to l2 and the maximum number of iterations to 100.

4. Results

The metrics used to evaluate the performance of the system are Precision, Recall and F1-score. The organizers of the shared task rank the participant systems by the average macro F1-score. Table 2 compiles the results obtained for each classification task, taking into account the four pipelines configured and the two datasets utilized for the development and evaluation phase.

Table 2 shows the results obtained in each pipeline. The table distributes the results by considering the development and evaluation datasets of the challenging task and the two Machine Learning models trained with two different feature distributions, with or without stopwords. Regarding the development dataset, in the gender identification task, the best results were achieved by the LR without stopwords in the male and female classification task. Still,

Table 2
Results achieved by the pipelines proposed

			Development			Evaluation		
			Precision	Recall	F1-score	Precision	Recall	F1-score
SVM (Stopwords)	gender	male	0.27	0.75	0.4	0.51	0.69	0.59
		female	0.89	0.5	0.64	0.8	0.65	0.72
	pofession	journalist	0.56	0.83	0.67	0.78	0.72	0.75
		politician	0.91	0.71	0.8	0.91	0.94	0.92
	ideology	left	0.56	1	0.71	0.77	0.82	0.8
		right	1	0.73	0.85	0.77	0.71	0.74
	ideology_multiclass	left	0.5	1	0.67	0.5	0.48	0.49
		moderate_left	0.5	0.75	0.6	0.59	0.64	0.61
		moderate_right	0.75	0.6	0.67	0.72	0.58	0.64
		right	0.75	0.6	0.67	0.57	0.7	0.63
SVM (Without Stopwords)	gender	male	0.16	0.5	0.25	0.54	0.69	0.6
		female	0.75	0.37	0.5	0.81	0.69	0.75
	pofession	journalist	0.75	0.5	0.6	0.75	0.84	0.79
		politician	0.81	0.93	0.87	0.95	0.91	0.93
	ideology	left	0.5	0.8	0.61	0.8	0.84	0.82
		right	0.92	0.73	0.81	0.8	0.75	0.77
	ideology_multiclass	left	0.5	1	0.67	0.55	0.76	0.64
		moderate_left	0.4	0.5	0.44	0.63	0.72	0.67
		moderate_right	0.62	0.5	0.56	0.68	0.42	0.52
		right	0.8	0.8	0.8	0.62	0.58	0.6
LR (Stopwords)	gender	male	0.18	0.5	0.27	0.57	0.69	0.62
		female	0.78	0.44	0.56	0.82	0.72	0.77
	pofession	journalist	0.75	0.5	0.6	0.77	0.8	0.78
		politician	0.81	0.93	0.87	0.94	0.92	0.93
	ideology	left	0.44	0.8	0.57	0.81	0.86	0.84
		right	0.91	0.67	0.77	0.82	0.77	0.79
	ideology_multiclass	left	0.5	1	0.67	0.55	0.76	0.64
		moderate_left	0.43	0.75	0.54	0.67	0.72	0.69
		moderate_right	0.67	0.4	0.5	0.71	0.48	0.58
		right	0.8	0.8	0.8	0.62	0.59	0.61
LR (Without Stopwords)	gender	male	0.3	0.75	0.43	0.49	0.64	0.55
		female	0.9	0.56	0.69	0.77	0.65	0.71
	pofession	journalist	0.71	0.83	0.77	0.82	0.72	0.76
		politician	0.92	0.86	0.89	0.91	0.95	0.93
	ideology	left	0.5	1	0.67	0.79	0.84	0.81
		right	1	0.67	0.8	0.79	0.73	0.76
	ideology_multiclass	left	0.5	1	0.67	0.52	0.57	0.54
		moderate_left	0.43	0.75	0.54	0.65	0.67	0.66
		moderate_right	0.75	0.6	0.67	0.75	0.58	0.65
		right	1	0.6	0.75	0.57	0.7	0.63

there are no inflated results considering other approaches. In the professional identification task, the worst results were obtained by two different models, which have obtained similar results, the SVM without stopwords and the LR with stopwords. LR obtained the best results without stopwords and SVM with stopwords, where the former has harvested the best results for classifing journalists and politicians. In the ideology classification task, the best results were obtained by the SVM with stopwords variant in both tasks, reaching a notable difference from other models. Finally, in the multilabel classification, the results are tight from two different models with two opposite variants, the SVM with stopwords and the LR without stopwords are obtained close results, where the ones obtained from moderate_left and right tasks are the different ones, where SVM performed better in moderate_left and worse task and LR showed an opposite behaviour.

On the other hand, regarding the evaluation dataset, if we compare the results obtained from both datasets, it can be noted that the results obtained in the evaluation dataset have been the subject of a notable increment. In the gender classification task, for instance, the precision of the male classification task has been increased almost 0.4 points in SVM without stopwords and LR with stopwords, increasing from 0.16 to 0.55 and 0.18 to 0.57, respectively. The best performance in the profession classification task, where SVM has obtained the highest results in both variants. For the ideology classification task, the LR with stopwords performed slightly better than the SVM without stopwords, a range of 0.2 points. Finally, in the multilabel classification task, the two variants of LR has gathered the best and worst result from the variant stopwords and without stopwords, respectively. Given the analysis conducted, we believe that one of the first triggerings of these results is the unbalanced set of samples provided for the task. Furthermore, the variability of the obtained results does not give us clear evidence of which is the best pipeline constructed to cover the variety of tasks provided since, during the analysis conducted, there was a clear model candidate that made a significant difference between other opponents. Therefore, we selected the SVM with a stopwords approach to competing in the challenge since it reached a high position in the development stage. However, it was not as good as we expected at the evaluation stage and dropped to the 17th position in the ranking.

To conclude the analysis, we add the confusion matrix of each classification task to visualise the proposed model's performance (see Figure 1). The matrix contrasts the instance of the actual class vs the samples predicted in this class. In the experiments, the worst results were obtained in the gender classification task, where the algorithm missed classifying 20 and 12 male and female instances, respectively. Conversely, the lowest mistakes quantity was obtained in the profession classification task, where only 8 and 5 cases were badly classified, considering the number of instances to classify, since the multilabel classification task has a lower amount. Still, the number of instances to classify are lower too. Therefore, we do not consider it a relevant behaviour to point out.

5. Conclusions

This paper describes the system designed by the URJC-Team at the PoliticEs task at the IberLEF 2022 evaluation campaign. We have employed two supervised Machine Learning techniques: Logistic Regression and Support Vector Machine with a linear kernel for the classification

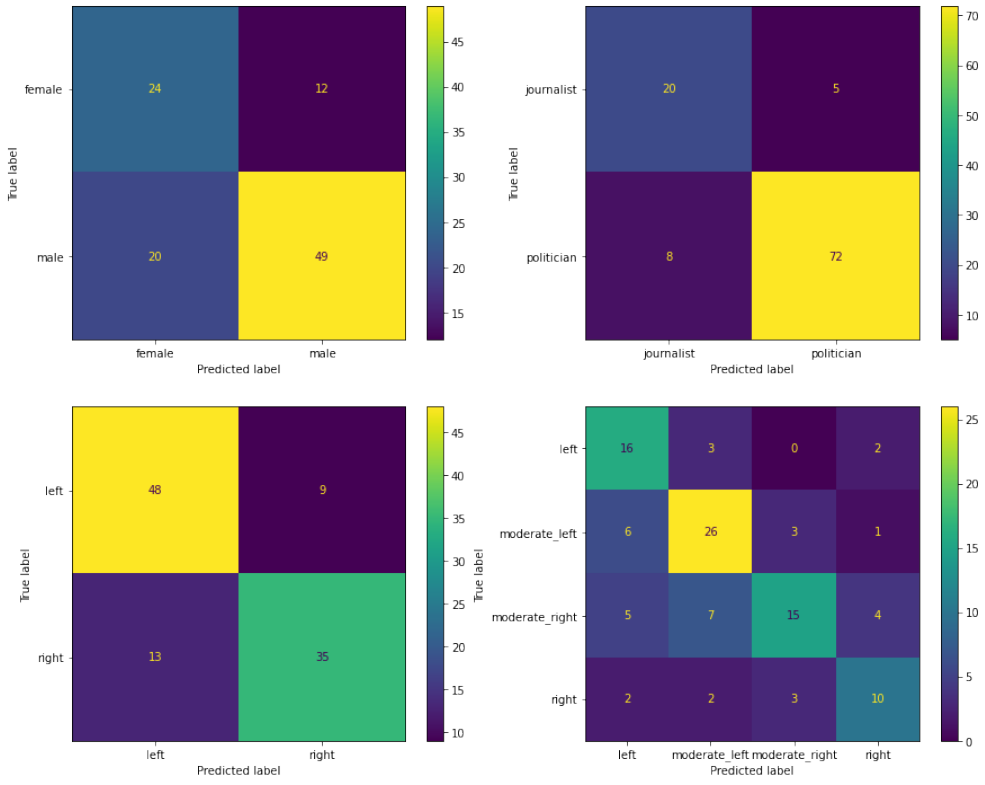

Figure 1: Confusion matrix of the different classification tasks.

tasks. The features used for training the models are extracted previously from preprocessed and cleaned tweets. The best results obtained were for the ideology binary and multiclass in the case of the development dataset. However, the best results were obtained in the evaluation dataset for the ideology and profession classification subtasks.

As future work, there are several research lines that we would like to explore. Although we tried to employ them for this challenge, we would like to incorporate new models at the classification stage. For instance, we think of using neural networks and transformers models to analyse their accuracy. Secondly, given the unbalanced dataset, we believe that this is a significant drawback for the system to achieve better performance. Therefore, we consider incorporating augmentation techniques to enrich both datasets and resolve the problem. Finally, another essential issue to analyse in detail is the feature distribution generated from natural language. Currently, there are numerous ways of extracting and representing features. We think it would be interesting to incorporate new feature extraction mechanisms into the system for analysing not only the behaviour of the current model but also the ones that we would integrate.

Acknowledgments

This work has been partially supported by projects DOTT-HEALTH (PID2019-106942RB-C32, MCI/AEI/FEDER, UE), grant "Programa para la Recualificación del Sistema Universitario Español 2021-2023", and the Community of Madrid, through the Young Researchers R+D Project. Ref. M2173 – SGTRS (co-funded by Rey Juan Carlos University).

References

[1] D. Bamman, N. A. Smith, Open Extraction of Fine-Grained Political Statements, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 76–85.

[2] W. Chen, X. Zhang, T. Wang, B. Yang, Y. Li, Opinion-aware Knowledge Graph for Political Ideology Detection, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), volume 17, 2017, pp. 3647–3653.

[3] Z. Jelveh, B. Kogut, S. Naidu, Detecting Latent Ideology in Expert Text: Evidence From Academic Papers in Economics, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1804–1809.

[4] M. Cardaioli, P. Kaliyar, P. Capuozzo, M. Conti, G. Sartori, M. Monaro, Predicting Twitter Users' Political Orientation: An Application to the Italian Political Scenario, in: Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2020, pp. 159–165.

[5] 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).

[6] S. Bhatia, D. P, Topic-Specific Sentiment Analysis Can Help Identify Political Ideology, in: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2018, pp. 79–84.

[7] Y. Sim, B. D. Acree, J. H. Gross, N. A. Smith, Measuring Ideological Proportions in Political Speeches, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 91–101.

[8] A. Ahmed, E. P. Xing, Staying Informed: Supervised and Semi-Supervised Multi-view Topical Analysis of Ideological Perspective, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1140–1150.

[9] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, L. Ungar, Beyond Binary Labels: Political Ideology Prediction of Twitter Users, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 729–740.

[10] 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.

[11] F. Rangel, A. Giachanou, B. H. H. Ghanem, P. Rosso, Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter, in: CEUR Workshop Proceedings, volume 2696, Sun SITE Central Europe, 2020, pp. 1–18.

[12] M. Nieuwenhuis, J. Wilkens, Twitter text and image gender classification with a logistic regression n-gram model, in: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), 2018.
[13] J.-P. Posadas-Durán, I. Markov, H. Gómez-Adorno, G. Sidorov, I. Batyrshin, A. Gelbukh, O. Pichardo-Lagunas, Syntactic n-grams as features for the author profiling task, Working Notes Papers of the CLEF (2015).
[14] 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.