UMUTEAM at DIPROMATS 2024: Feature Integration for Detecting Fine-grained Propaganda and Narrative

José Antonio García-Díaz[1], Ronghao Pan[1], Andreu Rodilla Lázaro[2], Camilo Cristancho[2] and Rafael Valencia-García[1]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*
[2]*Dep. de Ciència Política, Dret Constitucional i Filosofia del Dret, Universitat de Barcelona*

Abstract

These notes describe our participation in the 2nd edition of the DIPROMATS shared task, held at IberLEF 2024. This edition repeated the fine-grained detection of propaganda techniques in politics and added an additional subtask for narrative detection, which consists of a multiclass and multi-label classification problem to classify a set of predefined narratives of international actors using few-shot learning. Both tasks are multilingual. For the first task, we propose an approach similar to the one used in the previous edition, combining linguistic features and sentence embeddings using ensemble learning and knowledge integration. For Task 1, we obtain our best result by applying knowledge integration. For the Task 2, we evaluate TuLu and Zephyr, but our results fall below the proposed baseline based on Mixtral 8x7B.

Keywords

Propaganda Identification, Feature Engineering, Transformers, Feature Integration, Few-shot Learning, Natural Language Processing

1. Introduction

As defined in [1], propaganda encompasses *a continuously evolving set of techniques and mechanisms designed to facilitate the dissemination of ideas and actions.* To facilitate its spread, propaganda often uses rhetorical devices. The analysis of these techniques is detailed in [2]. Propaganda is usually perceived as persuasive communication that uses manipulative practices to persuade and has historically been associated with totalitarian regimes. This characterization implies a negative connotation of propaganda as a threat the principles of public debate [3]. However, persuasion is a central component of political debate in democratic contexts. As such, propaganda can also be considered a legitimate political communication strategy used by political actors in their everyday interactions and appeals to their followers and their adversaries.

Identifying the elements that make up propaganda in the political context is crucial to understanding the extent of its legitimate use within a democratic rationale. Communication practices based on microtargeting in the context of political campaigns are probably the most prominent case. Cases such as the Brexit referendum exemplify propagandistic interference in democratic decision-making processes [4]. However, these are only prominent cases at the extreme end of communication strategies that take place across multiple political arenas, such as diplomatic communication.

Everyday political interaction is based on narratives that emphasize group identities and exploit emotional rhetoric. Narratives convey political messages and worldviews to express particular political positions and to take distance themselves from opposing perspectives, and political opponents. Political actors thus seek to control the narrative in order to shape political processes according to their own

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
†These authors contributed equally.
joseantonio.garcia8@um.es (J. A. García-Díaz); ronghao.pan@um.es (R. Pan); rodilla.lazaro@ub.edu (A. R. Lázaro); camilo.cristancho@ub.edu (C. Cristancho); valencia@um.es (R. Valencia-García)
0000-0002-3651-2660 (J. A. García-Díaz); 0009-0008-7317-7145 (R. Pan); 0000-0002-1381-7664 (A. R. Lázaro); 0000-0003-1794-4457 (C. Cristancho); 0000-0003-2457-1791 (R. Valencia-García)

interests and strategic intentions [5]. Dissecting the constitutive elements in the communication strategies of political actors is crucial to describing the relationships between them.

The second edition of the DIPROMATS 2024 challenge [6], held at IberLEF [7], focuses on identifying propaganda techniques by analyzing the language used by official authorities in social networks. To this end, the organizers of the challenge have repeated the previous edition [8], with a dataset of micro-posting messages on Twitter in Spanish and English from diplomatic profiles of China, Russia, the United States and the European Union. Specifically, two subtasks are proposed. The first subtask is a binary classification in which participants have to decide when a text contains propaganda. The second subtask is to categorize the techniques used to spread propaganda. This categorization is done in two ways: a multi-classification approach and a multi-label approach.

In this edition, the organizers have added a novel multi-label classification task focused on narrative identification. Narratives are at the heart of propaganda because they consist of sequences of events, linked by cause and effect, that are selected and judged to be significant to a particular audience [9]. Because narratives reduce complex political processes and political values to simple descriptions, they are central to defining sociopolitical realities, and to the way in which individuals understand politics [10]. One of the most important issues is the power of narratives to promote or otherwise undermine the trust in the social relations that underpin everyday compromise and representation in political systems.

Our team participated in both tasks, achieving their best results in Task 1 using feature integration based on knowledge integration and evaluating two Large Language Models (LLMs), TuLu and Zephyr, for Task 2. However, we fall below the baseline based on Mixtral 8x7B and only one other participant submitted results for Task 2. Therefore, the limited limited number of participants in Task 2 prevents us us from making better comparison of our results.

2. Dataset

According to the organizers, the DIPROMATS' dataset consists of Spanish and English tweets by diplomats from four different countries, namely China, Russia, the United States and the European Union. The authorities are government accounts, embassies, ambassadors and other diplomatic profiles. The collected tweets were published between January 1, 2020 and March 11, 2021, with the last day coinciding with the first anniversary of the declaration of the COVID-19 pandemic. Specifically, the Spanish dataset contains 9,591 tweets and the English dataset contains 12,012 tweets from 619 agencies. The data was split with a 70/30 ratio using a temporal criterion, where the training split the oldest tweets and the test split the newest.

The tweets were labeled using a similar criteria as in previous edition, using the taxonomy proposed in [2] but removing some techniques used in the previous edition. In this edition they keep (1) Appeal to Commonality, which Ad populum and Flag Waving; (2) Discrediting the Opponent, with Name Calling, Appeal to Fear, Undiplomatic Assertiveness, and Doubt; and (3) loaded language, which refers to the use of hyperbolic language, metaphors and expressions with strong emotional implications. It is worth noting that the main category, Appeal to Authority has been removed.

Table 1 shows the statistics of the Spanish and English partitions of the DIPROMATS 2024 task. As can be seen, the dataset is very unbalanced. Furthermore, there are no instances of documents marked as *appeal to false authority* in the English partition and for *bandwagoning* in the Spanish partition.

For Task 2, the dataset is a subset of the Task 1 dataset. Since Task 2 is a few-shot learning task, the authors include the narrative description along with two or three examples for training.

3. Methodology

3.1. Task 1. Propaganda identification and characterization

As in the previous edition, we focus on the 3rd subtask for Task 1, since treating it as a multi-label problem also solves the subtasks of binary propaganda identification (subtask 1) and the Propaganda characterization, coarse-grained (subtask 2). This strategy reduces the number of models we need to train, thus saving time and effort.

Our proposal for solving Task 1 is based on feature integration of linguistic features (LFs) and sentence embeddings from some state-of-the-art LLMs. For the LFs, we rely on UMUTextStats [11], while for the sentence embeddings, we rely on feature extraction from the fine-tuned models BETO [12], MarIA [13], DeBERTa, Twitter XLM [14], and RoBERTalex [15] for Spanish; and BERT [16], XLM, BoBERTa [17], DeBERTa, Twitter XLM [14], and Legal BERT [18]. Compared with our proposal in 2023, we removed from our pipeline the lightweight models of ALBERT (and ALBETO) and DistilBERT (and DistilBETO), as well as BERTIN, and multilingual BERT. In fact, the only model added is RoBERTalex, which is a Spanish LLM trained on the Spanish Legal Domain Corpora, with a total of 8.9GB of text.

For each LLM, we obtain its sentence embeddings, since a fixed representation of the data simplifies the task of combining the LLM with the linguistic features. In order to identify the best configuration for each LLM, we train 10 models for each LLM for Spanish and English, evaluating different learning rates, training epochs, batch sizes, warm-up steps and weight decay. This step is done using RayTune [19] with Distributed Asynchronous Hyperparameter Optimisation (HyperOptSearch) with the Tree of Parzen Estimators (TPE) algorithm [20] and the ASHA scheduler (because it favors parallelism). Table 2 shows the best configuration found for each LLM for Spanish and English for subtask 3. It can be observed that all the models require a larger number of training epochs, between 4 and 5, with the only exception of BETO in Spanish. Regarding the batch size, almost all LLMs prefer smaller batch sizes (8), with the only exception of XLM in English. Regarding the warm-up steps, Spanish usually requires smaller steps compared to English.

The next step for our pipeline is to obtain the contextual sentence embeddings from the classification token, as suggested in [21]. This fixed representation of each document in the corpus allows us to more easily apply in feature combination strategies between the LLMs and with the LFs.

Once we have extracted the embeddings, we train a different neural network model for each LLM, but using Keras and a multi-input neural network that uses the LFs and the embeddings. This strategy is called Knowledge Integration (KI). With Keras, we also evaluate different network shapes, including the depth of the network and its shape. The learning rate, batch size, and dropout mechanism are also evaluated.

The results of the hyperparameter optimization with Keras are shown in the table 3. In the case of LFs, both models (Spanish and English) require a shallow neural network with one hidden layer, but only 16 neurons in the case of Spanish and 256 in the case of English. This can be explained by the fact

Table 1
Dataset statistics for the Spanish and English partitions of the DIPROMATS 2024 shared task

		Spanish				English			
Category	Label	train	val	test	total	train	val	test	total
1	ad populum	40	19		59	56	16		72
1	flag waving	181	53		234	429	116		545
2	doubt	19	8		27	55	19		76
2	fear appeals	44	17		61	39	18		57
2	name calling	64	26		90	164	49		213
2	undiplomatic assertiveness	447	122		430	536	145		681
3	loaded language	302	87		389	723	190		913
	total	1097	332			2002	553		

that the LFs extracted by UMUTextStats are best focused on Spanish, requiring a smaller number of neurons for the best evaluated parameters. Compared to the LFs, the sentence embedding models also required simpler but larger neural networks (one or two hidden layers but a large number of neurons). For KI, both languages required two hidden layers with 512 neurons.

Apart from the KI strategy, we build the ensemble learning models based on combining the outputs of the models trained with the sentence embeddings for each LLM and the LFs. Specifically, we evaluate three strategies for combining the outputs: (1) highest probability, as we choose the maximum probability for each label; (2) averaging the probabilities of each model in the ensemble, and (3) the mode of each label in the predictions.

Table 2
Hyperparameter tuning of the LLM before obtaining the sentence embeddings

LLM	learning rate	train epochs	batch size	warmup steps	weight decay
			Spanish		
BETO	4.7e-05	3	8	250	0.27
MARIA	2.1e-05	4	8	500	0.26
MDEBERTA	2.8e-05	5	8	250	0.079
ROBERTALEX	3.9e-05	5	8	500	0.3
XLMTWITTER	4.9e-05	5	8	500	0.19
			English		
BERT	3.4e-05	4	8	1000	0.032
LEGALBERT	4.7e-05	4	8	1000	0.11
MDEBERTA	3.1e-05	5	8	500	0.033
ROBERTA	3.8e-05	4	8	250	0.25
XLM	4.1e-05	4	16	1000	0.071
XLMTWITTER	2.9e-05	4	8	500	0.26

Table 3
Best hyperparameters for the deep learning models for Spanish (left) and English (right)

	hidden layers	neurons	dropout	lr	batch size	activation
feature-set			Spanish			
LF	1	16	-	0.001	64	linear
BETO	2	37	0.2	0.001	64	tanh
MARIA	1	64	0.1	0.001	64	relu
MDEBERTA	1	512	-	0.001	64	linear
ROBERTALEX	1	256	0.2	0.001	32	relu
XLMTWITTER	2	128	0.3	0.001	64	sigmoid
KI	2	512	0.3	0.01	64	linear
feature-set			English			
LF	1	256	-	0.001	64	linear
BERT	2	512	-	0.001	64	sigmoid
LEGALBERT	2	512	-	0.001	64	tanh
MDEBERTA	1	128	-	0.001	64	tanh
ROBERTA	1	37	0.3	0.001	64	linear
XLMTWITTER	1	256	0.1	0.001	32	tanh
KI	2	512	0.2	0.001	64	linear

3.2. Task 2. Automatic detection of narratives from diplomats of majors powers

Task 2 is a multi-class and multi-label classification problem that aims to determine the narrative to which the tweets belong, given a set of predefined narratives for each international actor. For the implementation, the organizers provided us with the description of each narrative and some examples of English and Spanish tweets corresponding to each narrative. Since this is a multi-label problem, a tweet can be associated with one, several or none of the narratives.

For Task 2, we have used the few-shot learning approach of different LLM models to determine the narrative of the tweets. LLMs are mainly neural language models based on the Transformer architecture, which contain tens to hundreds of billions of parameters and are pre-trained on massive text data. These models exhibit more robust language understanding and generation capabilities, as well as emergent capabilities not found in smaller-scale language models. These emergent capabilities include context learning, instruction following, and multi-step reasoning and among others. The in-context learning capability enables LLMs to generate more coherent and contextually relevant responses, making them suitable for interactive and conversational applications. In addition, this capability allows LLMs to quickly adapt to a new task by using examples in the input, without the need for retraining or model adaptation. Few-shot learning is a technique where a model can effectively generalize to new tasks using only a few training examples at the LLMs prompt. In this case, it would be the set of predefined narratives of each international actor.

The models evaluated for Task 2 are Zephyr-7b-beta and Tulu-2-dpo-7b. Zephyr-7B-beta [22] is the second model in the Zephyr series of language models designed to serve as useful assistants. It is a tuned version of the Mistral-7B-v0.1 [23] model, trained with a Direct Preference Optimization (DPO) approach using a mixture of public and synthetic datasets. Tulu-2-dpo-7b [24] is a language model developed as part of the Tulu series as a useful assistant in various natural language processing tasks. This model is a tuned version of the Llama 2 [25] (Llama-2-7b-hf) model, and has been trained using a technique called DPO.

For Zephyr-7b, prompts must be structured with designated fields: “System”, “User”, and “Assistant”. The “System” field provides instructions or guidance to the model. The “User” field contains the user’s intent and the item to be classified, while the “Assistant” field is the output indicator.

The DPO fine-tuned iteration of the Tulu model (Tulu-7b-dpo) requires that the model input consist of two fields: The “user” and the “assistant”. The “User” field is used to specify instructions and the instance to be classified by the model, while the “Assistant” field acts as an output indicator. It’s important to note that a new line must be added after each field, as this can significantly affect the quality of the generation.

4. Results and discussion

4.1. Task 1

To evaluate the performance of the models, we use a custom validation split. The results are shown in Table 4 for Spanish (left) and English (right). These results are for the third subtask, i.e., the multi-label task.

For Spanish, the best model is obtained with the KI strategy, with a macro f1-score of 54.467%, outperforming the individual models. This model achieved a very good recall but other models based on Ensemble Learning achieved better precision (EL based on the mode), and recall (based on highest probability). The LFs achieved limited results, but better in terms of recall and f1-score compared with our previous edition. In the case of English, the best result is obtained with ROBERTA in isolation (f1-score of 58.083%), but the EL based on the highest probability obtained a better recall but a very limited precision. The result obtained with English was more similar to our performance in the previous edition, where individual models achieved better performance than feature integration. The results for the Spanish are in line with other work carried out by our research group [26, 27, 28, 29].

For the competition, we sent a total of 5 runs. Three for Task 1 and two for Task 2. All three runs for

task 1 were based on feature integration. One for the KI strategy and the rest were two ensembles, one based on highest probability and the other in mode. All runs for Task 2 were based on FSL with TuLu and Zephyr.

For Task 1, we report the results of the official leaderboard. The metric used to compare the systems is the ICM-Hard [30]. The table 5 shows the official leaderboard of the DIPROMATS 2024 shared task. In the tables, we have published only one run per competitor, as we believe this is the fairest comparison. Our best results were with our third run, based on KI. We ranked 4th in the binary classification task, with an ICM hard of 0.1667. For the second and third subtasks, we ranked 3rd, with an ICM hard of -0.0832 and -0.33883, respectively.

Next, Table 6 shows the results obtained for each run on the test set with the macro F1-score. These runs are based on feature integration. The first run is based on ensemble learning on highest probability, the second run is based on ensemble learning based on mode, and the third run is based on KI.

As noted when reviewing the results per run, the best performing strategy is KI, which outperforms the rest with the F1-score except in Subtask 3 (English). In general, for subtasks 2 and 3 the results for English were better than Spanish. A possible interpretation for this is that the Spanish dataset is more complex and propaganda messages are harder to detect.

4.2. Task 2

Table 7 shows the results obtained with the LLMs on the test set for Task 2. Three metrics were used to evaluate the performance of the models: F1 Strict, F1 Lenient and F1 Avg. These metrics provide a comprehensive evaluation of the accuracy of the models in different aspects of classification. In addition, the results are presented for two languages: Spanish and English.

For the Spanish language, the baseline model scores highest on all metrics, followed by Zephyr and

Table 4
Results of the custom validation split of the DIPROMATS 2024 shared task for Spanish (left) and English (right)

Spanish				English			
feature-set	precision	recall	f1-score	feature-set	precision	recall	f1-score
LF	28.322	33.652	28.295	LF	17.425	31.776	20.841
BETO	63.496	40.014	46.525	BERT	63.383	51.708	56.391
MARIA	68.851	37.047	45.644	LEGALBERT	61.449	50.508	54.710
MDEBERTA	44.130	24.817	26.280	MDEBERTA	65.206	48.096	53.076
ROBERTALEX	67.399	32.479	41.127	ROBERTA	72.187	56.558	**58.083**
XLMTWITTER	61.669	27.030	33.935	XLMTWITTER	64.222	44.151	50.501
KI	70.041	48.796	**54.467**	KI	63.527	52.921	56.111
EL (HIGHEST)	33.078	**56.158**	39.965	EL (HIGHEST)	34.882	**76.522**	46.104
EL (MEAN)	**89.442**	30.138	39.725	EL (MEAN)	78.807	46.908	54.681
EL (MODE)	89.040	28.405	38.442	EL (MODE)	**79.635**	41.196	50.410

Table 5
Official leaderboard for the 3 subtasks of Task 1 of DIPROMATS 2024 shared task

Subtask 1		Subtask 2		Subtask 3	
Team	ICM-Hard	Team	ICM-Hard	Team	ICM-Hard
Victor Vectors	0.2048	DSHacker	-0.0074	DSHacker	-0.1144
DSHacker	0.2018	Victor Vectors	-0.0425	Victor Vectors	-0.1144
PropaLTL	0.1667	UMUTEAM (03)	-0.0832	UMUTEAM (03)	-0.1383
UMUTEAM (03)	0.1646	UC3M-LCPM	-0.2205		-0.6844
UC3M-LCPM	0.1268				

Table 6
Results per run using the test set with the macro F1-score for Task 1 of DIPROMATS 2024.

	Subtask 1		Subtask 2		Subtask 3	
run	Spanish	English	Spanish	English	Spanish	English
01	64.55	49.17	31.13	34.31	33.15	25.47
02	77.20	78.77	34.90	55.03	24.53	41.64
03	78.13	81.17	41.34	58.77	38.84	40.80

Tulu. For English, Zephyr outperforms both Tulu and the baseline on all metrics except the baseline's F1 Lenient.

In summary, the Mixtral-8x7B model performs better in Spanish, outperforming the other models evaluated such as Zephyr and Tulu on all metrics. However, the baseline performance in English is worse than Zephyr, with an F1 Avg of 0.4161, which is 0.46% lower, despite being an 8x larger model.

Table 7
Results obtained with LLMs in the test set for Task 2 of DIPROMATS 2024.

	Spanish			English		
run	F1 Strict	F1 Lenient	F1 Avg	F1 Strict	F1 Lenient	F1 Avg
Zephyr	0.3046	0.4427	0.3737	0.3149	0.5265	0.4207
TuLu	0.2729	0.3976	0.3353	0.2303	0.4441	0.3372
Baseline	0.3769	0.5278	0.4524	0.2875	0.5446	0.4161

5. Conclusions

In this paper we have presented our approach to solving the DIPROMATS 2024 shared task. We focused on propaganda characterization in a multi-label way, since models trained for this task can also solve the propaganda identification and propaganda characterization task using a multi-classification approach. Our approach evaluated linguistic features and sentence embeddings from several LLMs, including models specific to English, Spanish and other multilingual models. We achieved competitive results in all tasks and we are very satisfied with the results.

The task of identifying the constitutive elements in narratives, such as references to social symbols, loaded language, and emotional cues is central to understanding the communication processes that shape political identities and attitudes. Group appeals, and their contrasts with outgroups or adversaries, contain value-charged claims and emotional rhetoric that have important effects in attitude formation processes and behavioral intentions. Evidence for the mobilizing effects of anger and hope, or conversely for the inaction produced by fear and doubt, confirms the need to correctly identify the constituent elements of political narratives.

The 2024 task is particularly relevant because narratives not only have the power to influence the climate of public opinion, but also benefit from inflamed political contexts to effectively disseminate their messages. Thus, the implications of how diplomats' narratives are structured in the context of the COVID pandemic can shed light on how everyday political interactions are intimately connected to the context in which they are produced. In a media landscape dominated by fast-moving, profit-driven social media platforms and global media conglomerates, the ability to identify the structuring elements of political narratives that constitute propaganda or persuasive communication is a crucial matter. Identifying propaganda and the elements that constitute its underlying narratives allows us to understand the behavior of political actors and their media landscapes. This is a first, but very important step in observing political communication processes that are central to democracy.

In future work, we plan to compare our results in Task 1 with alternative results if we had trained

a model focused on propaganda identification. In addition, our evidence suggests that the results of models based on BERT and BETO outperform more sophisticated approaches that have been effective in other collaborative tasks. Accordingly, we will provide a detailed error analysis for each propaganda technique. In addition, we will compile audio and video of politicians and examine propaganda with audio features, similar to [31].

Acknowledgments

This work has been supported by projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way of making Europe, LT-SWM (TED2021-131167B-I00) funded by MICIU/AEI/10.13039/ 501100011033 and by the European Union NextGenerationEU/PRTR, "Services based on language technologies for political microtargeting" (22252/PDC/23) funded by the Autonomous Community of the Region of Murcia through the Regional Support Program for the Transfer and Valorization of Knowledge and Scientific Entrepreneurship of the Seneca Foundation, Science and Technology Agency of the Region of Murcia. Mr. Ronghao Pan is supported by the Programa Investigo grant, funded by the Region of Murcia, the Spanish Ministry of Labour and Social Economy and the European Union - NextGenerationEU under the "Plan de Recuperación, Transformación y Resiliencia (PRTR)".

References

[1] C. Sparkes-Vian, Digital propaganda: The tyranny of ignorance, Critical sociology 45 (2019) 393–409.

[2] G. Da San Martino, S. Yu, A. Barrón-Cedeno, R. Petrov, P. Nakov, Fine-grained analysis of propaganda in news article, in: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 5636–5646.

[3] A. Godber, G. Origgi, Telling Propaganda from Legitimate Political Persuasion, Episteme 20 (2023) 778–797.

[4] Q. Cassam, Bullshit, post-truth, and propaganda, Political epistemology (2021) 49–63.

[5] S. Groth, Political narratives/narrations of the political: An introduction, Narrative Culture 6 (2019) 1–18.

[6] P. Moral, J. Fraile, G. Marco, A. Peñas, J. Gonzalo, Overview of DIPROMATS 2024: Detection, characterization and tracking of Propaganda in messages from diplomats and authorities of world powers, Procesamiento del Lenguaje Natural 73 (2024).

[7] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[8] Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de-Albornoz, Iván Gonzalo-Verdugo, Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers, Procesamiento del Lenguaje Natural 71 (2023).

[9] C. K. Riessman, Narrative Methods for the Human Sciences, Sage, 2008.

[10] B. McLaughlin, J. A. Velez, J. A. Dunn, The political world within: How citizens process and experience political narratives, Annals of the International Communication Association 43 (2019) 156–172.

[11] J. A. García-Díaz, P. J. Vivancos-Vicente, A. Almela, R. Valencia-García, UMUTextStats: A linguistic feature extraction tool for Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 6035–6044.

[12] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, Pml4dc at iclr 2020 (2020) 1–10.
[13] A. G. F. no, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, MarIA: Spanish Language Models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.
[14] F. Barbieri, L. E. Anke, J. Camacho-Collados, Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 258–266.
[15] A. G.-F. no, J. Armengol-Estapé, A. Gonzalez-Agirre, M. Villegas, Spanish Legalese Language Model and Corpora, 2021. arXiv:2110.12201.
[16] 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. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.
[17] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.
[18] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The Muppets straight out of Law School, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 2898–2904.
[19] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, I. Stoica, Tune: A research platform for distributed model selection and training, arXiv preprint arXiv:1807.05118 (2018).
[20] J. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-parameter optimization, Advances in neural information processing systems 24 (2011).
[21] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410. doi:10.18653/v1/D19-1410.
[22] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, T. Wolf, Zephyr: Direct Distillation of LM Alignment, 2023. arXiv:2310.16944.
[23] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7B, arXiv preprint arXiv:2310.06825 (2023).
[24] H. Ivison, Y. Wang, V. Pyatkin, N. Lambert, M. Peters, P. Dasigi, J. Jang, D. Wadden, N. A. Smith, I. Beltagy, H. Hajishirzi, Camels in a Changing Climate: Enhancing LM Adaptation with Tulu 2, 2023. arXiv:2311.10702.
[25] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[26] J. A. García-Díaz, F. García-Sánchez, R. Valencia-García, Smart analysis of economics sentiment in Spanish based on linguistic features and transformers, IEEE Access 11 (2023) 14211–14224.
[27] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, Complex & Intelligent Systems (2022) 1–22.
[28] J. A. García-Díaz, R. Valencia-García, Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers, Complex & Intelligent Systems 8 (2022) 1723–1736.
[29] J. A. García-Díaz, G. Beydoun, R. Valencia-García, Evaluating Transformers and Linguistic Features

integration for Author Profiling tasks in Spanish, Data & Knowledge Engineering 151 (2024) 102307.

[30] E. Amigo, A. Delgado, Evaluating Extreme Hierarchical Multi-label Classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399. doi:10.18653/v1/2022.acl-long.399.

[31] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, R. Valencia-García, Spanish MEACorpus 2023: A multimodal speech–text corpus for emotion analysis in Spanish from natural environments, Computer Standards & Interfaces 90 (2024) 103856.