UMUTeam at FLARES@IberLEF 2024: Enhancing Disinformation Detection with 5W1H Techniques and Transformer Models

Ronghao Pan, José Antonio García-Díaz*, Francisco García-Sánchez and Rafael Valencia-García

Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain

Abstract

Disinformation, including fake news and hoaxes, is a significant problem, aggravated by digital media and social networks, which requires automatic detection tools. The FLARES shared task at IberLEF 2024 uses the 5W1H technique to assess the reliability of language in news. The task is divided into two subtasks: (1) 5W1H identification and (2) 5W1H based reliability. This paper presents UMUTeam's participation in both tasks. For Task 1, we developed a Named Entity Recognition (NER) model to identify 5W1H elements using fine-tuned transformer models such as BERT and MarIA, incorporating Part-Of-Speech (POS) and Syntactic Dependency (Dep) features. Our BETO + POS + Dep model achieved the second-best result with a score of 56.778%. For Task 2, which focused on assessing the reliability of 5W1H entities, our approach based on contextual fine-tuning of the MarIA model achieved the best result with a score of 65.820%.

Keywords

Disinformation, Named Entity Recognition, Fine-tuning, Pretrained Language Models, Transformers

1. Introduction

Disinformation, which includes the whole issue of fake news and hoaxes, is considered a type of "information disorder" in which false information is created or disseminated with the express intent to cause harm [1]. Technological advances, particularly in digital media and social networks, have created a more connected world. However, with greater connectivity comes the potential for misuse and rapid dissemination, which can cause problems in areas such as public health, social welfare, or identity. The rapid and viral spread of vast amounts of information requires automated detection tools necessary, as it is impossible to manually process and verify this volume of data. The automatic detection of misinformation is a complex task to solve from an engineering point of view, and the research community is approaching this task from different perspectives, such as position detection, polarization, credibility or automated fact checking.

In addition, since systems need annotated examples to learn from the experts' comments and to justify the decision made, existing automatic recognition algorithms require the intervention of human experts. Obtaining this huge amount of annotated data is a very costly task, both in terms of resources and time [1]. Therefore, in today's digital media landscape, assessing the reliability of the language used in news writing is becoming increasingly important. Thus, identifying specific segments of a news story to assess its linguistic credibility could provide a more nuanced understanding of the veracity of the message.

Recent studies have highlighted the importance of analyzing style, tone, and language structure to identify misleading content. In particular, style and language are features that have been shown to be

IberLEF 2024, September 2024, Valladolid, Spain

*Corresponding author.

ronghao.pan@um.es (R. Pan); joseantonio.garcia8@um.es (J. A. García-Díaz); frgarcia@um.es (F. García-Sánchez); valencia@um.es (R. Valencia-García)

0009-0008-7317-7145 (R. Pan); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2667-535 (F. García-Sánchez); 0000-0003-2457-1791 (R. Valencia-García)

valuable in distinguishing between false and true articles, and specific linguistic features have been shown to be valuable in indicating possible bias or misrepresentation in online content [2].

In this context, the shared task FLARES [3] at IberLEF 2024 [4], which uses the "5W1H" technique, a technique commonly used by journalists to clearly and explicitly present the key information of a notation, in order to systematically assess the reliability of language. The task is divided into two subtasks: (1) 5W1H identification and (2) 5W1H-based reliability.

In this paper, we describe the participation of the UMUTeam in the two FLARES tasks. For Task 1, we have evaluated different approaches based on the Named Entity Recognition (NER) approach, which consists of a model developed based on the recognition of entities (in this case "5W1H" elements) within the text. For this purpose, we have evaluated from fine-tuning different pre-trained models of Transformer languages such as BETO [5] and MarIA [6] for NER implementation to adding Part-Of-Speech (POS) features to the pre-trained models to improve their performance in the NER task. For Task 2, we evaluated the context fine-tuning of the pre-trained language models to determine the reliability of the 5W1H elements.

The rest of the paper is organized as follows. Section 2 presents the task and the provided dataset. Section 3 describes the methodology of our proposed system to address subtask 1 and subtask 2. Section 4 presents the obtained results. Finally, section 5 concludes the paper and suggests future work.

2. Task description

Task 1 (5W1Hs Identification) aims to determine the essential content by identifying and annotating the answers to the 5W1H questions in the text. The 5W1Hs are key questions that help break down the information in a message into its basic components:

- **What**. Identify the main fact or event.
- **Who**. Identify the person or persons involved.
- **When**. Identify the time when the event occurred.
- **Where**. Identify the location of the event.
- **Why**. Identify the cause or reason for the event.
- **How**. Identify the way the event occurred.

Task 2 aims to classify the reliability of the language used in each of the identified 5W1H elements. In this case, it is a multi-class classification problem, where reliability is divided into three levels: confiable, semiconfiable, and non-confiable.

The organizers provided us with a dataset containing approximately 9,034 5WH1 annotations, distributed over 190 news articles [7]. The 5WH1 dataset was divided into 70% (6,934 5W1H annotations) for training and 30% (2,100 5W1H annotations) for testing. The corpus contains Spanish texts manually annotated with an extraction technique known as 5W1H, with the initial and final position of each 5W1H entity, and the reliability of the language in each of the 5W1H items.

The distribution of the dataset at the 5W1H element level and the reliability distribution are shown in Table 1 and Table 2. We can see that we have divided the dataset into a training set and a validation set in an 80-20 ratio, with the goal of using the validation set to train the model and evaluate different hyperparameter configurations.

Table 1
Distribution of the training dataset of the Task 1.

Dataset	What	Who	When	Where	Why	How
Train	2,213	1,466	602	626	191	449
Validation	498	377	176	175	47	114

Table 2
Distribution of the training dataset of the Task 2.

Dataset	Confiable	Semiconfiable	No confiable
Train	3,796	1,044	707
Validation	969	232	186

3. Methodology

3.1. Task 1. 5W1Hs identification

For Task 1, we used an approach based on pre-trained Transformers-based language models to create a custom Named Entity Recognition (NER) model capable of identifying and classifying specific 5W1H entities in a text. A pre-trained language model is a Transformers-based model that has been trained on large amounts of text before being adapted (or tuned) for specific NLP tasks, allowing it to transfer knowledge to the specific task during the training or tuning phase.

In order to train a NER model based on Transformers models, it is crucial to perform a data preparation process in IOB (Inside-Outside-Beginning) format beforehand. The IOB format labels each token in a text as follows: B-<ENTITY>, indicating that the token is the beginning of an entity; I-<ENTITY>, indicating that the token is part of an entity and not the beginning; and O, indicating that the token is not part of any entity. For example, the text "*Hace 5 años, su vida era un calvario permanente.*" would be represented at the token level as follows: `['Hace', '5', 'años', ',', 'su', 'vida', 'era', 'un', 'calvario', 'permanente', '.']`. If we classify the 5WH1 elements in IOB format, we would get `['B-WHEN', 'I-WHEN', 'I-WHEN', 'I-WHEN', 'O', 'O', 'O', 'O', 'B-WHAT', 'I-WHAT', 'I-WHAT', 'I-WHAT', 'O']`. This technique has been applied to similar sequence labeling problems, such as extracting financial entities [8].

Figure 1 shows the general architecture of the system. Unlike other pre-trained models, we have created two additional inputs for the model, in addition to the input_ids and token_type_ids (Tokens) obtained by the model tokenizer. The additional inputs are Part-Of-Speech (POS) and dependency (Dep), which contain information about the part-of-speech and syntactic dependency of the text. These additional inputs are added as an additional embedding layer, i.e. the embeddings of the tokens, token types, POS and Dep tags are added together to obtain a text embedding for the pre-trained models. Once we have the new embeddings for the pre-trained models, we can perform the fine-tuning process to obtain a token classification model. We can see from the Figure 1 that in fine-tuning, the hidden state layer obtained by the pre-trained models is used as input for the classification head. The classification head, in the context of fine-tuning, refers to the part of a neural network that is specifically added or adjusted to adapt the pre-trained model to a token-level classification task.

It should be noted that an additional complexity in using this approach is that Transformer-based pre-trained models such as BERT and RoBERTa are based on subword tokens rather than words. This means that a word like "Arizona" can be tokenized to `['Ari', '##zone']`. So we need to provide POS and Dep tags at the token level and solve the subword problem. For this, we have created a dictionary that maps each POS tag to its corresponding integer and another one for Dep. It is also important to consider the special tokens of pre-trained models, such as [CLS], [SEP], among others. In this case, we have assigned them a value of -100 in tokens, 18 in POS and 31 in Dep during training, so that they are ignored by the loss function. In addition, we assigned the same token_ids of the first token to other subtokens of the same word to improve the performance of the model.

The language models evaluated for this task are BETO [5], MarIA [6] and MarIA-large [6]. All three models are pre-trained with a large amount of Spanish texts, but have different configurations and base architectures. On the one hand, BETO is a language model based on BERT (Bidirectional Encoder Representations from Transformers), specifically trained on a large Spanish corpus. It is essentially a version of BERT adapted to understand and process the Spanish language. BETO has been trained using the Whole Word Masking technique and is available in both lowercase and uppercase versions.

On the other hand, MarIA is a transformer-based language model specifically designed for Spanish. It is based on the RoBERTa base model and has been pre-trained on a very large Spanish text corpus collected from web crawls performed by the National Library of Spain between 2009 and 2019. Thus, MarIA-large is an extended version of MarIA.

We used the same hyperparameters to fine-tune the three models. The hyperparameters used are: learning rate of 2e-5, epoch of 20, training batch size of 8, and epoch as evaluation strategy.

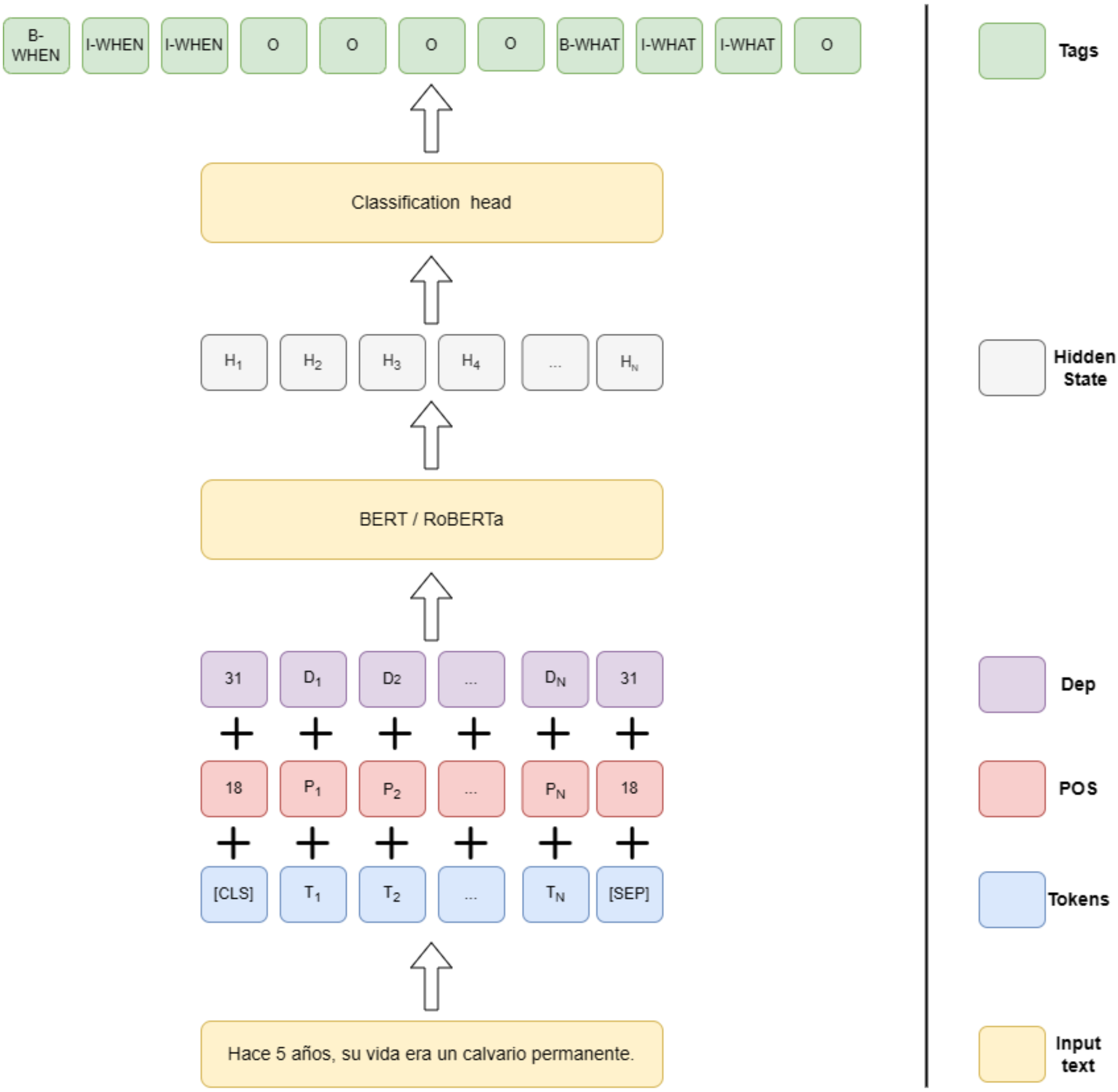

Figure 1: Overall system architecture for Task 1.

3.2. Task 2. 5W1H-based reliability classification

Task 2 aims to determine the reliability of the language used in each 5W1H entity. Thus, it is a multi-class classification problem with a total of three labels: reliable, semi-reliable, and unreliable. Figure 2 shows the architecture used. For this task, we have used a Spanish pre-trained model fine-tuning approach such as BETO [5], MarIA [6], BERTIN [9], ALBETO [10] and DistilBETO [5] for the classification task. We chose these models because they performed well in the financial news classification task in [11] and [12]. Unlike other fine-tuning approaches, we have used the entity text (5W1H) and the whole sentence as context as model input.

The entity text and context are concatenated and tokenized using the tokenization scheme of the pre-trained model. Tokenization converts words into tokens, which are numerical representations that the model can understand. Once generated, the tokens are processed by the BERT or RoBERTa model, which produces representations in the form of hidden states for each token. These hidden states are vectors containing contextual information about the tokens based on the input text. Finally, the hidden states are passed to a classification head, which takes these vectors and transforms them into a probability or logit indicating the reliability of the entity text.

We used the same hyperparameters as the Task 1 to fine-tune the three models. The hyperparameters used are: learning rate of 2e-5, epoch of 20, training batch size of 8, and epoch as evaluation strategy.

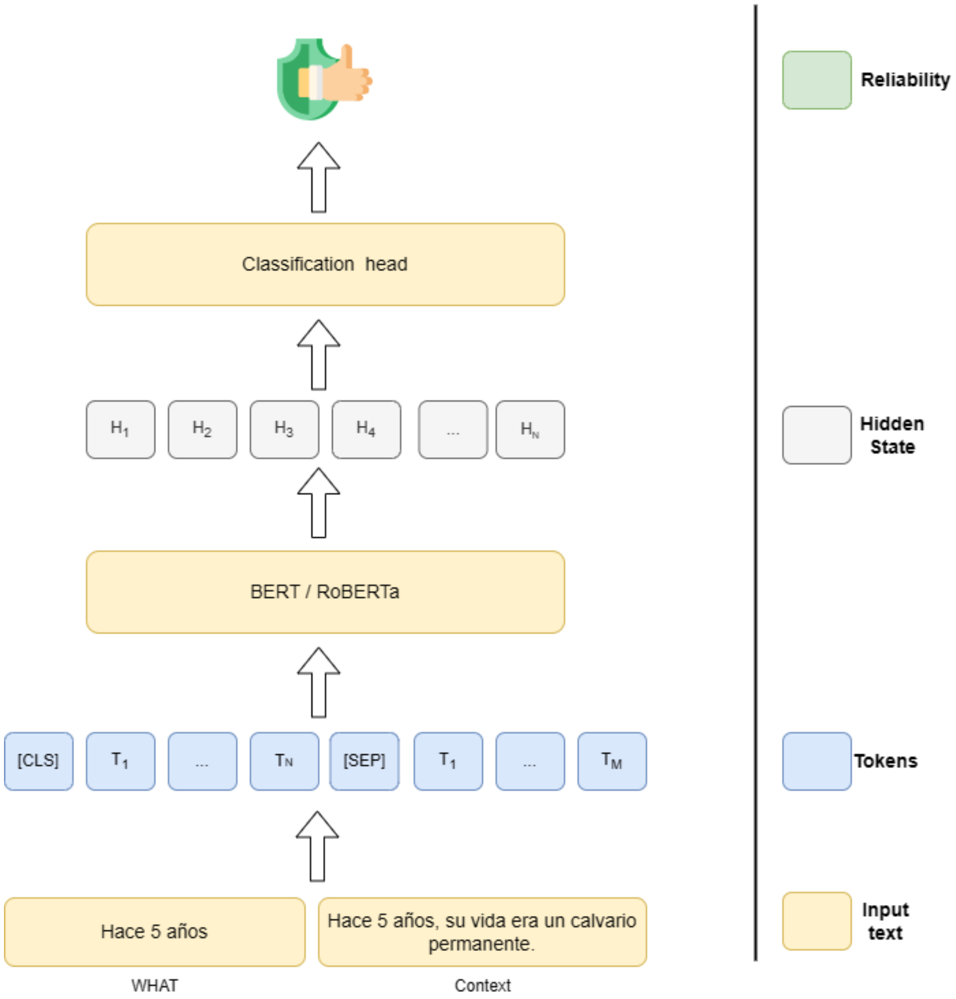

Figure 2: Overall system architecture for Task 2.

4. Results

For Task 1, which aims to determine the essential content by identifying and annotating the answers to the 5W1H questions in the text, we evaluated different configurations of the approach based on using pre-trained Transformers language models to create a custom NER model capable of identifying and classifying specific 5W1H entities in a text. In this case, we evaluated everything from fine-tuning a pre-trained Spanish model to adding different textual features such as part-of-speech and syntactic dependency to the model for fine-tuning.

Table 3 shows the results obtained with different combinations of pre-trained models and textual features. We can see that the combination of BETO with POS and Dep features obtained achieved the highest score (0.56778), outperforming configurations without these additional features. The BETO base model obtained a score of 0.55690, while BETO with POS alone obtained 54.881, showing a slight decrease when POS alone is added. The MarIA model alone obtained a score of 0.54838, and when POS was added, the score increased slightly to 0.55094. The combination of MarIA with POS and Dep significantly improved the performance, reaching a score of 0.55953. The MarIA-large model initially showed lower performance (0.51047) compared to BETO and MarIA. However, when POS was added, the score increased to 0.56373. The best configuration for MarIA-large was with POS and Dep, reaching a score of 0.56712, very close to the best score obtained by BETO with POS and Dep.

In conclusion, the use of additional features such as part-of-speech and syntactic dependencies improves the performance of pre-trained models in Spanish, especially for the task of identifying and classifying of specific 5W1H entities. Moreover, with the BETO+POS+Dep model, we obtained the second best result after the first one, which has a score of 65.957%, which represents a difference of 9.281% with respect to our approach.

Table 3
Results obtained in the ranking table with different configurations of the applied approach.

Approach	Score
BETO	55.690
BETO + POS	54.881
BETO + POS + Dep	**56.778**
MarIA	54.838
MarIA + POS	55.094
MarIA + POS + Dep	55.953
MarIA-large	51.047
MarIA-large + POS	56.373
MarIA-large + POS + Dep	**56.712**

For Task 2, which is a multi-class classification task with the goal of identifying the reliability of 5W1H entities, we evaluated the contextual fine-tuning approach of pre-trained models in Spanish. Unlike normal fine-tuning, where only the text of the 5W1H entities is added as input, here the full sentence is also added as context, which helps the model to better understand the entity and its reliability. In addition to BETO and MarIA, we also evaluated models such as BERTIN, ALBETO, and DistilBETO for this task. BERTIN is a RoBERTa-based model for Spanish, trained from scratch on the Spanish part of mC4 using Flax. ALBETO is a version of ALBERT, which in turn is a lightweight version of BERT, pre-trained only on documents written in Spanish. DistilBETO is a reduced version of BERT, trained using distillation techniques to transfer the BETO weights to a new model with fewer layers and less complexity.

Table 4 shows the results obtained from different models pre-trained with the contextual fine-tuning approach. We can see that the MarIA model shows the best overall performance, with the highest score in Macro F1 (69.1013%). BETO also shows a solid performance, closely followed by BERTIN and DistilBETO. However, ALBETO, being a lighter model, has the worst performance in terms of Macro F1, with the lowest score (51.4926%).

From Table 5, which shows MarIA's classification report, we can see that it has a solid performance in classifying the *confiable* class, with an M-F1 of 87. However, it has difficulties in correctly classifying the *no confiable* and *semiconfiable* classes, obtaining an M-F1 of 65 and 55 respectively. This may be because there are fewer examples for the *semiconfiable* and *no confiable* classes in the training set, which has led to a bias towards the *confiable* class.

Table 6 shows the official ranking. We can see that our approach, based on the contextual fine-tuning of the MarIA model, obtained the best result with a score of 65.82, beating the other teams by a difference of more than 5%.

Table 4
Results obtained in validation split with different approaches based on context fine-tuning of pre-trained language models. Macro Precision (M-P), Macro Recall (M-R) and Macro F1 (M-F1) are displayed and M-F1 is used as a reference metric.

Model	M-P	M-R	M-F1
BETO	70.1515	63.1256	65.6713
MarIA	**70.8529**	**69.8686**	**69.1013**
BERTIN	69.8333	66.6160	66.6160
ALBETO	62.2044	55.5012	51.4926
DistilBETO	67.7425	66.7431	67.2153

Table 5
Classification report of MarIA model in validation split.

	Precision	Recall	F1-score
confiable	91.4648	83.2109	87.1429
no confiable	75.0000	56.9832	64.7619
semiconfiable	46.0938	69.4118	55.3991
macro avg	**70.8529**	**69.8686**	**69.1013**
weighted avg	**80.9985**	**77.2891**	**78.4184**

Table 6
Official leaderboard for Task 2.

#	Team Name	Score
1	**UMUTeam**	**65.820**
2	Michael Ibrahim	59.658
3	Elena	42.917
4	Jiayun	29.063

5. Conclusion

This article summarizes UMUTeam's participation in the FLARES shared task at IberLEF 2024. This task focuses on the use of the "5W1H" technique, a technique commonly used by journalists to clearly and explicitly present the key information of a news item, in order to systematically assess the reliability of language. The task is divided into two sub-tasks: 5W1H identification and 5W1H-based reliability.

In this shared task, we participated in Task 1 and Task 2. For Task 1, which aims to identify and annotate answers to the 5W1H questions in the text, we evaluated various configurations of an approach using pre-trained Transformers language models to develop a custom NER model capable of classifying specific 5W1H entities in the text. We created two additional inputs for the model: Part-Of-Speech (POS) and dependency (Dep), which contain information about the part-of-speech and syntactic dependency of the text. These additional inputs are added as an additional embedding layer to improve the performance of the NER model. With the BETO+POS+Dep model, we obtained the second-best result after the first with a score of 56.778. For Task 2, which is a multi-class classification task aimed at identifying the reliability of 5W1H entities, we evaluated the contextual fine-tuning approach of pre-trained models in Spanish. In the test split, the MarIA model achieved the best result with a score of 65.820.

As a future line of research, we propose to incorporate Conditional Random Fields (CRF) into the architecture of our system and evaluate other pre-trained language models for Task 1. For Task 2, we plan to add features such as source credibility, clarity, and precision, among others, to improve the performance of the model. Besides, future work could explore the relationship between the 5Ws and author profiling. This approach will help identify specific patterns in how different authors present the 5Ws, revealing potential biases, thematic preferences, and characteristic writing styles of each author profile [13].

Acknowledgments

This work is part of the research projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way of making Europe and LT-SWM (TED2021-131167B-I00) funded by MICIU/AEI/10.13039/ 501100011033 and by the European Union NextGenerationEU/PRTR, and "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] A. Bonet-Jover, R. Sepúlveda-Torres, E. Saquete, P. Martínez-Barco, A. Piad-Morffis, S. Estevez-Velarde, Applying Human-in-the-Loop to construct a dataset for determining content reliability to combat fake news, Engineering Applications of Artificial Intelligence 126 (2023) 107152. URL: https://www.sciencedirect.com/science/article/pii/S0952197623013362. doi:https://doi.org/10.1016/j.engappai.2023.107152.

[2] 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.

[3] R. Sepúlveda-Torres, A. Bonet-Jover, I. Diab, I. Guillén-Pacho, I. Cabrera-de Castro, C. Badenes-Olmedo, E. Saquete, M. T. Martín-Valdivia, P. Martínez-Barco, L. A. Ureña-López, Overview of FLARES at IberLEF 2024: Fine-Grained Language-based Reliability Detection in Spanish News, Procesamiento del Lenguaje Natural 73 (2024).

[4] 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.

[5] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish Pre-Trained BERT Model and Evaluation Data, in: PML4DC at ICLR 2020, 2020.

[6] A. G. Fandiño, 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.

[7] A. Bonet-Jover, R. Sepúlveda-Torres, E. Saquete, P. Martínez-Barco, M. Nieto-Pérez, RUN-AS: a novel approach to annotate news reliability for disinformation detection, Language Resources and Evaluation (2023) 1–31.

[8] J. A. Garcia-Díaz, Á. Almela, F. García-Sánchez, G. Alcaraz-Mármol, M. J. Marín, R. Valencia-García, Overview of FinancES 2023: Financial Targeted Sentiment Analysis in Spanish, Procesamiento del Lenguaje Natural 71 (2023) 417–423.

[9] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[10] J. Canete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, ALBETO and DistilBETO:

Lightweight Spanish Language Models, in: International Conference on Language Resources and Evaluation, 2022. URL: https://api.semanticscholar.org/CorpusID:248266434.

[11] R. Pan, J. A. García-Díaz, F. García-Sánchez, R. Valencia-García, Evaluation of transformer models for financial targeted sentiment analysis in Spanish, PeerJ Comput. Sci. 9 (2023) e1377. URL: https://doi.org/10.7717/peerj-cs.1377. doi:10.7717/PEERJ-CS.1377.

[12] 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.

[13] 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.