Syntax Savants-UA at IberLEF 2024: Leveraging FLAN-T5-XXL for Automatic 5W1H Identification in Texts

Eduardo Grande[1], Ahmed Begga[2]

[1]Department of Software and Computing Systems, University of Alicante, Spain
[2]Department of Computer Science and AI, University of Alicante, Spain

Abstract

The 5W1H technique involves formulating six questions to briefly capture the basic information of a news text: What, When, Where, Who, Why, and How. This method enables an easy understanding of the news. This work, part of the FLARES competition at the Iberian Languages Evaluation Forum 2024, presents a Natural Language Processing (NLP) model, specifically FLAN-T5-XXL, designed to automatically extract 5W1H elements with an F1-score of 0.543. Our methodology includes creating task-specific templates and prompts and then fine-tuning the FLAN-T5-XXL model using annotated Spanish news articles provided by the competition organizers. We leverage Low-Rank Adaptation (LoRA) for efficient fine-tuning and conduct comprehensive hyperparameter optimization to enhance model performance and output generation. The evaluation demonstrates that FLAN-T5-XXL significantly outperforms other models. This study underscores the potential of Generative NLP models in automating information extraction from news texts.

Keywords

Natural Language Processing, Large Language Models, Information extraction, 5W1H

1. Introduction

The advent of digital media has led to a dramatic surge in the volume of news articles [1]. These articles contain a wealth of pertinent information such as events, people, reasons, times, places, and methods, among others [2, 3]. Consequently, the mining of news text is increasingly being utilized in various domains due to its practical applications in credibility prediction, bridging the gap between unstructured and structured information to provide richer datasets, and identifying potential biases or misrepresentations [4, 5].

In journalism, numerous studies have demonstrated that narrative text is more expressive, allowing journalists to accurately describe events [6]. Therefore, there is a demand to develop tools that process this valuable narrative text and extract useful knowledge to assist journalists [7]. In this regard, knowledge extraction is a subdomain of Natural Language Processing (NLP) used computationally to analyze this unstructured textual data and extract structured information [8]. This technique aims to identify related journalistic concepts, which is a cornerstone in identifying journalistic information, organizing data into suitable representation for efficient analysis, and concentrating the usable knowledge dispersed by journalists into various format files like news reports and articles [9].

In this context, FLARES proposes a challenge [10] inside the IberLEF 2024 competition [11]: a set of shared sub-tasks that demands the development of tools for Fine-Grained Language-based Reliability Detection in Spanish News. In this work, we focus on the first sub-task of 5W1Hs identification, which aims to locate and classify the WHAT, WHO, WHY,

IberLEF 2024, September 2024, Valladolid, Spain
eduardo.grande@ua.es (E. Grande); ahmed.begga@ua.es (A. Begga)
https://cvnet.cpd.ua.es/curriculum-breve/es/grande-ruiz-eduardo/327690 (E. Grande); https://github.com/AhmedBegggaUA (A. Begga)
0000-0002-4894-3943 (E. Grande); 0009-0000-8733-2072 (A. Begga)

WHEN, WHERE, and HOW elements within a text. The proposed system leverages the FLAN-T5-XXL [12] model, a state-of-the-art deep learning neural network architecture [13], to perform this task.

The remainder of the manuscript is organized as follows. Section 2 describes the FLARES's sub-task of 5W1Hs identification. Section 3 details the FLAN-T5-XXL-based system proposed for the challenge. Section 4 presents and discusses the results achieved in the challenge. Finally, Section 5 summarizes the harvested findings to address this challenge.

2. Task and dataset description

Assessing the reliability of the language used in news writing is becoming increasingly crucial in today's digital media landscape. Identifying specific segments of a news article to gauge linguistic credibility offers a more nuanced understanding of the message's truthfulness. This approach not only enhances our grasp of information presentation but also paves the way for developing more effective techniques in spotting fake or misleading news.

Recent studies [14] have highlighted the importance of analyzing language style, tone, and structure in identifying deceptive content. Style and language have proven valuable in distinguishing between fake and true articles, and specific linguistic features have indicated potential biases or misrepresentations in online content. These studies underscore the emerging significance of leveraging linguistic analysis to discern trustworthy news in the digital age.

In this context, we propose harnessing the "5W1H" technique, commonly employed by journalists to clearly present the key information of a news item in an explicit way. This method focuses on identifying the WHAT, WHO, WHY, WHEN, WHERE, and HOW elements within a piece of news. By applying this technique, we can systematically evaluate the reliability of the language across these dimensions. Analyzing the presence of these fundamental journalistic issues offers a structured approach to calibrate the linguistic integrity and possible biases of the content. Moreover, our challenge will utilize texts in Spanish, aiming to progress on the development of techniques of this nature specifically tailored for this language [15]. This integration of journalistic methodology with linguistic analysis not only provides a comprehensive framework but also could pave the way for enhancing the authenticity and trustworthiness of information in the Spanish digital media landscape.

The FLARES dataset [10] consists of Spanish news articles sourced from various digital media outlets, covering a wide range of topics such as politics, economy, health, technology, and culture. This diversity ensures a broad spectrum of linguistic styles and structures, providing a rich resource for analyzing language reliability.

Each article in the dataset is annotated with the "5W1H" elements, highlighting the WHAT, WHO, WHY, WHEN, WHERE, and HOW aspects of the content. Additionally, annotations include markers of linguistic credibility, such as tone, style, and indicators of potential bias. These annotations support the development and testing of models aimed at detecting the fine-grained linguistic reliability of news content.

To ensure high quality and consistency, the articles are curated from reputable sources and manually reviewed by experts proficient in Spanish. This review process guarantees that the annotations accurately reflect essential journalistic elements and credibility indicators. The dataset serves as a valuable resource for advancing research in linguistic reliability detection and for fostering the creation of more reliable and trustworthy Spanish digital media content.

Table 1 presents the distribution of the 5W1H elements in the training. Each text can contain multiple instances of the 5W1H elements, and also more than one specific 5W1H.

Table 1
Distribution of 5W1H Elements in the Training.

Element	Training Set
Number of texts	1,585
WHAT	2,711
WHERE	1,024
WHEN	778
WHO	533
HOW	563

This comprehensive annotation approach helps in systematically evaluating and improving the reliability of news content in Spanish.

3. Methodology

To assess the linguistic reliability of Spanish news articles, we implemented a comprehensive methodology that integrates template creation, prompt engineering, and fine-tuning a Large Language Model (LLM).

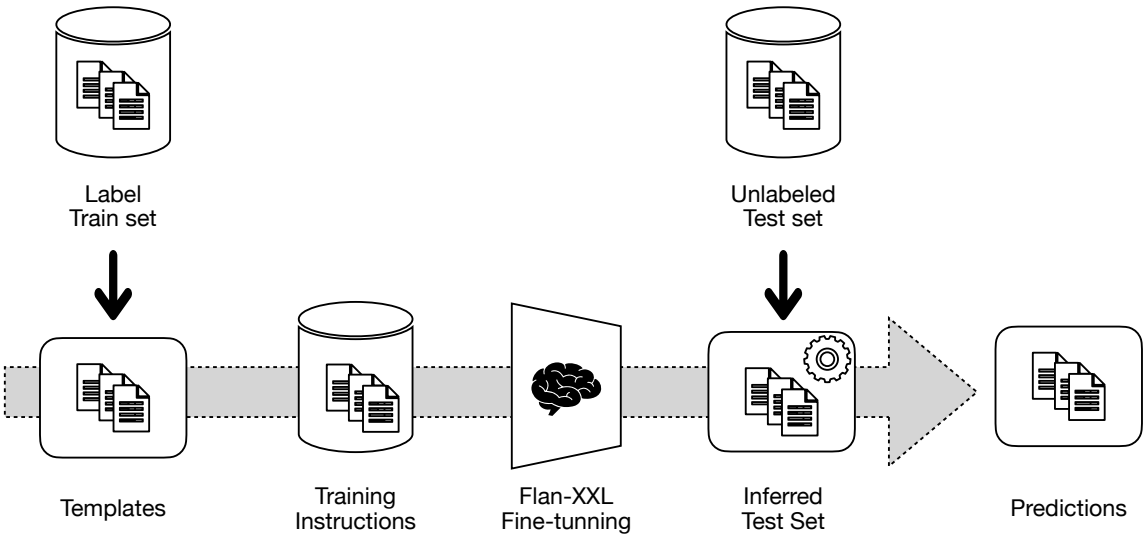

Figure 1: Overview of the Methodology for Fine-Grained Language-based Reliability Detection.

The individual steps detailed shown in Figure 1 are explained below.

3.1. Template creation

Following the procedure established by Google when creating Flan [12], we defined a set of templates for this specific task of extracting 5W1H information.

As will be commented under the subsequent sections, we created 6 models, one per category. So, when defining a template, we defined the same one as a set of 6 templates. In total, we generated both manually and using ChatGPT 3.5 [16] a total of 25 sets of templates.

```
{
    "templates": {
        "1":
        ["Dado el texto: \"{text}\", la W What es: \"{W_what}\"",
        "Dado el texto: \"{text}\", la W Who es: \"{W_who}\"",
        "Dado el texto: \"{text}\", la W Where es: \"{W_where}\"",
        "Dado el texto: \"{text}\", la W When es: \"{W_when}\"",
        "Dado el texto: \"{text}\", la W Why es: \"{W_why}\"",
        "Dado el texto: \"{text}\", la H How es: \"{W_how}\""],
        ...
        ...
}
```

Figure 2: Example of a set of a template.

El método 5W + 1H es una técnica de análisis que se utiliza para obtener una comprensión profunda de un problema o situación, identificando seis preguntas clave: qué (What), quién (Who), dónde (Where), cuándo (When), por qué (Why) y cómo (How). Al responder estas preguntas de manera exhaustiva, podemos obtener una visión completa del problema y establecer una base sólida para su resolución. El What describe la acción o evento principal que se está investigando o describiendo, proporcionando la base de la comprensión. Es la parte central que responde a la pregunta sobre qué está ocurriendo o qué se está discutiendo. Dado el texto: "Dos días, exactamente han pasado dos días desde que Sánchez compareciera en rueda de prensa en la Moncloa afirmando que a España llegarían, entre abril y septiembre, un total de 87 millones de vacunas para darnos cuenta de que las mentiras de Sánchez hacen bueno ese refrán que dice que 'la mentira tiene las patas muy cortas'", la W What es:

Figure 3: Example of a full constructed prompt with pre-prompt, description of the W What and a template having mixed it with a piece of data.

These templates will then be mixed with the provided data to generate a prompt that will be the input for the model. Figure 2 shows one template set example.

In order to test more strategies, some more templates were defined. The defined templates were not just for mixing them with data and creating input prompts for the model. Several templates were created to be appended before the main prompt so that the prompts could finally contain more information about the task to perform. These templates are:

- Pre-prompts: We call a pre-prompt to a general template that explains what the 5W1H are. The aim is that by explaining to the model this it will improve the results.
- Description of W: Whereas the pre-prompts were a general description of the 5W1H method, the descriptions are a concrete description of a single W/H. They describe what type of answer should respond to a corresponding W/H.

Figure 3 shows a complete example of a prompt that contains a pre-prompt, a description, and a template, mixed with an element of the data.

3.2. Input data preparation

Once all the templates have been defined, the input data preparation process consists of merging those templates with the given annotated corpus.

The prompts generated are zero-shot, meaning that a single input prompt contains just the information of the templates and the given text, without any examples of answers, as already shown in Figure 3.

For merging the data, the templates are iterated over and over until each piece of data has a single template. If pre-prompts and/or descriptions are appended, these are chosen randomly among the defined ones.

The prepared input prompts were saved into a *json* file that could be easily loaded when training the models.

3.3. Model fine-tuning

The fine-tuning of the FLAN-T5-XXL [17] model was conducted using specific parameters tailored to optimize performance on the given task. The fine-tuning process was aimed at adapting the model to effectively answer the 5W1H questions.

To achieve this, we configured the following hyperparameters: a learning rate of 1×10^{-3} was selected to control the adjustment of the model weights during training. We set the batch size to 32, allowing the model to process 6 samples per batch. The training process was conducted over 1 epoch, which was sufficient to allow the model to learn the patterns in the training data without overfitting. All these parameters were selected after performing a hyperparameters search so that the values selected are the best configuration set for improving the model performance. The maximum input length was capped at 1024 tokens to ensure that the model could handle lengthy input sequences, and the maximum output length was set to 400 tokens to accommodate detailed responses.

To further enhance the model's performance, we implemented Low-Rank Adaptation (LoRA) [18]. LoRA is an efficient fine-tuning technique that injects trainable rank decomposition matrices into each layer of the transformer's architecture. This method helps in optimizing the model without drastically increasing the computational load. For our task, we set the rank of the decomposition matrices (r) to 8 and used a scaling factor ($lora\_alpha$) of 16 for the LoRA layers. We targeted the query (q) and value (v) matrices within the transformer layers, as these are critical components in the model's attention mechanism. A dropout rate of 0.05 was applied to the LoRA layers to prevent overfitting and enhance the model's generalizability. The task type for the LoRA configuration was specified as sequence-to-sequence language modelling, which is appropriate for generating text responses based on input prompts.

These configurations were instrumental in adapting the pre-trained FLAN-T5-XXL model to our specific requirements, ensuring that it could effectively process and respond to 5W1H questions.

3.4. Inference and output parsing

During the inference phase, we conducted a comprehensive hyperparameter search to determine the optimal settings for generating the best possible results. The hyperparameters we varied included the number of beams for beam search, the use of early stopping, whether to sample, the top-p and top-k sampling parameters, and the temperature. The specific values tested are summarized in the Table 2.

By systematically varying these parameters, we aimed to find the combination that produced the most accurate and coherent answers to the 5W1H questions. The number of beams controlled the breadth of the search space, with higher values providing more candidate sequences but also increasing computational complexity. Early stopping was tested to determine if halting the search process early when a complete sequence was found would yield better or faster results. Sampling parameters such as do_sample, top_p, and top_k were adjusted to balance between deterministic outputs and diverse, creative

Table 2
Hyperparameters and their values used in the inference phase.

Hyperparameter	Values
num_beams	3, 4, 5, 6, 7, 8
early_stopping	True, False
do_sample	True, False
top_p	0.5, 0.7, 0.9, 1
top_k	50, 100, 200
temperature	0.5, 0.6, 0.7, 0.8, 0.9, 1

responses. The temperature parameter controlled the randomness of predictions, with lower values making the model more conservative and higher values introducing more variability.

Once the inference was complete, the results were parsed to isolate the positions of the 5W1H answers. This was a crucial step because our evaluation metric relied on accurately identifying the spans of the text where the answers were located. For instances where no valid 5W1H answer was found, we formatted the response as an empty string instead of the placeholder "No answer". This ensured consistency and allowed for seamless integration with our evaluation pipeline.

By recording the start and end positions of each 5W1H answer within the generated text, we were able to precisely evaluate the model's performance. This approach allowed us to not only assess the accuracy of the responses but also understand how well the model could locate relevant information within a given context. The detailed hyperparameter tuning and meticulous output parsing were essential to optimize the model's capability to provide reliable and contextually appropriate answers to the 5W1H questions.

4. Results

To evaluate the performance of our model, we used the metrics proposed by [19]. This evaluation method defines five categories of matches for span classification: correct, partial, missing, incorrect, and spurious matches.

- **Correct matches**. Reported when a text in the predicted file matches exactly with a corresponding text span in the gold file, including the start and end index, and the 5W1H label. Only one correct match per entry in the gold file can be counted.
- **Incorrect matches**. Reported when the start and end indices match, but not the 5W1H type.
- **Partial matches**. Reported when two intervals [start, end] have a non-empty intersection and match the 5W1H label. A partial phrase will only be matched against a single correct phrase.
- **Missing matches**. Those that appear in the gold file but not in the predicted file.
- **Spurious matches**. Those that appear in the predicted file but not in the gold file.

These metrics were used to ensure a comprehensive and accurate evaluation of the model's ability to correctly identify and classify the spans related to the 5W1H questions.

The results obtained by following the pipeline outlined in Figure 1(with the exception that we tested other LLM models apart from FLAN-T5-XXL) are shown in Table 3. Our best result on the test set was achieved with the FLAN-T5-XXL model, with a score of 54.256%. Other models such as FLAN-T5-Large [12], Bloom-7B1 [20], Bloomz-7B1 [21] and Llama-3-8B-Instruct [22] were also tested for comparison.

Table 3
Performance of various LLM models on the 5W1H Task.

Model	Score (%)
FLAN-T5-XXL	54.256
FLAN-T5-Large	44.222
Bloomz-7B1	43.875
Bloom-7B1	29.338
Llama-3-8B-Instruct	13.606

These results demonstrate that the FLAN-T5-XXL model significantly outperformed the other models in this task, showcasing its ability to effectively understand and respond to 5W1H questions.

5. Conclusions

In this study, we explored the fine-tuning and evaluation of various LLMs for the task of answering 5W1H questions. Our approach involved detailed data preparation, systematic hyperparameter tuning during inference, and precise output parsing to ensure robust evaluation.

The FLAN-T5-XXL model emerged as the top performer, significantly surpassing other tested models, including FLAN-T5-Large, Bloom-7B1, Bloomz-7B1, and Llama-3-8B-Instruct. This superior performance underscores the efficacy of advanced fine-tuning techniques and the importance of hyperparameter optimization in enhancing model capabilities.

By using the comprehensive evaluation metrics proposed in the previous section, we ensured a rigorous assessment of our models, considering various match types such as correct, partial, missing, incorrect, and spurious. This thorough evaluation methodology provided clear insights into the strengths and limitations of each model.

The results demonstrate the potential of LLMs, particularly FLAN-T5-XXL, in effectively addressing complex NLP tasks that require a deep understanding and generation of text based on specific prompts. Future work could focus on further refining these models, exploring additional fine-tuning techniques, and extending the evaluation to broader datasets and more diverse question types.

Overall, this study highlights the critical role of fine-tuning and careful evaluation in maximizing the performance of large language models, paving the way for more accurate and contextually aware AI systems.

Acknowledgments

This publication of the author Eduardo Grande is part of the grant PRE2022-101573, funded by MCIN/AEI/10.13039/501100011033 and the ESF+.

The author Ahmed Begga is funded by the project PID2022-142516OB-I00 of the Spanish Government.

References

[1] J. Lugea, Linguistic Approaches to Fake News Detection, 2021, pp. 287–302.

[2] B. D. Horne, S. Adali, This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, CoRR abs/1703.09398 (2017). URL: http://arxiv.org/abs/1703.09398. arXiv:1703.09398.

[3] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, W. Quattrociocchi, Science vs conspiracy: Collective narratives in the age of misinformation, PLoS ONE 10 (2014). URL: https://api.semanticscholar.org/CorpusID:2274379.
[4] F. Zollo, P. Kralj Novak, M. Del Vicario, A. Bessi, I. Mozetic, A. Scala, G. Caldarelli, W. Quattrociocchi, Emotional dynamics in the age of misinformation, PLoS ONE 10 (2015). doi:10.1371/journal.pone.0138740.
[5] A. Bessi, A. Scala, L. Rossi, Q. Zhang, W. Quattrociocchi, The economy of attention in the age of (mis)information, Journal of Trust Management 1 (2014). doi:10.1186/s40493-014-0012-y.
[6] A. Noain-Sánchez, Addressing the impact of artificial intelligence on journalism: the perception of experts, journalists and academics, Communication Society 35 (2022) 105–121. doi:10.15581/003.35.3.105-121.
[7] M. Túñez López, C. Fieiras-Ceide, M. Vaz-Álvarez, Impact of artificial intelligence on journalism: transformations in the company, products, contents and professional profile, Communication Society 34 (2021) 177–193. doi:10.15581/003.
[8] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen, Y. Zhao, S. Sohn, H. Liu, Clinical concept extraction: A methodology review, Journal of Biomedical Informatics 109 (2020) 103526. URL: https://www.sciencedirect.com/science/article/pii/S1532046420301544. doi:https://doi.org/10.1016/j.jbi.2020.103526.
[9] J. Templeton, I. Timmis, A Flexible Framework for Integrating Data-Driven Learning, 2023, pp. 39–58. doi:10.1007/978-3-031-11220-1_3.
[10] 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).
[11] 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.
[12] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, 2022. arXiv:2210.11416.
[13] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2023. arXiv:2303.18223.
[14] B. D. Horne, S. Adali, This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, 2017. arXiv:1703.09398.
[15] 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.
[16] OpenAI, ChatGPT 3.5, https://openai.com/chatgpt/, 2022. Accessed: June 7, 2024.
[17] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022. URL: https://openreview.net/forum?id=gEZrGCozdqR.
[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. arXiv:2106.09685.
[19] A. Piad-Morffis, Y. Gutiérrez, H. Cañizares-Diaz, S. Estévez-Velarde, R. Muñoz, A. Mon-

toyo, Y. Almeida-Cruz, Overview of the ehealth knowledge discovery challenge at iberlef 2020, 2020.

[20] B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).

[21] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf, et al., Crosslingual generalization through multitask finetuning, arXiv preprint arXiv:2211.01786 (2022).

[22] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.