Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 369-379 recibido 01-07-2024 revisado 19-07-2024 aceptado 29-07-2024

Overview of FLARES at IberLEF 2024: Fine-grained Language-based Reliability Detection in Spanish News

Resumen de FLARES en IberLEF 2024: Detección detallada de la confiabilidad en el lenguaje de noticias en español

Robiert Sepúlveda-Torres,[1] Alba Bonet-Jover,[1] Isam Diab,[2] Ibai Guillén-Pacho,[2] Isabel Cabrera-de Castro,[3] Carlos Badenes-Olmedo,[2] Estela Saquete,[1] M.Teresa Martín-Valdivia,[3] Patricio Martínez-Barco,[1] L.Alfonso Ureña-López,[3]
[1]Department of Software and Computing Systems, University of Alicante, Spain
[2]Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
[3]Computer Science Department, University of Jaén, Spain
{alba.bonet,rsepulveda,stela,patricio}@dlsi.ua.es
{carlos.badenes,ibai.guillen,isam.diab}@upm.es
{iccastro,maite,laurena}@ujaen.es

Abstract: This paper presents FLARES, a shared task organised in the framework of the evaluation campaign of Natural Language Processing systems in Spanish and other Iberian languages, IberLEF 2024. FLARES aims to detect patterns of reliability in the language used in news that will allow the development of effective techniques for the future detection of misleading information. To this end, the 5W1H journalistic technique for detecting the relevant content of a news item is proposed as a basis, as well as an annotation guideline designed to detect linguistic reliability. Two subtasks are proposed: the first focusing on the identification of the 5W1H elements and the second focusing on the detection of reliability. A total of 7 participants registered in the shared task, of which 3 participated in the first subtask and 4 in the second. The teams proposed various approaches, especially based on fine-tuning of encoding models and adjustment of instructions in decoding models.
Keywords: Natural Language Processing, Reliability Detection, Quality Information, Language Models.

Resumen: Este artículo presenta FLARES, una tarea compartida organizada en el marco de la campaña de evaluación de sistemas de Procesamiento del Lenguaje Natural en español y otras lenguas ibéricas, IberLEF 2024. FLARES tiene como objetivo detectar patrones de confiabilidad en el lenguaje utilizado en las noticias que permita desarrollar técnicas eficaces para la futura detección de información engañosa. Para ello, se propone como base la técnica periodística de las 5W1H para detectar el contenido relevante de una noticia, así como una guía de anotación diseñada para detectar la confiabilidad lingüística. Se proponen dos subtareas: la primera centrada en la identificación de los elementos 5W1H y la segunda en la detección de la confiabilidad. Un total de 7 participantes se registraron en la tarea compartida, de los cuales 3 participaron en la primera subtarea y 4 en la segunda. Los equipos propusieron diversos enfoques, especialmente basado en el ajuste de modelos de codificación y en el ajuste de instrucciones en modelos de decodificación.
Palabras clave: Procesamiento del Lenguaje Natural, Detección de Confiabilidad, Calidad de la Información, Modelos de Lenguaje.

1 Introduction

The disinformation phenomenon has become a global social problem and a challenge for the research community. As stated by Saquete et al. (2020), "since assessing the veracity of a news story is complex from an engineering point of view, the research community is approaching this task from different perspectives". Among them, it is worth highlighting four approaches based on: i) content

ISSN 1135-5948 DOI 10.26342/2024-73-28

(Zhou and Zhang, 2008), ii) context (Bani-Hani et al., 2020), iii) knowledge (Vlachos and Riedel, 2014) and iv) hybrid approaches (Seddari et al., 2022). The task presented in this paper addresses the content-based approach with the objective of analysing the influence and intention of language (its reliability) when presenting information in text, i.e. the ability of the way in which content is expressed to make it more or less credible.

Assessing the reliability of the language used in news writing is becoming increasingly crucial in today's digital media landscape, since it is a metric that allows to measure the quality of information. Identifying specific segments of a news article to gauge the 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 the development of more effective techniques in spotting fake or misleading news.

Recent studies have delved into this approach, highlighting the importance of analysing language style, tone, and structure in identifying deceptive content (Lugea, 2021). Style and language are features that have proven valuable in distinguishing between fake and true articles (Horne and Adali, 2017), and specific linguistic features have proven valuable in indicating 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 order to analyse the reliability of the language and to detect linguistic patterns that allow for the automatic detection of misleading information, we propose a shared task which harness the 5W1H technique commonly employed by journalists to clearly present the key information of a news item in an explicit way (Zhang, Chen, and Ma, 2019). This method focuses on identifying the What, Who, Why, When, Where, and How elements within a text.

By applying this technique, we can systematically evaluate the reliability of the language across these dimensions. Analysing the presence of these fundamental journalistic questions offers a structured approach to gauge the linguistic integrity and potential biases within the content. Moreover, our challenge will utilise texts in Spanish, aiming to advance techniques of this nature specifically tailored for this language. This integration of journalistic methodology with linguistic analysis not only provides a comprehensive framework but can also pave the way for enhancing the authenticity and trustworthiness of information in the Spanish digital media landscape.

2 Task description

The shared task proposed is entitled "FLARES: Fine-grained Language-based Reliability Detection in Spanish News"[1] and is organised by members of three groups specialised in Natural Language Processing (NLP): the GPLSI group (Language and Information Systems Group) from the University of Alicante; the OEG group (Ontoly Engineering Group) from Universidad Politécnica de Madrid; and the SINAI group (Intelligent System for Information Access) from the University of Jaén. Regarding the expected target community, this task is addressed to NLP researchers, media professionals, journalists, and policymakers, since the initiative offers tools to enhance content quality and accuracy.

The FLARES shared task is divided into two subtasks and managed on the Kaggle Platform[2]. Task description, teams, evaluation metrics, and other details can be consulted in the following Kaggle links:

- FLARES: Subtask 1 (5W1H identification)[3];
- FLARES: Subtask 2 (Reliability classification)[4].

2.1 Subtask 1: 5W1H identification

In journalism, there is a technique that allows to communicate a news item in a precise and complete way according to quality standards. This strategy is known as the 5W1H and consists of answering six key questions: Who?, What?, When?, Where?, Why? and How?.

[1] https://sites.google.com/gcloud.ua.es/flares/

[2] Kaggle is an online platform specialised in data science and machine learning competition.

[3] https://www.kaggle.com/competitions/flares-subtask-1-5w1hs-identification

[4] https://www.kaggle.com/competitions/flares-subtask-2-5w1hs-classification

The 5W1H "clearly describe key information of news in an explicit manner" (Zhang, Chen, and Ma, 2019) and "represents the semantic constituents of a sentence which are comparatively simpler to understand and identify" (Chakma et al., 2020). Information following this technique may be a reliability pattern, since a news item communicated in a complete and accurate way may have a higher degree of credibility than a news item that communicates the information with more opacity. Some research in literature has already used this approach specially applied to event extraction and semantic role labelling tasks, such in the cases of Keith, Horning, and Mitra (2020), Chakma and Das (2018) and Khodra (2015).

Following this approach, the objective of Subtask 1 is to identify all the elements proposed by the 5W1H technique in order to be able to subsequently analyse the reliability language of the segmented content. Thus, participants were provided with a text, and they had to determine and segment the essential content by annotating the answers to the 5W1H questions of the document.

2.2 Subtask 2: 5W1H-based reliability

This work focuses on assessing misleading information from a content and linguistic approach based on the reliability concept. The disinformation phenomenon is a concept that has been widely used in recent years in different contexts and tasks. Although the term "disinformation" has been used as a common term to refer to deceptive information, numerous nomenclatures have been used when addressing this problem, such as "toxicity", "information disorder", "hoax", "fake news", "veracity", "reliability", among others, all closely related concepts that allude to disinformation, but with different shades of meaning. For example, the fake news concept is often considered a politicised concept, "often used by the public, politicians, and the news media to attack news, journalists, and news outlets deemed to be problematic" (Grieve and Woodfield, 2023).

As explain in the Introduction, the FLARES task aims to address the reliability concept from a linguistic approach. Unlike the veracity concept, which is usually used in verification tasks (Vosoughi, Roy, and Aral, 2018), the reliability concept is more often related to the credibility of the source (Zhou and Zafarani, 2020) and to textual indicators such as ambiguity, opacity, emotion, intention or stylistic features.

Taking into account the relevance of the reliability concept for the FLARES research, Subtask 2 seeks to automatically detect the reliability of the content elements. To that end, participants had to determine if the language used in each item was "*Confiable*" (Reliable), "*Semiconfiable*" (Partially reliable) or "*No confiable*" (Unreliable) for each 5W1H detected. This had to be achieved by following an annotation guideline that takes into account linguistic aspects based on the reliability concept, such as the accuracy or the neutrality of the content (Bonet-Jover et al., 2024).

2.3 Evaluation and metrics

For the evaluation of both tasks, the F_1 score has been used, which requires first calculating the *Precision* and *Recall*. These last two metrics are implemented differently for each subtask.

Subtask 1: For the 5W1H span identification we employ the evaluation proposed by Piad-Morffis et al. (2020). Five span matching scenarios form the basis of this evaluation: *correct* (C), when the span and the label match the gold standard[5]; *incorrect* (I), when the span matches but the label does not; *partial* (P), when the label is correct and the span overlaps the gold standard, i.e. the intersection between spans is neither null nor exact; *missing* (M), when the annotation is present in the gold standard but not in the prediction; and *spurious* (S), when the annotation is present in the prediction but not in the gold standard.

Then, these cases are used to form the metrics *Recall* and *Precision*, shown in eq. (1) and eq. (2), respectively. And finally, with these values, the F_1 is calculated to compare the performance of the different approaches.

$$Recall = \frac{C + \frac{1}{2}P}{C + I + P + M} \quad (1)$$

$$Precission = \frac{C + \frac{1}{2}P}{C + I + P + S} \quad (2)$$

[5]Only one correct match can be reported for each annotation.

Subtask 2: The 5W1H label reliability classification has three candidate labels: "Confiable" (Reliable), "Semiconfiable" (Partially reliable), or "No confiable" (Unreliable). As this is a common classification task, the *Precision* and *Recall* metrics are calculated in a regular way with a confusion matrix to then compute the F_1 score.

3 Data

3.1 Human assessment

The data provided for this task was collected through an annotation process involving three experts from two fields: 2 linguists and 1 sociologist, all three specialists in NLP and with knowledge in the annotation guideline.

The guideline used was the RUN-AS annotation guideline (Bonet-Jover et al., 2024), which enables the detection of the essential parts of a news item together with the reliability of its semantic elements, as well as other linguistic elements of interest that allow finding linguistic patterns of reliability in text, without using external knowledge. The goal of this annotation proposal is to analyse content based on a purely linguistic analysis to find out whether how a news item is structured or written influences its reliability. To find out whether a news item presents objective information and follows journalistic standards, this guidelines enables a three-level annotation: Structure (Inverted Pyramid hypothesis), Content (5W1H technique), and Elements of Interest (typography, quotes). However, for the FLARES task, only the second level was used, that is, the 5W1H level, to focus on the essential content and its reliability.

The labels proposed for annotating this task are What (fact), Who (subject), When (time), Where (place), Why (cause), and How (manner). Using the 5W1H technique, we can break down the sentence "The arrest of the Italian scientist took place by force yesterday in Milan for selling an unauthorised vaccine" as follows:

- What: The arrest
- Who: Italian scientist
- How: by force
- When: yesterday
- Where: in Milan
- Why: for selling an unauthorised vaccine

Along with these labels, the attribute "reliability" is used to classify reliability in language with the values "Reliable", "Partially reliable" and "Unreliable", depending on the two following criteria:

- Accuracy: considers aspects such as vagueness, ambiguity, or lack of evidence.
  - Example: "A long time ago" (inaccurate = unreliable) vs. "On Friday 19 march" (accurate = reliable)
  - Example: "A scientist" (inaccurate = unreliable) vs. "The European Medicines Agency" (accurate = reliable)
- Neutrality: considers personal remarks, emotionally charged content, author's stance, or the objectivity of the title.
  - Example: "In my opinion" (subjective = unreliable) vs. "According to the rector of the University of Alicante" (objective = reliable)
  - Example: "This news item can save your life" (subjective = unreliable) vs. "This news item talks about the positive effects" (objective = reliable)

3.2 Agreements

As can be seen, several labels were considered in the RUN-AS guidelines, so the agreement was calculated separately, label by label. To obtain the agreement, two metrics were considered: Fleiss' kappa and Krippendorff. The first one was used to calculate the level of agreement among the annotators in the choice of labels, while the second was used to test the level of agreement between labelled text spans among annotators. So, originally, to calculate the agreement of structure labels, reliability attributes, and labels of interest, Fleiss' kappa was used. For the content labels (5W1H), both Fleiss' kappa and Krippendorff were used to calculate the agreement. Table 1 shows the agreement obtained for structure labels:

Regarding the content labels, the agreement reached is shown in Table 2:

In the case of the reliability attribute, the agreement obtained can be found in Table 3:

Annotators	Fleiss' kappa
Annot. 1-2	1
Annot. 2-3	1
Annot. 3-1	1
Average	**1**

Table 1: Agreement structure labels.

Annotators	Fleiss' kappa	Krippendorff
Annot. 1-2	1	0.96
Annot. 2-3	1	0.98
Annot. 3-1	1	0.97
Average	**1**	**0.97**

Table 2: Agreement content labels.

Finally, Table 4 shows the agreement obtained with the labels of interest:

3.3 Dataset

The dataset contains 190 news items in Spanish manually annotated with an extraction technique based on the journalistic technique called 5W1H, obtaining approximately 9,034 annotations. This dataset has been split into a proportion of 70% (6,934) for training and 30% (2,100) for testing. Furthermore, the dataset has been divided according to the two proposed tasks. In the first split, only the content labels (5W1H) were taken into account, while in the second split, dedicated to the reliability classification, the reliability attribute assigned to each content label was taken into account. The distribution of the labels 5W1H in the dataset is shown in Table 5. As can be seen, most of the labels annotated are What, followed by Who. This is not surprising since there is an act and, in most cases, a subject that carries out that act. On the other hand, Table 6 shows the distribution of the labels "Reliable", "Partially reliable", and "Unreliable". As can be seen, the values are somewhat unbalanced, predominating the label "Reliable".

The reason why Spanish has been chosen as the language for this task is because, although it is the fourth most widely spoken language in the world and the second mother tongue in the world in terms of number of speakers[6], there are few corpora built in Spanish to address the automatic reliability detection task in NLP.

[6] https://cvc.cervantes.es/lengua/anuario/anuario_23/informes_ic/p01.htm

Annotators	Fleiss' kappa
Annot. 1-2	0.65
Annot. 2-3	1
Annot. 3-1	0.78
Average	**0.81**

Table 3: Agreement reliability attribute.

Annotators	Fleiss' kappa
Annot. 1-2	1
Annot. 2-3	1
Annot. 3-1	1
Average	**1**

Table 4: Agreement labels of interest.

4 Systems and results

4.1 Baselines

This subsection presents two baselines created to evaluate the performance of simple models to address each of the proposed subtasks.

In recent years, language models based on Transformer architecture (Vaswani et al., 2017) have proven effective in addressing NLP tasks such as summarization (Zhao, You, and Liu, 2020) or opinion mining (Abas et al., 2020), among others. For this reason, for the baselines shown in this section, a fine-tuning of the Transformers model was performed. Both baselines can be replicated using this GitHub repository[7].

4.1.1 Subtask 1: 5W1H identification baseline

Subtask 1 consists of extracting key information (a continuous span of words) from a sentence responding to the 5W1H labels. Following this definition we have found the similarity between this task and that of Named Entity Recognition (NER). The difference lies in the fact that entities labelled are not the traditional ones of Person, Location, Organization, and Date.

To create the baseline we selected a fine-tuned model for the Spanish-language NER task (bert-spanish-cased-finetuned-ner[8]). This model was trained on the CONLL corpus in Spanish, obtaining a 90.17 in the F1 metric, which ranks it among the best models in this language to address this task. Subsequently, a second fine-tuning was carried out with the training corpus of Subtask 1. Taking as a starting point the model trained for

[7] https://github.com/rsepulveda911112/Flares_baseline

[8] https://huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner

	WHAT	WHO	WHEN	WHERE	WHY	HOW
Training	2,711	1,843	778	801	238	563
Test	736	482	182	215	61	167
Total	3,447	2,325	960	1,016	299	730

Table 5: Dataset of FLARES: Subtask 1 –5W1H identification–.

	Reliable	Partially Reliable	Unreliable
Training	4,765	1,276	893
Test	806	243	162
Total	5,571	1,519	1,055

Table 6: Dataset of FLARES: Subtask 2 –Reliability classification–.

the NER task allows us to transfer part of its knowledge to the execution of Subtask 1.

We used the common tagging format BIO (Beginning, inside, and outside) for tagging tokens in a chunking task in computational linguistics (e.g. NER). The B- prefix before a label indicates that that label is the beginning of a chunk, and an I- prefix before a label indicates that the label is inside a chunk. An O label indicates that a token belongs to no entity/chunk. Each sentence of the dataset is pre-processed and their annotation is converted to this format.

The pre-processed training data set with the explained format above was used to perform fine-tuning. The following hyperparameter settings were used: learning rate of 1e-5, batch size of 8, maximum sequence length of 512, weight decay of 2.07e-6, and during 5 training epochs.

4.1.2 Subtask 2: Reliability classification baseline

Subtask 2 consists of classifying phrases representing 5W1H entities according to their reliability ("Reliable", "Partially reliable", and "Unreliable"). This task can be approached as a text classification task where data from a sentence is classified into the three aforementioned classes. For them, a simple classifier based on fine-tuning of a model based on Transformers architecture was used.

The context from which the phrase to be classified was extracted is not used for this baseline. We used the RoBERTa model in Spanish (roberta-base-bne)[9] that obtained good results in similar tasks. We fine-tuned the RoBERTa models and a linear layer (with three neurons) was added to perform the classification. This linear layer uses Softmax as the activation function and cross-entropy as the loss function. The following hyperparameter settings were used: learning rate of 2e-5, batch size of 16, maximum sequence length of 512, and during 2 training epochs.

[9]Available for download at `https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne`

4.2 Submitted approaches

Three participants enrolled in Subtask 1 and four in Subtask 2. However, in the second subtask, only two teams presented their working notes. Participating teams will be referred to by the name they used in the competitions published on Kaggle. In general, to address both subtasks, the submitted approaches have based their methodologies on the so-called subsymbolic Artificial Intelligence (AI), making use of techniques based on Machine Learning (ML), specifically Large Language Models (LLMs).

4.2.1 Subtask 1: Approaches

Syntax Savants UA's Solution (Grande and Begga, 2024): This team implemented the FLAN-T5-XXL model (Chung et al., 2022), focusing on the automatic extraction of 5W1H elements from news texts. They created task-specific templates and prompts, followed by extensive fine-tuning of the model using Spanish news articles annotated by the competition organisers. The use of LoRA for efficient fine-tuning and comprehensive hyperparameter optimisation were key aspects of their approach. Their model achieved an F1 score of 0.543, demonstrating its capability in extracting structured information from unstructured news text.

KFloeg's Solution (Pardo et al., 2024): The proposal consisted of a modification of the original K-Adapter (Wang et al., 2020), approaching it as a NER task. Two adapters were tested to enrich the fine-tuning process, one based on the Wikidata relations of the entities, and one based on relations between labels (5W1H) inside each text from the same training corpus. Both adapters were used independently and then jointly, with the 5W1H yielding the best results, reaching a 0.6613 score in the test set, and achieving the winning result of the challenge.

UMUTeam's Solution (Pan et al., 2024):

They developed a NER model to identify 5W1H elements using fine-tuned transformer models such as BETO (Cañete et al., 2023) and MarIA (Gutiérrez-Fandiño et al., 2021), incorporating Part-of-Speech (PoS) and Syntactic Dependency (Dep) features. The model achieved the second-best result with a score of 0.567.

4.2.2 Subtask 2: Approaches

CUFE's Solution (Ibrahim, 2024): This team utilised the Llama-3 model (AI@Meta, 2024) fine-tuned with Parameter-Efficient Fine-Tuning (PEFT) techniques (Mangrulkar et al., 2022), specifically Low-Rank Adaptation (LoRA). They employed the 5W1H technique to annotate news articles and focused on detecting the reliability of news based on language and style without external knowledge. Their methodology involved freezing the Llama-3 weights and updating only the LoRA matrices, supplemented by quantisation and mixed-precision training to optimise resource use. The solution achieved a Macro F1 score of 0.5965, indicating a high level of accuracy and efficiency in classifying news reliability.

UMUTeam's Solution (Pan et al., 2024): Their strategy to address this subtask involved contextual fine-tuning of the MarIA model (Gutiérrez-Fandiño et al., 2021). They used the entity text (5W1H) and the whole sentence as context as model input. In this approach, 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. This approach attained the highest score of 0.6536.

4.3 Results and discussion

Tables 7 and 8 show the results extracted from the competitions published on Kaggle. All participants are included, even if they did not sent working notes. We downloaded the files with the best team submissions from Kaggle and calculated each system's F1 score by class and Macro F1. In addition, the results of the baseline systems are included. The best results are highlighted in bold.

Subtask 1: 5W1H Identification (Figure 1): The 5W1H identification task required teams to accurately identify and classify six key elements in news articles: What, Who, When, Where, Why, and How. The performance of the teams varied significantly across these categories.

KFloeg emerged as the leading team in this subtask, achieving an impressive overall Macro F1 of 0.6613. Their model demonstrated superior performance in five out of the six categories, leading in the identification of What, When, Where, Why, and How elements with F1 scores of 0.6292, 0.7154, 0.7099, 0.5431, and 0.4485 respectively. This indicates that KFloeg's approach was robust and well-rounded, capable of accurately extracting various elements from news texts.

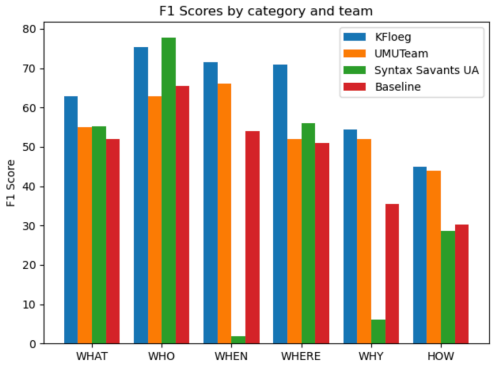

Figure 1: F1 scores by category and team for Subtask 1.

UMUTeam followed with a Macro F1 of 0.567. While they did not lead in any specific category, their performance was consistent across all elements, with F1 scores ranging from 0.4388 to 0.6605. This consistency suggests that their model was well-calibrated and effective across different aspects of the 5W1H framework.

Syntax Savants UA had a mixed performance, achieving a Macro F1 score of 0.5366. They excelled in the Who category with an outstanding F1 score of 0.7782, the highest among all teams for this element. However, their model struggled with the When and Why categories, scoring only 0.0183 and 0.0606 respectively. This indicates a specialisation in identifying Who elements, but a need for improvement in other areas.

In comparison, the **Baseline** system had an overall Macro F1 of 0.529. Both **KFloeg** and **UMUTeam** outperformed the **Baseline**, while **Syntax Savants UA** showed slightly better performance overall but significantly underperformed in specific categories.

Subtask 2: Reliability Classification (Figure 2): The reliability classification task involved determining the reliability of news articles by classifying them as "Reliable", "Partially reliable", and "Unreliable". The teams' performances in this subtask also varied, highlighting different strengths and weaknesses.

UMUTeam stood out with the highest overall Macro F1 score of 0.6536. Their model excelled in classifying "Partially reliable" and "Unreliable" news, with F1 scores of 0.4956 and 0.6344 respectively. This indicates a balanced and effective approach to distinguishing between different levels of reliability.

CUFE performed best in classifying "Reliable" news, achieving a F1 score of 0.8436. However, their overall Macro F1 score was 0.5938, as they struggled with the "Partially reliable" category, scoring only 0.3275. This suggests that while **CUFE** was highly accurate for "Reliable"

	F1 score						Macro F1
	What	**Who**	**When**	**Where**	**Why**	**How**	
KFloeg	**0.6292**	0.7539	**0.7154**	**0.7099**	**0.5431**	**0.4485**	**0.6613**
UMUTeam	0.5491	0.6286	0.6605	0.5196	0.5203	0.4388	0.5670
Syntax Savants UA	0.5523	**0.7782**	0.0183	0.5603	0.0606	0.2868	0.5366
Baseline	0.5194	0.6552	0.5398	0.5089	0.3544	0.3017	0.5290

Table 7: System results of FLARES: Subtask 1 –5W1H identification–.

	F1 score			Macro F1
	Reliable	**Partially reliable**	**Unreliable**	
UMUTeam	0.8308	**0.4956**	**0.6344**	**0.6536**
CUFE	**0.8436**	0.3275	0.6102	0.5938
Elena	0.8105	0.2789	0.4879	0.5258
KFloeg	0.7955	0	0.0909	0.2954
Baseline	0.7989	0.4761	0.5741	0.6164

Table 8: System results of FLARES: Subtask 2 –Reliability classification–.

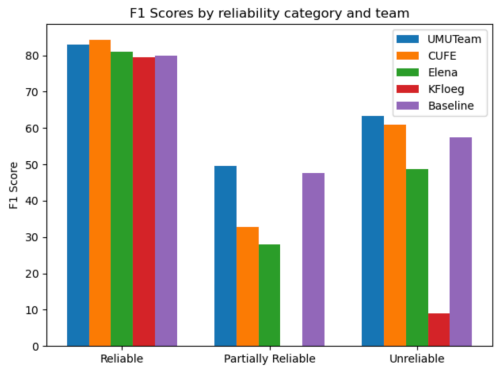

Figure 2: F1 scores by reliability category and team for Subtask 2.

news, it had difficulty discerning "Partially reliable".

ElenaA's model and **KFloeg** underperformed relative to the baseline. **Elena** achieved a Macro F1 score of 0.5258, showing moderate success in the "Reliable" category but low performance in "Partially reliable" and "Unreliable" classifications. **KFloeg** had the lowest overall Macro F1 score of 0.2954, indicating significant challenges across all categories, particularly in classifying "Partially reliable" news where the model scored zero.

The **Baseline** system had a Macro F1 score of 0.6164. **UMUTeam** and **CUFE** both surpassed the **Baseline**, demonstrating the effectiveness of their approaches in reliability classification. The **Baseline** performance highlights the challenge of this task and the need for refined models to improve classification accuracy.

Overall Analysis: The results of the competition reveal several key insights into the current state of NLP models for news classification. **KFloeg**'s and **UMUTeam**'s robust methodologies led to strong performances in their respective tasks, showcasing the potential of advanced fine-tuning techniques and consistent model training. **Syntax Savants UA**'s specialised success in the Who category highlights the importance of targeted model improvements.

Overall, these findings underscore the importance of balancing specialisation and consistency in model development. Future research should focus on hybrid approaches, leveraging the strengths of multiple models to achieve more comprehensive and accurate results across all elements of news classification.

The radar chart (Figure 3) illustrates the overall performance of different teams in both Subtask 1 (5W1H identification) and Subtask 2 (Reliability classification) of the FLARES competition. Each axis represents a specific evaluation category from both subtasks: What, Who, When, Where, Why, How, Reliable, Partially Reliable, and Unreliable.

KFloeg showed strong performance in Subtask 1, particularly excelling in the categories of Who and When, but did not participate in Subtask 2. **UMUTeam** demonstrated balanced performance across both subtasks, with notable strength in the "Reliable" and "Unreliable" categories of Subtask 2. **Syntax Savants UA** excelled in identifying the Who category in Subtask 1 but did not participate in Subtask 2. **CUFE** showed strong performance in Subtask 2, especially in the "Reliable" category, but did not

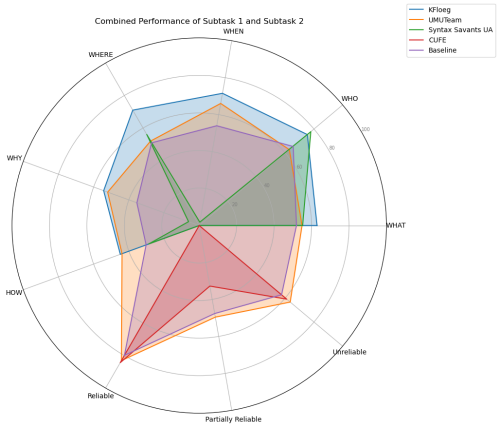

Figure 3: Combined Performance of Subtask 1 and Subtask 2.

participate in Subtask 1. The **Baseline** performance is provided for comparison, highlighting areas where the competing teams outperformed standard benchmarks.

5 Conclusions and future work

This paper presents the description of the first shared task on Fine-grained Language-based Reliability Detection in Spanish News (FLARES at IberLEF 2024). Two subtasks are proposed with 7 participants.

The use of LLMs like Llama-3 and FLAN-T5-XXL, combined with advanced fine-tuning techniques such as LoRA and quantization, showcases significant progress in the automated classification of news reliability. Participants' solutions highlight the effectiveness of these models in handling language-specific tasks, particularly in the context of Spanish news.

Furthermore, some approaches demonstrates the advantages of parameter-efficient fine-tuning, achieving high performance with optimised resource use, while others show comprehensive methodologies, for example involving template creation and extensive hyperparameter tuning, that underscores the importance of task-specific customisation in enhancing model performance.

Future advancements in this line of research could focus on further improving model efficiency and accuracy, potentially exploring hybrid models that combine the strengths of multiple LLMs. Additionally, expanding the dataset diversity and incorporating more nuanced linguistic features could enhance the robustness of news classification systems.

Overall, the comparative analysis of these solutions reveals that while significant strides have been made, there remains substantial scope for innovation and improvement in the automated detection of news reliability using advanced NLP techniques.

Acknowledgments

This research work is part of the R&D&I projects: COOLANG.CONSENSO/TRIVIAL (PID2021-122263OB-C21/PID2021-122263OB-C22) funded by MCIN/AEI/10.13039/501100011033/ and by ERDF A way of making Europe; SOCIAL-FAIRNESS.SOCIALTOX/SOCIALTRUST (PDC2022-133146-C21/PDC2022-133146-C21C22), funded by MCIN/AEI/10.13039/501100011033/ and by the European Union NextGenerationEU/PRTR; NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/21) funded by the Generalitat Valenciana; Project MODERATES (TED2021-130145B-I00); and Public Procurement Assessment in the Healthcare Sector - ProCure (101128437) funded by the European Union NextGenerationEU/PRTR. It is also supported by the Predoctoral Grants PIPF-2022/COM-25947 and PIPF-2022/COM-25762 of the Consejería de Educación, Ciencia y Universidades de la Comunidad de Madrid, Spain.

References

Abas, A. R., I. El-Henawy, H. Mohamed, and A. Abdellatif. 2020. Deep learning model for fine-grained aspect-based opinion mining. *IEEE Access*, 8:128845–128855.

AI@Meta. 2024. Llama 3 model card.

Bani-Hani, A., O. Adedugbe, E. Benkhelifa, M. Majdalawieh, and F. Al-Obeidat. 2020. A semantic model for context-based fake news detection on social media. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEEE.

Bonet-Jover, A., R. Sepúlveda-Torres, E. Saquete, P. Martínez-Barco, and M. Nieto-Pérez. 2024. Run-as: A novel approach to annotate news reliability for disinformation detection. *Language Resources and Evaluation*, 58(2):609–639.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2023. Spanish pre-trained bert model and evaluation data. *arXiv preprint arXiv:2308.02976*.

Chakma, K. and A. Das. 2018. A 5w1h based annotation scheme for semantic role labeling of english tweets. *Computación y Sistemas*, 22(3):747–755.

Chakma, K., S. D. Swamy, A. Das, and S. Debbarma. 2020. 5w1h-based semantic segmentation of tweets for event detection using bert. In *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*, pages 57–72. Springer.

Chung, H. W., L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, D. Valter, S. Narang, G. Mishra, A. W. Yu, V. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. hsin Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Grande, E. and A. Begga. 2024. Syntax Savants-UA at IberLEF 2024: Leveraging FLAN-T5-XXL for Automatic 5W1H Identification in Texts. In *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), CEURWS.org.*

Grieve, J. and H. Woodfield. 2023. *The language of fake news.* Cambridge University Press.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, and M. Villegas. 2021. Maria: Spanish language models. *arXiv preprint arXiv:2107.07253.*

Horne, B. and S. Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Ibrahim, M. 2024. Fine-Grained Language-based Reliability Detection in Spanish New with Fine-Tuned Llama-3 Model. In *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), CEURWS.org.*

Keith, B., M. Horning, and T. Mitra. 2020. Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization. *Computational Journalism C+ J.*

Khodra, M. L. 2015. Event extraction on indonesian news article using multiclass categorization. In *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5. IEEE.

Lugea, J. 2021. Linguistic approaches to fake news detection. *Data science for fake news: Surveys and perspectives*, pages 287–302.

Mangrulkar, S., S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Pan, R., J. A. García-Díaz, F. García-Sánchez, and R. Valencia-García. 2024. UMUTeam at FLARES@IberLEF 2024: Enhancing Disinformation Detection with 5W1H Techniques and Transformer Models. In *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), CEURWS.org.*

Pardo, J., J. Liu, V. Ramón-Ferrer, E. Amador-Domínguez, and P. Calleja. 2024. K-Flares: A K-Adapter Based Approach for the FLARES Challenge. In *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), CEURWS.org.*

Piad-Morffis, A., Y. Gutiérrez, Y. Almeida-Cruz, and R. Muñoz. 2020. A computational ecosystem to support ehealth knowledge discovery technologies in spanish. *Journal of Biomedical Informatics*, 109:103517.

Saquete, E., D. Tomás, P. Moreda, P. Martínez-Barco, and M. Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert systems with applications*, 141:112943.

Seddari, N., A. Derhab, M. Belaoued, W. Halboob, J. Al-Muhtadi, and A. Bouras. 2022. A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media. *IEEE Access*, 10:62097–62109.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vlachos, A. and S. Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Vosoughi, S., D. Roy, and S. Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Wang, R., D. Tang, N. Duan, Z. Wei, X. Huang, G. Cao, D. Jiang, M. Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808.*

Zhang, H., X. Chen, and S. Ma. 2019. Dynamic news recommendation with hierarchical attention network. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1456–1461. IEEE.

Zhao, S., F. You, and Z. Y. Liu. 2020. Leveraging pre-trained language model for summary generation on short text. *IEEE Access*, 8:228798–228803.

Zhou, L. and D. Zhang. 2008. Following linguistic footprints: Automatic deception detection in online communication. *Communications of the ACM*, 51(9):119–122.

Zhou, X. and R. Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.