

K-Flares: A K-Adapter Based Approach for the FLARES Challenge

Joel Pardo^{1,*†}, Jiayun Liu^{1†}, Virginia Ramón-Ferrer^{1,†}, Elvira Amador-Domínguez^{1,2,†} and Pablo Calleja^{1,†}

¹*Ontology Engineering Group, Departamento de Inteligencia Artificial, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain*

²*Departamento de Sistemas Informáticos, ETSI Sistemas Informáticos, Universidad Politécnica de Madrid, 28031 Madrid, Spain*

Abstract

We refer as 5W1H method to the technique used to extract comprehensive information from textual data answering the questions Who, What, When, Where, Why, and How. Identifying these instances within a text can provide a structured approach to better understand the details presented in a text, which can help to assess the reliability of the information displayed there. In this context, we present K-Flares, an adaptation of the original work of K-Adapter for 5W1H Instances Recognition in the context of the challenge of FLARES 2024. In this challenge the aim is to detect 5W1H instances from text and assess the reliability of the information displayed in each of them. We focus on the first part of the task, the detection of 5W1H instances.

Keywords

K-adapter, Covid-19, News, NER

1. Introduction

With the increasing use of the Internet, there has been a growing emergence of news providers that deliver information to society. However, this growth in information sources has also raised questions about the reliability of the information being disseminated. Many of the current online sources lack fact-checking and editorial standards that traditional media, such as newspapers or television news broadcasts, used to maintain, leading to the growing spread of non-factual information and biased reporting. The amount of information continuously generated prevents the correct manual monitoring of the veracity of each and every piece of information. This has led to a growing effort to research new approaches to be able to monitor the reliability of this type of information without human supervision and its derived tasks. An example of a task related to said line of research is the detection of fake news, where, among other approaches, the use of Natural Language Processing (NLP) techniques, such as Global Vectors (GloVe) [1], has been repeatedly combined with Deep Learning (DL) methods, such as Long Short-Term Memory networks (LSTMs) [2], to tackle said problem [3].

In this context, the 5W1H approach poses an interesting choice for text reliability detection. The 5W1H method is a technique used to extract comprehensive information from textual data answering the questions Who, What, When, Where, Why, and How. The identification of these key elements within a text can provide a structured approach to understand better the details presented in said text, which can help in the assessment of the reliability of the information displayed. For example, the identification of the source of a statement presented, that is, a "who" in a text, can greatly impact the human perception of the credibility of the information relied upon. If a statement is reported to come from a specific commonly trusted source, such as "The Organisation for Economic Co-operation and Development (OECD) has reported...", for example, it will likely be more believable. On the other hand,

IberLEF 2024, September 2024, Valladolid, Spain

*Corresponding author.

†These authors contributed equally.

✉ joel.pardof@upm.es (J. Pardo); jiayun.liu@upm.es (J. Liu); virginia.ramon@upm.es (V. Ramón-Ferrer);

elvira.amador@upm.es (E. Amador-Domínguez); p.calleja@upm.es (P. Calleja)

ORCID 0000-0001-8064-0128 (J. Pardo); 0009-0008-0740-4936 (J. Liu); 0009-0009-7197-1676 (V. Ramón-Ferrer);

0000-0001-6838-1266 (E. Amador-Domínguez); 0000-0001-8423-8240 (P. Calleja)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

if the source has a more vague identity, such as "Experts have reported...", it can lead to a distrust in the information originating from the lack of a precise information source.

The mistrust originating from the lack of precision in the information can be present in all the 5W1H elements in a text, which motivated the introduction of FLARES 2024 (Fine-Grained Language-based Reliability Detection in Spanish News) [4], a shared task introduced in IberLEF 2024 [5], whose goal is to assess the reliability of the language used in Spanish news writing, basing their efforts on the 5W1H method. They divide the main task into two subtasks: Subtask 1, where, when provided with a text, the goal is to find and annotate all the instances of 5W1H, and Subtask 2, where, for each 5W1H element detected in the previous subtask, the goal is to determine if the language used in each item is "reliable", "semi-reliable" or "unreliable" ("*confiable*", "*semiconfiable*" or "*no confiable*" in Spanish). To tackle these tasks, a manually annotated dataset consisting of 190 news items is provided, containing over 9,000 5W1H annotations. Annotations consist on the 5W1H of each piece of news, as well as the reliability label of each text snippet, on each case. A single sample has none, one, or multiple instances of the same 5W1H label.

When analysing the FLARES dataset, we find the difficulty of not having a clear syntactical consistency in the labelling of the text. For example, we find that some annotations refer to nominal chunks, while others do not. We also have that nominal chunks that are usually the direct object of the verb have different annotation labels among themselves. These inconsistencies make it difficult to tackle this problem using more traditional approaches, such as dependency parsing. On the other hand, we do find a regular interconnection in the 5W1H instances in each individual text, which we can exploit to generate a graph-like structure that can be used to inject information to the model chosen to tackle the task.

In this work, we present the KFloeg team submission for Subtask 1 in FLARES 2024. Our proposal consists of a modification of the original K-Adapter [6] for the 5W1H task approaching it as a Named Entity Recognition (NER) task. The adapters used to enrich the fine-tuning process are based on the Wikidata relations of the entities and relations between labels (5W1H) inside each text from the same training corpus.

For Subtask 2, we conducted a simple fine-tuning of the Mistral-7B [7] model. Our approach involved using an instruction prompt adapted to the specific task and fine-tuning this model using its causal language model architecture to generate the expected output in an easy-to-process format. However, due to time constraints and poor results (precision lower than 0.30), we did not pursue further improvements or explore new approaches. Our primary efforts and contributions were concentrated on Subtask 1.

Experimental results show that our approach obtains the best results using just one adapter and a low number of epochs to avoid overfitting. Moreover, our proposal ranked 1st in the first task of the challenge.

This work is structured as follows: Section 2 shows the work related to our approach, followed by section 3, where we present our approach, the data and its pre-processing, and the experimental setup used. Finally, in sections 4 and 5 we present the results obtained from our experiments and the conclusions extracted from them.

2. Related work

Pretrained language models have proven to be a powerful approach for various NLP tasks. However, language models usually tend to have hallucination problems or lack of factuality. In this regard, knowledge injection appears to offer a potential solution to this problem by including factual knowledge, extracted from structured databases, or knowledge graphs, to boost model performance.

One of the first works in this area is represented by K-BERT [8], which uses a specific domain KG in the inference process. The knowledge triples coming from the graph are injected into the input sentence. Another relevant work is K-former[9], which is a variation of the RoBERTa architecture[10], in which knowledge is included in the network in the final encoder layers of the model. This work was

evaluated in question answering tasks.

K-adapter [6] is another method that integrates linguistic and factual knowledge into language models using adapters. These adapters, which are trainable networks inserted in the middle of the transformer layer, allow the injection of knowledge without modifying the existing parameters of the language model. These adapters operate independently from each other and can be trained in parallel, providing a flexible and efficient approach to fuse different types of knowledge into language models. This approach has been evaluated in the tasks of entity typing, relation extraction, and question answering.

In this context, K-adapter suits perfectly the main problems identified and faced in this challenge. First, the training data available is scarce, and there is no possibility to rely on external knowledge to reinforce the learning process. The concept of adapters perfectly aligns with this purpose: the possibility of fine-tuning the model for a task including other aspects of the same corpora not reflected directly in the natural language text. Second, the labels of the 5W1H instances annotated in the corpus contain a lot of inconsistencies and do not present a clear linguistic common pattern. However, these instances do show that they are interconnected inside the context of the document (e.g., the WHAT labels with the WHO labels in document), so we can generate a graph in which elements are interconnected for the reinforcement of the learning process, which is the specific purpose of adapters.

3. Methodology

Along this section we present the approach we followed to detect 5W1H entities in Spanish texts. Specifically, we present the dataset used for our task and the modifications made in the original K-Adapter model in sections 3.1, and 3.2 respectively, followed by the explanation of the data preparation we followed in section 3.3 and the experimental setup we used in section 3.4.

3.1. FLARES dataset

The FLARES 2024 dataset [4] is used for our experiments. It consists of approximately 9,034 5W1H annotations across 190 news articles. The dataset is divided into 70% for training (6,934 annotations) and 30% for testing (2,100 annotations). These annotations are based on the 5W1H journalistic technique, which labels entities in the text corresponding to the questions: WHAT, WHO, WHERE, WHEN, WHY, and HOW. The annotations were performed by two linguists and one sociologist specialized in NLP following specific annotation RUN-AS guidelines [11] established for this dataset. Figure 1 depicts an example of annotated text provided, where the position of each detected 5W1H instance with its 5W1H label and reliability label (in Spanish) are identified.

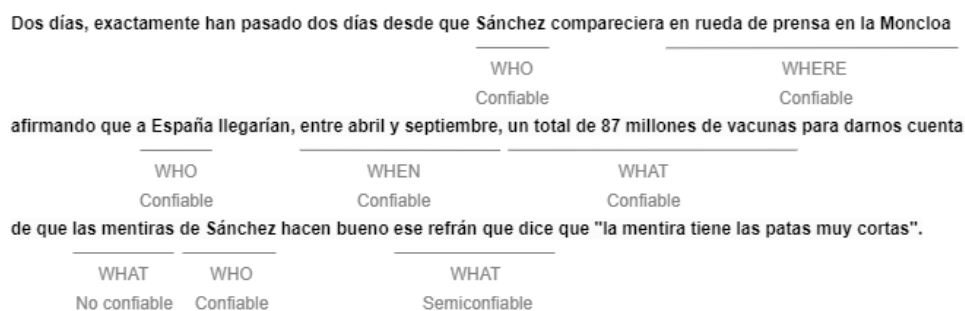


Figure 1: Example of FLARES annotation [4].

The specific data for Subtask 1 is divided in two subsets: training and testing. A trial dataset is also provided, but since it is duplicated from the training data, it is dismissed. The training set consists of 1,585 samples extracted from news with a total of 6,934 annotations. In Table 2 we can see a big gap in the number of annotation instances of each 5W1H label: roughly 40% of these annotations are "What" instances and over 25% are instances of "who", leaving a proportion of about 35% of annotations for the

other four labels. Test data consists of 402 non-annotated texts, which were to be evaluated using the Kaggle challenge provided by FLARES 2024¹ [4].

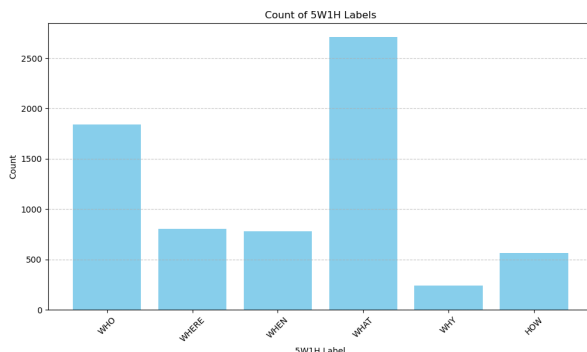


Figure 2: Distribution of 5W1H Labels from the Subtask 1 training Dataset. The bar plot shows the count of each 5W1H label (Who, What, When, Where, Why, How) across the training data.

3.2. K-Flares model

For our approach², we have taken the original K-Adapter [6] implementation³ and adapted it to our needs, as shown in Figure 3. The first change we made was motivated by the language. The original K-Adapter was designed to run on English data, not Spanish. Consequently, we first substituted the pre-trained language model, the original RoBERTa-large [12], with a Spanish fine-tuned version of said model, specifically RoBERTa-large-bne [13], a fine-tuned RoBERTa-large model pre-trained with a corpus extracted from the National Library of Spain (Biblioteca Nacional de España)⁴ with a size of over 500GB. The second change was related to the adapters used. Originally, the K-Adapter used

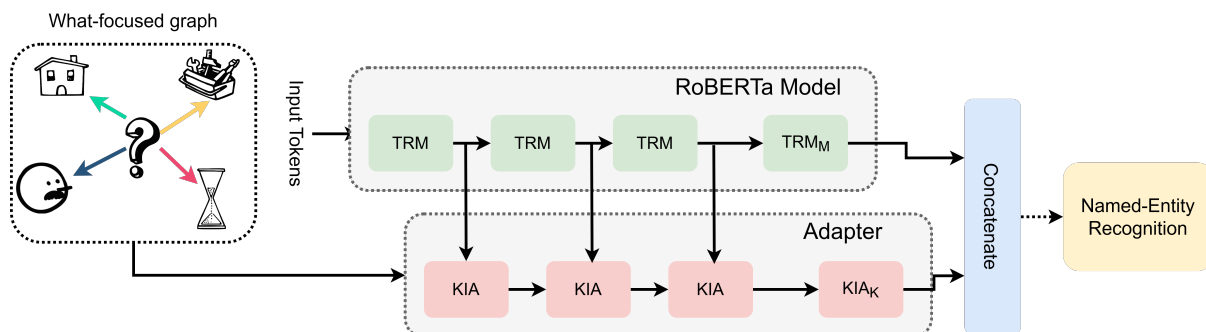


Figure 3: Overview of the K-Flares proposal.

two adapters: a factual adapter, which injects factual knowledge from the relationships among entities in natural language; and a linguistic adapter, which injects linguistic knowledge from dependency relationships among words in natural language text. In our case, we decided to use only one adapter, a factual adapter that maintains some syntactical information related to the data represented. We initially tested the use of two factual adapters, but, on the basis of the results obtained, we decided to only maintain one of them. Adapters were trained using FLARES 2024 training data, presented in Section 3.1. This data was transformed from its original state to a triple-like format, that is, relationships of entities, where we used the WHAT elements of the texts as central nodes. In Section 3.3, we further explain the data preparation made for this training.

¹<https://kaggle.com/competitions/flares-subtask-1-5w1hs-identification>

²<https://github.com/oeg-upm/k-flares>

³<https://github.com/microsoft/k-adapter>

⁴<https://www.bne.es/en>

Once the adapter is trained, the model is fine-tuned for the Named Entity Recognition (NER) task. The fundamental purpose of NER is to identify people, organisations, and locations in text, but it can also be used to identify other entities. Given that our task is fundamentally the recognition of 5W1H instances in the text, our goal can be devised as a variation of the classical NER approach in which the entities detected correspond to the 5W1H labels.

3.3. Data preparation

Triples are needed to train the adapter, which are composed by a subject, a relation, and an object. The subject is the entity performing the action, the relation represents the action being performed, and the object is the entity that receives the action. These triplets follow the format: subject-relation-object. Different approaches can be employed to extract triplets from text depending on what is the goal of the adapter; in this case, we have trained two different adapters: one based on the relations between the labels 5W1H inside the document (5W1H labels factual adapter), and the other based on the relations that the entities have in Wikidata (Wikidata factual adapter).

For the 5W1H labels factual adapter, we used a rule-based approach in which the action is represented by the mix of the labels of the subject and the object. The subject always corresponds to the WHAT, as it appears to be the core of each sentence and its the most common label. If there are two annotations containing a WHAT, the first is related to the rest, followed by the second becoming the subject and being related to the rest. If a sample lacks of the WHAT annotation, the subject defaults to the first presented label and is then related to the remaining labels. For an example, see the 'relation' field in the data structure below. Figure 4 shows the graph generated from the sample of Figure 1.

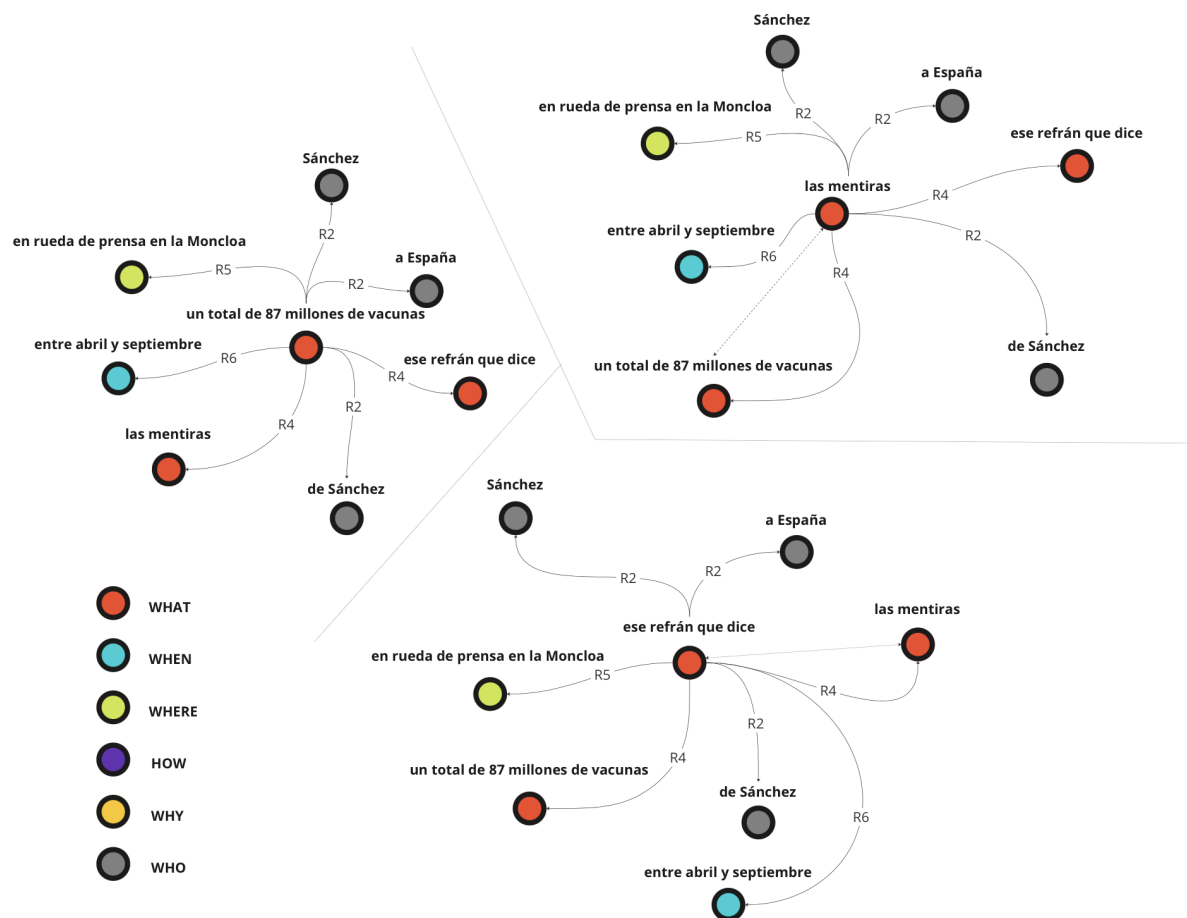


Figure 4: Overview of the triplets extraction output for the 5W1H labels factual adapter.

For the Wikidata factual adapter, we used the mREBEL (multilingual REBEL) model [14]. This autoregressive method employs a seq2seq model that, given a text input, directly outputs triplets. It is trained with around 400 relation types. For both of them, the data is formatted by the following structure and stored in json files:

- **docid:** This field represents a unique identifier for the document from which the data is extracted. It is an integer stored as a string, ensuring each document can be distinctly recognized and referenced.
- **token:** This is a list of strings, where each string represents a token or a word in the document. The tokens include words, punctuation marks, and other elements, reflecting the segmentation of the text. This list is made by a SpaCy model named `es_core_news_lg`.
- **relation:** This field denotes the type of relationship identified between two segments in the document. It is represented by a code (e.g., "R1", "R2") that maps to a specific type of relationship between two specific entity types. For example, "R1" refers to the relationship between "WHAT" and "HOW" entities, "R23" refers to the relationship between a "WHY" and "WHERE" entities. This set of relations corresponds to all possible combinations of the 5W1H labels. This encoding was used for the 5W1H labels factual adapter. For the Wikidata factual adapter, we assess the variety of relations extracted by the mREBEL model and enumerate them using the same nomenclature this model returns (e.g. the subject is labeled as "concept" and the object is labelled as "loc", the relationship according to the model would be "country"). Both are contemplated in two different json files.
- **subj_start:** An integer indicating the starting index of the subject token in the text in letter level. This index marks the beginning of the segment identified as the subject in the relationship.
- **subj_end:** An integer indicating the ending index of the subject token in the text in letter level. This index marks the end of the segment identified as the subject in the relationship.
- **obj_start:** An integer indicating the starting index of the object token in the text in letter level. This index marks the beginning of the segment identified as the object in the relationship.
- **obj_end:** An integer indicating the ending index of the object token in the text in letter level. This index marks the end of the segment identified as the object in the relationship.
- **subj_label:** A string that labels the subject segment, categorizing it according to its role in the relationship. It includes for the 5W1H labels factual adapter the six labels (WHAT, WHO, WHERE, WHEN, HOW, WHY) presented in the original data. For the Wikidata adapter we have as many labels as the model has output.
- **obj_label:** A string that labels the object segment, categorizing it according to its role in the relationship. It includes for the 5W1H labels factual adapter the six labels (WHAT, WHO, WHERE, WHEN, HOW, WHY) presented in the original data. For the Wikidata adapter we have as many labels as the model has output.

In addition, the dataset is provided in IOB format for the NER task. Each document is identified by a unique "docid". For each token within a document's token list, a corresponding label is assigned. These labels denote whether a token is at the Inside, Outside, or Beginning of a particular span. These nuances are prefixed accordingly with "I-", "O-", or "B-", respectively.

3.4. Experimental setup

All the experiments were conducted on a remote server, equipped with one NVIDIA A100-PCIE-40GB GPU. To obtain the final model, we tried a series of configurations combining the two different adapters, the factual adapter created with the 5W1H labels and the factual adapter created with Wikidata, to determine the optimal architecture. For each adapter and the final model, specific training hyperparameters were used to optimize the training process while avoiding overfitting. The following Table 1 detail the hyperparameters that yielded the best results.

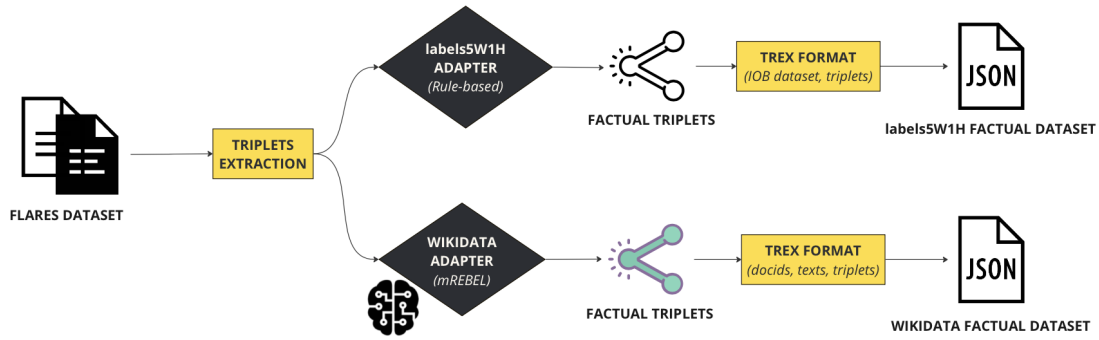


Figure 5: Overview of the preprocessing workflow of the Flares dataset for both adapters.

Table 1

Comparison of hyperparameters across 5W1H labels and Wikidata Factual Adapters, and Complete K-Adapter configurations.

Parameter	5W1H labels adapter	Wikidata adapter	Complete K-Adapter
epochs	5	10	8
model	roberta-large	Roberta-large	roberta-large
per_gpu_train_batch_size	64	32	8
per_gpu_eval_batch_size	64	8	8
max_seq_length	64	64	512
scheduler	WarmupLinearSchedule	WarmupLinearSchedule	WarmupLinearSchedule
optimizer	AdamW	AdamW	AdamW
learning_rate	5e-5	2e-5	5e-5
warmup_steps	1200	500	120
freeze_adapter	False	False	True
adapter_size	768	768	768
adapter_list	"0,11,22"	"0,11,22"	"0,11,22"
adapter_transformer_layers	2	2	-
fusion_mode	-	-	add

4. Evaluation and Results

Once our model was fine-tuned for the NER (Named Entity Recognition) task, it generated predictions that identified each token as the appropriate 5WH1 label. To fulfill the submission requirements of the challenge, it was essential to convert the identified textual fragments into a structured format. This format required the inclusion of both starting and ending indices of each identified text fragment under the 5WH1 label. This was achieved by locating the exact match of the fragment within the original text and extracting their positional indices. The final submission format, comprises the unique identifier along with the indices and labels of the tags, which are submitted to the Kaggle challenge.

We submitted three different configurations of our model, each differing in its use of adapters and hyperparameters, the comparative results are detailed in Table 2. The model labeled "K-adapter KG",

which utilised only the 5W1H labels factual adapter, was trained in 8 epochs, and achieved the 1st ranking in the competition. This underscores the effectiveness of this factual adapter in enhancing model performance for the specified task. In the second model, "K-adapter KG 2A" the model used the both adapters (5W1H and Wikidata), did not perform well. The noise produced by the external relations of Wikidata do not improve the acquired knowledge for the NER task. Finally, the "K-adapter KG+" is the same as the first one but trained with 12 epochs, which exhibited signs of overfitting during the training process, achieving perfect score in training set, which likely caused the poorer performance compared to the first configuration.

Table 2

Scores obtained in the submissions on Kaggle platform. Table shows the submission of the model, the obtained ranking, the used adapters and the evaluation results. The group has submitted three different types of models, being the first one the best of them.

Submitted (Ranking)	Adapters	30% split test set	70% split test set
K-adapter KG (1st)	5W1H	0.65957	0.66544
K-adapter KG 2A (3rd)	5W1H + Wikidata	0.39649	0.40070
K-adapter KG+ (3rd)	5W1H	0.39578	0.39572

5. Conclusions

This paper presents K-Flares, a K-Adapter-based approach for the FLARES challenge. This challenge is composed of two different subtasks. The first subtask comprises the identification of the 5W1H labels within a set of textual data extracted from pieces of news. In the second task, the goal is to classify each 5W1H according to their reliability level. K-Flares addresses the first subtask.

For this task, two different adapters were devised, using a Spanish version of roBERTa as the baseline language model. The task was devised as a Named-Entity Recognition problem, in which the entities to be detected correspond to the the 5W1H labels. The first of the adapters, the 5W1H adapter, focuses on the idea that knowledge graphs can be generated for each sample considering the WHAT as the core and relating it with the rest of the elements in the sample. For the second adapter, the Wikidata factual adapter, the mREBEL model was used to generate triples from the text samples according to their knowledge in Wikidata. Both adapters were used independently and then conjunctively, with the 5W1H yielding the best results. This approach reached a 65.95% accuracy in the test set, achieving the winning result of the challenge.

The results show that the inclusion of knowledge of how the labels interact between them reinforce the learning process of the NER task. Moreover, experiments also showed that the inclusion of factual knowledge of external knowledge graphs such as Wikidata decreases the performance of the model. Additionally, it has been demonstrated that treating the problem as a Named-Entity Recognition problem in combination with external knowledge is a feasible and effective approach for this challenge.

Acknowledgments

This work has been partially founded by INESData (<https://inesdata-project.eu/>) project, funded by the Spanish Ministry of Digital Transformation and Public Affairs and NextGenerationEU, in the framework of the UNICO I+D CLOUD Program - Real Decreto 959/2022.

References

- [1] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in

- Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: <https://aclanthology.org/D14-1162>. doi:10.3115/v1/D14-1162.
- [2] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (1997) 1735–1780. URL: <https://doi.org/10.1162/neco.1997.9.8.1735>. doi:10.1162/neco.1997.9.8.1735.
- [3] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, M. S. Rahman, A comprehensive review on fake news detection with deep learning, *IEEE Access* 9 (2021) 156151–156170. doi:10.1109/ACCESS.2021.3129329.
- [4] 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).
- [5] 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.
- [6] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, J. Ji, G. Cao, D. Jiang, M. Zhou, K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters, in: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, Online, 2021, pp. 1405–1418. URL: <https://aclanthology.org/2021.findings-acl.121>. doi:10.18653/v1/2021.findings-acl.121.
- [7] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, *Mistral 7b*, 2023. URL: <https://arxiv.org/abs/2310.06825>. arXiv:2310.06825.
- [8] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, P. Wang, K-bert: Enabling language representation with knowledge graph, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020, pp. 2901–2908.
- [9] Y. Yao, S. Huang, L. Dong, F. Wei, H. Chen, N. Zhang, Kformer: Knowledge injection in transformer feed-forward layers, in: *Natural Language Processing and Chinese Computing*, Springer International Publishing, Cham, 2022, pp. 131–143.
- [10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, *CoRR* (2019).
- [11] 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.
- [12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.
- [13] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodríguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, *Procesamiento del Lenguaje Natural* (2022) 39–60. URL: <https://doi.org/10.26342/2022-68-3>. doi:10.26342/2022-68-3.
- [14] P.-L. H. Cabot, S. Tedeschi, A.-C. N. Ngomo, R. Navigli, Red^{fm}: a filtered and multilingual relation extraction dataset, *arXiv preprint arXiv:2306.09802* (2023).