Fine-Grained Language-based Reliability Detection in Spanish New with Fine-Tuned Llama-3 Model

Michael Ibrahim

Computer Engineering Department, Cairo University, 1 Gamaa Street, 12613, Giza, Egypt

Abstract

In today's digital media landscape, it's essential to evaluate the credibility of news language. Analyzing certain parts of a news article enhances the assessment of its linguistic trustworthiness. This method strengthens the ability to identify deceptive news and improves the comprehension of news content. Large language models' latest advancements offer a more effective solution to the news reliability classification problem. For Spanish news reliability classification, this paper suggests fine-tuning the 8B-parameter Llama 3 model recently launched by Meta GenAI. In the IberLEF 2024 - Flares subtask 2, the proposed method ranked second for Spanish news reliability classification with a Macro F_1 score of 0.59658.

Keywords

Text Classification, Reliability Detection, Llama3, PEFT, LoRA

1. Introduction

Billions of internet users depend on social media for news consumption, however, large amounts of intentionally false information, such as fake news, are disseminated online via social media due to its weak supervision. Fake news serves as an illustration of disinformation. The pervasive distribution of false information can seriously harm individuals and society. Fake news can undermine readers' trust in the news ecosystem. [1]

To detect fake news, it's essential to perform fact-checking, however, this fact-checking requires external knowledge that can be hard to acquire especially for an ongoing event. IberLEF 2024 [2] - FLARES - subtask 2 [3] aimed to develop systems that can effectively distinguish between unreliable and trustworthy online articles based on the style and language of the news without any external knowledge. The "5W1H" journalist method was used to annotate the data, this method determines the What, Who, Why, When, Where, and How aspects of a text. This technique allows for a systematic evaluation of language reliability across various dimensions. Assessing these core journalistic questions provides a framework for evaluating the linguistic accuracy and potential biases in the content.

In NLP, for text classification, the transformer architecture has driven the growth of massive self-supervised neural networks, each possessing tens or hundreds of billions of parameters and trained on trillions of tokens. Recent models based on the transformer architecture include GPT series [4], LLaMA [5], and others. As of late 2024, the release of LLaMA-3 sparked the growth of a thriving open-source community, resulting in the development of near-state-of-the-art models like Mistral [6] and Mixtral [7].

With minimal finetuning, these LLMs can adjust to new tasks and instructions. The NLP community employs PEFT techniques over full-parameter fine-tuning for cost savings. These PEFT techniques can substantially minimize a model's computational requirements [8]. Low-Rank Adaptation (LoRA) [9] balances fine-tuning's performance and efficiency via low-rank matrix approximations.

Other PEFT techniques, include LLM model weights compression through quantization techniques, cutting down storage and memory usage during both training and inference. By employing NormalFloat

IberLEF 2024, September 2024, Valladolid, Spain

michael.nawar@eng.cu.edu.eg (M. Ibrahim)

www.linkedin.com/in/michael-ibrahim-90 (M. Ibrahim)

0000-0003-2340-8917 (M. Ibrahim)

Quantization [10], model performance is preserved while significantly reducing LLMs' resource requirements, enabling resource-efficient fine-tuning through integration with PEFT methods like LoRA.

The rest of the paper is organized as follows. The related work is summarized in Section 2. The dataset used for training and validation was detailed in Section 3. In Section 4, the system is presented. Section 5 summarizes the study's key findings.

2. Related Work

2.1. Reliability Detection

In [11], a news reliability assessment system was introduced. The system employs Natural Language Processing (NLP) techniques like Term Frequency-Inverse Document Frequency (TF-IDF), Phrase Detection, Cosine Similarity, and Latent Semantic Analysis (LSA) simultaneously. 9203 articles, half from reliable and half from unreliable sources, were gathered. The dataset was randomly divided into a training set and a testing set. The precision, recall, and accuracy of the final results were 81.87%, 86.95%, and 73.33% respectively.

In [12], multiple machine learning and deep learning techniques were used to detect the reliability of Spanish news articles. The machine learning and deep learning were trained twice, once with the news text, and another with the news text and 42 extra features. The best results are attained with the Decision Tree using the extra 42 features, obtaining a macro F_1 score of 0.948. It is noteworthy that when using the whole document annotated without external features, the best macro F_1 score is obtained by the deep learning approach based on the Spanish Bert transformer (BETO) [13] with a macro F_1 score of 0.80, followed by AdaBoost with a macro F_1 score of 0.748.

LogicDM[14] is a logic-based neural network model for multimodal misinformation detection that uses interpretable logic clauses to combine NeuralSymbolic AI's neural network learning with the explainability of symbolic reasoning. The learning process was more effective by representing symbolic logical elements neurally and generating/evaluating logic clauses automatically. LogicDM is applicable to a range of misinformation sources, due to the incorporation of five meta-predicates, each instantiable with varying correlations. LogicDM proves effective and adaptable based on evaluations on Twitter, Weibo, and Sarcasm datasets.

Bayesian graph local extrema convolution (BLC) was used for misinformation detection [15]. In BLC, the emphasis is placed on the unreliable relationships and uncertainties inherent in the propagation structure, while the differences between nodes and their neighbors are highlighted through their attributes. A new strategy focusing on long-tail users is proposed to prevent over-concentration on high-degree nodes in graph neural networks for the global social network. The misinformation detection model was assessed using two Twitter datasets, and the results show that the long-tail strategy enhances the performance of existing graph-based methods for detecting misinformation.

2.2. LLM Applications

Utilizing LLMs for a wide range of NLP tasks is currently popular. The most effective and efficient utilization of these models is yet to be determined. Three primary methods exist for constructing applications utilizing LLMs.

- **Zero-shot prompting** It is querying prompts that aren't in LLM's training data to elicit responses. These prompts often include detailed instructions and a primary question. Crafting precise prompts is essential for maximizing the effectiveness of large language models.
- **Few-shot learning** Few-shot learning involves giving LLMs a limited number of examples to produce appropriate responses. Zero-shot prompting refers to a method without providing any examples. In few-shot learning, examples are incorporated into the prompt template to guide the model's response.

- **Fine-tuning**. The two methods above enable task adaptation without the requirement for additional training on the LLMs while fine-tuning necessitates further training of the LLMs with task-specific data. When tailored datasets are available, this method is especially advantageous.

According to [16], LLMs perform effectively in most NLP tasks based on their examination of 'use cases' and 'no use cases' for particular downstream tasks using the mentioned methods.

2.3. LLM for Reliability Detection

LLMs can be used to identify misleading news headlines [17], ChatGPT-4 closely mirrors human decisions in clear-cut cases, however, it shows discrepancies in performance when there is a mixed human consensus, revealing the complexity of misleading headline detection. This study recommends further research on LLM-generated explanations and multimodal content expansion can bridge the gap between AI and human judgment, paving the way for reliable, ethical, and effective misinformation combat tools.

In [18] the effectiveness of large language models in detecting fake news and optimally utilizing their advantages for enhanced performance was investigated. While underperforming the task in comparison to small LMs like BERT, the large LM (GPT-3.5) excels at providing informative rationales in news understanding. This study shows that LLMs, despite their superior analyzing skills, may fail to fully utilize their own abilities. Uncovering their full potential may necessitate innovative prompts and a more intricate grasp of its workings.

DELL [19] (**D**iverse Reaction Generation; Explainable Proxy Tasks; and **LL**M-Based Expert Ensemble) integrates LLMs into the fake news identification pipeline. DELL generates news reactions from diverse perspectives using LLMs and model user-news networks accordingly. Six explainable tasks were developed to enable LLMs to identify misinformation within news articles and generate corresponding expert explanations. Three strategies were presented for merging task-specific experts within LLMs for an overall prediction. The three tasks across seven datasets show that DELL's misinformation detector outperforms others with improved calibration and diverse perspectives.

3. Data

RUN-AS [12] (Reliable and Unreliable News Annotation Scheme) is a fine-grained annotation for classifying news reliability solely through linguistic and textual analysis. The purpose of this annotation is to assess the contribution of individual elements towards the overall credibility of a news report. The dataset collected contains approximately 8,145 5W1H annotations. 85% (6,934 instances) of the 5W1H dataset are used for training, with the remaining 30% (1,211 instances) utilized for testing. Those 8,145 annotations are from 190 news items collected manually and via web crawling, gathered from various digital newspapers including ABC, BBC News, CNN Spanish, El Espectador, El Financiero, El Mundo, El PaÃŋs, Huffpost, Marca, La Jornada, El Diestro, Eje21, Periodista Digital, Ok Diario, 20 Minutos, and la Vanguardia. The corpus contains text in Spanish manually annotated with an extraction technique known as 5W1H, which is a journalistic technique consisting of annotating all the entities present in a text related to the questions What, Who, Where, When, Why and How. Three NLP experts with backgrounds in linguistics and sociology annotated the text. Table 1 shows the total number of 5W1H labels in the dataset, and table 2 shows the distribution of the classification labels in the training data.

As a preprocessing to this dataset, the 5W1H labels were transformed to other labels as described in table 3. Those new labels will be used in the training and testing of the proposed system instead of the 5W1H labels.

4. Methodology

The Llama-3 model was fine-tuned with LoRA keeping the Llama-3 weight frozen and updating only the LoRA matrices using the following formatted prompt-response quadruplets: 5W1H label [label], tag text

Table 1
Dataset description (5W1H labels).

Label	Training	Testing
WHAT	2711	482
WHO	1843	305
WHEN	778	143
WHERE	801	150
WHY	238	35
HOW	563	96

Table 2
Classification label distribution in the training data.

Label	Confiable	Semiconfiable	Noconfiable
WHAT	1722	549	440
WHO	1395	316	132
WHEN	571	165	42
WHERE	659	102	40
WHY	151	39	48
HOW	267	105	191

Table 3
Dataset new labels.

Label	New Label
WHAT	fact
WHO	subject
WHEN	time
WHERE	place
WHY	cause
HOW	manner

[tag], the text [target], and the reliability label [reliability]. The dev set, like the training set, is formatted similarly. This prompt customizes the model for the DA-MSA machine translation assignment. The prompt is formatted as follows:

[s] [INST] [SYS] You are a Spanish news fact checker! [/SYS]
Consider the following Spanish news item and determine whether the language used to describe the [label] ([tag]) is confiable or semiconfiable or no confiable and return the answer as "confiable", "semiconfiable" or "no confiable". [text] [/INST]
[reliability] [/s]

For example, the first annotation in the training set will have the following prompt.

[s] [INST] [SYS] You are a Spanish news fact checker! [/SYS]
Consider the following Spanish news item and determine whether the language used to describe the subject (a una diputada del Partido Popular) is confiable or semiconfiable or no confiable and return the answer as "confiable", "semiconfiable" or "no confiable". Se llama JosÃľ TomÃĺ, es presidente de la DiputaciÃşn de Lugo, del PSOE y un machista que se ha dedicado a denigrar a una diputada del Partido Popular en la DiputaciÃşn por su forma de vestir. [/INST]
confiable [/s]

The learning rate for finetuning the Llama-3 model was set to 1e-4 with the Adam optimizer used during training. The evaluation is conducted every 100 steps with a batch size of 16. The low-rank

approximation rank is set to 64, and its scaling factor for adaptation is set to 16 in the LoRA configurations. The model trainable parameters are all linear layers: "q_proj", "up_proj", "o_proj", "k_proj", "down_proj", "gate_proj", and "v_proj". A 0.05 dropout is applied in the LoRA layer. The model's weights are quantized to 4-bit precision and mixed-precision training with float16 and float32 is enabled to reduce memory requirements and accelerate the training process. This model was trained on Google Colab with a single NVIDIA A100 GPU with 40GB of memory. The finetuning of the model and the generation of the test results tooks almost one hour. The code used for training and generating the results can be found on the following GitHub repository. [1]

Finally, to generate the label for each record in the test set, we give the fine-tuned LLM the formatted prompt without the label, the reliability label, and the LLM generates it. For example, this is the prompt for the first record in the test set (document with "Id" 157).

> *[s] [INST] [SYS] You are a Spanish news fact checker! [/SYS]*
> *Consider the following Spanish news item and determine whether the language used to describe the fact (el objetivo fundamentalmente legÃŋtimo) is confiable or semiconfiable or no confiable and return the answer as "confiable", "semiconfiable" or "no confiable". Las medidas que no se basan en pruebas [âĂę] son inadecuadas para alcanzar el objetivo fundamentalmente legÃŋtimo que persiguen, a saber, evitar la sobrecarga del sistema sanitario o reducir la incidencia de la infecciÃşn por el SRAS-CoV-2. [/INST]*

The LLM will generate the text " *no confiable* [/s]", then the label will be "no confiable".

5. Conclusion

Assessing news credibility is crucial in the current digital news environment. Analyzing certain parts of a news article enhances the assessment of its linguistic trustworthiness. This method strengthens the ability to identify deceptive news and improves the comprehension of news content. The latest advancements in large language models effectively address the news reliability classification problem. This study recommends fine-tuning the recently released 8B-parameter Llama 3 model by Meta GenAI for Spanish news reliability classification. In the IberLEF 2024 - Flares subtask 2, the proposed method achieved a second-place ranking and a Macro F_1 score of 0.59658 for Spanish news reliability classification.

References

[1] K. Shu, S. Wang, D. Lee, H. Liu, Mining disinformation and fake news: Concepts, methods, and recent advancements, Disinformation, misinformation, and fake news in social media: Emerging research challenges and opportunities (2020) 1–19.

[2] 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.

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

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[1] https://github.com/MichaelIbrahim-GaTech/FLARES2024-subtask2.git

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[6] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
[7] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al., Mixtral of experts, arXiv preprint arXiv:2401.04088 (2024).
[8] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for nlp, in: International conference on machine learning, PMLR, 2019, pp. 2790–2799.
[9] 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, arXiv preprint arXiv:2106.09685 (2021).
[10] J. Chen, A. Zhang, X. Shi, M. Li, A. Smola, D. Yang, Parameter-efficient fine-tuning design spaces, arXiv preprint arXiv:2301.01821 (2023).
[11] G. Xiaoning, T. De Zhern, S. W. King, T. Y. Fei, L. H. Shuan, News reliability evaluation using latent semantic analysis, TELKOMNIKA (Telecommunication Computing Electronics and Control) 16 (2018) 1704–1711.
[12] 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.
[13] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, arXiv preprint arXiv:2308.02976 (2023).
[14] H. Liu, W. Wang, H. Li, Interpretable multimodal misinformation detection with logic reasoning, arXiv preprint arXiv:2305.05964 (2023).
[15] G. Zhang, S. Zhang, G. Yuan, Bayesian graph local extrema convolution with long-tail strategy for misinformation detection, ACM Transactions on Knowledge Discovery from Data (2024).
[16] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, X. Hu, Harnessing the power of llms in practice: A survey on chatgpt and beyond, ACM Transactions on Knowledge Discovery from Data 18 (2024) 1–32.
[17] M. M. U. Rony, M. M. Haque, M. Ali, A. S. Alam, N. Hassan, Exploring the potential of the large language models (llms) in identifying misleading news headlines, arXiv preprint arXiv:2405.03153 (2024).
[18] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, P. Qi, Bad actor, good advisor: Exploring the role of large language models in fake news detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 22105–22113.
[19] H. Wan, S. Feng, Z. Tan, H. Wang, Y. Tsvetkov, M. Luo, Dell: Generating reactions and explanations for llm-based misinformation detection, arXiv preprint arXiv:2402.10426 (2024).