Improving Explainable Fact-Checking via Sentence-Level Factual Reasoning

Francielle Vargas
University of São Paulo
francielleavargas@usp.br

Isadora Salles
Federal University of Minas Gerais
isadorasalles@dcc.ufmg.br

Diego Alves
Saarland University
diego.alves@uni-saarland.de

Ameeta Agrawal
Portland State University
ameeta@pdx.edu

Thiago Pardo
University of São Paulo
taspardo@icmc.usp.br

Fabricio Benevenuto
Federal University of Minas Gerais
fabricio@dcc.ufmg.br

Abstract

Most existing fact-checking systems are unable to explain their decisions by providing relevant rationales (justifications) for their predictions. It highlights a lack of transparency that poses significant risks, such as the prevalence of unexpected biases, which may increase political polarization due to limitations in impartiality. To address this critical gap, we introduce ***Se**ntence-**L**evel **Fa**ctual **R**easoning* (SELFAR)[1], aimed at improving explainable fact-checking. SELFAR relies on fact extraction and verification by predicting the news source reliability and factuality (veracity) of news articles or claims at the sentence level, generating post-hoc explanations using SHAP/LIME and zero-shot prompts. Our experiments show that unreliable news stories predominantly consist of subjective statements, in contrast to reliable ones. Consequently, predicting unreliable news articles at the sentence level by analyzing impartiality and subjectivity is a promising approach for fact extraction and improving explainable fact-checking. Furthermore, LIME outperforms SHAP in explaining predictions on reliability. Additionally, while zero-shot prompts provide highly readable explanations and achieve an accuracy of 0.71 in predicting factuality, their tendency to hallucinate remains a challenge. Lastly, this paper also presents the first study on explainable fact-checking in the Portuguese language.

1 Introduction

While journalism is tied to ethical standards, including truth and fairness, it often strays from impartial facts (Mastrine, 2022). As a result, low credibility news may be produced and spread on modern media ecosystem. Nowadays, fact-checking organizations have manually provided lists of unreliable articles and media sources, however it is a very time-consuming task, needs to be updated faster and relies on specific expertise (Baly et al., 2018a).

Towards addressing this issue, fact-checking systems have classified claims of unknown veracity (factuality), identifying evidence and predicting whether they support or refute the claims (Glockner et al., 2023; Guo et al., 2022). Nevertheless, as low credibility news or claims may comprise multiple sentences containing facts, media bias, and fake information, fact-checking at scale should be able to accurately predict both news source reliability and factuality at a fine-grained level. Table 1 shows an example of low credibility news segmented into sentences and classified according to its reliability (biased/unbiased) and factuality (fake and fact).

Furthermore, the veracity of claims can be verified using metadata (Augenstein et al., 2019), Wikipedia (Thorne et al., 2018), social networks (Hardalov et al., 2022), scientific assertions (Wadden et al., 2020), manually checked-claims from social media provided by fact-checking organizations (Wang, 2017; Couto et al., 2021), the language used in claims (Sheikh Ali et al., 2021), LLMs (Lee et al., 2021; Zhang and Gao, 2023), generating justifications for verdicts on claims (Atanasova et al., 2020a). For example, the FEVER (Thorne et al., 2018), SciFact (Wadden et al., 2020), LIAR (Wang, 2017) and Check-COVID (Wang et al., 2023) are widely used datasets for this setting.

In recent years, there has been significant progress in the area of fact-checking e.g., new comprehensive datasets (Yang et al., 2018; Wang, 2017; Hanselowski et al., 2019; Reis et al., 2020), high performance of deep learning models (Ribeiro et al., 2022), different domains aside from political (Naderi and Hirst, 2018; Kotonya and Toni, 2020b; Arana-Catania et al., 2022; Chamoun et al., 2023; Vladika and Matthes, 2024). However, while justifying the verification of a claim's veracity is the most important part of the manual process, most existing fact-checking systems are unable to explain their decisions, which could assist human fact-checkers and help mitigate the lack of transparency (Baly

[1]The SELFAR datasets, models and code are publicly available: https://github.com/franciellevargas/SELFAR

N.	Sentence-level news article	Label
S1	President Jair Bolsonaro **touch a sore point of** Europeans when he pointed out that the increased use of fossil fuels is a **serious** environmental setback, in his opening speech at the UN General Assembly, Tuesday (20).	Biased
S2	"The St. Francisco River transposition was completed during my government", said Bolsonaro at the UN.	Fake
S3	"Brazil was a pioneer in the implementation of 5G in Latin America", Bolsonaro said at the UN.	Fact
S4	Bolsonaro signed measures favouring to environmental protection during the 4 years of the Brazilian government.	Fake
S5	The Bolsonaro also requested for reform of the UN Security Council.	Fact
S6	However, there is **a huge difference** between speaking at the UN and being heard at the UN.	Biased

Table 1: Example of low credibility news segmented into sentences extracted from the FactNews (Vargas et al., 2023) and FACTCK.BR (Moreno and Bressan, 2019) datasets. Note that the low credibility news may comprise a mix of complex content such as media bias (unreliable) (S1, and S6), fake (S2 e S4), and facts (S3 and S5).

et al., 2018b). Therefore, automated fact-checking should also be capable to provide justifications in the form of post-hoc explanations for model outputs or by incorporating explanation methods directly into these models (Kotonya and Toni, 2020a).

Explainable Artificial Intelligence (XAI) methods provide the causes of a single prediction, a set of predictions, or all predictions of a model by identifying parts of the input, model, or training data that are most influential on the model outcome (Balkir et al., 2022). Hence, transparency and explainability are related to the notion of "explanations" (Guidotti et al., 2018). In particular, XAI methods are commonly categorized into two aspects: (i) whether they provide *local* or *global* explanations, and (ii) whether they are *self-explaining* or *post-hoc explaining* (Guidotti et al., 2018). Local explanations are provided for individual instances, while global explanations apply to the model's behavior across any input (Balkir et al., 2022). Self-explaining methods rely on the internal structure of the prediction model, making these methods often specific to the model type. Conversely, post-hoc explaining (also know as model-agnostic) methods do not rely on knowledge of the to-be-explained model, but rather only input-output pairs (Balkir et al., 2022).

The most commonly used model-agnostic explainable methods are LIME (*Local Interpretable Model-Agnostic Explanations*) (Ribeiro et al., 2016) and SHAP (*SHapley Additive exPlanations*) (Lundberg and Lee, 2017). The LIME provides local explanations for predictions by perturbing the input data and observing the resulting changes in the model's predictions. On the other hand, the SHAP measures the contribution of each feature to the prediction by considering all possible combinations of features. Unlike LIME, SHAP can be used to generate both local and global explanations. Lastly, recent approaches to automated fact verification have also taken advantage of the high performance achieved through In-Context Learning (ICL) [2] to generate post-hoc explanations for veracity prediction (Zeng and Gao, 2023, 2024).

Here, we introduce the ***Se**ntence-**L**evel **Fa**ctual **R**easoning* (SELFAR) aims to improve explainable fact-checking. It covers the entire fact-checking pipeline, generating post-hoc explanations for each task. Specifically, SELFAR predicts news source reliability and factuality of claims or news articles at the sentence-level for fact extraction and verification, respectively. It then generates post-hoc explanations using SHAP and LIME for fact extraction and zero-shot prompts for fact verification. Based on our findings, the sentence-level prediction of unreliable news by analyzing impartiality and subjectivity is promising for fact extraction and improving explainable fat-checking. Additionally, LIME is better than SHAP in explaining predictions on reliability. Finally, although zero-shot prompts provided high readable explanations, and achieved an accuracy of 0.71 in predicting veracity, their tendency to generate hallucinations remains a challenge.

Our contributions are summarized as follows:

- We study an under-explored and relevant problem: explainable automated fact-checking.
- We introduce the SELFAR, a sentence-level factual reasoning that relies on fact extraction and verification by predicting news source reliability and factuality of a news article or claim at the sentence-level, generating post-hoc explanations using SHAP/LIME and zero-shot prompts. The datasets, models and code are available, which may boost future research.
- We propose the first study and baselines for explainable fact-checking in Portuguese.

[2] *In-context learning* refers to generative model's ability to understand and generate responses based on information provided in the context of the conversation or task at hand (Brown et al., 2020).

2 Related Work

2.1 Explainable Fact-Checking

Explainability in fact-checking systems refers to the ability of models to provide a rationale for their decisions. Regarding the explainable fact-checking pipeline, Kotonya and Toni (2020a) suggest a set of tasks, as shown in Figure 1. Note that the explainable fact-checking pipeline includes both fact extraction and fact verification tasks, along with the generation of suitable explanations related to the system's inputs.

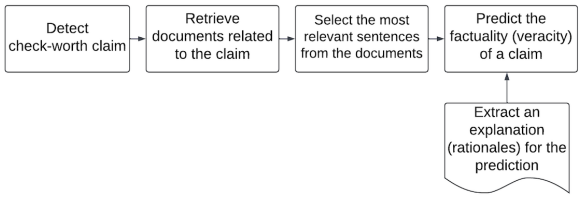

Figure 1: Explainable fact-checking pipeline.

Most existing explainable fact-checking methods produce explanations that consist of the most relevant portions of the system input (Kotonya and Toni, 2020a). Specifically, there are (i) *attention-based explanations*, which rely on the form of some type of visualization of neural attention weights, for example, using LSTM and DNN-based methods with attention mechanisms to extract explanations (Thorne et al., 2019; Popat et al., 2017; Thorne et al., 2019); (ii) *explanation as rule discovery*, that uses rules-based approaches and knowledge graphs to provide explanations (Gad-Elrab et al., 2019; Ahmadi et al., 2019); (iii) *explanation as summarization*, that formulate the automatic generation of explanations as a text summarization problem: extractive text summarization (Atanasova et al., 2020a), or both extractive and abstractive text summarization (Kotonya and Toni, 2020b); (iv) *adversarial claims justification*, that generates adversarial claims (e.g. method that uses a GPT-2 based model) for robust fact-checking (Thorne et al., 2019; Niewinski et al., 2019; Atanasova et al., 2020b); and (v) *retrieved evidence as justifications* that consists of the task of generating justifications based on robust evidence retrieved from data sources (Zeng and Gao, 2024; Wang et al., 2023) or based on prompt engineering enabled by in-context learning (Brown et al., 2020) using zero-shot prompting (Zeng and Gao, 2024; Wang et al., 2023; Zeng and Gao, 2024) or few-shot prompting (Zarharan et al., 2024).

2.2 News Credibility Verification

Estimating the reliability of a news source is relevant not only when fact-checking a claim (Popat et al., 2016); however, it also contributes significantly to tackling article-level tasks such as fake news detection (De Sarkar et al., 2018; Yuan et al., 2020; Reis et al., 2019; Pan et al., 2018; Vargas et al., 2022; Dong et al., 2015). News credibility verification methods have primarily focused on measuring the reliability of news reporting (Pérez-Rosas et al., 2018; Hardalov et al., 2016), the entire media outlet (Baly et al., 2018a; Horne et al., 2018; Baly et al., 2019), and content and user accounts on social media platforms (Castillo et al., 2011; Mukherjee and Weikum, 2015; Iftene et al., 2020) to mitigate various types of harmful strategies. For instance, Yuan et al. (2020) proposed a jointly news credibility and fake news detection structure-aware multi-head attention network (SMAN), which combines the news content, publishing, and reposting relations of publishers and users. Similarly, Long et al. (2017) proposed a new approach to validate the credibility of news articles by analysing a multi-perspective speaker profiles. Iftene et al. (2020) implemented a real-time application based on networks to identify both fake users and fake news over countries and continents in Twitter. Bhattarai et al. (2022) proposed an explainable framework using the Tsetlin[3] that learns linguistic features to distinguish between fake and true news and provides a global interpretation of fake news. In this paper, we estimate the reliability of news sources for fact extraction.

2.3 Fact Verification with Language Models

Large Language Models (LLMs) have been used to provide evidence for fact-checking. For instance, Lee et al. (2021) explored the few-shot capability to assess a claim's veracity based on the perplexity of evidence-conditioned claim generation. Zhang and Gao (2023) proposed a prompt engineering-based method for fact verification that leverages LLMs to separate a claim into sub-claims and then verify each of them through multiple progressive question-answering. Additionally, the reasoning capabilities of LLMs have also been used to address misinformation. For example, Press et al. (2023); Jiang et al. (2023) concluded that LLMs' reasoning capabilities, combined with external knowledge, are promising for a wide range of NLP tasks, including fact extraction and fact verification tasks.

[3]A Tsetlin machine is an AI algorithm based on propositional logic.

3 The Proposed Approach

3.1 Sentence-Level Factual Reasoning

Building on the explainable fact-checking pipeline proposed by Guo et al. (2022), this paper introduces a new method called SELFAR to enhance explainable fact-checking. SELFAR encompasses three main tasks: *Fact Extraction (FE), Fact Verification (FV)*, and *Explanation Generation (EG)*, as shown in Figure 2, and described in detail as follows.

Fact Extraction (FE): According to Guo et al. (2022), fact extraction relies on predicting the most relevant claims to be checked. Therefore, we propose an approach for *sentence-level news source reliability estimation* using a fine-tuned mBERT model. In the context of misinformation, unreliable news and media outlets are targets of a substantial amount of misleading content, often presented as evidence in the form of hyper-partisan or subjective language (Kotonya and Toni, 2020a). Hence, the main hypothesis is that accurately estimating source reliability can be achieved by analyzing the subjectivity and impartiality of text at the sentence level. In particular, our model classifies each sentence into two categories: *reliable* and *unreliable*. Reliable sentences are presented impartially and focus on objective facts. Conversely, unreliable sentences are presented with partiality and therefore focus on subjective interpretations. Table 1 shows examples of biased (unreliable) sentences.

Fact Verification (FV): According to Guo et al. (2022), fact verification relies on finding appropriate evidence and predicting whether that evidence supports or refutes the claim given as input. Since the required evidence can often be unrefined or unavailable, either due to gaps in the knowledge sources (Alhindi et al., 2018), we propose a model for *sentence-level factuality prediction* using LLMs. This model predicts whether a sentence is *fact* or *fake* using retrieved evidence from LLMs, which are trained on a large number of diverse data repositories. It checks whether the evidence of veracity for the sentence is refuted or supported. As example of sentences classified according to their veracity, Table 1 shows examples of fake content and facts.

Explanation Generation (EG): According to Kotonya and Toni (2020a), explainable fact-checking must include the task of extracting an explanation for the prediction. Instead of generating explanations solely for fact verification, we propose the post-hoc explanation generation for both fact extraction and fact verification tasks.

Explanation generation for fact extraction: We used LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) methods to generate post-hoc explanations for fact extraction. These methods produce explanations based on a vector of tokens, where the coefficients represent the most relevant features for predicting a class. In particular, we measure the performance of LIME and SHAP in generating post-hoc explanations for sentence-level news source reliability estimation. Figure 3 shows examples of explanations generated by LIME and SHAP. Note that for each sentence given as input to these methods, they assign a value for a set of tokens. The red bars show the value assigned to the most relevant features to predict the class *unreliable*, and the blue bars show the value assigned to the relevant features to predict the class *reliable*.

Explanation generation for fact verification: We proposed a set of zero-shot prompts using ChatGPT 4.0 (OpenAI et al., 2024) to generate post-hoc explanations for factuality (veracity) prediction at the sentence level. Zero-shot prompting is a technique in which specific examples for that task are not required. Instead, the model generalizes from examples of other related tasks. Table 2 shows post-hoc explanations generated by the zero-shot prompts.

4 Experimental Setup

4.1 Model Architecture and Settings

We propose an approach for fact extraction using a fine-tuned mBERT model, a second approach for fact verification using retrieved evidence from LLMs, and two approaches for post-hoc explanation generation using LIME/SHAP and zero-shot prompts. We describe these approaches as follows.

Fine-Tuned mBERT: We used the fine-tuned mBERT model proposed by Vargas et al. (2023). In essence, this model classifies news article sentences as *reliable* or *unreliable*. It was trained on the FactNews dataset (Vargas et al., 2023), which comprises 6,191 annotated sentences in Portuguese.

Retrieved-Evidence from LLM: Due to the success of ICL across NLP benchmarks, we proposed a set of zero-shot prompts and manually assessed them using ChatGPT 4.0 to recover evidence. The proposed prompts are shown in Table 2. Moreover, to predict factuality, we considered a set of spans described in Table 4 provided as recovered evidence. For this task, we utilized the checked claims from fact-checking organizations in the FACTCK.BR dataset (Moreno and Bressan, 2019) in Portuguese.

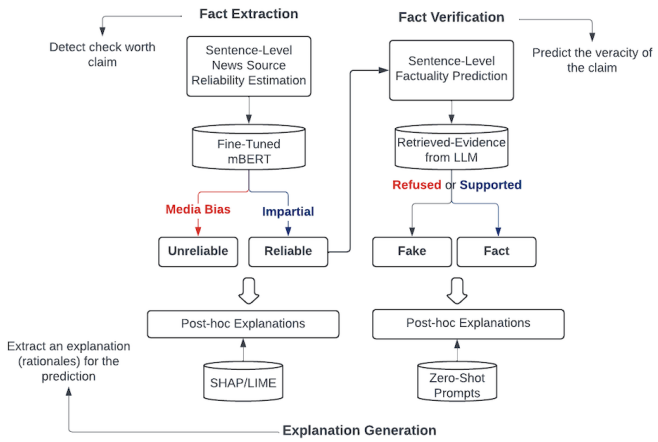

Figure 2: The ***Sentence-Level Factual Reasoning*** (SELFAR) for explainable fact-checking.

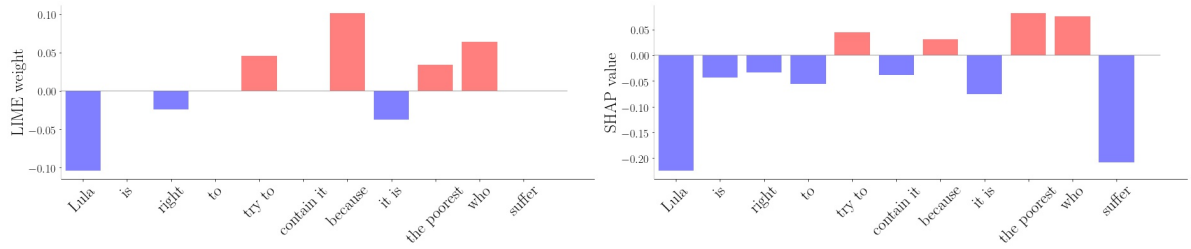

Figure 3: Explanations generated at sentence-level by LIME and SHAP for the unreliable class.

Claims	Zero-Shot Prompts	Sentences	Explanations	Type
Fake	Is it (fake | lie) that <sentence>? Show evidence	Vaccines cause autism.	**Yes, it is fake that** vaccines cause autism. This idea is based on a fraudulent 1998 study led by Andrew Wakefield that suggested a link between the MMR vaccine (measles, mumps, and rubella) and autism.	Supported
		Brazil places tanks on border with Venezuela.	As of my latest update, there haven't been reports confirming that Brazil has placed tanks on its border with Venezuela	Refused
True	Is it (fact | true) that <sentence>? Show evidence	40% of crimes against women occur at home.	**Yes, it is true that** around 40% of crimes against women in Brazil occur at home. According to recent data from the IBGE and other sources that monitor feminicides, most of these crimes occur in domestic settings, which reflects the impact of intra-family violence. This pattern of domestic violence reinforces the need for public policies focused on protecting women in their own homes (Correio Braziliense).	Supported
		Brazil spends R$3 per person per day on everything from vaccinations to transplants in the SUS.	The specific amount can vary depending on the year and the data source, but it's clear that the figure of R$3 per person per day is likely inaccurate	Refused

Table 2: Explanations (justifications) generated at sentence-level by ChatGPT using zero-shot prompts.

LIME and SHAP Post-hoc Explanations: We proposed a post-hoc explanation method using SHAP and LIME for fact extraction. We randomly selected 510 sentences from the FactNews dataset, equally labeled as unreliable and reliable. Then, we asked a linguist, who is an NLP expert, to annotate rationales for the sentences classified as unreliable. An example of the annotated rationales is shown in bold in Table 1. Note that the rationales were annotated by an expert and consist of segments that justify the classification of sentences as unreliable.

Zero-Shot Prompt Post-hoc Explanations: We proposed a set of zero-shot prompts using ChatGPT 4.0 to generate explanations for fact verification. We randomly extracted an average of 400 claims from the FACTCK.BR dataset, equally classified as fake and true. Then, we segmented them into sentences, totaling 510 sentences. The proposed prompts and their generated explanations are shown in Table 2. It should be noted that we used the same number of instances (510 sentences) to evaluate both proposed explainability methods.

5 Evaluation and Results

5.1 Evaluation of Models

We evaluated our models using F1-score, as shown in Table 3. The results are available on GitHub[4].

	FE	FV	SELFAR
class	F1	F1	F1
0	0.85	0.61	0.60
1	0.82	0.81	0.85
Avg	0.85	0.71	**0.72**

Table 3: Evaluation for FE, FV and SELFAR. Note that as shown in Figure 2, For FE, the classes are reliable (0) and unreliable (1). Conversely, for FV and SELFAR, the classes are fact/true (0) and fake (1).

For the FE evaluation, we reported the prediction results obtained from the fine-tuned mBERT model. We also conducted a ROC error analysis, as shown in Figure 4. Note that the FE model achieved a high F1-Score of 0.85 and an AUC of 0.92, which corroborates our hypothesis that analyzing subjectivity and impartiality in text at the sentence level is promising for predicting news source reliability.

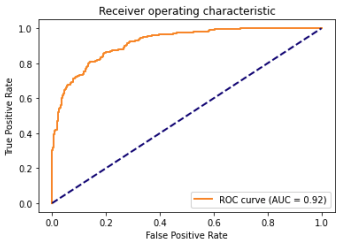

Figure 4: ROC curves for fine-tuned mBERT model.

For the FV evaluation, we assessed the ability to predict whether a sentence is fake or fact/true using recovered evidence from a set of zero-shot prompts shown in Table 2. Specifically, we classified as *supported* recovered evidence included any of the spans described in Table 4. Otherwise, it was classified as *refused* (see examples in Table 2).

For the FV and SELFAR evaluations, we used 510 sentences extracted from the FACTCK.BR. Specifically for SELFAR, we first applied the FE model, which predicts whether a sentence is reliable or unreliable. We then selected only the sentences classified as *reliable* and used them as input for the FV model. Finally, we computed the F1-Score for factuality prediction using our retrieved-evidence from LLMs method. As shown in Table 3, the FV model performs poorly in predicting true claims, indicating that the prompts designed for fake claims may be more effective for predicting veracity.

[4] https://github.com/franciellevargas/SELFAR

	Fake	**True**
Spans	<Yes>; <Yes, it is a lie>; <Yes, it's fake/false>; <Yes, that/this statement is a lie>; <There is no evidence>; <There is no reliable evidence or records>; <Yes, that seems to be a lie/fake>; <It can be considered fake>; <It is not true that>; <Yes, the statement <sentence>is fake/lie>; <Yes, the statement <sentence>is fake>.	<Yes>; <Yes, it is true that>; <Yes, that/this statement is true/fact>; <Yes, there is evidence>; <It is consistent with the available data>; <The evidence suggests>; <The evidence points to true>; <It is true that>; <Yes, the statement <sentence>is true/fact>;<The available evidence confirms>.

Table 4: Spans used to predict factuality by retrieved-evidence from LLM using zero-shot prompts.

Finally, we observed that ChatGPT can report inaccurate or false information. For example, in the prompt, *Is it true that Rodrigo Maia (a Brazilian politician) was not born in Brazil?*, the verdict was, "No, Rodrigo Maia was born in Brazil". However, Rodrigo Maia was actually born in Chile[5]. Similarly, in the prompt, *Is it true that the law regulating the profession of translator and interpreter of Brazilian Sign Language (Libras) was created by Maria do Rosário?*, the verdict was, "This law was proposed by Otávio Leite". However, the fact is that the Brazilian politician Maria do Rosário is the one who created this law[6].

5.2 Evaluation of Explanations

5.2.1 Metrics

We evaluated the EG methods using *faithfulness*, *plausibility* and *readability*. Theses metrics focus on different aspects of the quality of these explanations. For instance, faithfulness measures whether the explanation accurately captures the real relationships between the input features and the model's output. On the other hand, plausibility measures whether the explanation is understandable and intuitive from a human perspective, particularly for domain experts. Finally, readability measures how easily a human can understand the explanations.

Plausibility: We report the IOU (Intersection-Over-Union) F1-score, and as token-level Precision, Recall, and F1-score metrics (DeYoung et al., 2020) to measure plausibility. These scores are computed at the token level, comparing the model's rationales against tokenized human-annotated ones.

IOU F1-score is proposed on a token level rationales (DeYoung et al., 2020), in which the IOU is

[5] https://lupa.uol.com.br/jornalismo/2019/03/25/verificamos-maia-chile-brasileiro

[6] https://lupa.uol.com.br/jornalismo/2019/01/02/verificamos-bolsonaro-libras/

given by overlap of tokens in two sets divided by the size of their union, as shown in Equation 1.

$$\text{IOU-F1} = \frac{1}{N}\sum_{i=1} \text{Greater}(IOU_i, 0.5)$$
$$\text{where } IOU_i = \frac{M_i \cap H_i}{M_i \cup H_i} \quad (1)$$

where M_i and H_i represent the rationale set of the i-th instance provided by the model and human respectively; N is the number of instances.

Token-level F1-score is defined in Equation 2, which is also computed on a token level by the overlap of the rationales tokens predicted by the models with the human-annotated ones. To measure the Token-level F1 score, we measured the Token-level Precision (P_i) and Recall (R_i) and also reported both metrics.

$$\text{Token-F1} = \frac{1}{N}\sum_{i=1}^{N}(2 \times \frac{P_i \times R_i}{P_i + R_i})$$
$$\text{where } P_i = \frac{M_i \cap H_i}{M_i} \text{ and } R_i = \frac{M_i \cap H_i}{H_i} \quad (2)$$

Faithfulness: We report two metrics: *comprehensiveness* and *sufficiency* (DeYoung et al., 2020) to measure faithfulness.

Comprehensiveness measures whether the tokens necessary for making a prediction were selected. To calculate rationale comprehensiveness, for each instance x_i, we construct a contrasting example $\tilde{x}_i$, which is x_i without the predicted rationales r_i[7]. Let $m(x_i)_j$ be the original prediction provided by a model m for the predicted class j for the instance x_i. We then define $m(x_i \backslash r_i)_j$ as the predicted probability of $\tilde{x}_i$ by the model m for class j. The comprehensiveness score is shown in Equation 3. A high comprehensiveness value implies that the rationales are influential in the prediction.

$$\text{Comp} = \frac{1}{N}\sum_{i=1}^{N}(m(x_i)_j - m(x_i \backslash r_i)_j) \quad (3)$$

Sufficiency measures the degree to which the predicted rationales are adequate for a model to make a prediction. The sufficiency score is shown in Equation 4. Where $m(r_i)_j$ is defined as the prediction probability of giving only the predicted rationales r_i to a model m for class j. A low sufficiency implies the rationales are sufficient to make a prediction.

$$\text{Suff} = \frac{1}{N}\sum_{i=1}^{N}(m(x_i)_j - m(r_i)_j) \quad (4)$$

Readability: We applied *Flesch Reading Ease* (Flesch, 1948) and *Szigriszt-Pazos Index* (Pazos, 1993), both of which are applicable to Portuguese, to evaluate zero-shot prompt post-hoc explanations.

[7]We select the top k tokens from the rationales to remove, where k is defined as the average length of the token sets predicted by each explainability model.

5.2.2 Results

Tables 5 and 6 present the evaluation results of explanations generated by LIME, SHAP, and zero-shot prompt methods from the perspectives of plausibility and faithfulness for LIME and SHAP, and readability for the zero-shot prompts. Our evaluation revealed that for class 0 (the reliable sentences), both SHAP and LIME yielded poor results. One possible explanation is that the words used to identify unreliable sentences, which are predominantly subjective, have a much greater impact on predicting unreliable sentences compared to those used to identify reliable sentences. Additionally, the zero-shot prompt post-hoc explanations achieved high readability. We also observed that the prompts proposed for fake claims generated more readable explanations compared to those for true claims.

Quantitative Analysis: When examining unreliable sentences, the rationales highlight the tokens that contribute to media bias. Removing these tokens from the sentence would make the remaining text appear less unreliable, thus altering the classification probability. This effect does not occur with reliable sentences, so we cannot observe similar effects when computing comprehensiveness and sufficiency metrics for this class. In Table 5, We observe that LIME performs better on faithfulness metrics, while SHAP excels in plausibility metrics. However, the number of tokens returned as rationales by each method differs significantly. LIME, by default, returns a maximum of 10 tokens, whereas SHAP returns more. The plausibility metrics are computed by comparing these tokens against human-annotated rationales, which are often more complex and contextually rich, such as entire phrases. Consequently, the intersection between LIME's tokens and human-annotated rationales is generally smaller than SHAP's, leading to lower metric scores for LIME. Despite this, the token-level precision is higher for LIME because this metric is calculated as the intersection divided by the total number of tokens retrieved by the method (SHAP or LIME). Since LIME retrieves fewer tokens than SHAP, it achieves a higher precision. However, LIME's recall performance is significantly worse. When examining the performance of both methods on faithfulness metrics, the situation is reversed, with LIME showing superior results for both comprehensiveness and sufficiency metrics. One possible explanation is that the words selected by LIME have a more significant impact on the model's prediction. Since LIME selects fewer words than SHAP, it may focus

Methods	Plausibility				Faithfulness	
	IOU F1 ↑	Token Precision ↑	Token Recall ↑	Token F1 ↑	Comp. ↑	Suff. ↓
BERT-LIME	0.1098	**0.4378**	0.3913	0.3698	**0.2961**	**-0.0546**
BERT-SHAP	**0.1529**	0.4312	**0.5111**	**0.4285**	0.2868	-0.0491

Table 5: Evaluation of post-hoc explanations generated by LIME and SHAP methods.

Method	Readability			
	Flesch Reading Ease		Szigriszt Pazos Index	
Zero Shot Prompts	True	Fake	True	Fake
	0.77	**0.84**	-1519.48	**-1361.47**

Table 6: Evaluation of post-hoc explanations generated from zero-shot prompts.

more on the most critical words for the prediction, thus improving the faithfulness metrics. This observation aligns with the qualitative analysis conducted by a specialist and described below. In Figure 5, we present the top 20 most important words predicted by LIME and SHAP. This includes 10 words most important for predicting the unreliable class (red bars) and 10 most important for predicting the reliable class (blue bars). This analysis is based on all 510 selected sentences. We observed that the vocabulary on the right side of the graphs, representing unreliable words, tends to be more subjective and includes more adjectives. This observation also aligns with the qualitative analysis conducted by a specialist, which is described as follows.

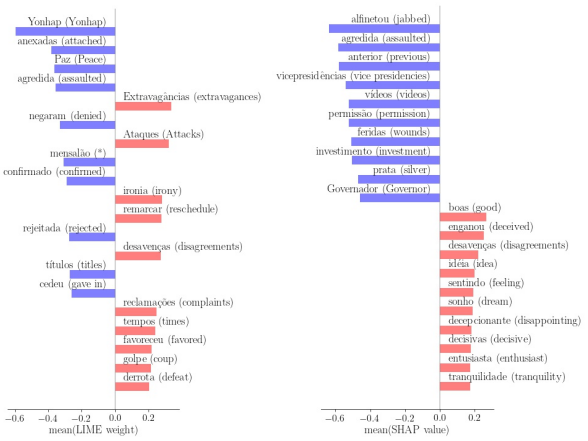

Figure 5: Most relevant features provided by LIME and SHAP to predict each class (reliable/unreliable) for sentence-level news source reliability estimation.

Qualitative Analysis: We also conducted a qualitative analysis with a linguist, comparing LIME and SHAP scores with human-annotated rationales for the most impactful tokens in determining whether a sentence is unreliable. Regarding the agreement between the LIME and the human rationals, abstract verbs that involve subjective interpretation, where the author projects an action or feeling onto the subject (e.g., "attack", "deceive"), were frequently identified as indicative of media bias. Another critical feature of unreliable sentences was the presence of adjectives (e.g., "prudent", "useful") and adverbs (e.g., "negatively"). Regarding disagreements between LIME and human-annotated rationales, LIME often identified articles and prepositions as indicators of bias. In many cases, specific nouns (e.g., "history") were also flagged, although their potential for bias depends on the context in which they are used. Finally, SHAP identified a higher number of articles, prepositions, and possessive pronouns as indicators of bias compared to LIME. However, these terms alone do not necessarily influence the degree of bias in the sentences. Regarding the agreement between SHAP and human-annotated rationales, we observed the same types of terms as with LIME. However, SHAP tended to identify a larger number of nouns and proper nouns as being linked to bias. Thus, there seems to be a greater cohesion between the LIME method and human-annotated rationales for this specific task.

6 Conclusions

This paper introduces a new method to improve explainable fact-checking. The SELFAR predicts reliability and the factuality of news articles or claims at the sentence level, generating post-hoc explanations using LIME/SHAP and zero-shot prompts. Our experiments showed that unreliable news stories are comprised mostly of subjective words, in contrast to reliable ones. Thus, predicting unreliable news stories by analyzing text impartiality and subjectivity is promising for fact extraction and improving explainable fact-checking. In addition, LIME outperforms SHAP in explaining reliability predictions. Lastly, while zero-shot prompts provide highly readable explanations and achieve an accuracy of 0.71 in predicting factuality, their tendency to hallucinate presents a challenge. We also present baselines for explainable fact-checking in Portuguese.

Limitations

Although the proposed method for explainable fact-checking has addressed relevant gaps in providing more accurate and transparent automated fact-checking, the method for retrieving evidence from LLMs for factuality (veracity) prediction may present limitations due to the potential for LLMs to hallucinate. Therefore, as future work, we aim to mitigate this limitation by extracting evidence from multiple and diversified data sources.

Ethics Statement

The data resources and artifacts used in this paper are open source and have been anonymized.

Acknowledgements

The first author is grateful to Google for financial support. This project was partially funded by FAPESP, CNPq, FAPEMIG, CAPES, and the Ministry of Science Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

References

Naser Ahmadi, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming. In *Proceedings of the Conference for truth and trust Online*, pages 1–9, London, UK.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Miguel Arana-Catania, Elena Kochkina, Arkaitz Zubiaga, Maria Liakata, Robert Procter, and Yulan He. 2022. Natural language inference with self-attention for veracity assessment of pandemic claims. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1496–1511, Seattle, United States. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020a. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020b. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, Online. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen Fraser. 2022. Challenges in applying explainability methods to improve the fairness of NLP models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 80–92, Seattle, U.S.A. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018a. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2109–2116, Minneapolis, Minnesota. Association for Computational Linguistics.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics.

Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. 2022. Explainable tsetlin machine framework for fake news detection with credibility score assessment. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4894–4903, Marseille, France. European Language Resources Association.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, page 675–684, New York, United States.

Eric Chamoun, Marzieh Saeidi, and Andreas Vlachos. 2023. Automated fact-checking in dialogue: Are specialized models needed? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16009–16020, Republic of Singapore, Singapore.

João Couto, Breno Pimenta, Igor M. de Araújo, Samuel Assis, Julio C. S. Reis, Ana Paula da Silva, Jussara Almeida, and Fabrício Benevenuto. 2021. Central de fatos: Um repositório de checagens de fatos. In *Anais do III Dataset Showcase Workshop*, pages 128–137, Porto Alegre, Brasil. SBC.

Sohan De Sarkar, Fan Yang, and Arjun Mukherjee. 2018. Attending sentences to detect satirical fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3371–3380, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 3(32):221–233.

Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. Exfakt: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, page 87–95, New York USA.

Max Glockner, Ieva Staliunait, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2023. Ambifc: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 1:37–53.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5).

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, page 493–503, Hong Kong, China.

Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry Ilvovsky, and Preslav Nakov. 2022. Crowd-Checked: Detecting previously fact-checked claims in social media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 266–285, Online only. Association for Computational Linguistics.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2016. In search of credible news. In *17th International Conference on Artificial Intelligence: Methodology, Systems, and Application*, pages 172–180, Varna, Bulgaria.

Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Proceedings of the The Web Conference 2018*, page 235–238, Geneva, Switzerland.

Adrian Iftene, Daniela Gifu, Andrei-Remus Miron, and Mihai-Stefan Dudu. 2020. A real-time system for credibility on Twitter. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6166–6173, Marseille, France. European Language Resources Association.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore, Singapore.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, Online. Association for Computational Linguistics.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 252–256, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Julie Mastrine. 2022. *How to Spot 16 Types of Media Bias*. AllSides: Don't be fooled by media bias and misinformation, California, United States.

João Moreno and Graça Bressan. 2019. Factck.br: a new dataset to study fake news. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, page 525–527, Rio de Janeiro, Brazil. Association for Computing Machinery.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, page 353–362, New York, United States.

Nona Naderi and Graeme Hirst. 2018. Automated fact-checking of claims in argumentative parliamentary debates. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 60–65, Brussels, Belgium. Association for Computational Linguistics.

Piotr Niewinski, Maria Pszona, and Maria Janicka. 2019. GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 20–26, Hong Kong, China. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff,

Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *The Semantic Web - ISWC 2018*, pages 669–683.

Manuel Martín Szigriszt Pazos. 1993. Índices de legibilidade formulados para a língua espanhola. *SEECI, Revista de la Facultad de Ciencias de la Información*.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, page 2173–2178, New York, United States.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, page 1003–1012, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore, Singapore.

Julio C. S. Reis, Andre Correia, Fabricio Murai, Adriano Veloso, and Fabricio Benevenuto. 2019. Explainable machine learning for fake news detection. In *Proceedings of the 11th ACM Conference on Web Science*, pages 17–26, Massachusetts, United States.

Julio C. S. Reis, Philipe Melo, Kiran Garimella, Jussara M. Almeida, Dean Eckles, and Fabrício Benevenuto. 2020. A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. In *Proceedings of the 14th International AAAI Conference on Web and Social Media*, pages 903–908, Held Online.

Manoel Horta Ribeiro, Savvas Zannettou, Oana Goga, Fabrício Benevenuto, and Robert West. 2022. Can online attention signals help fact-checkers to fact-check? In *Workshop Proceedings of the 17th International AAAI Conference On Web and Social Media*, pages 1–10, Atlanta, United States.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. AraFacts: The first large Arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2944–2953, Hong Kong, China. Association for Computational Linguistics.

Francielle Vargas, Jonas D'Alessandro, Zohar Rabinovich, Fabrício Benevenuto, and Thiago Pardo. 2022. Rhetorical structure approach for online deception detection: A survey. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5906–5915, Marseille, France. European Language Resources Association.

Francielle Vargas, Kokil Jaidka, Thiago Pardo, and Fabrício Benevenuto. 2023. Predicting sentence-level factuality of news and bias of media outlets. In *Proceedings of the 14th International Conference on ecent Advances in Natural Language Processing*, pages 1197–1206, Varna, Bulgaria.

Juraj Vladika and Florian Matthes. 2024. Comparing knowledge sources for open-domain scientific claim verification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2103–2114, St. Julian's, Malta.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. Check-COVID: Fact-checking

COVID-19 news claims with scientific evidence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14114–14127, Toronto, Canada. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2020. Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5444–5454, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Majid Zarharan, Pascal Wullschleger, Babak Behkam Kia, Mohammad Taher Pilehvar, and Jennifer Foster. 2024. Tell me why: Explainable public health fact-checking with large language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 252–278, Mexico City, Mexico. Association for Computational Linguistics.

Fengzhu Zeng and Wei Gao. 2023. Prompt to be consistent is better than self-consistent? few-shot and zero-shot fact verification with pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4555–4569, Toronto, Canada. Association for Computational Linguistics.

Fengzhu Zeng and Wei Gao. 2024. JustiLM: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics*, 11:334–354.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 996–1011, Nusa Dua, Bali.