Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5787

Comparing Explainable AI Techniques In Language Models: A Case Study For Fake News Detection in Portuguese

Jéssica Vicentini [**Unesp - São José do Rio Preto** | ***jessica.vicentini@unesp.br***]
Rafael Bezerra de Menezes Rodrigues [**Unesp - Rio Claro** | ***rafael.rodrigues@unesp.br***]
Arnaldo Candido Junior [**Unesp - São José do Rio Preto** | ***arnaldo.candido@unesp.br***]
Ivan Rizzo Guilherme [**Unesp - Rio Claro** | ***ivan.guilherme@unesp.br***]

Universidade Estadual Paulista, R. Cristóvão Colombo, 2265 - Jardim Nazareth, São José do Rio Preto, SP, 15054-000, Brazil

Received: 30 March 2025 • **Accepted:** 29 July 2025 • **Published:** 21 January 2026

Abstract Language models are widely used in natural language processing, but their complexity makes interpretation difficult, limiting their adoption in critical decision-making. This work explores Explainable Artificial Intelligence (XAI) techniques, such as LIME and Integrated Gradients (IG), to understand these models. The study evaluates the effectiveness of BERTimbau in classifying Portuguese news as true or fake, using the FakeRecogna and Fake.Br Corpus datasets. In the experiments, LIME proved to be easier to interpret than IG, and both methods showed limitations when applied to texts, as they focus only on the morphological and lexical levels, ignoring other important levels.

Keywords: Explainable artificial intelligence, BERTimbau, Local interpretable model-agnostic explanations, Integrated gradients, Natural language processing, Machine learning, Deep learning, Language models, Transformer.

1 Introduction

Advancements in Artificial Intelligence (AI) and its incorporation into applications increasingly widespread among people resulted in the widely adoption of natural language processing (NLP) models and machine learning algorithms, especially language models. But the transparency of these models, often referred to as "black boxes", raises questions about their reliability and understandability. These models are capable of providing answers and solving problems, but they do not detail how exactly they arrived at a particular conclusion [Shevskaya, 2021].

Furthermore, in regulated sectors such as healthcare and finance, explainability is a requirement to prevent discrimination and unfair practices. Therefore, explainability plays a key role in the development and application of AI-powered systems [Ahmed *et al*., 2022]. Therefore, understanding the behavior of a model is crucial to evaluating its reliability. Additionally, with a better understanding of the model, insights can be gained that can be used to improve its ability to generate reliable information.

In response to this challenge of improving interpretability, the emerging field of Explainable Artificial Intelligence (XAI) has been developed to clarify how models make decisions and explain those decisions in a way that can be easily understood by users. These tools have the potential to increase the transparency and reliability of these models, enabling continuous improvement.[Gohel *et al*., 2021].

This work has the objective of investigating whether decisions made by models for natural language processing are reliable, particularly in the context of fake news detection for Brazilian Portuguese. Our proposal presents four main contributions:

- A qualitative comparison between two widely used XAI methods, Local Interpretable Model-Agnostic Explanations (LIME) [Ribeiro *et al*., 2016] and Integrated Gradients (IG) [Sundararajan *et al*., 2017]. Our analysis favors LIME in terms of easiness of interpretation, which is consistent with related work.
- Identification of limitations in these explanation methods regarding textual data. We show that they are too restricted to morphological and lexical features, lacking clear ways to express other linguist levels information, which also plays a relevant role in fake news classification (ex.: semantics for name entity present in the text; rhetoric for text length and structure and syntactic for text adherence to grammar rules).
- A snapshot containing a fake news profiling of fake news from Brazilian political scenario in the period of 2016 – 2021. Our results points an evolving picture of fake news profile, moved mainly by political matters, regarding public figures and organizations.
- A quantitative analysis of which explanations are more robust: those from LIME or those from IG. This is done by altering the input texts, removing the words identified as more important according to each method and then evaluating the impact on the classifier. Our findings show that even in the lexical level, the explanations have limitations, as removing words detected as important have limited impact in the confidences of the fake news classifier.

Additionally, we also provide a performance analysis of a BERTimbau-based model [Souza *et al*., 2020] for fake news detection in Portuguese, and we also evaluate whether stopwords have a relevant role in the classification process given

some fake news are poorly written and have a bellow than average textual structure.

This article is structured as follows: Section 2 presents related works; Section 3 describes the XAI methods used, the dataset, materials employed, and how the experiments to train fake news detectors were conducted. Section 4 focus on the XAI analysis, along with a case study over explanations generated by LIME and IG for selected texts. Section 5 evaluates how the classification model behaves in out-of-distribution scenarios and the corresponding impact on the explainability methods. Section 6 provides statistics of the dataset, exploring at which extent linguistic levels not covered by the explanations impacts on classifier performance. Section 7 discusses the main findings. Finally, Section 8 presents the conclusions and future work.

2 Related Work

Related works are centered on three main axes: XAI method comparison; XAI Applied to fake news detection; XAI focusing on the Portuguese Language. Regarding XAI comparison methods, we focus on LIME and IG algorithms instead of other XAI techniques because these methods are widely adopted. However, related works also include other methods. For example, Moradi and Samwald [2021] present a comparison of several methods. The authors examine the fidelity of the scores obtained in explanations using their own proposal, BioCIE, specialized in medical texts, and compare it with LIME, MUSE [Lakkaraju *et al*., 2019], and Greedy and Random baselines [Moradi and Samwald, 2021]. BioCIE obtained the best results; however, LIME came in second place, proving to be a flexible solution for obtaining insights about language models.

In the study by Mersha *et al*. [2025], the authors investigate five different methods for XAI for the task of sentiment mining in texts using IMDb as the reference dataset. The methods investigated were LIME, IG, Layer-wise Relevance Propagation and Attention Mechanism Visualization. In total, five LLMs (Large Language Models) were evaluated using following metrics: consistency, and contrastivity, robustness and human-reasoning agreement. LIME showed consistently good scores across all evaluated metrics, achieving the best human-reasoning agreement and ranking second among the methods on the remaining metrics.

Regarding the axis of XAI for fake news detection, Pendyala and Hall [2024] explored the use of LLMs for misinformation detection, investigating their ability to verify new information based on the knowledge learned during training. In their study, the authors analyzed the models LLama, Orca, Falcon, and Mistral, assessing their effectiveness across multiple datasets. To interpret the results, they used explainability techniques such as LIME, SHAP, and Integrated Gradients, in addition to asking the LLMs themselves to explain their classifications. The findings highlighted that the effectiveness of these models in detecting misinformation heavily depends on the quality and scope of the data they were trained on. Furthermore, the study reinforces the need for explainable techniques to understand how these models reach their decisions, especially in the context of misinformation containment. Among the evaluated techniques, SHAP provided the most varied and detailed attributions, making it useful for identifying key words that influence the model's decision. Integrated Gradients excelled in identifying critical tokens for classification, while LIME offered localized and easily interpretable explanations.

Desai *et al*. [2024] presented an approach for detecting fake news and hate speech, utilizing machine learning models to identify typical patterns in these phenomena. The authors also incorporated explainable artificial intelligence (XAI) techniques, such as LIME and SHAP, to increase the transparency of their models. By applying these XAI methods, the study aims to provide clearer insights into the decision-making process, enabling a deeper understanding of how specific features influence the classification of news articles as fake or genuine, and hate speech or not. In their analysis, the authors experimented with different XAI methods to explain the predictions made by their machine learning models. They focused on evaluating the performance of logistic regression models in detecting fake news and hate speech, using datasets with labeled examples. The results highlighted that explainable models could not only improve model performance but also help fine-tune decision parameters, addressing the trade-off between accuracy and interpretability. The authors concluded that the method has the potential to bridge the gap between model accuracy and transparency, enhancing the trustworthiness of AI systems in tackling complex issues like misinformation.

Regarding our third axis of investigation, although XAI techniques are popular, there are not many works applying these techniques to taks involving Brazilian Portuguese datasets, to the best of our knowledge. One of the few works in Portuguese is by Oliveira *et al*. [2023], who carried out a study with the aim of proposing and evaluating approaches for estimating the cohesion of essays in Portuguese and English. They used the SHapley Additive exPlanations (SHAP) technique to examine the explainability of the approaches SHAP [Lundberg and Lee, 2017]. The authors found that SHAP provides better explanation for traditional machine learning algorithms when compared to deep learning-based models.

Lima *et al*. [2024] also focused on Portuguese. The authors applied Integrated Gradients to investigate the interpretability of the T5 model in the task of punctuation restoration for Brazilian Portuguese. The technique was used to highlight the most relevant tokens in correct predictions, with a particular focus on student-written essays. Their analysis showed that the model captured grammatical patterns such as the use of commas in enumerations and the influence of specific verbs in determining punctuation. These results suggest that IG can effectively reveal how the model internalizes linguistic rules, even in texts that are structurally inconsistent or contain common errors. Moreover, the authors emphasize the pedagogical potential of explainability techniques like the IG. By making the model's decisions more transparent, especially in educational contexts, IG can support the development of automated feedback tools that not only correct but also justify their suggestions. This work reinforces the relevance of applying XAI methods in underexplored linguistic settings and demonstrates that IG is a valuable tool for interpreting

token-level contributions in transformer-based models.

Other lines of investigation include works such as Moraliyage *et al*. [2025], who examine XAI techniques to respond to adversarial attacks on language models. For this, they use IG tool to detect adversarial attacks in text classifiers. The work is based on the hypothesis that the most relevant words for the model's decisions, as identified by IG, exhibit different patterns in original versus adversarially perturbed texts. To explore this, the authors train a secondary classifier that uses the importance scores generated by IG to distinguish between clean and adversarial inputs. Their study evaluates various types of attacks, including HotFlip and TextFooler, across multiple datasets, demonstrating the robustness of the proposed approach. The results show that this method can effectively detect adversarial examples, achieving high accuracy and low false positive rates, outperforming traditional statistical baselines. Moreover, the use of IG adds an interpretability layer, helping to understand model decisions and how perturbations affect the input. This work highlights the potential of XAI techniques not only to explain model predictions but also as active tools for improving security in NLP. The approach is particularly relevant in sensitive domains, such as misinformation detection or educational applications, in which adversarial attacks can compromise trust and fairness in automated systems.

3 Model Training

In this work, we carried out experiments to analyze a language model with two XAI techniques, LIME and IG. They are applied to a language model trained for text classification in Brazilian Portuguese using a set of selected samples. For this purpose, we trained a version of BERTimbau model [Souza *et al*., 2020], namely BERTimbau-Base.

The first XAI technique, LIME [Ribeiro *et al*., 2016], is based on a method to explain the predictions of classification or regression models. Instead of trying to understand the entire model, LIME focuses on explaining how the model arrived at a specific prediction for a data instance (such as an image or text). LIME generates a simple model, called an interpretation model, which is trained locally (i.e., only for the data instance in question) and can be easily interpreted, presenting textual and visual artifacts that provide a qualitative understanding of the relationship between the components of the sample.

The second XAI technique, IG [Sundararajan *et al*., 2017], is a method widely used for interpreting and assigning importance to the characteristics of a machine learning model. As LIME, IG is also a local method, i.e., it provides an explanation for a single data instance given its prediction. The main objective of this method is to attribute individual contributions to the model input features, providing a quantitative measure of the importance of each feature in relation to the final prediction. When applied to BERTimbau, IG returns a saliency score for each subtoken in the input. However, for better visualization and comparison against LIME, one score per token (not subtoken) was considered in this work. The maximum saliency amongst all subtokens of a token was taken to represent the whole token.

3.1 Dataset

We used two datasets in Brazilian Portuguese to train BERTimbau and evaluate its predictions using LIME and IG: Fake.Br Corpus [Santos *et al*., 2018] and FakeRecogna [Garcia *et al*., 2022]. The dissemination of fake news is a significant problem worldwide, especially on social media. To combat this practice, researchers in the field of computer science have developed techniques and tools capable of identifying and combating the spread of these news. Fake.Br Corpus contains news from 2016 to 2018, while and FakeRecogna texts range from 2019 to 2021. We used 7,200 samples from Fake.Br and 11,901 samples from FakeRecogna. We divided both datasets into training and testing sets using 80 and 20% of the samples, respectively.

3.2 Training and Testing

We trained BERTimbau for both Fake.Br and FakeRecogna, in two settings: with and without stop words. We used this strategy due to an initial hypothesis that part of the fake news lack good textual structure due to being poorly written. In the version without stop words, we also removed numbers. Table 1 shows time spent training both models, was well as the model accuracy and loss.

We used the Adam optimizer with the learning rate set to 1e-06. Both models were trained over 10 epochs, with batch size set to 12. Our experiments are based on BERTimbau-Base model provided by the Hugging Face Transformers[1] library. The tokenizer was also loaded from the pre-trained model, used to convert the input text into a sequence of tokens that the model can understand. For training we used the TensorFlow library [Abadi *et al*., 2015] on a premium Google Colaboratory[2] GPU, specifically the NVIDIA A100-SXM4 with 40 GB of virtual RAM and 80 CUDA multiprocessors.

Regarding the testing results, Model 1, which includes stop words, performed better compared to Model 2, which does not contain them, as can be seen in Table 1, despite our initial hypothesis. The pre-training on well strucutured text may have a role in this result. Model 1 achieved an accuracy of 0.9 and a loss rate of 0.01, while Model 2 obtained an accuracy of 0.95 and a loss rate of 0.14. The same result was observed among the models trained on the Fake.Br Corpus dataset, with Model 3, which includes stop words, outperforming Model 4, which does not include them.

3.3 Sampling

In order to analyze these models, we selected forty samples from the FakeRecogna dataset and the Fake.Br Corpus. From these samples, we selected 20 with the following distribution: 5 true positives (TP); 5 true negatives (TN); 5 false positives (FP); and 5 false negatives (FN). These 20 samples had predictions with high confidence values, while the remaining 20 samples were selected randomly. However, due to high accuracy of the trained models, it was not possible to recover FPs from Model 1 and FNs from Model 3. As result, 37 samples were analyzed from Model 1, 39 from Model 3, while the

[1] https://huggingface.co/

[2] https://colab.research.google.com/.

Table 1. Models and Metrics

Model	Dataset	Stop Words	Training time (mins.)	Test Accuracy	Test Loss
Model 1	FakeRecogna	True	59.22	**0.9945**	**0.0190**
Model 2	FakeRecogna	False	60.86	0.9551	0.1417
Model 3	Fake.Br Corpus	True	**36.15**	0.9910	0.0379
Model 4	Fake.Br Corpus	False	38.82	0.9604	0.1142

Table 2. All experiments conducted with LIME and IG.

Exp.	Method	Model	Dataset	Stop words	Exp.	Method	Model	Dataset	Stop words
1	LIME	Model 1	FakeRecogna	True	5	IG	Model 1	FakeRecogna	True
2	LIME	Model 2	FakeRecogna	False	6	IG	Model 2	FakeRecogna	False
3	LIME	Model 3	Fake.Br	True	7	IG	Model 3	Fake.Br	True
4	LIME	Model 4	Fake.Br	False	8	IG	Model 4	Fake.Br	False

analysis of models 2 and 4 were complete with 40 samples. In total, 156 samples were analyzed using both LIME and IG, resulting in 312 explanations being generated.

Table 2 presents how the experiments were organized. Eight experiments were conducted so that all four models had the selected samples explained by both the LIME and IG.

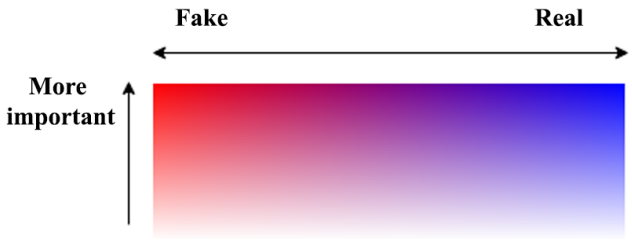


Figure 1. Color scaling of saliency maps.

4 XAI Analysis

From the 312 explanations generated by LIME and IG, we conducted an in-depth manual analysis of 32 explanations to examine which features were highlighted by each method. Due to space constraints, we present only eight examples [3], selected based on recurrent patterns observed across the analyzed cases. Two examples are from Experiment 1 (Table 2), where LIME was used to explain predictions from Model 1; two from Experiment 3, which also employed LIME but for predictions made by Model 3; two from Experiment 5, which used IG to explain predictions from Model 1; and finally, two from Experiment 7 (Table 2), where IG was applied to predictions from Model 3. These examples were selected to enable a direct comparison between the explanations produced by LIME and IG, focusing on the models that achieved the best performance.

Figures 2a, 3a, 4a and 5a contain explanations generated by IG and Figures 2b, 3b, 4b and 5b contain examples of explanations generated by LIME. For LIME, the words are highlighted in two colors: blue for words with negative weight (i.e., they contribute to classifying the news as false, which is also represented by 0) and orange for words with positive weight (i.e., they contribute to classifying the news as real, represented by 1).

For IG, explanations use a color scale to highlight the weight of words (as detailed in Figure 1). The redder the word, the more it contributed to the classification as fake news and the bluer the more it contributed to the news being classified as real, and the more saturated the color, the more important the weight of the word for classification.

As observed in Figures 2a, 3a, 4a and 5a, LIME explanations tend to be more straightforward to interpret. This occurred mainly due to two reasons. First, the number of highlighted words are much smaller in LIME compareed to IG, allowing an analysis more focused on important words. Second, while IG highlights almost all words with saturated colors, it also choose mostly purple HUEs, indicating that these words are equally important for classification as fake and real news, which difficults the analysis.

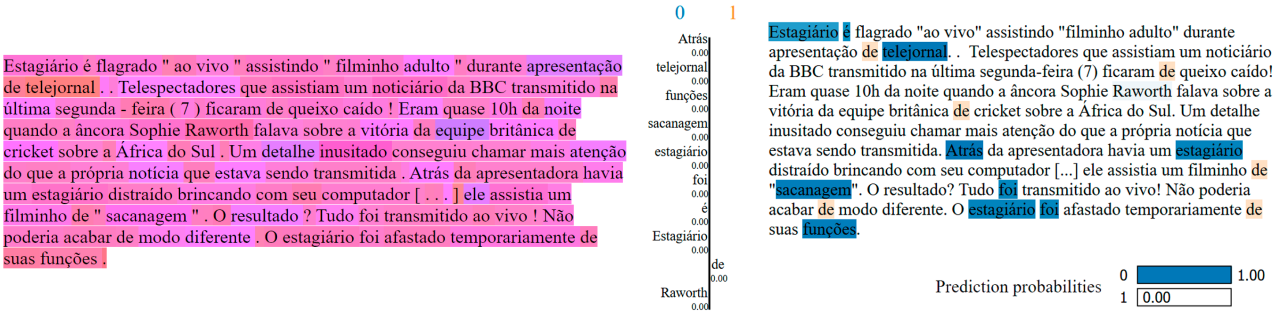
4.1 Fake news classified as fake

Analyzing Figure 2a from IG for a fake news sample classified by Model 3, we observe that the words show similar levels of importance with corresponding colors. Notably, ***Telejornal*** (newscast) received a negative score, possibly indicating that the model associates this term with sources previously discredited for spreading misinformation.

Other elements that received significant negative attributions include common function words such as ***de*** (of), ***suas*** (your), and ***funções*** (functions), as well as punctuation marks like **]** and **period**. While function words and punctuation typically carry limited semantic content, their high importance scores may indicate that the model is either using textual struccuture into the decision process or is relying on superficial patterns rather than deep linguistic understanding.

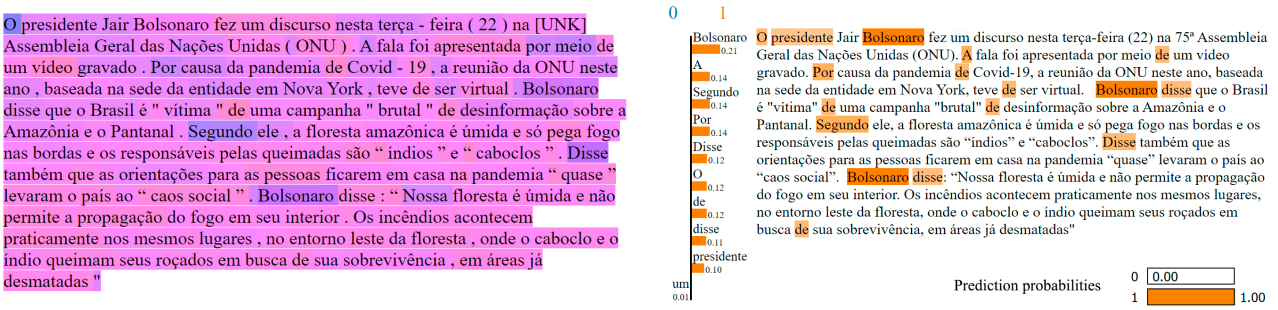
Regarding the LIME explanation for the same sample in the Figure 2b, the word ***Telejornal*** (newscast) once again received a negative attribution, consistent with the result from the IG explanation. However, some differences are evident: the words ***estagiário*** (intern), ***é*** (is), and ***atrás*** (behind) also received negative weights, while ***de*** (of) was assigned a positive score. This contrasts with the IG explanation, in which ***de*** received a negative attribution. These differences illustrate how distinct interpretability methods may highlight different aspects of the model's internal reasoning, emphasizing the value of employing multiple explanation techniques to achieve a more comprehensive interpretation of model behav-

[3]The 32 analyses are available on our GitHub; the link can be found in the Availability of data and materials section



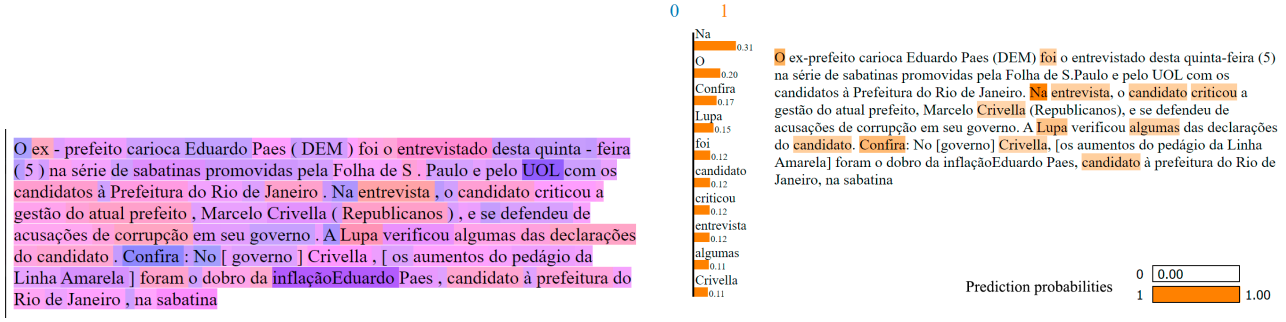
(a) IG: Fake news classified as fake by Model 3 **(b)** LIME: Fake news classified as fake by Model 3

Figure 2. Explanations generated by IG and LIME for a fake news sample classified as fake by Model 3



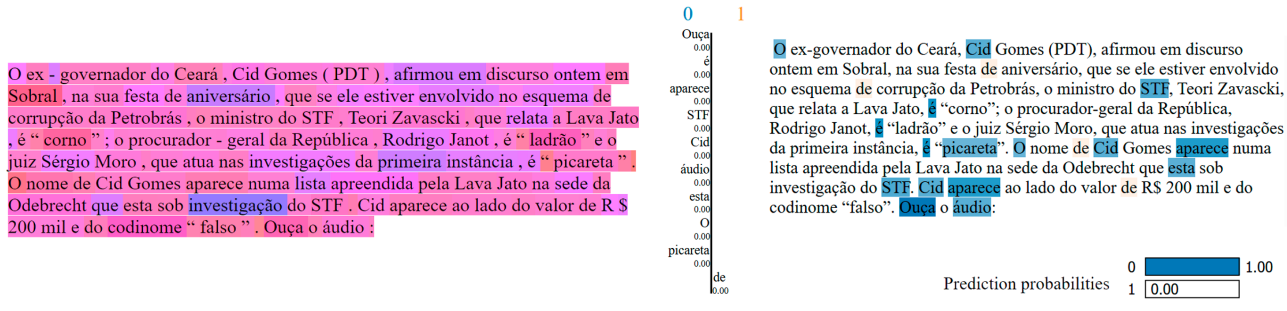
(a) IG: Fake news classified as real by Model 1 **(b)** LIME: Fake news classified as real by Model 1

Figure 3. Explanations generated by IG and LIME for a fake news sample classified as real by Model 1



(a) IG: Real news classified as real by Model 1 **(b)** LIME: Real news classified as real by Model 1

Figure 4. Explanations generated by IG and LIME for a real news sample classified as real by Model 1



(a) IG: Real news classified as fake by Model 3 **(b)** LIME: Real news classified as fake by Model 3

Figure 5. Explanations generated by IG and LIME for a real news sample classified as real by Model 3

ior.

4.2 Fake news classified as real

Analyzing the IG explanation in Figure 3a for a fake news sample misclassified as real by Model 1, it is possible to observe that the word ***O*** (the) received a strong positive attribution. Other words such as ***disse*** (said), ***segundo*** (according to), ***Bolsonaro*** (former president of Brazil), and ***meio*** (means) also contributed positively to the prediction. Many of these tokens are contextually associated with reported speech and the attribution of statements, elements frequently found in political news. The model's positive attribution to these words may indicate a difficulty in distinguishing politically charged narratives from factual.

The LIME explanation for the same sample, shown in Figure 3b, highlights the words ***A*** (the), ***O*** (the), and ***de*** (of) as having positive weights. Although these are common function words in Portuguese, their relevance in the explanation suggests that the model may associate them with more formal or structured writing. This observation points to a possible influence of writing style on the model's decision-making process. Additionally, ***Bolsonaro*** and ***presidente*** (president) are also positively weighted, potentially indicating a learned association between political figures and real news. In contrast, the word ***um*** (a) received a slightly negative attribution, although its impact on the final prediction appears negligible.

4.3 Real news classified as real

Figure 4a presents the IG explanation for a real news article correctly classified by Model 1. In this example, the word ***UOL*** (well-known news portal) received a strong positive attribution, suggesting that the model associates reliable media sources with truthful content. Other words such as ***Confira*** (check), ***No*** (in/on), and ***A*** (the) also received positive contributions, likely due to their role in structuring the sentence and presenting verifiable information. Interestingly, the words ***entrevista*** (interview) and ***Lupa*** (fact-checking agency) received negative attributions, which may indicate some inconsistency in how the model interprets fact-checking contexts or reported speech.

Figure 4b shows the LIME explanation for the same sample. In this case, the word ***Lupa*** receives a positive score, suggesting that the model considers the presence of a fact-checking source as a signal of credibility. This contrast with the IG result highlights differences in how these explanation methods evaluate feature importance. The term ***candidato*** (candidate) contributes positively by providing essential context about the subject of the news piece, while ***Crivella*** (a political figure) also receives a positive attribution, potentially due to the political relevance of the content.

Additionally, the contribution of common function words was examined, based on the hypothesis that real news texts tend to exhibit more coherent and well-structured writing. Words such as ***Na*** (in), ***O*** (the), and ***foi*** (was) received positive scores, likely because of their role in building grammatical structure and temporal coherence.

4.4 Real news classified as fake

According to Figure 5a, the IG explanation for a real news article classified as fake is as follows: ***Sobral*** (Brazilian city) received a negative score, this may be due to a bias in the training data, a spurious correlation, or the existence of related fake news involving this city. ***Investigação*** (investigation) scored positively, suggesting that its presence indicates a responsible approach to uncovering accurate information, thereby enhancing confidence in the news's truthfulness. Conversely, ***ladrão*** (thief) received a negative score, which may reflect its common use in fake news to unjustly accuse individuals, although it is also frequently reported in crimes.

Figure 5b shows the LIME explanation for the same instance. The word ***Ouça*** (listen) was assigned a strong negative weight, potentially due to its frequent appearance in clickbait headlines or misleading content that prompts user engagement without offering substantive information.

The presence of ***áudio*** (audio) and ***Ouça*** together may have triggered associations with multimedia content often used in disinformation strategies to appear more persuasive or authentic. Such content can be used out of context, making it more difficult to verify. The negative attribution to ***STF*** (Brazilian Supreme Court) and ***Cid*** (a Brazilian politician) may reflect learned associations between institutional or political entities and controversial topics frequently featured in fake news narratives, especially when names of public figures are combined with emotionally charged accusations or language.

The term ***picareta*** (crook), a strong pejorative used to discredit individuals, is typical of defamatory or sensationalist language. Its negative score may stem from its frequent use in fake content that seeks to provoke outrage or moral judgment. Even ***é*** (is), although a common verb, may be penalized for appearing in assertive or accusatory statements often found in misleading headlines that present unverified claims as facts. In contrast, only the word ***de*** (of) received a positive attribution. These divergent attributions between IG and LIME suggest that the model's misclassification may stem from complex interactions between lexical cues and learned biases.

We can observe from the explanations generated by LIME and IG that models trained with FakeRecogna predominantly had positive weights to words related to politicians. On the other hand, the political-related words highlighted by LIME contributed to classifying the news as fake for the models trained using Fake.Br. An explanation for the phenomenon may be a political view drift in Brazil during the studied period, were Fake.Br captured news more of the later yers regarding of the old rulling part and FakeRecogna capture mostly initial years of the new rulling part. Also, as Fake.Br is older (2016-2018) and FakeRecogna is newer (2019-2021), the texts in the latter may be more affected to the political use of social networks. It is also worth to note that in the period Brazil had a change in elected presidents. Additionally, in both datasets, it was possible to analyze that the name of organizations and institutions most of the time contributes to the classification of news as real.

5 Out-of-Distribution Analysis

5.1 Impact of Removing Words Highlighted by LIME and IG

After applying the explainability techniques LIME and IG to the selected samples, an experiment was conducted to evaluate the models' behavior when the most important words, as identified by these methods, were removed.

In this experiment, the model's **confidence scores**, on the predictions of each selected sample (the sampling procedure is described in Section 3.3) were recorded. Initially, the confidence was measured on the original samples (with no words removed). Then, the most relevant words were progressively removed according to each explainability method: first the most important word, then the top two, and so on, up to the top five most influential words.

Figure 6a presents the variation in model confidence when the words highlighted by **LIME** were removed. In contrast, 6b shows the same process, but based on the words identified as most relevant by the **IG** technique.

According to the graph in the Figure 6a, the Model 1 exhibited the most sensitive behavior: after a slight increase in confidence with the first few removals, its confidence dropped sharply from 0.9719 (after 3 words removed) to 0.4973 (after 4 words removed), remaining below 0.5 thereafter. In practical terms, this indicates that LIME explanation for Model 1 predictions are more reliable, as their removals had a deep impact on model performance. LIME did not perform as well in the other models, which was also the case for IG. These results strongly suggested that the prediction model is more holistic, considering the text as a whole instead of giving too much wait to specific keywords.

For LIME, models 2, 3, and 4 demonstrated greater stability. Model 3, for example, maintained values close to its original confidence even after all removals (from 0.9679 to 0.9597), indicating a more balanced distribution of word importance and a higher capacity for generalization. These results highlight the importance of explainability techniques not only for interpreting models but also for exposing their vulnerability or resilience to input perturbations.

The confidence scores of the models after progressively removing the most important words identified by the IG method reveal interesting patterns regarding their robustness like show the graph in Figure 6b. Models 1, 3, and 4 consistently maintain high confidence levels, with values mostly above 0.92 even after removing multiple key words.

Model 2 shows a small but noticeable decline in confidence as more important words are removed, dropping from an initial 0.93 to values close to 0.90 after several removals. On the one hand, this may indicate that Model 2 is more sensitive to the removal of crucial input features at some degree, but not as strong as Model 1 as observed in the LIME analysis. On the other hand, statistical noise may also be in action here.

Interestingly, the confidence does not always decrease monotonically with each additional word removal, which suggest some statistical noise in the analysis. For example, some fluctuations in confidence are observed in Models 1 and 3.

In addition to confidence scores, the metrics of **accuracy** and **loss** function were also analyzed to evaluate the models' behavior under the removal of the most relevant words. The accuracy results for both techniques are presented together in Figure 7, while the loss results are shown in Figure 8. The same patterns observed for model confidence are also present in loss and accuracy results.

These findings reinforce the importance of combining multiple evaluation metrics, such as confidence, accuracy, and loss with explainability methods, in order to assess not only model performance, but also its robustness and sensitivity to the removal of critical input information.

5.2 Cross-Dataset Analysis

Cross-dataset evaluation and input perturbations were applied to assess model robustness. First, each model was tested on a dataset different from the one used during training, enabling the analysis of generalization capacity to out-of-distribution data. For instance, models trained on the FakeRecogna dataset were evaluated on FakeBrCorpus, and vice versa, simulating real-world scenarios where textual characteristics vary across sources.

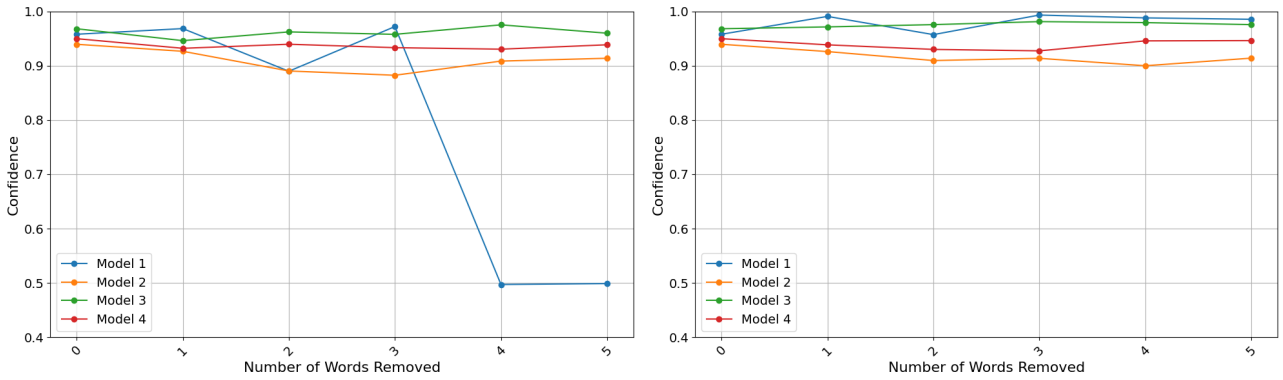
A second level of analysis was centered on systematically modifying the input data by altering the presence of stop words. For each training–testing combination, models were evaluated under two preprocessing conditions: with and without stop words. This made it possible to observe the models' sensitivity to minor linguistic changes. The experimental design included three types of perturbation: (i) changing only the attack dataset; (ii) changing only the stop word configuration; and (iii) changing both. The results of these evaluations are visualized in Figures 9a and 9b, which show how model accuracy and loss varied under different perturbation scenarios, revealing distinct levels of robustness and sensitivity across models.

The results reveal contrasting behaviors among the models. Model 2, trained on FakeRecogna without stop words, showed the highest resilience, achieving the best accuracy across perturbation scenarios and maintaining relatively low loss values. This suggests that removing stop words during training may help the model focus on more meaningful patterns, improving generalization even under adverse conditions. In contrast, Model 1, trained on FakeRecogna with stop words, was more sensitive to perturbations, with accuracy dropping to close to 0.50 (Figure 9a), practically random guessing, and loss (Figure 9b) values increasing sharply, especially when both the dataset and preprocessing were changed.

Models 3 and 4, trained on FakeBrCorpus, consistently showed poor performance in most scenarios, frequently producing accuracy values close to 0.50 regardless of the perturbation applied, which indicates a lack of robustness. Although no overfitting were detected in testing sets, models failed to generalize to different textual domains or preprocessing settings.

Overall, these findings highlight the importance of evaluating model behavior under distributional shifts and subtle linguistic perturbations, as such analyses can expose critical limitations in real-world applications like misinformation detection.

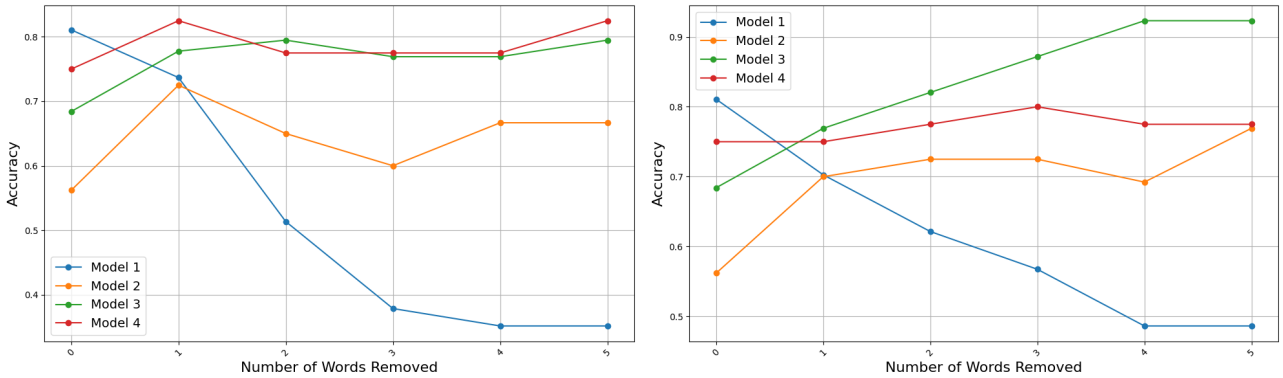
Changing stop-words preprocessing impacted the most,



(a) LIME: Model Confidence vs. Number of Words Removed

(b) IG: Model Confidence vs. Number of Words Removed

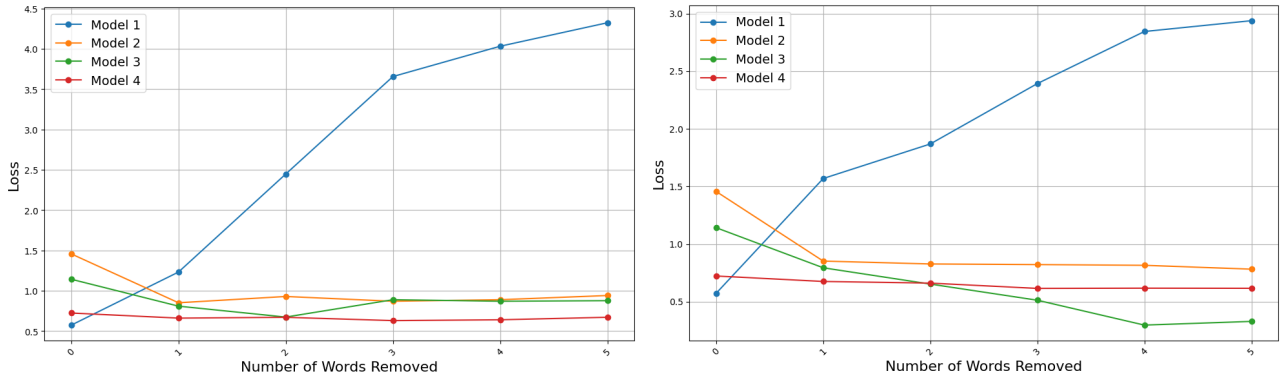
Figure 6. Comparison of model confidence as a function of the number of words removed using LIME and IG



(a) LIME: Accuracy vs. Number of Words Removed

(b) IG: Accuracy vs. Number of Words Removed

Figure 7. Comparison of model accuracy as a function of the number of words removed using LIME and IG



(a) LIME: Loss vs. Number of Words Removed

(b) IG: Loss vs. Number of Words Removed

Figure 8. Comparison of model loss as a function of the number of words removed using LIME and IG

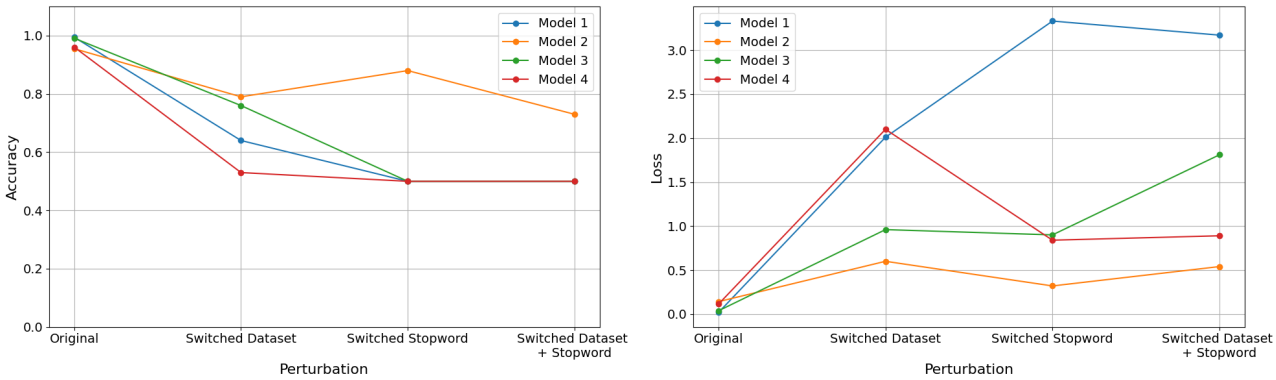
since accuracy droped to 50%, for three models (1, 3 and 4 Figure 9a). Changing the dataset also affected models. This is most notable in model 4, which reduced the accuracy from 96% to 52%. Furthermore, all models reduced their performance by at least 16%.

6 Linguistic Level Analysis

LIME and IG have some limitations when used with text, as their explanations cover mostly morphological and lexical information. Other linguistic levels are also important, such as the morphosyntactic level, which could, for example, to identify Part-of-Speech tags that appear more frequently in real and fake news. Similarly, semantic information could enhance explanations, for example, showing that some entities (like organizations) are distributed differently among real and fake news. Furthermore, rhetorical data can be useful for comparing the structure and length of texts in both news classes.

Considering these limitations, we performed an analysis on these linguistic levels. We created charts to understand whether some of the hypotheses we raised in Section 4. We used SpaCy[4] library to perform these extra analysis. The hypothesis are: (a) real news articles have more citations

[4]https://spacy.io/



(a) Model accuracy under different perturbations

(b) Model loss under different perturbations

Figure 9. Model accuracy and loss under different perturbations.

of organizations, such as public health agencies, while fake news articles cite individuals; (b) real news articles are better structured while fake news aim to be simpler and easy to understand, which may result in differences in stop words distribution or frequency across both classes. We quantified entities, grammatical classes, and the most frequent words from the texts.

Figures 10 - 12 show the results normalized by the number of words present in the news articles of each class. It is important to emphasize that real news articles are, on average, longer than fake ones. The average word count for fake news is 91.90, while for real news, it is 119.69 (average of both datasets combined).

According to the chart of entities identified in the news in Figure 10a of FakeRecogna dataset, the assumption that real news mentions more organizations is proven true. It is possible to observe that this distribution is not limited to organizations only, but also includes all other types of entities, with a larger distribution difference in entities of type LOC and ORG.

In contrast, the Fake.Br Corpus dataset (Figure 10b) reveals a different pattern: fake news articles contain a significantly higher proportion of person entities (PER), suggesting that this type of news tends to focus more on individuals, possibly to personalize content or appeal to emotions. The other entity categories (LOC, ORG, and MISC) show very similar distributions between real and fake news.

According to the grammatical class chart in Figure 11a, the number of proper nouns (PROPN) is higher in real news within the FakeRecogna dataset. The same chart also shows that real news makes greater use of adjectives and verbs.

In contrast, in the Fake.Br Corpus dataset, as shown in Figure 11b, proper nouns (PROPN) appear more frequently in fake news, which highlights a difference in linguistic patterns between the two datasets.

Regarding the list of the most frequent words (Figure 12), our results showed that the use of stop words in both real and fake news are similar in both datasets, leading to the conclusion that stop words are important for the correct prediction of news due to their usage in the text rather than their frequency.

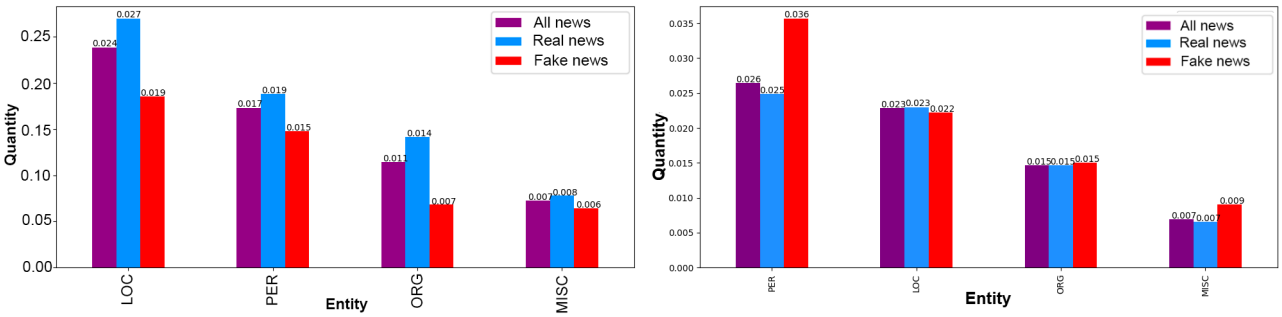
7 Discussion

This work focuses on four main topics: a qualitative comparison between LIME and IG; limitations of these methods when applied to text; fake news profiling during the period of FakeRecogna and Fake.br; quantitative analysis of both XAI methods.

Qualitative Comparison: Through the experiments, it was possible to notice that IG provided an explanation that was harder to interpret compared to LIME, due to the subtle differences in colors representing word weights. We focused on whole word, although it is possible to analyse both methods using word pieces (using methods such as Byte Pair Enconding). In that regard, IG output is given directly in word pieces, while LIME requires additional post-processing.

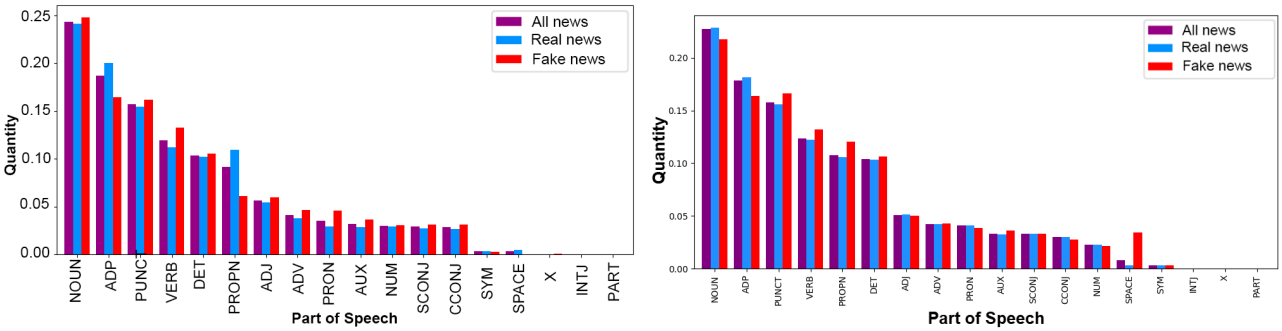
Method Limitation: LIME and IG have limitations when applied to texts, as they focus on the morphological and lexical levels. However, other levels are important: the morphosyntactic level allows identifying the frequency of labels and variations in grammatical categories between classes (as show in Figures 11a and 11b); for example, in the FakeBr corpus, proper nouns are more common in fake news, while the opposite occurs in FakeRecogna. The semantic level reveals how entities are distributed differently, as seen in the FakeBr corpus, where real news refer more to organization and fake news cite more person and family names. Regarding rhetorical level, real news have a bigger macro-structure, being longer than its counterparts. This happens because real news are more detailed to provide better information, whereas fake news is generally shorter and less detailed, making it easier to being understood and being easier to spread quickly. Additionaly, although not directly a limitation of the XAI methods, we also identified that classifiers for fake news detection are very sensitive to out-of-domain distribution, which, in turn, impacts how the XAI will behave.

Fake News Profiling: In the experiments, LIME and IG highlighted the importance of stop words for the correct classification of news, due to their distribution throughout the texts. Thus, Models 1 and 3, which retained the stop words, achieved best results. LIME, IG and SpaCy revealed that, in models trained with FakeRecogna, words related to politics have positive weights, indicating a classification as real news, while in the Fake.Br Corpus, these words are associated with fake news. This difference may be explained by the change



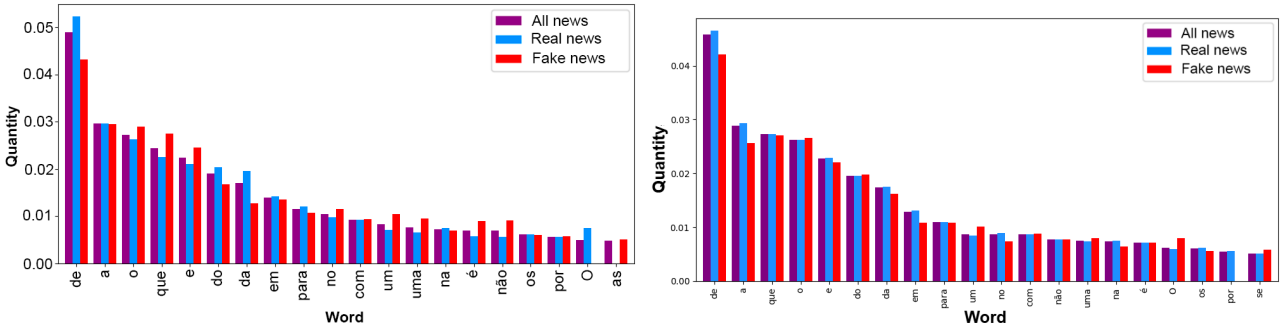
(a) FakeRecogna Dataset **(b)** Fake.Br Corpus Dataset

Figure 10. Quantity of entities with stop words normalized by the total number of words per class.



(a) FakeRecogna Dataset **(b)** Fake.Br Corpus Dataset

Figure 11. Quantity of parts of speech with stop words normalized by the total number of words per class.



(a) FakeRecogna Dataset **(b)** Fake.Br Corpus Dataset

Figure 12. Quantity of the most frequent words with stop words normalized by the total number of words per class.

in the political parties that rule the country during the data collection period. Furthermore, graphs showed that words related to health influence the prediction as real in FakeRecogna, possibly due to the context of the COVID pandemic, since collected news are from this period (2019-2021).

Quantitative Comparison: the removal of words detect as important by both LIME and IG had few to moderate impact on the classifiers confidences, showing limitations to the XAI methods, while identifying that the fake news classifier considers the text as whole unit to perform its predictions. It should be noted that in that regard LIME had an advantage over IG, specially due to the behavior of model 1, suggesting that LIME's explanations are more accurate.

8 Conclusion

In this study, we have qualitatively and quantitatively compared two methods for XAI, LIME and IG, in the domain of fake news texts in Portuguese. The analysis of results from these methods provided valuable insights into how these methods highlight patterns that make an impact in BERTimbau model decision process.

Although LIME and IG have similar behaviors (highlighting important words using colors), LIME generated explanations that were easier to interpret. This results are in line with the findings of Pendyala and Hall [2024]. We also noticed that LIME had a small performance gain compared to IG in accurately identifying important words in its explanations.

We identified limitations for the fake news classifier and for the XAI methods. The first presented small resilience when presented to data out of its training distribution, while XAI explanations lack important linguistic information regarding morphosyntactic, semantic and rhetorical structures, resulting in incomplete explanations. The removal of the most relevant words detected by both methods from the original texts had a moderate impact on the classifier's confidence, which also may suggest the explanations may be incomplete.

Finally, we also made a fake news profiling for Brazilian Portuguese texts from the political domain in the period of 2016 – 2021, noticing that organizations mentioned in the text are an important clue to detect the credibility of a news, while person names (mostly, politician names) had mixed effects on predictions due to politics volatility. Text structure also seems important, according to our analysis over the presence of stop-words in the texts. It is also important to note that there is a diachronic aspect of the use of fake news and that words and clues associated with fake news may change during the passage of time.

For future investigations, it would be interesting to explore additional explainability methods, such as SHAP, and conduct studies across diverse domains to understand how different methods perform in varied contexts. It would also be relevant to apply robust adversarial attack approaches, preferably across different datasets, to assess the resilience of models under adverse conditions. Furthermore, developing a new explainability framework that goes beyond word distribution could overcome the limitations of the methods examined.

Declarations

Acknowledgements

We would like to express our sincere gratitude to everyone who contributed in some way to the completion of this work, and to everyone who, directly or indirectly, contributed to the development of this work, our deepest appreciation.

Funding

No external funding was received for the completion of this work. The research was conducted independently, without the support of grants or sponsors.

Authors' Contributions

JV was the main contributor and writer of this manuscript, conducting the fine-tuning of the models and experiments with LIME. Meanwhile, RR contributed to the experiments with IG. Finally, AC and IG were responsible for evaluating the experiments.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

All the explanations generated by LIME and IG are available in the GitHub at this link: https://github.com/JessicaVicentini99/Comparing-Explainable-AI-For-Fake-News-Detection

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., *et al*. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Available at: `https://www.tensorflow.org/`.

Ahmed, I., Jeon, G., and Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, 18(8):5031–5042. DOI: 10.1109/tii.2022.3146552.

Desai, V., Gattani, A., and Dalvi, H. (2024). Explainable models for the detection of incidents of fake news and hate speech. In *Text and Social Media Analytics for Fake News and Hate Speech Detection*, pages 114–136. Chapman and Hall/CRC. DOI: 10.1201/9781003409519-6.

Garcia, G., Afonso, L., and Papa, J. (2022). *FakeRecogna: A New Brazilian Corpus for Fake News Detection*, pages 57–67. DOI: 10.1007/978-3-030-98305-5_6.

Gohel, P., Singh, P., and Mohanty, M. (2021). Explainable ai: current status and future directions. *arXiv preprint arXiv:2107.07045*. DOI: 10.48550/arxiv.2107.07045.

Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2019). Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–138. DOI: 10.1145/3306618.3314229.

Lima, T. B., Rolim, V., Nascimento, A. C., Miranda, P., Macario, V., Rodrigues, L., Freitas, E., Gašević, D., and Mello, R. F. (2024). Towards explainable automatic punctuation restoration for portuguese using transformers. *Expert Systems with Applications*, 257:125097. DOI: 10.1016/j.eswa.2024.125097.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. DOI: 10.48550/arxiv.1705.07874.

Mersha, M. A., Yigezu, M. G., Shakil, H., AlShami, A. K., Byun, S., and Kalita, J. (2025). A unified framework with novel metrics for evaluating the effectiveness of xai techniques in llms. DOI: 10.48550/arxiv.2503.05050.

Moradi, M. and Samwald, M. (2021). Explaining black-box models for biomedical text classification. *IEEE journal of biomedical and health informatics*, 25(8):3112–3120. DOI: 10.1109/jbhi.2021.3056748.

Moraliyage, H., Kulawardana, G., De Silva, D., Issadeen, Z., Manic, M., and Katsura, S. (2025). Explainable artificial intelligence with integrated gradients for the detection of adversarial attacks on text classifiers. *Applied System Innovation*, 8(1):17. DOI: 10.3390/asi8010017.

Oliveira, H., Ferreira Mello, R., Barreiros Rosa, B. A., Rakovic, M., Miranda, P., Cordeiro, T., Isotani, S., Bittencourt, I., and Gasevic, D. (2023). Towards explainable prediction of essay cohesion in portuguese and english. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 509–519. DOI: 10.1145/3576050.3576152.

Pendyala, V. S. and Hall, C. E. (2024). Explaining misinformation detection using large language models. *Electronics*, 13(9):1673. DOI: 10.3390/electronics13091673.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD interna-*

tional conference on knowledge discovery and data mining, pages 1135–1144. DOI: 10.18653/v1/n16-3020.

Santos, R. L., Monteiro, R. A., and Pardo, T. A. (2018). The fake. br corpus-a corpus of fake news for brazilian portuguese. In *Latin American and Iberian Languages Open Corpora Forum (OpenCor)*, pages 1–2. DOI: 10.5753/erbd.2023.229495.

Shevskaya, N. V. (2021). Explainable artificial intelligence approaches: challenges and perspectives. In *2021 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS)*, pages 540–543. IEEE. DOI: 10.1109/itqmis53292.2021.9642869.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer. DOI: 10.1007/978-3-030-61377-$8_2$8.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org. DOI: 10.48550/arxiv.1703.01365.