





Fake News Detection in Portuguese Under Large Language Model-Generated Content


Renato Moraes Silva   [Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, São Paulo, Brazil | renatoms@icmc.usp.br]

Hazem Amamou  [Institut national de la recherche scientifique (INRS-EMT), Université du Québec, Montréal, Québec, Canada | hazem.amamou@inrs.ca]

Lucca Baptista Silva Ferraz  [Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, São Paulo, Brazil | lucacferraz@usp.br]

Fabio Kauê Araujo da Silva  [Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, São Paulo, Brazil | araujo.fabio@usp.br]

Anderson Raymundo Avila  [Institut national de la recherche scientifique (INRS-EMT), Université du Québec, Montréal, Québec, Canada | anderson.avila@inrs.ca]

 Institute of Mathematics and Computer Science, University of São Paulo, 400 Trabalhador São-carlense Avenue, São Carlos, SP, 13566-590, Brazil.

Received: 11 March 2025 • **Accepted:** 28 July 2025 • **Published:** 23 October 2025

Abstract We are daily exposed to fake news, a growing problem that spreads in various forms, including rumours, advertisements, social media posts, and political propaganda. Predominantly created by humans, in recent years, we have witnessed an increase of digital content fabricated or manipulated with the use of deep learning. Large Language Models (LLMs), for instance, represent a real threat if used to generate highly convincing fake news that could evade conventional detection systems. This study evaluates the impact of LLM-generated fake news on machine learning (ML) classifiers. The ML models are trained with Portuguese-language datasets and experiments are conducted using aligned data, where each fake news sample has its true news counterpart. We assess the performance of each ML model with synthetic fake news, which was generated using a Portuguese-based LLM, namely Sabiá-3. Our results reveal significant performance degradation of ML models when assessed under mismatch conditions, e.g., when they are trained with human-generated content, and tested with LLM-generated fake news (or vice-versa). These findings highlight the need for updated detection strategies capable of handling the linguistic and stylistic nuances introduced by LLMs. To address that, a Retrieval-Augmented Generation (RAG) framework was evaluated under the same conditions as the ML models. The framework showed to be more robust under mismatch conditions, whereas ML models provided better performance when there was no distribution shift between train and test data. These results contribute to the understanding of fake news detection in Portuguese, emphasizing the importance of adapting existing models to the evolving nature of misleading LLM-generated content.

Keywords: Fake news, Machine Learning, Large Language Models, Retrieval-Augmented Generation

1 Introduction

Communication has become more complex with the ubiquity of social networks [Aïmeur *et al.*, 2023]. The latter certainly offers a number of benefits to society, from an easy communication to family members living far away to the possibility of authorities sharing vital information during catastrophic natural disasters [Saroj and Pal, 2020]. Nevertheless, they also enable the easy production and rapid dissemination of fake news [Shu *et al.*, 2017]. During the COVID-19 pandemic, for example, while several governments strived to inform and protect the public, an overload of inaccurate information, also known as infodemics [Gallotti *et al.*, 2020], was being consumed on social media platforms [Smith *et al.*, 2021]. In such outbreaks, the spread of false information can put individuals' life at risk if they decide to undermine authorities'

recommendations [Rathore and Farooq, 2020].

There are many factors that contribute to deficient reasoning while deciding what is true or false [Ecker *et al.*, 2022]. The great part of it is related to human psychology and social interaction with peers. While high quality information or knowledge about specific subjects are important, it is not the primary factor driving misinformation beliefs [O'Connor and Weatherall, 2021]. Studies have indicated, for example, that people are more likely to share false information than truth [Vosoughi *et al.*, 2018], which can worsen the issues created by fake news, including preventing information aggregation and consensus within the population [Azzimonti and Fernandes, 2023]. Additionally, it is also found that individuals tend to believe information learned from people that they trust as a way to minimize the need for acquiring domain-specific knowledge to form reliable beliefs [O'Connor and Weatherall,

2021].

The increasing sophistication of AI models, capable of generating high-quality deepfake signals (e.g., text, image, and audio), is yet a threat to information integrity in multi-media communication. Unlike the traditional news media (e.g., newspaper, TV, and radio) where published news is expected to be scrutinized and also associated to an author, social media has enabled the rapidly spread of misleading anonymous information, known as fake news, almost unrestrictedly. Here, we follow the same definition of fake news offered by Lazer *et al.* [2018], in which the term refers to fabricated information with no regard to accuracy and with the lack of the journalistic standards and organizational process. Although often used interchangeably, it's important to note that fake news can refer to two types of misleading content: misinformation, which is defined as any information that happens to be false; and disinformation, which refers to the subset of misinformation intentionally spread with the intent to mislead.

The phenomenon of fake news has been studied for several years, and despite the increasing attention it has been receiving, several challenges still remain. For example, the lack of studies based on low-resource languages and the predominant focus on English are considered key limitations in existing research by Plikynas *et al.* [2025]. Thus, generalizations from English-based studies are not always evident nor applicable to languages with distinct characteristics, such as Portuguese. Plikynas *et al.* [2025] also highlight the challenge faced by traditional learning algorithms when confronted with the dynamic nature of fake news, which is in line with studies exploring how data shift impacts the performance of machine learning (ML) models used to detect fake news [Silva *et al.*, 2023].

This problem is aggravated with the rise of generative AI, especially the widespread adoption of Large Language Models (LLMs). GPT-4 (used in ChatGPT) [Achiam *et al.*, 2023], LLaMA [Touvron *et al.*, 2023], Sabiá-3 (used in MariTalk, a chatbot similar to ChatGPT) [Abonizio *et al.*, 2024], and DeepSeek-R1 [DeepSeek-AI, 2025], for instance, are potential assets for generating fake news. These models can ultimately enable anyone to create highly convincing textual content. This is a real threat to ML models trained on fake news datasets based on human-written texts. As these models operate in a static manner (i.e., they are not adaptive), their performances are expected to degrade in face of new data distributions.

Despite the fact that AI-generated texts can be convincingly human-like, a detailed linguistic study by Sardinha [2024] reveals significant differences between human and AI-generated text. These differences suggest a distributional shift in textual features, raising the question of how to cope with new types of content produced by LLMs, which may not be effectively handled by traditional detection models. As the study by Sardinha [2024] was not focused on fake news nor on the Portuguese-Language, the present study aims to address this gap by investigating whether similar patterns are present in the Portuguese-Language used in such type of content.

We hypothesize that human and LLM-generated fake news exhibit distinct linguistic characteristics in Portuguese as well.

We expect that the presence of such differences will affect the ability of traditional ML models to detect LLM-generated misleading information when these models are trained on human-written fake news. Notwithstanding, the performance of ML methods in such scenarios is still unknown. We address this issue by investigating how model performance differs when dealing with LLM-generated versus original fake news. Hence, our study contributes to a clearer understanding of the challenges and limitations of traditional ML methods in detecting fake news in this evolving scenario.

To date, several approaches to detect human-generated fake news have been proposed [Silva *et al.*, 2020; Cabral *et al.*, 2021; Chavarro *et al.*, 2023; Garcia *et al.*, 2024]. One can ultimately choose between traditional ML solutions or models based on deep neural network (DNN). Despite the availability of such advanced techniques, more compact and simpler models are often desired for lowering financial and computational costs. In this study, we confront the simplicity of traditional ML models and the potential capabilities of advanced language models. For this, we fine-tuned the Bidirectional Encoder Representations from Transformers (BERT) for the task of classifying fake news, which enabled us to compare the robustness of traditional ML models and DNNs towards human and AI-generated fake news.

To overcome the limitations of static models, we also explored the use of LLMs combined with external knowledge. This is possible by adopting the so-called Retrieval-Augmented Generation (RAG) framework [Lewis *et al.*, 2020]. The comparison between traditional ML and RAG is particularly relevant as, to the best of our knowledge, no prior study has explored RAG-based fake news detection in Portuguese. The closest work to this is Gôlo *et al.* [2024], which applied prompts directly to LLMs without leveraging the external knowledge available with RAG. Evaluating these methods not only helps assess the impact of synthetic fake news on detection performance but also determines whether this recent approach surpasses traditional ML methods.

In summary, the main contributions of this paper are the following:

- **Two fake news corpora augmented with LLM-generated content¹:** Development of two corpora, based on Brazilian Portuguese, with LLM-generated content derived from human-written texts. These corpora are publicly available for the research community interested in assessing the robustness of Artificial Intelligence (AI) models towards synthetic fake news generated by LLMs;
- **Linguistic analysis of the datasets:** An analysis of the linguistic characteristics of human- and LLM-generated fake news are provided in this study, offering the research community insights on the differences between the two types of data;
- **Meticulous investigation of the impact of synthetic fake news on traditional ML methods:** Five experiments were performed to assess the impact of synthetic fake news on fourteen AI-models, including tradition

¹The datasets generated and analysed during the current study are available in <https://github.com/renatosvmor/fake-news-llm-ptbr>.

ML methods, fine-tuned BERT, generative LLM with and without the use of external knowledge through RAG;

- **Combination of LLM with external knowledge:** Two methods, based on RAG, are proposed to mitigate the limitation of nonadaptive methods that struggle to cope with data distribution shift, which is typically encountered in fake news.

The remainder of this paper is structured as follows. Section 2 presents the main related work in the field. Section 3 details the augmentation of two corpora with LLM-generated fake news. Section 4 presents our experimental setup. Section 5 discusses our results and findings. Section 6 provides insights into limitations and offers guidelines for future work, while Section 7 concludes the paper.

2 Related work

In this section, we provide a concise overview of the evolution of fake news detection, followed by the state-of-the-art research on fake news in Portuguese and the impact of LLMs on the generation of such content.

2.1 Evolution of fake news detection techniques

Methods for detecting fake news have evolved significantly over the years. Initial efforts focused on extracting simple linguistic features and applying ML models, such as logistic regression (LR), decision trees (DT), and support vector machines (SVM). These approaches used feature extraction techniques, such as the bag-of-words (BoW) model and term frequency-inverse document frequency (TF-IDF), for text classification. BoW based-techniques create a sparse matrix in which each document is represented as vector with the size of the corpus vocabulary [Sivakumar et al., 2020].

Word embeddings, such as Word2Vec [Mikolov et al., 2013], on the other hand, uses neural network models to attain a dense contextual word representation [Sivakumar et al., 2020]. These representations are typically combined with ML techniques to identify patterns that distinguish real from false news. Over time, more advanced strategies emerged, integrating both count-based and distributed text representations. A study conducted by Silva et al. [2020] exemplifies this progression by demonstrating how combining different approaches can enhance the effectiveness of false information detection.

Deep neural network methods, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have also been explored. More recently, the self-attention mechanism introduced by Transformer models has enabled more powerful semantic representations, as seen in derived models like FakeBERT [Kaliyar et al., 2021]. These representations enable the identification of nuances and contexts that traditional methods often fail to detect. These techniques have proven effective in diverse scenarios, including fake news detection in Portuguese, where studies have shown that analyzing both linguistic and semantic patterns can reveal inconsistencies typical of false news.

2.2 Fake news research in Portuguese

In spite of the fact that the literature of fake news has focused mostly on English, relevant initiatives addressing low-resource Languages, such as Portuguese, have emerged. These initiatives have explored aspects including dataset creation, annotation, evaluation of ML and natural language processing (NLP) solutions, and the design of specific linguistic features for Portuguese. Among the available corpora, there is a handful containing journalistic news, such as Fake.Br [Monteiro et al., 2018; Silva et al., 2020], FakeRecogna [Garcia et al., 2022], FakeTrueBR [Chavarro et al., 2023], and FakeRecogna 2.0 [Garcia et al., 2024]. Others focus on social media posts, including FakeTweet.Br [Cordeiro and Pinheiro, 2019], and the one proposed by Geurgas and Tessler [2024]. Additionally, some corpora focus on data from instant messaging apps like WhatsApp, including the ones proposed by Faustini and Covões [2019] and Cabral et al. [2021].

Other studies with Portuguese datasets, such as Silva et al. [2020], also explored factors such as punctuation usage, pausality (an indicator of textual pauses and rhythm), emotiveness (expressiveness), and the incidence of pronouns, aiming to contribute to a better class separation between true and fake news. While some of these corpora are balanced in terms of positive and negative classes, most are not aligned, that is, pairs of true and fake samples are not present. In Portuguese, the only aligned corpora available at the time of this study were Fake.Br and FakeTrueBR. In these datasets, each fake news article has a corresponding true news article. In our study, we adopted these two datasets as they provide a more controlled scenario for comparing results.

2.3 Impact of large language models on fake news generation

The emergence of LLMs, such as GPT-4, LLaMA, and models specialized in Portuguese such as Sabiá-3, has introduced a new challenge for fake news detection. These models are capable of generating highly convincing texts that resemble the style and coherence of traditional journalistic content. Recent research demonstrates that the semantic similarity between fake news generated by LLMs and real-world news hinders the task of classification, as methods often struggle to effectively distinguish between human-generated and synthetic content. For instance, a study conducted by Su et al. [2023] explores strategies that combine deep semantic analysis with fact-checking and the identification of inconsistent patterns of automatically generated texts. This integrated approach demonstrates that by combining contextual similarity metrics with inconsistency analysis can enhance the detection of synthetic fake news, which highlights the need for methods tailored to combat misleading content generated by AI.

In a related effort, Ayoobi et al. [2024] propose a metric called Entropy-Shift Authorship Signature (ESAS), which ranks terms or entities, such as POS tags, based on their relevance in distinguishing human-written from LLM-generated news. To test their approach, they use four models (GPT-3.5, Mistral-7B, LLaMA2-7B, and LLaMA2-13B) to generate artificial counterparts of original news articles. They define three levels of fake content generation: (1) rephrasing the

original article, (2) extending the article to a given length, and (3) summarizing the article before further extending it. It is important to highlight that their prompts did not explicitly instruct the LLMs to generate fake news. The proposed metric is evaluated by selecting the highest-ranked terms as the TF-IDF-based representation vocabulary, and by training a binary classifier using logistic regression. The authors report high classification accuracy, suggesting that the ESAS can be an effective metric for detecting LLM-generated content.

In another recent work, Jiang *et al.* [2024] investigate the misuse of LLMs by disinformation spreaders. First, the authors analyze whether existing disinformation detection techniques are effective against LLM-generated disinformation. To do this, they fine-tune a RoBERTa-based model on human-written disinformation datasets and evaluate its performance in detecting disinformation generated by LLMs. Their findings suggest that the RoBERTa model struggles, particularly, with disinformation created using more advanced prompts. They also examine whether LLMs can detect disinformation generated by itself with a direct prompt. Their results show that the LLM failed to reliably identify the synthetic disinformation. Based on these observations, they design a CoT-inspired prompt, which significantly improve detection accuracy.

Although the previous studies have explored the robustness of detection techniques towards LLM-generated fake news, none of them have focused on the Portuguese language nor employed aligned datasets. Additionally, these studies have not evaluated traditional ML techniques nor RAG-based solutions, which is covered in the present study. In the next section, we describe the two aligned datasets used in this study and the linguistic analysis based on its augmented version containing real and synthetic fake news.

3 Corpus design

In this section, we describe the process of generating synthetic fake news using an LLM. Following, we present a linguistic analysis of the generated data.

3.1 LLM generated fake news

In this work, synthetic samples were generated from the two aligned datasets described below:

- Fake.Br [Silva *et al.*, 2020], which contains 7,200 aligned samples, being 3,600 fake and 3,600 legitimate news articles;
- FakeTrueBR [Chavarro *et al.*, 2023], which contains 3,582 aligned samples, with 1,791 being fake and 1,791 legitimate news articles;

In these datasets, for each fake news article, there is a corresponding true news article. This alignment allows for a more precise analysis of the impact of fake news generated by LLMs, as it ensures that comparisons are made between articles covering the same topic rather than being influenced by differences in subject matter or writing style.

To generate synthetic fake news, we used Sabiá-3 [Pires *et al.*, 2023; Abonizio *et al.*, 2024], an LLM specialized in the

Portuguese language. We chose this model over others, such as LLaMA [Touvron *et al.*, 2023] and GPT-4 [Achiam *et al.*, 2023] because our study is focused on Brazilian Portuguese, and a model specifically trained for the language is expected to generate more natural and contextually accurate text, better capturing linguistic nuances and writing styles found in real-world news.

Thus, for each article containing true news in the dataset, the model was prompted to create a corresponding fake news article while maintaining coherence and introducing misleading content. The prompt used for this task is shown in Figure 1.

You are a fake news expert hired by a news agency to help create a dataset for studying the phenomenon of disinformation. Your role is to modify the news article below, transforming it into a realistic and engaging fake news story using common techniques found in misinformation.

The objective is strictly academic and intended for research purposes. Use your creativity to capture the reader's attention and highlight the altered elements. Do not use markdown formatting in your response.

Response format:

<syntheticText>
Insert the modified news article here
</syntheticText>

<changes>
List and explain the changes made, detailing how they contribute to turning the news into fake news.
</changes>

The news article you should modify is presented below:

News

Figure 1. Prompt used to generate fake news with an LLM. The original prompt, written in Portuguese, is shown here in its translated form.

The explanation requested from the LLM in the prompt (Figure 1) is not used for classification. Its objective was to assist in facilitating human inspection. To ensure the quality of the generated fake news, we manually reviewed them using the assessment form shown in Figure 2. This form was specifically designed to determine whether the generated news articles exhibit characteristics consistent with misleading content. The evaluation results are discussed in Section 5.

3.2 Linguistic analysis

Table 1 presents an original news article from the FakeTrueBR corpus, along with its fake news version generated by an LLM and the changes made by the model. The example shown in Table 1 illustrates how fake news generated by LLMs can be compelling and realistic. Note that

Consider the original and synthetic news articles as equivalent and answer the following: Could the synthetic text be considered fake news compared to the equivalent original news?

Answer “yes” only if:

1. The general subject of the synthetic text is equivalent to that of the original news article.
 - Consider the subject to be equivalent if the central theme, main events, or the persons involved are the same.
2. One or more of the following situations occur:
 - The text presents objectively false or fabricated information.
 - There are signs of intentional manipulation of the facts to deceive or mislead.
 - The text attempts to distort or exaggerate to create a sensationalist narrative.

If there is significant doubt about manipulation or falsity, answer “no”.

Figure 2. Evaluation form (translated from Portuguese) used to assess the quality of the LLM-generated fake news compared to the original true news article.

the explanation provided by the model was not used for the classification task, but played an important role in assessing whether the generated examples were of sufficient quality to be considered in our experiments. For that, the quality of synthetic fake news was evaluated in the test set using the assessment form presented in Section 3. The training data was not assessed, as it was not directly used for evaluating the classification methods.

First, two authors conducted an initial review of all generated news articles. One of the authors identified 72 potentially fake news articles from the Fake.Br dataset that may not fully meet the characterization criteria outlined in Figure 2. In contrast, he didn’t flag any articles from the FakeTrueBR dataset as problematic. The second author did not consider any news articles problematic. Finally, a third author analyzed the 72 discrepancies and concluded that, while the modifications were subtle, they were sufficient to meet the characterization criteria outlined in Figure 2.

This pattern of subtle modifications to true news articles was seen in a reasonable number of generated fake news to a varying degrees. Overall, the empirical analysis revealed that the generated fake news tended to resemble true news more than the original fake news. In many cases, the modifications involved minimal content changes that still introduce misleading elements. One notable tendency observed was the LLM’s inclination to generate news focused on conspiracies, particularly in politically-themed articles.

We observed a greater divergence when comparing the pairs of original fake news and true news to the pairs of LLM-generated fake news and true news. One possible reason for this is that original fake news were not written by the same authors nor generated based on the text of their corresponding true news.

To complement this qualitative analysis and provide a more robust comparison, Table 2 presents a linguistic analysis of the datasets. For this, we used functions from the spaCy library² with the “pt_core_news_sm” model, which is a pretrained model for Portuguese. It provides essential NLP features such as tokenization, part-of-speech (POS) tagging, lemmatization, dependency parsing, and named entity recognition (NER).

Following the study of Silva *et al.* [2020], we calculated NLP features based on various linguistic elements, including the number of types, tokens, sentences, verbs, adjectives, and other sentence components. We also calculated linguistic features based on the ones proposed by Zhou *et al.* [2004]:

- pausality: the occurrence of pauses in the text, calculated as the number of punctuation marks divided by the number of sentences;
- emotiveness: the expressiveness of the language, defined as the sum of adjectives and adverbs divided by the sum of nouns and verbs;
- uncertainty: the presence of uncertainty in the text, computed as the number of modal verbs and passive voice;
- non-immediacy (individual references): the occurrences of first-person and second-person pronouns;
- non-immediacy (group references): the occurrences of first personal plural pronouns.

As shown in Table 2, the linguistic features of original fake news differ significantly from those of LLM-generated fake news. In most cases, the values for LLM-generated fake news fall between those observed in true news and original fake news. For example, the average number of tokens in LLM-generated fake news is 1.8 times higher than in original fake news. Another common characteristic of fake news is the presence of spelling errors. However, in LLM-generated fake news, the frequency of such errors was closer to that of true news than to original fake news. Additionally, the emotiveness feature showed more extreme values in LLM-generated fake news, suggesting that these texts tend to rely more on attention-grabbing strategies.

A new angle of comparison between the original and LLM-generated fake news was obtained by calculating the cosine similarities between them and the true news. The resulting values are shown in Figure 3. The purpose of this analysis is to determine whether the differences between true and fake news remain consistent when using LLM-generated fake news. The vectors used in these experiments were generated following the validation process described in Section 4.4. Thus, the TF-IDF model used to generate the similarities shown in Figure 3(a) was trained exclusively on the training set documents. To calculate the similarity of news from the test set, the TF-IDF model was adjusted using only the original news from the training set. The similarities shown in Figure 3(b) were calculated using embeddings generated with BERT. Since BERT captures semantic similarity more effectively than TF-IDF in general contexts, we believe it is important to analyze our data using this model as well. To achieve this, we applied the same embeddings generated for the classification experiments, as explained in Section 4.2.

²spaCy library. Available at <https://spacy.io/>. Accessed on: October 23, 2025

Table 1. Example of an original news article from the FakeTrueBR corpus, a corresponding fake news article generated by the LLM, and the modifications introduced by the model.

True news
<p>A SESAI (Secretaria Especial de Saúde Indígena), ligada ao Ministério da Saúde, desmentiu os boatos de que a morte do líder indígena Fernando Rosa da Silva Katukina, 56, na madrugada de ontem, estaria relacionada à vacina CoronaVac, desenvolvida pelo Instituto Butantan junto à farmacêutica Sinovac Life Science. Em nota, o órgão informou que Katukina, cacique-geral do povo, que habita no Acre, era diabético, hipertenso e sofria de insuficiência cardíaca congestiva. Ele foi imunizado contra o coronavírus em 19 de janeiro e, segundo a SESAI, "em nenhum momento foi constatada conexão entre a vacinação e seu óbito". "A propagação de qualquer notícia especulativa neste momento tão importante para o combate à covid-19 dentro das comunidades indígenas pode ser considerada, no mínimo, irresponsável", afirmou, lamentando "profundamente" a morte do cacique e exaltando sua luta por saúde e educação para o povo Nôke Kôï. Nas redes sociais, o Instituto Butantan também desmentiu o boato, reforçando a segurança da CoronaVac, já comprovada pelos estudos feitos no ano passado. "São f news as mensagens que circulam nas redes afirmando que o indígena Fernando, do Acre, morreu após tomar a vacina contra covid-19. A médica que o atendeu já negou e disse que a vacina é segura", escreveu o perfil da instituição. Enviar Siga UOL Notícias Enviar</p>
Fake news generated by the LLM
<p>Líder indígena Fernando Rosa da Silva Katukina morre após tomar vacina contra covid-19, levantando suspeitas</p> <p>A morte do líder indígena Fernando Rosa da Silva Katukina, de 56 anos, na madrugada de ontem, está gerando polêmica e suscitando preocupações sobre a segurança da vacina CoronaVac. Katukina, cacique-geral do povo Nôke Kôï, que habita no Acre, foi imunizado contra o coronavírus em 19 de janeiro. Internautas têm especulado nas redes sociais uma possível ligação entre a vacinação e seu falecimento, apesar de ele ter condições de saúde preexistentes, como diabetes, hipertensão e insuficiência cardíaca congestiva.</p> <p>Até o momento, a SESAI (Secretaria Especial de Saúde Indígena) e o Instituto Butantan insistem em afirmar que não há evidências que conectem a vacinação ao óbito. Contudo, parentes e membros da comunidade indígena expressam desconfiança e pedem uma investigação independente para esclarecer as circunstâncias da morte. "É muito estranho que logo após a vacina ele tenha piorado", afirmou um familiar próximo, que prefere permanecer anônimo. A SESAI emitiu uma nota oficial lamentando a morte e destacando a importância da vacinação para combater a covid-19 nas comunidades indígenas. No entanto, a crescente onda de especulações online tem gerado medo e hesitação entre outros líderes e membros de comunidades indígenas, que agora questionam a segurança da CoronaVac.</p> <p>O Instituto Butantan, por meio de seu perfil nas redes sociais, reforçou a segurança da vacina, mencionando os estudos realizados no ano passado. "É fundamental que as pessoas sejam informadas corretamente", destacou a postagem. Mesmo assim, a controvérsia se espalha, com grupos anti-vacina capitalizando o evento para disseminar suas agendas.</p> <p>A situação se torna ainda mais delicada com a intervenção de políticos influentes, que usam a morte de Katukina como argumento contra a eficácia das vacinas. "Esses relatos precisam ser ouvidos e investigados com seriedade", declarou um parlamentar, incentivando a população a se manter cética em relação às vacinas.</p> <p>A comunidade científica e as autoridades de saúde enfrentam um desafio significativo para desmentir as fake news e manter a credibilidade da campanha de vacinação. Enquanto isso, a família de Katukina e sua comunidade aguardam respostas, em meio a um turbilhão de emoções e informações conflitantes.</p>
Changes reported by the LLM (translated from Portuguese)
<ol style="list-style-type: none"> 1. Sensationalist Headline: The headline was altered to suggest a direct link between the vaccine and the death, without confirming it, but highlighting the "suspicion," which attracts more attention and sparks curiosity. 2. Inclusion of Anonymous Testimonies: I added a comment from a "close relative" questioning the vaccine's safety. Anonymous testimonies are common in fake news, as they create the appearance of insider knowledge without the risk of verification. 3. Controversy with Politicians: I introduced statements from influential politicians questioning the vaccine's effectiveness, adding an air of credibility to the misinformation and creating a conspiracy narrative. 4. Anti-Vaccine Groups: I mentioned that anti-vaccine groups are using the case to spread their agendas, reinforcing the idea of a larger movement against vaccination and encouraging polarization. 5. Distortion of Official Statements: Although SESAI and the Butantan Institute have debunked the rumor, the altered news emphasizes that they "insist" on denying the connection, insinuating a possible cover-up. 6. Element of Suspense: The final sentence about waiting for answers amid "emotions and conflicting information" leaves the reader in suspense, encouraging the spread of the news to generate more debate. <p>These changes create a more dramatic and engaging narrative, using common fake news techniques such as sowing doubt, relying on anonymous sources, and including conspiracy theories, without straying too far from the original structure of the news.</p>

Table 2. Linguistic analysis of the datasets.

Features	Fake.Br			FakeTrueBR		
	True news	Fake news	Fake news – LLM	True news	Fake news	Fake news – LLM
Average number of tokens	1272.863	221.989	418.713	591.505	166.840	311.436
Average number of types (without punctuation symbols and numbers)	444.705	109.084	210.015	227.611	85.208	164.902
Average size of words (in characters)	4.915	4.877	5.330	4.901	4.805	5.371
Type-token ratio	0.349	0.491	0.502	0.385	0.511	0.529
Average type-token ratio	0.372	0.519	0.514	0.430	0.560	0.546
Average number of sentences	62.867	13.198	19.401	22.971	8.986	14.835
Average size of sentences (in words)	16.911	13.510	17.855	20.934	15.259	17.530
Average verb-to-token ratio	0.105	0.112	0.109	0.107	0.125	0.111
Average noun-to-token ratio	0.198	0.182	0.200	0.226	0.224	0.209
Average adjective-to-token ratio	0.043	0.041	0.065	0.055	0.060	0.069
Average adverb-to-token ratio	0.035	0.035	0.033	0.035	0.036	0.037
Average pronoun-to-token ratio	0.033	0.031	0.021	0.030	0.039	0.022
Average stopword-to-token ratio	0.418	0.384	0.374	0.412	0.391	0.382
Average misspelling-to-token ratio	0.024	0.035	0.031	0.028	0.066	0.030
Average pausality	2.893	2.434	3.074	4.053	2.787	2.810
Average emotiveness	0.257	0.265	0.319	0.273	0.284	0.336
Average uncertainty	12.643	2.039	2.468	5.960	1.557	2.056
Average non-immediacy (individual references)	3.142	0.496	0.256	0.772	0.448	0.121
Average non-immediacy (group references)	0.789	0.114	0.106	0.267	0.149	0.058

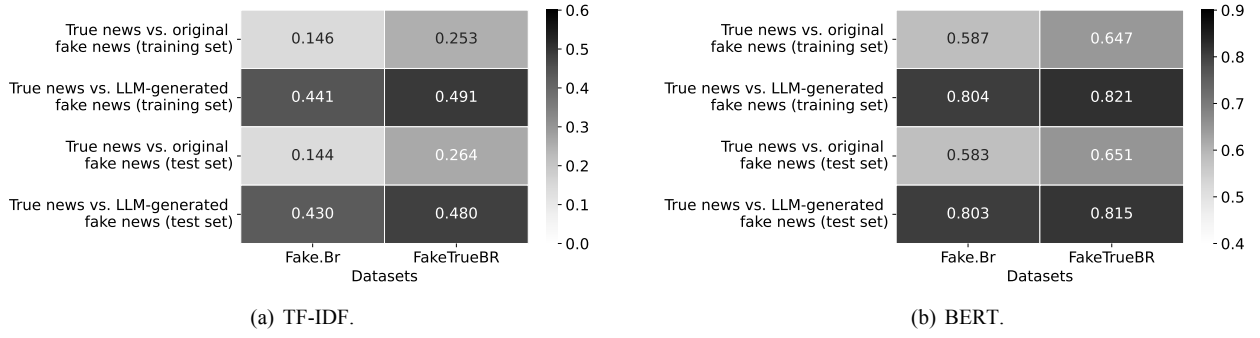


Figure 3. Cosine similarity between true news and fake news.

In both Figures 3(a) and 3(b), we present the similarity between the news in the training set and the similarities in the news from test set. As observed in the figures, in both representations, the similarities in FakeTrueBR are higher. We also noticed that the similarities obtained with BERT embeddings were often higher than those computed with the TF-IDF representation. This was expected, as BERT produces dense and context-aware representations that place semantically related texts close together in the embedding space. We also found that the similarities between true news and LLM-generated fake news are higher than those obtained between true news and original fake news, in both the TF-IDF and BERT analyses. This suggests that distinguishing true news from synthetic fake news is more challenging than distinguishing them from the original fake news.

4 Experimental setup

In this section, we outline the experimental setup adopted in this work.

4.1 Benchmark models

This section details the benchmark models used in our experiments. To provide a comprehensive comparative analysis, we selected models from distinct paradigms, including traditional ML classifiers, transformer-based models, and generative language models.

4.1.1 Machine learning classifiers

We performed experiments with the following established classification ML methods: logistic regression (LR) [Yu et al., 2011], support vector machines (SVM) [Boser et al., 1992; Cortes and Vapnik, 1995], decision trees (DT) [Breiman et al., 1984], and random forest (RF) [Breiman, 2001]. These methods were chosen because many fake news detection studies have been based on traditional ML approaches. Therefore, it is crucial to verify whether the conclusions drawn in these studies still hold in face of fake news generated by LLMs. Additionally, more advanced methods tend to be computationally expensive and often require paid access, making traditional ML techniques a valuable alternative in resource-constrained environments or scenarios with limited financial resources. However, these models may now face the challenge of han-

dling LLM-generated news, making it essential to study their robustness in this scenario.

All methods were implemented using the `scikit-learn` library [Pedregosa et al., 2011]. The experiments with SVM were evaluated using a linear kernel because its computational cost is lower than RBF and polynomial. As the performance of SVM, and RF can be highly affected by the choice of parameters, we performed a grid search using hold-out cross-validation to find the best values for their main parameters. For the regularization parameter of SVM, the following range of values were evaluated: $\{2^{-5}, 2^{-3}, 2^{-1}, \dots, 2^{15}\}$. For the number of estimators used in RF, the following range of values were tested: $\{10, 30, 50, \dots, 110\}$. For the other methods, we set their parameters to the default values.

4.1.2 Bidirectional encoder representations from Transformers (BERT)

Based on the Transformer architecture, BERT [Devlin et al., 2019] is a family of language models designed with the ability to capture context bidirectionally, outperforming contemporary models in understanding complex semantic and syntactic relationships. Building on this foundation, BERTimbau [Souza et al., 2020] is a variant specifically developed for Brazilian Portuguese, pre-trained on large-scale Portuguese corpora, and has shown superior performance on downstream tasks compared to multilingual BERT, while also being more efficient in terms of resources.

In this work, we assess a fine-tuned BERT model as a benchmark, using the BERTimbau architecture as the backbone for our fake news classification task. To optimize the hyperparameters for the experiments, we used Optuna³ to define the learning rate $\{1e-6, 1e-4\}$ and the weight decay $\{1e-2, 0.3\}$. Also, as suggested by Zhang et al. [2020], we reinitialized the top layer of the pre-trained model and opted for AdamW as the optimizer, since it mitigates the instability of the fine-tuning process on small datasets.

4.1.3 Retrieval-augmented generation (RAG)

LLMs are typically finetuned for specific NLP tasks by updating their parameters using a training recipe (e.g., back-propagation) and labelled data. An alternative mechanism

³Optuna. Available at <https://optuna.org/>. Accessed on: October 23, 2025.

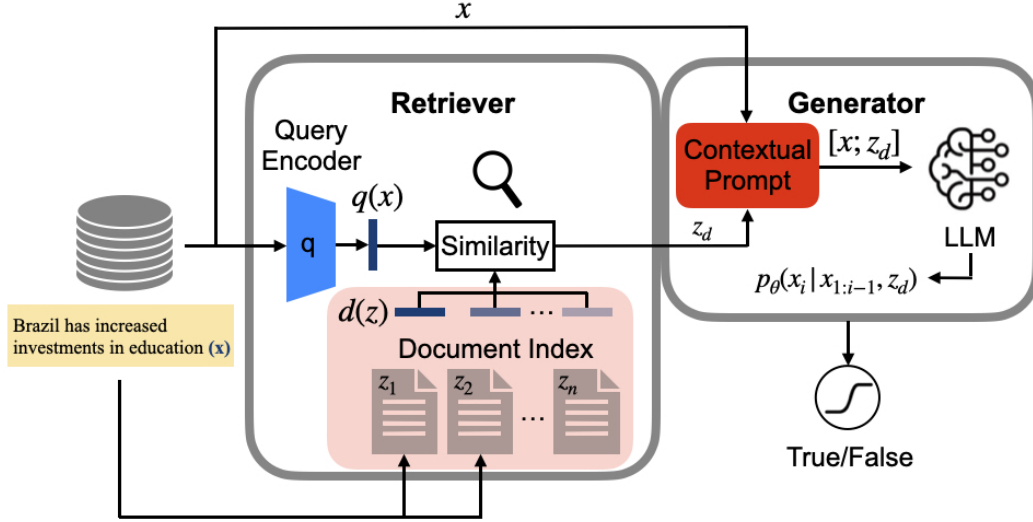


Figure 4. Retrieval-based question answering with two stages: **retrieval** which fetches relevant documents from the vector database, and the **generation** that provides answers based on contextual prompt. Query samples are collected from the fake news test set, and relevant documents are based on the training set.

to leverage LLMs without parameter optimization was introduced by Lewis *et al.* [2020]. The method combines LLMs’ encoded knowledge with an external knowledge for language generation. This method, referred to as RAG, consists of two stages, as depicted in Figure 4: **retrieval**, which returns relevant documents from the external knowledge, usually stored in a vector database; and **generation**, in which the LLM generates answers given a contextual prompt. For fake news detection, our retriever keeps a collection of indexed documents, Z , based on the training set. If a new query, x , is received, the retriever returns the top- k documents using similarity scores between the encoded query, $q(x)$, and the indexed documents, $d(z)$. Note that the queries are based on the test set whereas the indexed documents come from the training set. The generator receives x and the retrieved documents, z_d , where $d = 1, \dots, k$, to form the contextual prompt. The LLM, then, generates the output, $p_\theta(x_i | x_{1:i-1}, z_d)$, by predicting the next word, x_i , conditioned to the given context, that is, $[x_{1:i-1}, z_d]$.

Unlike traditional LLMs that rely solely on the pre-trained parameters, θ , the RAG framework dynamically retrieves relevant information from an external knowledge base, allowing for more accurate and context-aware text classification and fact-checking. In this work, we implemented a RAG-based system using LangChain⁴, a library that provides a modular interface for integrating LLMs with retrieval mechanisms, enabling efficient query processing and response generation. Our system utilizes the FAISS [Douce *et al.*, 2025] vector database to store and retrieve external knowledge, ensuring efficient nearest-neighbor search for relevant documents. We used the default similarity measure in FAISS, which calculates the cosine similarity between the user’s question and the embeddings of the stored documents. To improve retrieval quality, the documents were split into chunks of 500 tokens and an overlap of 200 tokens. For each query, the system retrieves the top 4 most relevant chunks as context for the LLM. The embedding model used for document indexing and retrieval is detailed in Section 4.2. The language model

employed for generating responses and classifying texts was the Sabiá-3, the same LLM used to generate the synthetic fake news, as explained in Section 3.

To evaluate the effectiveness of RAG for fake news detection, we designed two distinct prompting strategies, as depicted in Figure 5, and described below:

- **RAG_{EX}**: The LLM is instructed to use only the external knowledge for classification, without relying on its encoded knowledge from the pretraining phase.
- **RAG_{IN_EX}**: The model leverages both the encoded and external knowledge. The context is retrieved from the FAISS vector database, which contains indexed documents, and used to create the contextual prompt, as shown in Figure 4. In this work, we refer interchangeably to internal and encoded knowledge as the knowledge acquired by the LLM during training.

The motivation for incorporating RAG-based experiments was to analyze its performance on different data conditions. We hypothesize that in scenarios where the test data distribution closely matches the training data (i.e., no significant domain shift), traditional ML models may perform adequately without external knowledge retrieval. Notwithstanding, in presence of distributional shift — such as the emergence of new misleading content patterns or unseen claims — RAG could help mitigate the mismatch between train and test data, which is possible by storing and retrieving relevant new information from the external knowledge. This capability suggests the potential of RAG to enhance model robustness towards fake news landscapes, demonstrating its importance in fact-checking tasks where static training data alone may be insufficient.

4.1.4 Generative large language model

In addition to analyzing the results from traditional ML, BERT-like, and RAG-based solutions, we also considered a generative LLM for classifying fake and true news. For this, we considered the same pre-trained model, Sabiá-3, which was also used for generating synthetic fake news and within

⁴LangChain. Available at <https://www.langchain.com/>. Accessed on: October 23, 2025.

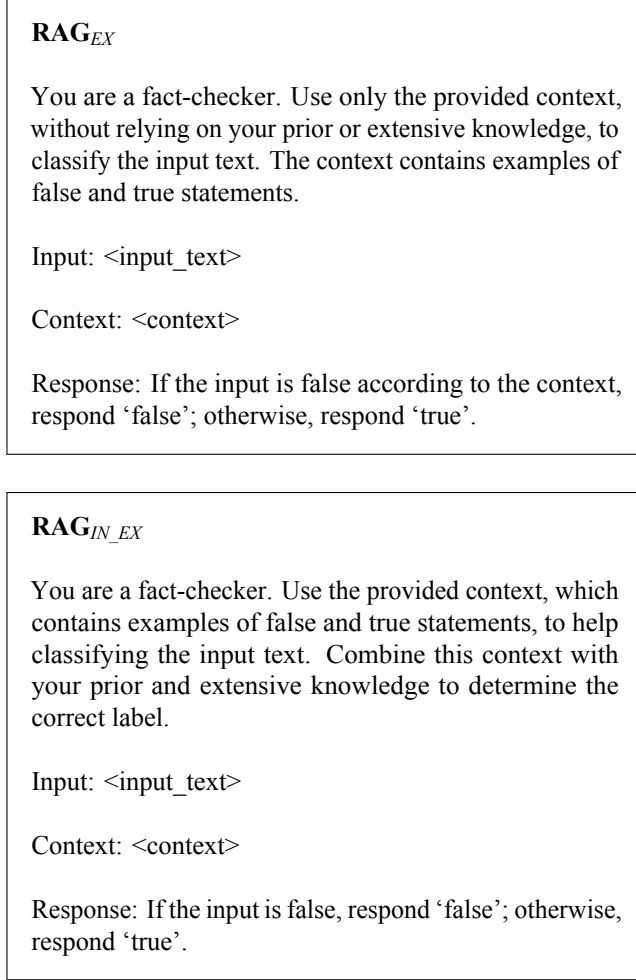


Figure 5. Prompt used to classify the input text as false or true. The top row represents the RAG framework based on its encoded knowledge (i.e., data seen during the pretraining phase) and its external knowledge, that is, the training data within the fake news database. The row on the bottom represents the RAG framework using only the external knowledge. The prompts were translated from Portuguese.

the RAG framework. Classification is performed by providing the LLM with context represented by the news text. It, then, generates the output, $p_{\theta}(x_i|x_{1:i-1})$, by predicting the next word, x_i , conditioned to the given context, $[x_{1:i-1}]$, which is similar to the RAG approach, but without the retrieved information, z_d . We used the prompt shown in Figure 6. This approach was crucial to evaluate how much prior knowledge regarding fake news was already encoded into the LLM during pretraining. Additionally, it also enables a fairer assessment to the contribution of external knowledge provided through the RAG framework. It is important to note that in our experiments, the LLM had no access to the fake news training data and was evaluated on the same test sets as the other methods.

4.2 Text representation

In our experiments, two text representations were considered: BoW and BERT embeddings [Devlin et al., 2019]. In the experiments with BoW, we applied the TF-IDF technique to adjust token weights in each document. Some studies, such as Silva et al. [2020] and Garcia et al. [2024], in the context of fake news detection in Portuguese, have shown

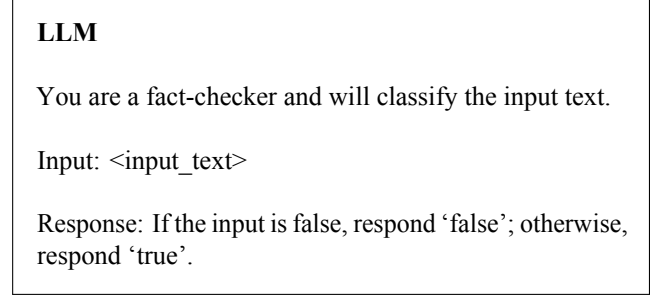


Figure 6. Prompt used to classify the input text as fake or true, without any contextual information. This prompt was translated from Portuguese.

that, for traditional machine learning methods, TF-IDF yields better results than static word embeddings, such as Word2Vec [Mikolov et al., 2013] and FastText [Bojanowski et al., 2017]. The vectors were generated using a TF-IDF model adjusted exclusively on the training data, which was split from the test set following the validation process described in Section 4.4. Besides the lexical representation based on TF-IDF, we also explored semantic embeddings using BERTimbau, which is expected to capture semantic similarity more effectively than TF-IDF and to provide a contrast between lexical and contextual approaches.

As traditional ML algorithms require fixed-size vector representation, the [CLS] token embedding provided by BERT models is typically used as fixed input representing an entire sequence of words or subwords. This approach can be suitable for tasks based on sentence-level representations, but fall short to provide semantically meaningful representation for longer texts. To mitigate this problem, we employed the *bert-large-portuguese-cased-sts2* model, available on Hugging Face⁵, which was derived from BERTimbau and maps sentences and paragraphs into a 1024-dimensional dense vector space. This model was trained following the approach proposed by Reimers and Gurevych [2019], which modifies BERT to generate semantically meaningful sentence embeddings. We used this model to create a single embedding for each news article. This BERT-based model was also used to create the embeddings for document indexing and retrieval in the RAG approach explained in Section 4.1.3.

4.3 Preprocessing

To ensure that traditional classification models and BERT did not rely on the length of the text as a distinguishing factor, we conducted experiments under two conditions. In the first condition, all texts were truncated to 200 tokens, following Silva et al. [2020]. This prevents the models from exploiting length differences between true and fake news — an issue that could be easily bypassed by deliberately adjusting the article length. Conversely, a second set of experiments used the full, untruncated text to assess model performance without this specific length constraint. As detailed in Section 3.2, our analysis confirms that fake news, whether human-written or LLM-generated, tends to be shorter than true news, reinforcing the importance of considering both processing methods.

⁵Model: *bert-large-portuguese-cased-sts2*. Available at <https://huggingface.co/rufimelo/bert-large-portuguese-cased-sts>. Accessed on: October 23, 2025

However, we did not apply the truncation in the RAG-based experiments. Since RAG retrieves external knowledge to enrich its responses, restricting the input length could limit the amount of contextual information available during retrieval and generation. Moreover, truncating the text before retrieval would be methodologically inconsistent, as the documents are already processed into overlapping chunks, as described in Section 4.1.3. Preserving the full text ensures that the retrieval mechanism can leverage as much information as possible, improving the system’s ability to retrieve relevant evidence.

Another important step of preprocessing was the text normalization before generating vector representations. Prior to computing the feature vectors with TF-IDF, all instances were converted to lowercase. Then, numerals, URLs, and emails were normalized into the dummy features ‘0’, ‘URL’, and ‘EMAIL’, respectively. After that, we tokenize the documents based on whitespaces and punctuation marks.

It is important to note that this normalization have drawbacks for fake news detection, as exaggerated or misleading numerical values can be an important evidence in determining whether a news article is fake. However, keeping numbers in their original form would increase the dimensionality of the feature space, making the representation sparser. Moreover, specific numbers often have low frequency in the dataset, meaning that TF-IDF would naturally assign them lower importance.

This preprocessing step was not applied when generating text representations with BERT. Since BERT relies on its own tokenizer, designed and optimized during pretraining, it is essential to maintain consistency with its tokenization process to preserve the model performance.

4.4 Validation

For model evaluation, we applied a holdout validation, with 80% of the data used for training and 20% for testing. The data split was performed randomly. However, we ensured that aligned pairs of true and fake news articles remained in the same set. This approach preserves the integrity of the alignment and prevents information leakage between training and testing sets.

4.5 Performance measures

To compare the results, we employed the following well-known performance measures for classification:

- accuracy: proportion of correct predictions;
- false positive rate (FPR): proportion of true news incorrectly labeled as fake news;
- recall: proportion of fake news correctly identified;
- precision: proportion of news classified as fake and that truly belong to the fake class;
- F-measure: harmonic average of the precision and recall.

For FPR, lower values indicate better performance in the classification. On the other hand, for accuracy, recall, precision, and F-measure, higher values represents better results.

5 Results

In this section, we present results from five experiments conveying different training and test sets configurations. The first experiment consists of training and testing models with original fake news only. The second one relies on training models with original fake news and testing them with synthetic samples. The third experiment inverts the order, training models with synthetic fake news and testing them with original samples. The fourth experiment is based on training and testing models with synthetic fake news. Finally, in the fifth experiment, models are trained with varying amount of synthetic fake news in the training data.

In the following sections, results are presented in Tables 3, 4, 5 and 6, where the first column refers to the method being used for classification, and the subscript designates the text representation adopted. For example, Method_{TF-IDF} or Method_{BERT} represent a method based on the TF-IDF and BERT representations, respectively. The subscript for the last two rows describes the RAG framework based on external knowledge (RAG_{EX}) or the combination of internal and external knowledge (RAG_{IN_EX}).

5.1 Experiment I: original fake news only

In this experiment, we evaluate the performance of several models on the task of fake news detection. For that, these models were trained and tested with original fake news only. Results are presented in Table 3, where performances in Tables 3(a) and 3(b) are based on truncated texts, as described in Section 4.3, whereas the results in Tables 3(c) and 3(d) are based on full texts. In such scenario, fine-tuned BERT achieved the highest overall performance for both the Fake.Br and FakeTrueBR datasets. This trend was observed for both truncated and full text approaches, with the latter achieving slightly higher performances. Using the Fake.Br dataset, fine-tuned BERT and all the traditional machine learning methods, except for the DT_{BERT} , outperformed the LLM-based approaches in terms of accuracy, recall, precision, and F-measure. The F-measure for the RAG_{EX} , for instance, was about 20% lower than the score achieved by the BERT model. Additionally, the FPR obtained by RAG_{IN_EX} was as high as 13%, which can be considered significant as it compromises the credibility of true news.

For the FakeTrueBR dataset (Tables 3(b) and 3(d)), the fine-tuned BERT is still providing the best performance for the full text approach. Note that for the FakeTrueBR dataset, the LLM-based approaches achieved higher recall than most other methods, suggesting that the dataset conveys relevant differences compared to the Fake.Br. For instance, it may include lexical or stylistic patterns, topic distribution, or annotation characteristics that are better captured by the generative nature or retrieval capabilities of the LLM-based models, but at the cost of failing to maintain balanced performance across all metrics.

The results obtained with the full text approach were, in most cases, superior to those achieved with the truncated texts. This is in line with the findings of Silva et al. [2020], which show that classifiers may rely on text length to leverage their ability to distinguish between true and fake documents. This

Table 3. Trained and tested with the original fake news.

(a) Fake.Br (truncated text).						(b) FakeTrueBR (truncated text).					
Method	Accuracy	FPR	Recall	Precision	F-measure	Method	Accuracy	FPR	Recall	Precision	F-measure
DT _{TF-IDF}	0.795	0.214	0.804	0.790	0.797	DT _{TF-IDF}	0.875	0.114	0.864	0.883	0.873
DT _{BERT}	0.696	0.303	0.694	0.696	0.695	DT _{BERT}	0.831	0.189	0.852	0.818	0.835
LR _{TF-IDF}	0.924	0.069	0.918	0.930	0.924	LR _{TF-IDF}	0.944	0.061	0.950	0.939	0.945
LR _{BERT}	0.945	0.060	0.950	0.941	0.945	LR _{BERT}	0.969	0.033	0.972	0.967	0.969
SGD _{TF-IDF}	0.938	0.053	0.928	0.946	0.937	SGD _{TF-IDF}	0.965	0.033	0.964	0.966	0.965
SGD _{BERT}	0.923	0.117	0.963	0.892	0.926	SGD _{BERT}	0.964	0.033	0.961	0.966	0.964
RF _{TF-IDF}	0.908	0.096	0.911	0.905	0.908	RF _{TF-IDF}	0.955	0.070	0.981	0.934	0.957
RF _{BERT}	0.876	0.132	0.885	0.870	0.877	RF _{BERT}	0.942	0.075	0.958	0.927	0.942
SVM _{TF-IDF}	0.938	0.049	0.924	0.950	0.937	SVM _{TF-IDF}	0.967	0.031	0.964	0.969	0.966
SVM _{BERT}	0.951	0.051	0.954	0.949	0.952	SVM _{BERT}	0.960	0.039	0.958	0.961	0.960
BERT	0.997	0.003	0.996	0.997	0.997	BERT	0.997	0.000	0.994	1.000	0.997
Sabiá-3	0.750	0.055	0.940	0.680	0.790	Sabiá-3	0.560	0.005	0.990	0.530	0.700
RAG _{EX}	0.770	0.119	0.880	0.720	0.790	RAG _{EX}	0.550	0.002	1.000	0.530	0.690
RAG _{IN_EX}	0.780	0.137	0.860	0.740	0.800	RAG _{IN_EX}	0.560	0.008	0.990	0.530	0.690

(c) Fake.Br (full text).						(d) FakeTrueBR (full text).					
Method	Accuracy	FPR	Recall	Precision	F-measure	Method	Accuracy	FPR	Recall	Precision	F-measure
DT _{TF-IDF}	0.905	0.110	0.919	0.893	0.906	DT _{TF-IDF}	0.929	0.064	0.922	0.935	0.928
DT _{BERT}	0.813	0.190	0.817	0.811	0.814	DT _{BERT}	0.876	0.150	0.903	0.857	0.879
LR _{TF-IDF}	0.959	0.058	0.976	0.944	0.960	LR _{TF-IDF}	0.948	0.067	0.964	0.935	0.949
LR _{BERT}	0.976	0.014	0.965	0.986	0.975	LR _{BERT}	0.975	0.022	0.972	0.978	0.975
SGD _{TF-IDF}	0.967	0.044	0.978	0.957	0.967	SGD _{TF-IDF}	0.971	0.039	0.981	0.962	0.971
SGD _{BERT}	0.965	0.007	0.938	0.993	0.964	SGD _{BERT}	0.976	0.019	0.972	0.980	0.976
RF _{TF-IDF}	0.952	0.071	0.975	0.932	0.953	RF _{TF-IDF}	0.942	0.092	0.975	0.914	0.943
RF _{BERT}	0.928	0.057	0.914	0.941	0.927	RF _{BERT}	0.937	0.086	0.961	0.918	0.939
SVM _{TF-IDF}	0.969	0.035	0.974	0.966	0.970	SVM _{TF-IDF}	0.968	0.039	0.975	0.962	0.968
SVM _{BERT}	0.977	0.010	0.964	0.990	0.977	SVM _{BERT}	0.982	0.019	0.983	0.981	0.982
BERT	0.999	0.001	0.999	0.999	0.999	BERT	0.999	0.000	0.997	1.000	0.999
Sabiá-3	0.750	0.055	0.940	0.680	0.790	Sabiá-3	0.560	0.005	0.990	0.530	0.700
RAG _{EX}	0.770	0.119	0.880	0.720	0.790	RAG _{EX}	0.550	0.002	1.000	0.530	0.690
RAG _{IN_EX}	0.780	0.137	0.860	0.740	0.800	RAG _{IN_EX}	0.560	0.008	0.990	0.530	0.690

gap in terms of performance is more pronounced in the experiments based on the Fake.Br dataset. This finding is consistent with the statistics presented in Table 2, which indicate that the average number of tokens for fake and true news differs more in the Fake.Br compared to the FakeTrueBR dataset. For BERT, however, the difference was less pronounced, likely because its attention mechanism builds contextual embeddings mainly from the first tokens up to a fixed limit, reducing the impact of additional content that exceeds this limit.

Focusing the analysis on the performance of traditional machine learning methods, most of them achieved an F-measure that outperformed Sabiá-3 and the RAG approaches. This is particularly interesting considering the inherently more robust and advanced nature of the latter. Among the traditional models, SVM was the most effective method. The F-measure was comparable between methods employing TF-IDF and BERT representations, with the only method showing a relevant difference being the DT, where TF-IDF yielded better results.

5.2 Experiment II: testing with synthetic fake news

In this experiment, original fake news are used for training whereas synthetic fake news are used solely for evaluation. Table 4 presents the results, which show that the LLM-based models (i.e., Sabiá-3, RAG_{EX} and RAG_{IN_EX}) have improved their performance in the Fake.Br experiments compared to the experiments discussed in Section 5.1. For the FakeTrueBR

dataset, however, their results were similar. On the other hand, ML methods had their performance decreased in respect to those found in Section 5.1 for both datasets, with lower accuracy, recall, precision, and F-measure. This suggests that ML models can be less robust to LLM-generated fake news. This pattern was found in both the Fake.Br and FakeTrueBR datasets.

It is important to notice that, although ML models were affected by the data distribution shift, few ML algorithms are still performing well. For the Fake.Br dataset, for instance, RF_{TF-IDF} achieved the highest accuracy of 0.833 using the full text approach. SGD_{TF-IDF}, on the other hand, reached an accuracy of 0.623 on the FakeTrueBR dataset also based on full texts. The RAG_{IN_EX} achieves an accuracy of 0.820 for the same dataset. Note that the ability of ML methods to identify fake news (i.e., true positives) is severely impacted, which lower the scores attained for recall. This led the F-measure provided by all ML algorithms to be inferior to the ones achieved by the LLM-based models. Thus, in terms of F-measures, RAG_{IN_EX} provided the best scores for the Fake.Br (0.840), while Sabiá-3 achieved the best values for the FakeTrueBR (0.700). LLM-based approaches, on the other hand, showed lower precision than most of the other methods for both datasets, thus producing more false positives. These results indicate that the use of external knowledge (i.e., RAG_{IN_EX}) can be suitable to identify synthetic fake news, although it may struggle to mitigate false positives.

While there is no evidence that during its pretraining phase the LLM had access to articles conveying the facts present in the test set, it could be argued its ability to detect synthetic

Table 4. Trained with original and tested on synthetic fake news

(a) Fake.Br (truncated text).						(b) FakeTrueBR (truncated text).					
Method	Accuracy	FPR	Recall	Precision	F-measure	Method	Accuracy	FPR	Recall	Precision	F-measure
DT _{TF-IDF}	0.573	0.214	0.360	0.627	0.457	DT _{TF-IDF}	0.547	0.114	0.209	0.647	0.316
DT _{BERT}	0.576	0.303	0.456	0.601	0.518	DT _{BERT}	0.532	0.189	0.253	0.572	0.351
LR _{TF-IDF}	0.666	0.069	0.401	0.853	0.546	LR _{TF-IDF}	0.546	0.061	0.153	0.714	0.252
LR _{BERT}	0.731	0.060	0.522	0.897	0.660	LR _{BERT}	0.506	0.033	0.045	0.571	0.083
SGD _{TF-IDF}	0.667	0.053	0.386	0.880	0.537	SGD _{TF-IDF}	0.582	0.033	0.198	0.855	0.321
SGD _{BERT}	0.762	0.117	0.642	0.846	0.730	SGD _{BERT}	0.506	0.033	0.045	0.571	0.083
RF _{TF-IDF}	0.619	0.096	0.333	0.777	0.466	RF _{TF-IDF}	0.535	0.070	0.139	0.667	0.230
RF _{BERT}	0.618	0.132	0.368	0.736	0.491	RF _{BERT}	0.524	0.075	0.123	0.620	0.205
SVM _{TF-IDF}	0.665	0.049	0.379	0.886	0.531	SVM _{TF-IDF}	0.577	0.031	0.184	0.857	0.303
SVM _{BERT}	0.723	0.051	0.497	0.906	0.642	SVM _{BERT}	0.506	0.039	0.050	0.562	0.092
BERT	0.765	0.008	0.539	0.985	0.697	BERT	0.601	0.005	0.341	0.634	0.443
Sabiá-3	0.770	0.043	0.960	0.700	0.810	Sabiá-3	0.570	0.008	0.990	0.540	0.700
RAG _{EX}	0.780	0.083	0.920	0.720	0.810	RAG _{EX}	0.540	0.011	0.990	0.520	0.690
RAG _{IN_EX}	0.820	0.072	0.930	0.760	0.840	RAG _{IN_EX}	0.560	0.005	0.990	0.530	0.690

(c) Fake.Br (full text).						(d) FakeTrueBR (full text).					
Method	Accuracy	FPR	Recall	Precision	F-measure	Method	Accuracy	FPR	Recall	Precision	F-measure
DT _{TF-IDF}	0.694	0.110	0.497	0.819	0.619	DT _{TF-IDF}	0.603	0.064	0.270	0.808	0.405
DT _{BERT}	0.596	0.190	0.382	0.667	0.486	DT _{BERT}	0.529	0.150	0.209	0.581	0.307
LR _{TF-IDF}	0.809	0.058	0.676	0.921	0.780	LR _{TF-IDF}	0.589	0.067	0.245	0.786	0.374
LR _{BERT}	0.740	0.014	0.494	0.973	0.656	LR _{BERT}	0.511	0.022	0.045	0.667	0.084
SGD _{TF-IDF}	0.813	0.044	0.671	0.938	0.782	SGD _{TF-IDF}	0.623	0.039	0.284	0.879	0.429
SGD _{BERT}	0.667	0.007	0.340	0.980	0.505	SGD _{BERT}	0.514	0.019	0.047	0.708	0.089
RF _{TF-IDF}	0.833	0.071	0.736	0.912	0.815	RF _{TF-IDF}	0.578	0.092	0.248	0.730	0.370
RF _{BERT}	0.607	0.057	0.271	0.826	0.408	RF _{BERT}	0.556	0.086	0.198	0.696	0.308
SVM _{TF-IDF}	0.793	0.035	0.621	0.947	0.750	SVM _{TF-IDF}	0.618	0.039	0.276	0.876	0.419
SVM _{BERT}	0.745	0.010	0.500	0.981	0.662	SVM _{BERT}	0.517	0.019	0.053	0.731	0.099
BERT	0.766	0.001	0.533	0.997	0.695	BERT	0.579	0.091	0.310	0.850	0.455
Sabiá-3	0.770	0.043	0.960	0.700	0.810	Sabiá-3	0.570	0.008	0.990	0.540	0.700
RAG _{EX}	0.780	0.083	0.920	0.720	0.810	RAG _{EX}	0.540	0.011	0.990	0.520	0.690
RAG _{IN_EX}	0.820	0.072	0.930	0.760	0.840	RAG _{IN_EX}	0.560	0.005	0.990	0.530	0.690

fake news could have been attained from specific information in the articles used to pretrain these models, potentially characterizing data leakage. Since the news articles in the Fake.Br and FakeTrueBR datasets were published before the LLM’s pretraining period ended, such argument cannot be entirely discarded. Notwithstanding, one evidence against such an assumption is the use of the RAG_{EX} approach, which explicitly instructs the LLM to not use its internal knowledge (see Section 4.1.3) during the inference step — although there is no guarantee that this constraint was fully respected.

It is important to note that the direct LLM approach, without RAG, performed worse than RAG_{IN_EX} in terms of accuracy, F-measure, and precision on the Fake.Br dataset. This shows that, even if some leakage occurred, RAG remains a promising approach. For the FakeTrueBR dataset, on the other hand, the direct prompt approach yielded better results.

The fine-tuned BERT model, although performing better than most traditional ML methods, was also impacted by LLM-generated fake news. Despite being a more sophisticated model, capable of capturing complex linguistic patterns, its performance decayed significantly compared to its results on original fake news, with nearly 30% and 54% decays, respectively, for the Fake.Br and FakeTrueBR datasets. This suggests that more sophisticated models equally struggle to remain robust in face of synthetic fake news, highlighting the need for enhanced strategies to attain detection robustness.

5.3 Experiment III: training with synthetic fake news

In this experiment, synthetic fake news are used solely for training while original fake news are used for evaluation. Table 5 presents a trend similar to the one observed in Table 4, where the mismatch between the training and test sets severely affects the performance of ML models. In general, accuracies and F-measures are lower for the ML algorithms compared to the ones found in Table 3. Similarly to the previous experiments, recall is the most affected by the mismatch condition (i.e., with training and test sets from different distributions). Among the ML models, RF_{TF-IDF} seems to be the most robust one, providing the best performance, in terms of accuracy and F-measure, 0.903 and 0.896, for the Fake.Br dataset (see Table 5(c)). Note that a similar trend is observed in Table 4(c). For the FakeTrueBR, we found the LR_{TF-IDF} providing the highest accuracy and F-measure, 0.777 and 0.718, respectively, in Table 5(c).

As mentioned before, the results attained by the ML algorithms were generally higher for the full text approach. Additionally, the ML methods with the highest scores seem to rely more on the lexical representation (i.e., TF-IDF), rather than on semantic embeddings. The LLM-based approaches outperformed the traditional ML methods and the fine-tuned BERT, in terms of recall, in all cases, and in terms of F-measure for the truncated instances. This can be observed in Tables 5(a) and 5(b). This reinforces the hypothesis that such models can be more robust to data distribution shifts, characterized by the mismatch between original and synthetic fake news. As previously discussed, we cannot rule out the

Table 5. Trained with synthetic and tested on original fake news.

(a) Fake.Br (truncate text).						(b) FakeTrueBR (truncate text).					
Method	Accuracy	FPR	Recall	Precision	F-measure	Method	Accuracy	FPR	Recall	Precision	F-measure
DT _{TF-IDF}	0.558	0.108	0.224	0.674	0.336	DT _{TF-IDF}	0.547	0.100	0.195	0.660	0.301
DT _{BERT}	0.593	0.232	0.418	0.643	0.507	DT _{BERT}	0.506	0.189	0.201	0.514	0.289
LR _{TF-IDF}	0.533	0.004	0.071	0.944	0.132	LR _{TF-IDF}	0.639	0.006	0.284	0.981	0.441
LR _{BERT}	0.574	0.010	0.157	0.942	0.269	LR _{BERT}	0.493	0.022	0.008	0.273	0.016
SGD _{TF-IDF}	0.520	0.004	0.044	0.914	0.085	SGD _{TF-IDF}	0.574	0.006	0.153	0.965	0.264
SGD _{BERT}	0.560	0.015	0.135	0.898	0.234	SGD _{BERT}	0.493	0.025	0.011	0.308	0.022
RF _{TF-IDF}	0.519	0.010	0.047	0.829	0.089	RF _{TF-IDF}	0.532	0.022	0.086	0.795	0.156
RF _{BERT}	0.590	0.062	0.242	0.795	0.371	RF _{BERT}	0.501	0.086	0.089	0.508	0.152
SVM _{TF-IDF}	0.520	0.003	0.043	0.939	0.082	SVM _{TF-IDF}	0.560	0.006	0.125	0.957	0.222
SVM _{BERT}	0.567	0.010	0.143	0.936	0.248	SVM _{BERT}	0.493	0.031	0.017	0.353	0.032
BERT	0.640	0.069	0.291	0.833	0.413	BERT	0.702	0.150	0.540	0.664	0.596
Sabiá-3	0.750	0.055	0.940	0.680	0.790	Sabiá-3	0.560	0.005	0.990	0.530	0.700
RAG _{EX}	0.770	0.145	0.850	0.740	0.790	RAG _{EX}	0.530	0.008	0.999	0.520	0.680
RAG _{IN_EX}	0.760	0.158	0.840	0.730	0.780	RAG _{IN_EX}	0.550	0.008	0.999	0.530	0.690

(c) Fake.Br (full text).						(d) FakeTrueBR (full text).					
Method	Accuracy	FPR	Recall	Precision	F-measure	Method	Accuracy	FPR	Recall	Precision	F-measure
DT _{TF-IDF}	0.728	0.076	0.532	0.874	0.661	DT _{TF-IDF}	0.557	0.086	0.201	0.699	0.312
DT _{BERT}	0.636	0.201	0.474	0.702	0.566	DT _{BERT}	0.526	0.128	0.181	0.586	0.277
LR _{TF-IDF}	0.828	0.015	0.671	0.978	0.796	LR _{TF-IDF}	0.777	0.014	0.568	0.976	0.718
LR _{BERT}	0.614	0.010	0.237	0.961	0.381	LR _{BERT}	0.496	0.019	0.011	0.364	0.022
SGD _{TF-IDF}	0.680	0.006	0.365	0.985	0.533	SGD _{TF-IDF}	0.625	0.000	0.251	1.000	0.401
SGD _{BERT}	0.608	0.015	0.231	0.938	0.370	SGD _{BERT}	0.493	0.042	0.028	0.400	0.052
RF _{TF-IDF}	0.903	0.035	0.840	0.960	0.896	RF _{TF-IDF}	0.666	0.006	0.337	0.984	0.502
RF _{BERT}	0.645	0.031	0.321	0.913	0.475	RF _{BERT}	0.501	0.089	0.092	0.508	0.156
SVM _{TF-IDF}	0.665	0.007	0.338	0.980	0.502	SVM _{TF-IDF}	0.609	0.000	0.217	1.000	0.357
SVM _{BERT}	0.602	0.006	0.210	0.974	0.345	SVM _{BERT}	0.497	0.025	0.019	0.438	0.037
BERT	0.619	0.025	0.262	0.913	0.408	BERT	0.669	0.140	0.472	0.734	0.575
Sabiá-3	0.750	0.055	0.940	0.680	0.790	Sabiá-3	0.560	0.005	0.990	0.530	0.700
RAG _{EX}	0.770	0.145	0.850	0.740	0.790	RAG _{EX}	0.530	0.008	0.999	0.520	0.680
RAG _{IN_EX}	0.760	0.158	0.840	0.730	0.780	RAG _{IN_EX}	0.550	0.008	0.999	0.530	0.690

possibility of data leakage from the LLM’s pretraining phase, which might be influencing these results.

In general, while all ML methods experienced a significant decay in performance compared to experiments in Table 3, the generative LLM-based approach maintained similar results, with RAG_{EX} and RAG_{IN_EX} outperforming the LLM without external knowledge in terms of accuracy on the Fake.Br dataset, which reiterates the importance of providing LLMs with context to improve its performance. These results suggest that training on LLM-generated fake news alone is not sufficient for traditional ML models to generalize effectively to real-world fake news. Thus, as models trained solely on synthetic data may struggle to perform well on real-world misleading content, developing strategies, such as new human-labeled datasets tailored for increasing the robustness of fake news detection methods, remains highly important.

5.4 Experiment IV: synthetic fake news only

In this experiment, we evaluate the performance of models trained and tested with synthetic fake news only. Results are presented in Table 6 and the scenario explored is characterized by a matching condition, with no shift between the training and test distributions. Similarly to our findings discussed in Section 5.1, most ML methods outperformed the LLM-based approaches. This corroborates with the idea that combining LLM with external knowledge is more suitable in cases where the training and test sets are not represented by the same distributions.

The results in Table 6 also suggest that ML models com-

bined with the lexical representation (i.e., TF-IDF) yields better results when compared with the results based on the semantic embeddings. This aligns with the linguistic feature analysis presented in Table 2, where we observed significant differences in certain linguistic characteristics between true news and LLM-generated fake news. Although TF-IDF does not directly consider class labels, it can highlight words that are statistically more frequent or distinctive within subsets of documents, which may correlate with different categories.

5.5 Experiment V: training with varying amount of synthetic fake news

In this experiment, models are trained by mixing the training set with original fake news and a varying amount of synthetic data. Our main motivation is to assess whether a mix of original and synthetic fake news in the training set could offer a better approach for detecting either human or synthetic fake news. We are particularly concerned with the fine-tuned BERT and traditional ML methods, which were more impacted by the mismatch between the training and test data distributions (see Sections 5.2 and 5.3), often providing lower performance when compared with their results under matched conditions (see Sections 5.1 and 5.5). For this, we varied the proportion of LLM-generated fake news added to the training set from 0% to 100%, with a 10% “step”. Figures 7(a) and 7(c) show the F-measure when the test set contains only original fake news, while Figures 7(b) and 7(d) display the F-measure when the test set contains only LLM-generated fake news. Thus, while the test set is always kept with pure

Table 6. Trained and tested on synthetic fake news.

(a) Fake.Br (truncated text).						(b) FakeTrueBR (truncated text).					
Method	Accuracy	FPR	Recall	Precision	F-measure	Method	Accuracy	FPR	Recall	Precision	F-measure
DT _{TF-IDF}	0.897	0.108	0.903	0.893	0.898	DT _{TF-IDF}	0.904	0.100	0.908	0.901	0.904
DT _{BERT}	0.754	0.232	0.740	0.761	0.751	DT _{BERT}	0.834	0.189	0.858	0.819	0.838
LR _{TF-IDF}	0.992	0.004	0.988	0.996	0.992	LR _{TF-IDF}	0.987	0.006	0.981	0.994	0.987
LR _{BERT}	0.990	0.010	0.989	0.990	0.990	LR _{BERT}	0.979	0.022	0.981	0.978	0.979
SGD _{TF-IDF}	0.996	0.004	0.996	0.996	0.996	SGD _{TF-IDF}	0.987	0.006	0.981	0.994	0.987
SGD _{BERT}	0.988	0.015	0.990	0.985	0.988	SGD _{BERT}	0.969	0.025	0.964	0.975	0.969
RF _{TF-IDF}	0.987	0.010	0.983	0.990	0.987	RF _{TF-IDF}	0.983	0.022	0.989	0.978	0.983
RF _{BERT}	0.927	0.062	0.917	0.936	0.926	RF _{BERT}	0.944	0.086	0.975	0.919	0.946
SVM _{TF-IDF}	0.997	0.003	0.996	0.997	0.997	SVM _{TF-IDF}	0.987	0.006	0.981	0.994	0.987
SVM _{BERT}	0.990	0.010	0.990	0.990	0.990	SVM _{BERT}	0.975	0.031	0.981	0.970	0.975
BERT	0.998	0.004	1.000	0.996	0.998	BERT	0.989	0.019	0.997	0.981	0.989
Sabiá-3	0.770	0.043	0.960	0.700	0.810	Sabiá-3	0.570	0.008	0.990	0.540	0.700
RAG _{EX}	0.800	0.090	0.910	0.740	0.820	RAG _{EX}	0.540	0.019	0.980	0.520	0.680
RAG _{IN_EX}	0.820	0.065	0.930	0.760	0.840	RAG _{IN_EX}	0.560	0.013	0.990	0.530	0.690

(c) Fake.Br (full text).						(d) FakeTrueBR (full text).					
Method	Accuracy	FPR	Recall	Precision	F-measure	Method	Accuracy	FPR	Recall	Precision	F-measure
DT _{TF-IDF}	0.929	0.076	0.935	0.924	0.930	DT _{TF-IDF}	0.909	0.086	0.905	0.913	0.909
DT _{BERT}	0.792	0.201	0.786	0.796	0.791	DT _{BERT}	0.875	0.128	0.877	0.873	0.875
LR _{TF-IDF}	0.992	0.015	0.999	0.985	0.992	LR _{TF-IDF}	0.989	0.014	0.992	0.986	0.989
LR _{BERT}	0.993	0.010	0.996	0.990	0.993	LR _{BERT}	0.985	0.019	0.989	0.981	0.985
SGD _{TF-IDF}	0.997	0.006	0.999	0.994	0.997	SGD _{TF-IDF}	0.999	0.000	0.997	1.000	0.999
SGD _{BERT}	0.990	0.015	0.996	0.985	0.990	SGD _{BERT}	0.975	0.042	0.992	0.960	0.975
RF _{TF-IDF}	0.981	0.035	0.996	0.966	0.981	RF _{TF-IDF}	0.990	0.006	0.986	0.994	0.990
RF _{BERT}	0.960	0.031	0.951	0.969	0.960	RF _{BERT}	0.951	0.089	0.992	0.918	0.953
SVM _{TF-IDF}	0.996	0.007	0.999	0.993	0.996	SVM _{TF-IDF}	1.000	0.000	1.000	1.000	1.000
SVM _{BERT}	0.995	0.006	0.996	0.994	0.995	SVM _{BERT}	0.982	0.025	0.989	0.975	0.982
BERT	0.997	0.006	0.999	0.994	0.997	BERT	0.997	0.000	0.994	1.000	0.997
Sabiá-3	0.770	0.043	0.960	0.700	0.810	Sabiá-3	0.570	0.008	0.990	0.540	0.700
RAG _{EX}	0.800	0.090	0.910	0.740	0.820	RAG _{EX}	0.540	0.019	0.980	0.520	0.680
RAG _{IN_EX}	0.820	0.065	0.930	0.760	0.840	RAG _{IN_EX}	0.560	0.013	0.990	0.530	0.690

original or synthetic samples, the training data is mixed, except for the extremes of 0% and 100%. For the benefit of visualization, we chose one single model as representative of the ML models' behave under such regime. As a result, our analysis is based on the SVM, which provided the best performance, in terms of F-measure, for the experiments where both the training and test sets contained only original fake news. Note that the results for Sabiá-3 were included solely for comparison purposes, since this approach does not rely on the training set, thus its performance remains unchanged.

The results indicate that the hybrid approach performs better for SVM compared to training with only LLM-generated fake news to classify human-generated fake news (as shown at 100% on the x-axis in Figures 7(a) and 7(c), which corresponds to Tables 5(a) and 5(b)). The F-measure obtained by SVM decreases as the percentage of LLM-generated fake news in the training set increases. This is because the training set becomes progressively more similar to the test set, which is composed only of human-generated fake news. Similarly, the hybrid approach also outperforms the inverse scenario, where the model is trained only on human-generated fake news to classify LLM-generated fake news (represented at 0% on the x-axis in Figures 7(b) and 7(d), corresponding to Tables 4(a) and 4(b)). In this case, the F-measure obtained by SVM increases as the proportion of LLM-generated fake news in the training set grows, because the training set becomes progressively more similar to the test set (composed only of LLM-generated fake news). These results suggest that simply mixing original and synthetic fake news is not sufficient to allow the methods to generalize effectively. Moreover, this in-

dicates that generating synthetic fake news from LLMs based on true news is not an optimal strategy for dataset augmentation, since even with less than 50% synthetic fake news in the training set, SVM performance dropped when tested against human-generated fake news. While generating synthetic fake news from human-generated fake news could yield better results, it is more challenging. In contrast, obtaining true news from reliable sources is relatively easier and eliminates the need for manual labeling.

Among the methods tested in the hybrid approach, BERT demonstrated greater robustness, exhibiting a smaller performance decline when the proportional difference in data types between the training and test sets was more pronounced. This suggests that, while the hybrid strategy may not be ideal for most methods, it could still be effective for BERT. For the RAG-based approaches, the hybrid approach had no significant impact on performance, neither positive nor negative, even in the experiments with the original test set. Furthermore, in most experiments with the hybrid approach, RAG outperformed the LLM on the Fake.Br dataset. Since the LLM does not utilize the training data, this consistent superiority of RAG suggests that external knowledge plays a crucial role in achieving better results. On the FakeTrueBR dataset, however, the direct LLM approach and RAG performed nearly identically.

6 Limitation and future work

Our evaluation considered only text-based classification, without incorporating multimodal features such as images or meta-

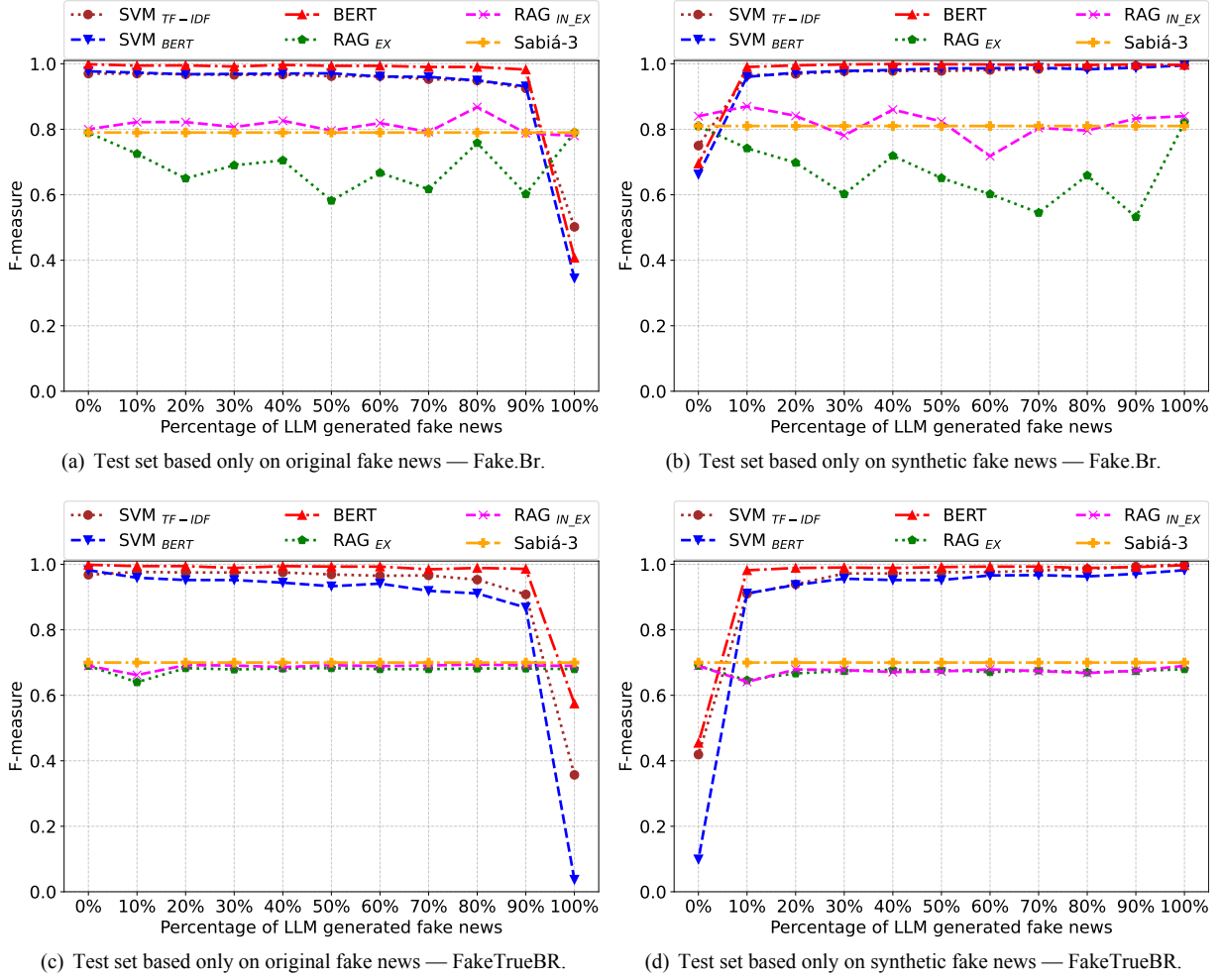


Figure 7. Performance in terms of F-measure given a varying percentage of fake news in the training set. Experiments are based on full texts.

data, which are often present in real-world fake news. In future research, we intend to investigate how incorporating multimodal features, such as images, videos, and social context, can improve fake news detection in both human and LLM-generated fake news. Furthermore, this work focused specifically on the Portuguese language, which has fewer NLP and ML resources compared to English. Future work could explore whether our findings can also be observed in other low-resource languages, such as Spanish, French, Hindi, and Indonesian, where similar challenges in fake news detection may arise due to limited labeled datasets and language-specific features.

7 Conclusion

In this study, we evaluated how LLM-generated fake news can impact classifiers' performance. We conducted extensive experiments using two Portuguese-language datasets with aligned news articles. That is, for each true news article, we generated a fake version using an LLM model. Experiments included traditional ML algorithms, a fine-tuned BERT, an LLM, and an LLM augmented with external knowledge (or RAG). These models were used to classify fake news written by humans versus those generated by LLMs. Results showed a performance decay in mismatch conditions, that is, when

the models were trained on human-written fake news and tested on LLM-generated fake news, or vice-versa. We observed that by mixing both types of fake news during training, i.e., original and synthetic, slightly improved performance.

Despite this general trend in the results, the classification strategies showed important differences. When the training and test datasets contained the same type of fake news (either human-written or LLM-generated), traditional ML classifiers, particularly SVM and SGD, along with the fine-tuned BERT, generally obtained the best results. However, when the data set contained mixed types, with LLM-generated fake news appearing only in the training or test set, generative LLM-based classifiers outperformed the other methods, especially the RAG approach that combined internal and external knowledge. The findings indicate that automated classification methods exhibit limited generalizability across diverse fake news types, exposing a critical weakness in existing detection systems. This underscores the necessity for more resilient and adaptive approaches, particularly in light of the increasing sophistication of LLM-generated content.

Declarations

Authors' Contributions

Renato Moraes Silva: Writing - review & editing, Project administration, Formal analysis, Conceptualization, Methodology, Data curation. **Lucca Baptista Silva Ferraz:** Writing - original draft, Validation, Formal analysis, Data curation. **Fabio Kauê Araujo da Silva:** Writing - original draft, Validation, Formal analysis, Data curation. **Hazem Amamou:** Writing - review & editing, Validation, Methodology, Formal analysis. **Anderson Raymundo Avila:** Writing - review & editing, Validation, Methodology, Formal analysis, Conceptualization.

Competing interests

The authors declare that they have no competing interests.

Funding

We gratefully acknowledge the support provided by São Paulo Research Foundation (FAPESP; grants #2024/17834-6 and #2025/13608-4) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – grant #444933/2024-7). Some researchers involved in this study are also affiliated with the Center for Artificial Intelligence of the University of São Paulo (C4AI – <http://c4ai.inova.usp.br/>), with support by FAPESP (grant #2019/07665-4), the IBM Corporation, and the Ministry of Science, Technology and Innovation, with resources from Law No. 8,248 of October 23, 1991, under the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

Availability of data and materials

The datasets generated and analysed during the current study are available in <https://github.com/renatosvmor/fake-news-llm-ptbr>.

Carbon consumption evaluation

To quantify the environmental impact of our computational experiments, we followed the methodology proposed by Jegham *et al.* [2025]. Specifically, we estimated the carbon emissions by calculating the average energy consumption per token and multiplying it by the Carbon Intensity Factor (CIF) associated with the data center's region. Our calculations yielded an estimated emission of 17.82 kgCO₂e for all experiments.

We adopt an emission factor of 0.000428 gCO₂e per token, based on Jegham *et al.* [2025], who measured average energy use per token for GPT-4o and applied a carbon intensity factor. Since we accessed Sabiá-3 solely via its API and no official information is available about its underlying computational infrastructure, we use this value as a conservative proxy. Given that GPT-4o is likely one of the largest models currently in production, this estimate may overstate the actual emissions.

We limited our estimate to the inference emissions resulting from fake news generation and LLM-based classification pipelines. We do not include the one-time carbon

cost of downstream classification steps (e.g., BERT fine-tuning or RAG retrieval), as these are considered negligible. Prior studies have shown that amortized training emissions and additional per-query overheads are orders of magnitude smaller than the direct inference emissions in comparable settings [Tomlinson *et al.*, 2024].

References

- Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 technical report. *arXiv preprint arXiv:2410.12049*. Available at: <https://arxiv.org/abs/2410.12049>.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., *et al.* (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. DOI: 10.48550/arxiv.2303.08774.
- Aïmeur, E., Amri, S., and Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30. DOI: 10.1007/s13278-023-01028-5.
- Ayoobi, N., Shahriar, S., and Mukherjee, A. (2024). Seeing through ai's lens: Enhancing human skepticism towards llm-generated fake news. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, HT'24, page 1–11, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3648188.3675136.
- Azzimonti, M. and Fernandes, M. (2023). Social media networks, fake news, and polarization. *European Journal of Political Economy*, 76:102256. DOI: 10.1016/j.ejpoleco.2022.102256.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. DOI: 10.1162/tacl_a00051.
- Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. pages 144–152. DOI: 10.1145/130385.130401.
- Breiman, L. (2001). Random forests. *ml*, 45(1):5–32. DOI: 10.1023/A:1010933404324.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, California, USA. Book.
- Cabral, L., Monteiro, J. M., da Silva, F., Wellington, J., Matos, C. L. C., and Mourão, P. J. C. (2021). Fakewhatsapp.br: Nlp and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages. In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1*, pages 63–74. DOI: 10.5220/0010446800630074.
- Chavarro, J., Carvalho, J., Portela, T., and Silva, J. (2023). Faketruebr: Um corpus brasileiro de notícias falsas. In *Anais da XVIII Escola Regional de Banco de Dados*, pages 108–117, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/erbd.2023.229495.
- Cordeiro, P. R. and Pinheiro, V. (2019). Um corpus de notícias falsas do twitter e verificação automática de rumores

- em língua portuguesa. In *Proceedings of the Symposium in Information and Human Language Technology*, pages 219–228. DOI: 10.5753/stil.2021.17796.
- Cortes, C. and Vapnik, V. N. (1995). Support-vector networks. *ml*, 20(3):273–297. DOI: 10.1007/BF00994018.
- DeepSeek-AI (2025). Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*. Available at: <https://arxiv.org/abs/2501.12948>.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423.
- Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. (2025). The faiss library. *arXiv preprint arXiv:2401.08281*. DOI: 10.1109/tbdata.2025.3618474.
- Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., and Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29. DOI: 10.1038/s44159-021-00006-y.
- Faustini, P. and Covões, T. F. (2019). Fake news detection using one-class classification. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems (BRACIS'19)*, pages 592–597, Salvador, BA, Brazil. IEEE. DOI: 10.1109/BRACIS.2019.00109.
- Gallotti, R., Valle, F., Castaldo, N., Sacco, P., and De Domenico, M. (2020). Assessing the risks of ‘infodemics’ in response to covid-19 epidemics. *Nature human behaviour*, 4(12):1285–1293. DOI: 10.1038/s41562-020-00994-6.
- Garcia, G. L., Afonso, L. C., and Papa, J. P. (2022). Fakerecogna: a new brazilian corpus for fake news detection. In *International Conference on Computational Processing of the Portuguese Language*, pages 57–67. Springer. DOI: 10.1007/978-3-030-98305-5_6.
- Garcia, G. L., Paiola, P. H., Jodas, D. S., Sugi, L. A., and Papa, J. P. (2024). Text summarization and temporal learning models applied to portuguese fake news detection in a novel brazilian corpus dataset. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 86–96, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics. Available at: <https://aclanthology.org/2024.propor-1.9>.
- Geurgas, R. and Tessler, L. R. (2024). Automatic detection of fake tweets about the covid-19 vaccine in portuguese. *Social Network Analysis and Mining*, 14(1):55. DOI: 10.1007/s13278-024-01216-x.
- Gôlo, M. P. S., Mori, A. L. V., Oliveira, W. G., Barbosa, J. R., Graciano-Neto, V. V., de Lima, E. A., and Marcacini, R. M. (2024). On the use of large language models to detect brazilian politics fake news. In *Proceedings of the 21st National Meeting on Artificial and Computational Intelligence (ENIAC'2024)*, pages 1–12, Belém, PA, Brazil. Brazilian Computer Society. DOI: 10.5753/eniac.2024.245119.
- Jegham, N., Abdelatti, M., Elmoubarki, L., and Hendawi, A. (2025). How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference. *arXiv preprint arXiv:2505.09598*. Available at: <https://arxiv.org/abs/2505.09598>.
- Jiang, B., Tan, Z., Nirmal, A., and Liu, H. (2024). Disinformation detection: An evolving challenge in the age of llms. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 427–435. SIAM. DOI: 10.1137/1.9781611978032.50.
- Kaliyar, R. K., Goswami, A., and Narang, P. (2021). Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788. DOI: 10.1007/s11042-020-10183-2.
- Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096. DOI: 10.1126/science.aao2998.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.48550/arxiv.2005.11401.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*, pages 3111–3119, Lake Tahoe, Nevada, USA. Curran Associates Inc.. DOI: 10.48550/arXiv.1310.4546.
- Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *13th International Conference on Computational Processing of the Portuguese Language (PROPOR'2018)*, pages 324–334, Canela, Rio Grande do Sul, Brazil. Springer International Publishing. DOI: 10.1007/978-3-319-99722-3_33.
- O'Connor, C. and Weatherall, J. O. (2021). Modeling how false beliefs spread. *The Routledge Handbook of Political Epistemology*, pages 203–213. DOI: 10.4324/9780429326769-25.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830. DOI: 10.5555/1953048.2078195.
- Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R.

- (2023). Sabiá: Portuguese large language models. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland. DOI: 10.48550/arXiv.2304.07880.
- Plikynas, D., Rizgelienė, I., and Korvel, G. (2025). Systematic review of fake news, propaganda, and disinformation: Examining authors, content, and social impact through machine learning. *IEEE Access*, 13:17583–17629. DOI: 10.1109/ACCESS.2025.3530688.
- Rathore, F. A. and Farooq, F. (2020). Information overload and infodemic in the covid-19 pandemic. *J Pak Med Assoc*, 70(5):S162–S165. DOI: 10.5455/jpma.38.
- Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. DOI: 10.18653/v1/d19-1410.
- Sardinha, T. B. (2024). Ai-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1):100083. DOI: 10.1016/j.acorp.2023.100083.
- Saroj, A. and Pal, S. (2020). Use of social media in crisis management: A survey. *International Journal of Disaster Risk Reduction*, 48:101584. DOI: 10.1016/j.ijdr.2020.101584.
- Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36. DOI: 10.48550/arXiv.1708.01967.
- Silva, R. M., de Sales Santos, R. L., Pardo, T. A. S., and Almeida, T. A. (2020). Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:1–48. DOI: 10.1016/j.eswa.2020.113199.
- Silva, R. M., Pires, P. R., and Almeida, T. A. (2023). Incremental learning for fake news detection. *Journal of Information and Data Management*, 13(6). DOI: 10.5753/jidm.2022.2542.
- Sivakumar, S., Videla, L. S., Kumar, T. R., Nagaraj, J., Itnal, S., and Haritha, D. (2020). Review on word2vec word embedding neural net. In *2020 international conference on smart electronics and communication (ICOSEC)*, pages 282–290. IEEE. DOI: 10.1109/icosec49089.2020.9215319.
- Smith, S. T., Kao, E. K., Mackin, E. D., Shah, D. C., Simek, O., and Rubin, D. B. (2021). Automatic detection of influential actors in disinformation networks. *Proceedings of the National Academy of Sciences*, 118(4):e2011216118. DOI: 10.1073/pnas.2011216118.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Proceedings of the 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23*, pages 403–417, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-61377-8_28.
- Su, J., Zhuo, T. Y., Mansurov, J., Wang, D., and Nakov, P. (2023). Fake news detectors are biased against texts generated by large language models. *arXiv preprint*. DOI: 10.48550/arXiv.2309.08674.
- Tomlinson, B., Black, R. W., Patterson, D. J., and Torrance, A. W. (2024). The carbon emissions of writing and illustrating are lower for ai than for humans. *Scientific Reports*, 14. DOI: 10.1038/s41598-024-54271-x.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. Available at: <https://arxiv.org/abs/2302.13971>.
- Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151. DOI: 10.1126/science.aap9559.
- Yu, H.-F., Huang, F.-L., and Lin, C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *ml*, 85(1-2):41–75. DOI: 10.1007/s10994-010-5221-8.
- Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., and Artzi, Y. (2020). Revisiting few-sample bert fine-tuning. *CoRR*, abs/2006.05987. Available at: <https://arxiv.org/abs/2006.05987>.
- Zhou, L., Burgoon, J., Twitchell, D., Qin, T., and Nuna-maker Jr., J. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20(4):139–165. DOI: 10.1080/07421222.2004.11045779.