Using Large Language Models for Identifying Satirical News in Brazilian Portuguese

Gabriela Wick-Pedro
Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
gabiwick@gmail.com

Cássio Faria da Silva
Rede Gonzaga de Ensino Superior (REGES)
cassiofs@gmail.com

Marcio Lima Inácio
Centre for Informatics and Systems of the University of Coimbra (CISUC)
Intelligent Systems Associate Laboratory (LASI)
mlinacio@dei.uc.pt

Oto Araújo Vale and **Helena de Medeiros Caseli**
Federal University of São Carlos (UFSCar)
{otovale, helenacaseli}@ufscar.br

Abstract

Satirical news is featured as texts grounded in actual events or information but which are presented in an exaggerated, humorous, and incongruous manner. An intriguing aspect is that satirical news can be mistaken for authentic by readers who fail to discern the intended humorous and ironic elements that satirical texts seek to convey. In this paper, we investigate if fine-tuned large language models are able to identify satirical news in Brazilian Portuguese. We found out that they can identify satirical news with 78-96% F-measure. Furthermore, we also investigate if they do that based on the same linguistic clues as humans do.

1 Introduction

Satirical news comprises fictional news stories that parody the news genre and encompass a wide range of topics, including social issues, politics, entertainment, sports, and others. Typically, these satirical news pieces are grounded in actual events or information but are presented in an exaggerated, humorous, and incongruous manner with the intention of critiquing or lampooning societal events. Furthermore, it is common for these satirical news items to be widely disseminated online, significantly influencing how individuals perceive their society. They transcend the conventional boundaries of media and are distributed through various channels and formats, ranging from magazines to television programs, websites, and even fictional web characters (Rubin et al., 2016; Ermida, 2012).

An intriguing aspect within this context is that satirical news can be mistaken for authentic by readers who fail to discern the intended humorous and ironic elements that satirical texts seek to convey. This is particularly attributable to the extensive sharing of such satirical content (Wick-Pedro et al., 2020; Santos et al., 2020). Often, this challenge of distinguishing between factual and humorous content arises because satirical news articles may incorporate genuine information and real events into their satirical narratives. This overlap of real and fictional information can perplex the reader.

Another significant aspect when it comes to dealing with satirical content is to understand how sources of satirical news "reinterpret" real events, as satire is often used to critique and convey subjective messages to the public. Therefore, identifying distinctive features that delineate these types of content can provide a strong foundation for distinguishing between satirical news and factual news. Consequently, automatically identifying satirical news can be a challenging task, given that satire can be subtle and often necessitates an understanding of the context and the author's intent (Rubin et al., 2016).

It is important to emphasize the need for a thorough assessment of the reliability of these automatic identification models before concluding that they are suitable for addressing specific challenges, particularly in highly complex tasks, such as fake news detection (Monteiro et al., 2018), humor, irony and sarcasm recognition (Inácio et al., 2023; Van Hee et al., 2016), among others. Therefore, it

becomes essential to question whether we are truly capable of understanding what the machine is learning and whether it is effectively capturing relevant information for the phenomenon under analysis.

In this particular context, we present a study on the recognition of satirical news, with a special focus on Large Language Models (LLMs). In addition to assessing the performance of the fine-tuned LLMs for this task, we also aim to check whether the linguistic elements identified by humans as indicative of satire are the same as those highlighted by the machine. To achieve these goals, we compare human annotations with the results obtained from SHAP (Lundberg and Lee, 2017), a machine learning explainability tool. It is worth noting that this work was entirely conducted for Brazilian Portuguese, a language considerably less developed in this task compared to languages like English.

Thus, this paper aims to answer two research questions:

RQ1 How well can fine-tuned LLMs identify satire?

RQ2 Is the knowledge that the machine uses to make such identification the same as that considered by humans?

The experiments and results discussed in this paper are all publicly available at `https://github.com/LALIC-UFSCar/satire-recognition`.

This paper is organized as follows. Section 2 describes some important work related to ours regarding satire identification and machine learning explainability. The methodology adopted for our experiments, including the corpus, the pre-trained LLMs, and the explainability tool, is described in Section 3. Section 4 brings the results that helped us to answer our research questions. Finally, Section 5 finishes this paper with some conclusions and proposals for future work.

2 Related Work

In this section, we briefly describe some important work related to ours regarding satire identification (2.1) and machine learning explainability (2.2).

2.1 Satire Identification

As previously mentioned, satire identification is a challenging task since satirical texts may incorporate genuine information and real events, and this overlap of real and fictional information can perplex the reader. Indeed, satirical news can turn into fake news by leading to deception when the satire is not recognized in its content. The use of ML and LLMs techniques has had a substantial impact on the identification and classification of fake news (Fischer et al., 2022; Low et al., 2022). Previous studies on fake news detection primarily relied on the analysis of linguistic features to generate relevant information (Silva et al., 2020; Alghamdi et al., 2022). Therefore, similar to the approach used for fake news and deceptive content, it is possible to apply methods to automatically identify satirical news (De Sarkar et al., 2018; Horvitz et al., 2020; Ionescu and Chifu, 2021). An alternative involves using ML and LLMs to analyze the news content, taking into account words or expressions that may suggest the satirical nature of the news.

Burfoot and Baldwin (2009) pioneered satire classification using SVMs with lexical and semantic features, focusing on headline attributes, offensive language, slang, and semantic analysis. They employed Named Entity Recognition (NER) for semantic validity. SVMs outperformed the baseline, especially when incorporating elements such as titles, puns, and profanity. The inclusion of validity features resulted in the highest F-score of 79.8%, which was statistically significant, but the additional gains were negligible due to the scarcity of satire cases. Despite a lower recall of 50%, the classifiers effectively identified satire, even in subtle articles.

Horvitz et al. (2020) introduced an innovative approach to satire analysis, creating a dataset of satirical headlines in English paired with factual context. They utilized transformer-based models, including BertSum (Liu, 2019) and BERT (Devlin et al., 2019), to generate satirical headlines. To accomplish this, the authors employed three primary fine-tuning schemes, resulting in the creation of three distinct context-based models: E-Context (which includes an encoder and decoder trained with specific learning rates), A-Context (involving a network trained on preprocessed contexts), and D-Context (wherein the decoder and encoder were trained with varying learning rates). As a result, the Decoder-Weighted-Context (D-Context) model attained the highest Funny rating[1] among all models at 9.4%, followed by the E-Context model at

[1]Human annotators were employed to evaluate the performance of different models in the satire generation task, answering the following questions: (1) Is the headline coherent? (2) Does the headline sound like The Onion? and (3) Is the headline funny?.

8.7%.

In languages other than English, Ionescu and Chifu (2021) focused on satire detection in a multi-source context in the French language, conducting a comparison between shallow and deep approaches that depended on low-level features and CamemBERT embeddings (Martin et al., 2020). Consequently, the authors observed that the CamemBERT model, based on embeddings, achieved superior results when dealing with complete true news. Meanwhile, the model relying on characters and n-grams demonstrated superior performance in the more challenging task of headline satire detection, attaining a maximum accuracy rate of 74.07%. For Portuguese, Carvalho et al. (2020) conducted a study on detecting irony in satirical headlines and discovered that the extraordinary nature of these headlines arises from the combination of terms from different conceptual domains. They noted that simple word-based linear classifiers are effective in distinguishing between fictional and real headlines, achieving an average F-measure of 85%. Furthermore, incorporating features to identify contrasts beyond the trained domain led to significant improvements, resulting in an F-measure of 91%. Additionally, in the realm of Portuguese, there are other noteworthy initiatives related to our work, such as the detection of irony in tweets (Vanin et al., 2013; Wick-Pedro and Vale, 2020) and the recognition of one-line jokes (Gonçalo Oliveira et al., 2020; Inácio et al., 2023).

2.2 Machine Learning Explainability

Modern Machine Learning (ML) systems generally lack interpretability, i.e. it is virtually impossible to understand qualitatively how their prediction is obtained. This aspect of the models raises questions about whether they are leveraging their decisions on meaningful information from the data (Ribeiro et al., 2016).

Additionally to traditional explainability methods — such as using inherently interpretable models (Ustun and Rudin, 2016) or evaluating attention weights (Xu et al., 2015) — researchers started developing approaches to obtain model-agnostic explanations of single input examples, as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). In general, such methods work by perturbating input units (features, tokens, pixels, etc.) and calculating the degree to which they alter the model's final prediction. Since they provide local explanations only, research usually relies upon visualization techniques or manual analysis of a range of examples to pursue global conclusions about the model's performance.

3 Methodology

As previously mentioned, our main focus in this paper is to not only evaluate the performance of fine-tuned large language models in the task of Satire Recognition but also assess the linguistic knowledge that the models resort to when doing such classification. Therefore, our methodology consists of three specific phases: defining the corpus and data to be used, training and evaluating the models, and, finally, using ML explainability techniques to understand the models' decisions.

3.1 Corpus

In this paper, we used a subset of a corpus of satirical news automatically extracted from Sensacionalista[2], a Brazilian website of satirical and humorous news. For the experiments presented in this paper, we selected 150 satirical news (Satirical) and their counter-part non-satirical (Real) ones. The collection process involved a manual approach, with an initial focus on keywords identified in the satirical news, followed by a manual search for each corresponding real article. For additional corpus details, please refer to Table 1.

News	Tokens	Types	Sentences
Satirical News	22,963	4.843	1,212
Real News	107,133	11,304	5,721

Table 1: Corpus characteristics

From a linguistic perspective, we understand that the number of words, sentences, and lexical diversity can serve as a distinguishing characteristic among different types of content. This becomes evident in the marked structural differences between real and satirical news, for instance, notably in the quantity and complexity of sentences employed.

In Table 2 we present an example of excerpts from a Satirical news[3] and its Real counterpart[4].

[2] https://www.sensacionalista.com.br/

[3] English version: *Marcela's dog threw itself into the lake because it had to live with Temer. First Lady Marcela Temer went into a pond at the Alvorada Palace two weeks ago, fully clothed, to rescue her dog Picolly. According to veterinarians at the Planalto, the dog had thrown itself into the lake because it was depressed about having to live with President Michel Temer. "It's not easy for him to live in the same house as*

Satirical news	Cachorro de Marcela se jogou no lago por ter que conviver com Temer. A primeira-dama Marcela Temer entrou de roupa e tudo há duas semanas em uma lagoa no Palácio da Alvorada, para resgatar seu cachorro Picolly. Segundo veterinários do Planalto, o cachorro teria se jogado no lago pois estava deprimido por ter que conviver com o presidente Michel Temer. "Não é fácil para ele viver na mesma casa que Temer."
Real news	Marcela Temer pula em lago para salvar seu cachorro e afasta segurança que não ajudou. A primeira-dama Marcela Temer pulou em um lago do Palácio da Alvorada, em Brasília, para resgatar seu cachorro, Picoly. O animal, da raça jack russell, se viu em apuros após se jogar nas águas do jardim do palácio e não conseguir sair. Assustada, a mulher do presidente Michel Temer ainda pediu auxílio a uma agente de segurança.

Table 2: Example of excerpts of around 400 characters of a Satirical news and its Real counterpart

The 150 headlines of this subset were annotated by three annotators. They were tasked with identifying which part of the headline contained satire in the sentence (delimited by <sat> and </sat> tags). Table 3 shows an example[5] of the annotations performed by them.

3.2 Classification Models

To answer our first research question, experiments were conducted by fine-tuning pre-trained transformer models, specifically BERTimbau[6] (Souza et al., 2020), RobertaTwitterBR[7], and Albertina PT-BR[8] (Rodrigues et al., 2023), all of them neural models for the Portuguese language. BERTimbau was pre-trained on BrWaC (Brazilian Portuguese Web as Corpus) (Wagner Filho et al., 2018), a substantial Portuguese corpus consisting of 2.7 billion tokens from 3.5 million documents gathered by web crawling across various websites. This corpus, as suggested by the authors, ensures a broad diversity of topics. RobertaTwitterBR was trained on a dataset of approximately 7 million Portuguese tweets. Albertina PT-BR, derived from DeBERTa (He et al., 2020), was also pre-trained using the brWaC dataset.

In this paper, we investigated how ML can be applied to classify satirical news in the domain of Brazilian politics. For this, different classifiers, with different hyperparameters, were tested in our corpus (see Section 3.1). From the 300 pairs of satirical and non-satirical news, 240 of them were used for fine-tuning the models and the remaining 60 were used for testing.

The news articles in the corpus span a variety of genres, including satire and non-satire, with varying lengths. The length of these articles varies from 69 characters to almost 20 thousand characters. To ensure consistency and standardization, we chose to truncate the news used for training and validation to a maximum of 400 characters. In this procedure, the news articles were truncated, focusing only on the first 400 characters, which include the headlines. This decision was made after finding that only two articles were less than 400 characters in length, making this limit an appropriate choice to maintain uniformity in the data set. Moreover, according to Table 1, real news typically exhibits greater length compared to satirical news pieces. The news articles used for testing were not truncated.

Figure 1 depicts the methodology, which unfolded in two distinct stages: (i) the fine-tuning of the neural model, and (ii) the model's evaluation on the test dataset, resulting in the generation of standard evaluation measures. It is worth mentioning that the same training and testing partitions were used, in a stratified manner, in all experiments.

The experiments were conducted on Google Colab Pro, using TPU, Tesla T4 GPU, V100-SXM2-16GB and NVIDIA A100-SXM4-40GB,

Temer."

[4]English version: *Marcela Temer jumps into a lake to save her dog and pushes away the security guard who didn't help. First Lady Marcela Temer jumped into a pond at the Alvorada Palace in Brasília to rescue her dog, Picoly. The Jack Russell terrier found itself in trouble after jumping into the waters of the palace garden and being unable to get out. Alarmed, the wife of President Michel Temer even asked for help from a security agent.*

[5]English version: *Temer has 5% and MDB confuses it with a bribe.*

[6]Available at: https://github.com/neuralmind-ai/portuguese-bert

[7]Available at: https://huggingface.co/verissimomanoel/RobertaTwitterBR

[8]Available at: https://huggingface.co/PORTULAN/albertina-900m-portuguese-ptbr-encoder-brwac

Original headline: Temer tem 5% e MDB confunde com propina.	
Annotator A	Temer tem 5% e MDB **<sat>confunde com propina</sat>**
Annotator B	Temer tem **<sat>5% e MDB confunde com propina</sat>**
Annotator C	Temer tem 5% e MDB confunde com **<sat>propina</sat>**

Table 3: Example of annotations for a satirical headline

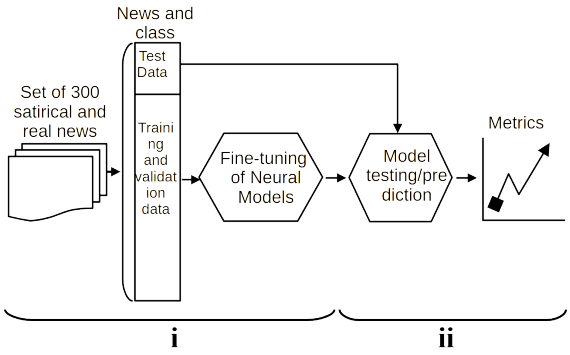

Figure 1: The proposed experimental configuration was divided into two steps: i. fine-tuning of neural model and ii. model's evaluation.

with 35.2GB of available RAM.

3.3 Model Explainability

A main concern we have regarding the ML models is if they are in fact learning the intended phenomenon, i.e. what is the information that the machine uses to reach its final classification decision?

To answer this question, we take advantage of SHAP[9] (Lundberg and Lee, 2017), which provides model-agnostic local explanations for ML models. Given a model — in our case, a transformer — and an input (a text), SHAP masks out different tokens to assess how they impact the final prediction scores, testing different combinations of the mask to account for interactions between features. Finally, SHAP returns a base value (the class probabilities when every token is masked) and additive values for each token in the input, representing the contribution of that specific token to the final prediction score[10].

SHAP values can be positive (when the token contributes to the class in question) or negative (when the token points out to another class). Since our classification task is binary, SHAP values for each class (satiric and real) are necessarily inverse. As our main focus is to understand if the model is capturing satire, we did our analyses for the satiric class.

Seeing that SHAP provides only explanations for single instances, we developed a method to better analyze if the model is associating the same text passages to satire as a human would. To this extent, we take advantage of the manual annotation of satiric news headlines, described in subsection 3.1. Since the corpus has an annotation of the exact excerpts in which humans consider the satire to be, we want to compare if SHAP values for tokens inside such passages are higher than for those tokens outside of the annotation tags, meaning that the model is associating the same pieces of information to the presence of satire as humans have done.

For our analysis, since the contribution of each token for the instance classification is different for each text, we first normalize the values according to Equation 1. Given a text of n tokens with SHAP values $\{s_1, \cdots, s_i, \cdots, s_n\} = S$, each token with a positive SHAP value has its value normalized to s'_i, which indicates how much this specific token contributes to the prediction score of the class of satire.

$$s'_i = \frac{s_i}{\sum_{s_j \in S, s_j > 0} s_j}, \forall s_i > 0 \in S \quad (1)$$

Only for positive SHAP values

Sum of positive SHAP values

Besides, we also used the SHAP values in a manual analysis as explained in Section 4.2.

4 Results

In this section, we present the results that helped us to answer our research questions regarding the performance of fine-tuned models (4.1), the explainability of their classification (4.2) and our manual analysis of some instances (4.3).

4.1 Classification Performance

In terms of the quantitative measures usually applied in the evaluation of computational models

[9] Available at: https://github.com/shap/shap

[10] In other words, the final class probability is equal to the base value summed with the SHAP values for each token.

— Accuracy, Precision, Recall, and F-measure — based on the values presented in Table 4, we concluded that the neural model generated by the fine-tuning of Albertina obtained the best values: 96.67% for all measures[11].

Table 5 and Figure 2 present the detailed results of the best-performing model obtained with the fine-tuning of Albertina. This model had excellent performance, correctly predicting 58 of the 60 instances present in the test corpus. As a result, the models achieved an accuracy of 96.67%.

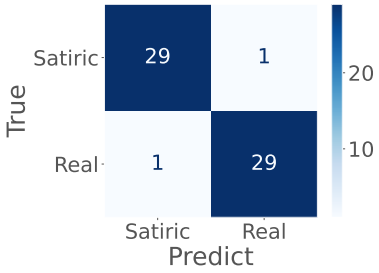

Figure 2: Confusion matrix of the model obtained with fine-tuning of Albertina.

Furthermore, the model presents a very consistent and reliable performance, with precision, recall, and F-measure values of 96.67% for both categories, indicating that it is effective in classifying satirical and real news. This is particularly relevant since identifying satirical news is crucial to preventing the spread of misleading information.

Thus, although the test corpus sample is small, these results provide evidence of the model's excellent assertiveness in predicting satirical news in the domain of Brazilian politics.

In addition, we carried out experiments with news headlines. The results (Table 6) obtained revealed significant differences in the performance of these models compared to previous results obtained when analyzing the same full news stories. Albertina achieved the best results, with an accuracy of 78.33% and a precision of 79.14%. RobertaTwitterBR achieved an accuracy of 71.67% and a precision of 76.68%, while BERTimbau, with an accuracy of 68.33%, was slightly behind in terms of precision (68.86%). The F-measure, which combines precision and coverage, corroborated the superiority of the model trained with Albertina with a score of 78.18%. This lower performance of the model fine-tuned with the full texts and tested in headlines can be explained, in part, by the very different syntactic structure of the headlines compared with the full texts.

The main motivation for the news headline-only experiments was based on evidence that headlines often contain linguistic cues that can indicate whether the text is satirical or not. Additionally, news headlines are the first or, in some cases, only piece of news that readers consume, which makes them especially important for identifying satirical content. Therefore, with the headline-only experiments, we were able to explore language models in identifying satirical content in limited, highly condensed texts. Furthermore, this may also have practical implications, as the ability to identify satire based on headlines alone may be useful in scenarios where readers have limited access to the full news content.

4.2 Explainability Results

As we mentioned in Section 3.3, we want to see if the normalized SHAP values inside manually annotated passages in satiric headlines are higher than the ones outside such excerpts. An overview analysis can be seen in Figure 3, in which we present, for each model, the general distribution of the normalized SHAP values of text passages. In the graph, values under "Inside tags" correspond to the total contribution of the annotated text passage, i.e. the sum of the normalized SHAP values of all tokens identified by at least one annotator. Conversely, values under "Outside tags" represent the contribution of tokens outside such tags.

In Figure 3, we can observe that generally, the models consider roughly at the same degree text passages inside and outside the annotation tags (medians revolve around 50%). This shows that the information the model uses does not match exactly with human perceptions of satiric content, although it uses the same knowledge to some extent.

These observations highlight that, even though the models often identify correctly satiric instances, their decisions sometimes rely on text passages that a human would not consider as the main point of the satire. This could mean that the models do not identify satire but rather other related or unrelated text characteristics. On the other hand, the machine might have identified subtle satiric characteristics that human eyes were not able to perceive at first, which requires a more detailed intrinsic analysis of

[11]The best results were achieved with the following optimized hyperparameters: number of epochs= 20; *batch size*= 8; *early stop*= 2; *learning rate*=1e-5. The same hyperparameters were used for Albertina, BERTimbau, and RobertaTwitterBR. To ensure the reliability of the results, we ran the LLMs with different training-test partitions of the data.

	Accuracy	Precision	Recall	F-measure
Albertina	96.67%	96.67%	96.67%	96.67%
RobertaTwitterBR	95.00%	95.45%	95.00%	94.99%
BERTimbau	85.00%	87.02%	85.00%	84.79%

Table 4: Values of evaluation measures obtained in neural models Albertina, RobertaTwitterBR and BERTimbau.

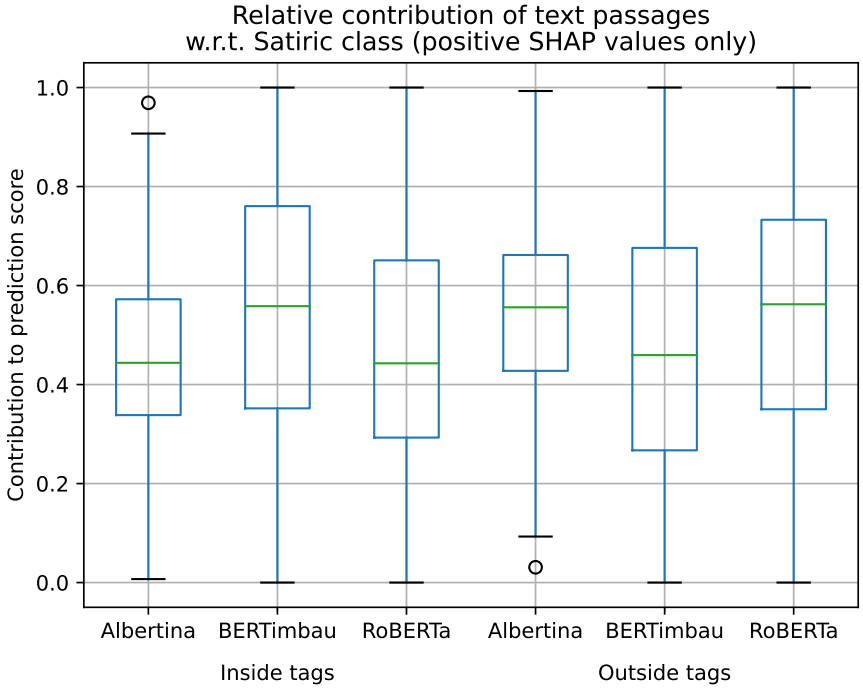

Figure 3: Extrinsic analysis of relative SHAP values of text passages inside and outside manual annotations.

	Precision	Recall	F-measure
Satirical	96.67%	96.67%	96.67%
Real	96.67%	96.67%	96.67%

Table 5: Detailed prediction results returned by the model obtained by fine-tuning Albertina.

the results to attest.

4.3 Manual analysis

In Table 7 we show two examples of instances correctly classified by the fine-tuned Albertina model. The Satirical news is the same as shown in Table 2. The SHAP scores for tokens that had a positive influence on the class are shown subscribed. Tokens with a score of at least 0.02 (empirically defined value) are highlighted in bold. It is worth noting that SHAP values are scattered across a longer text, which can make the values per individual word seem small. However, when you take into account the total contribution of all the words, the overall impact is still substantial. For instance, in the first example, the combined contribution of all the tokens is 0.598.

As we can notice, the words that most influenced the classification of the satirical news were: cachorro (dog), roupa_e_tudo_há (fully clothed), para (to) resgatar (rescue), and viver (live). These words indeed bring clues for the satirical feature of this text. On the other hand, the words that most influenced the classification of the real news were: quarta (Wednesday), confirmou (confirmed), que_uma (that a), Ipanema, na (in the), Zona (Zone), Sul_do (South of), Rio, na_manhã_desta quarta-feira_(15) (on Wednesday morning (15)) está (is)[12]. It seems to be that the words that most influenced the classification of the real news are

[12]English version of the real news: *Dead whale strands on Ipanema Beach in Rio. A biologist from Uerj confirmed the death of the animal, which appeared on the shore of the South Zone. The area has been isolated for removal, which will be done Wednesday night. The animal will be taken by Comlurb to the sanitary landfill in Seropédica. A biologist confirmed that a stranded whale on Ipanema Beach in the South Zone of Rio, on Wednesday morning (15), is dead.*

	Accuracy	Precision	Recall	F-measure
BERTimbau	68.33%	68.86%	68.33%	68.11%
RobertaTwitterBR	71.67%	76.68%	71.67%	70.27%
Albertina	78.33%	79.14%	78.33%	78.18%

Table 6: Values of evaluation measures obtained on headlines.

those that indicate facts such as dates and places.

It is important to emphasize that the SHAP visualization has a clear tendency to group sequential tokens that have a high level of interaction. This leads to the creation of chunks, such as "roupa_e_tudo_há" or "Sul_do," where the SHAP value corresponds to the sum of individual parts. However, these chunks, obtained through automatic hierarchical clustering methods, do not necessarily correspond to sensible linguistic chunks as in constituency parsing. Therefore, it is crucial to keep in mind these aspects of the SHAP visualization when interpreting the results.

We also took a look at the two test instances that Albertina's fine-tuned model classified incorrectly. They are presented in Appendix A.

Finally, in Table 8 we show the fine-tuned Albertina's SHAP scores for the headlines with a minimum threshold of 0.6[13] (empirically defined). Each headline is also accompanied by the human-annotated version as indicated by tags <X> ... </X> for annotators X. As we can notice, for the first two headlines there are intersections between human annotations and the best SHAP scores: "jogou" and "lago" in the first example and "porque", "mãe" and "eles" in the second one. However, in the third example, there is no intersection. For this last example, annotator C also didn't highlight any token as indicative of satire.

5 Conclusion

In this paper, we presented a work on the identification of satirical news in Brazilian Portuguese by fine-tuning and evaluating the performance of different LLMs. When classifying news texts, Albertina PT-BR (Rodrigues et al., 2023) had the best results, reaching 96.67% F-measure. Meanwhile, when evaluating on only news headlines, Albertina obtained 78.18% F-measure. From these values we can conclude that the performance of the best fine-tuned LLM for satire identification lies between 78 to 96% F-measure. These values allow us to answer our first research question pointing out that fine-tuned LLMs presented very promising performance on the task of satire identification.

Besides, we also provide an ML explainability analysis using a tool named SHAP (Lundberg and Lee, 2017) and compared its results with manual annotations. An overview analysis showed that the models consider pieces of information that humans associated with satire to roughly the same extent as those not used by the human annotators. Thus, we conclude that the knowledge taken into account by the fine-tuned model when doing satire identification is not always the same as considered by humans, answering our second research question. On the other hand, we highlight that these specific pieces of information may be unrelated to the problem in question (satire identification), bringing up two scenarios: (i) the model learned a different but correlated task (e.g. to identify the Sensacionalista's writing style), or (ii) the model did not learn anything and the results are due to statistical fluctuation. A third scenario is possible in which (iii) the model was able to capture further details that humans were not able to perceive at first during annotation. Further detailed analyses of these results and a thorough review of the corpus and linguistic theories about satire can be of great value to attest to our observations and decide which is the best-suiting scenario we observed in this paper.

In our manual analysis of the explainability results for the Albertina's fine-tuned model we were able to find interesting clues to classify satirical and real news, but an in-depth linguistic investigation is needed to allow some robust conclusions to be drawn. This is one of our future steps in this research.

As future work we also highlight two ways that would bring greater benefit to the proposals presented here: (i) additions of updated news from the Sensacionalista portal and other satirical news sources; and (ii) include other domains in the news

[13] As headline sizes are smaller, tokens scores tend to be higher, justifying the increase in our empirically defined threshold.

Satiric	**Cachorro**$_{0.053}$ de$_{0.014}$ Marcela$_{0.01}$ se$_{0.003}$ jogou$_{0.01}$ no$_{0.006}$ lago$_{0.009}$ por$_{0.007}$ ter$_{0.007}$ que$_{0.004}$ conviver_com_Temer$_{0.005}$. A primeira-dama Marcela Temer entrou de **roupa_e_tudo_há**$_{0.021}$ duas_semanas_em_uma_lagoa$_{0.017}$ no_Palácio_da$_{0.01}$ Alvorada$_{0.008}$, **para**$_{0.031}$ **resgatar**$_{0.027}$ seu$_{0.016}$ **cachorro**$_{0.029}$ Picolly$_{0.012}$. Segundo_veterinários$_{0.015}$ do Planalto$_{0.003}$, o_cachorro$_{0.01}$ teria_se$_{0.014}$ jogado_no$_{0.016}$ lago_pois$_{0.01}$ estava$_{0.011}$ deprimido$_{0.013}$ por$_{0.01}$ ter$_{0.007}$ que$_{0.009}$ conviver$_{0.011}$ com_o$_{0.014}$ presidente$_{0.009}$ Michel_Temer$_{0.009}$. “$_{0.044}$ Não é fácil$_{0.015}$ para$_{0.008}$ ele$_{0.014}$ **viver**$_{0.038}$ na mesma casa$_{0.019}$
Real	Baleia morta encalha na_Praia_de Ipanema, no Rio. Biólogo da Uerj confirmou morte do_animal, que apareceu na orla da_Zona_Sul . Área$_{0.012}$ foi$_{0.007}$ isolada$_{0.019}$ para_a$_{0.017}$ retirada que$_{0.003}$ será$_{0.007}$ feita$_{0.006}$ na$_{0.003}$ noite$_{0.012}$ desta$_{0.01}$ **quarta**$_{0.026}$ ·$_{0.008}$ Animal$_{0.009}$ será levado pela Comlurb_para$_{0.009}$ aterro sanitário_de$_{0.013}$ Seropédica$_{0.006}$ ·$_{0.026}$ Um$_{0.006}$ biólogo$_{0.007}$ **confirmou**$_{0.022}$ **que_uma**$_{0.024}$ baleia encalhada$_{0.011}$ na Praia de **Ipanema**$_{0.02}$,$_{0.012}$ **na**$_{0.022}$ **Zona**$_{0.023}$ **Sul_do**$_{0.039}$ **Rio**$_{0.021}$,$_{0.016}$ **na_manhã_desta**$_{0.036}$ **quarta-feira_(15),**$_{0.039}$ **está**$_{0.032}$

Table 7: Examples of instances correctly classified by the fine-tuned Albertina model. The SHAP scores for tokens that had a positive influence for the class are shown subscribed. Tokens with a score of at least 0.02 (empirically defined value) are highlighted in bold.

H	Cachorro de Marcela <A>se jogou no lago</A> <B>por ter que <A><C>conviver com Temer</C></A></B>
A	**Cachorro**$_{0.154}$ de$_{0.054}$ Marcela$_{0.016}$ se$_{0.046}$ **jogou**$_{0.065}$ no$_{0.03}$ **lago**$_{0.069}$ por$_{0.048}$ ter$_{0.026}$ que$_{0.053}$ conviver$_{0.059}$ com Temer$_{0.004}$
H	PSDB homenageou Gilmar Mendes ontem <B>porque ele é <A>uma <C>mãe</C> para eles</A></B>
A	PSDB homenageou$_{0.049}$ Gilmar$_{0.019}$ Mendes ontem$_{0.026}$ **porque**$_{0.112}$ ele$_{0.053}$ é$_{0.023}$ uma$_{0.036}$ **mãe**$_{0.099}$ para$_{0.001}$ **eles**$_{0.106}$
H	63 viagens de Rodrigo Maia pela FAB desencadeiam a <A><B>Operação Lava-Jatinho</B></A>
A	63$_{0.052}$ viagens de Rodrigo$_{0.008}$ **Maia**$_{0.066}$ pela **FAB**$_{0.081}$ **desencadeiam**$_{0.063}$ a Operação Lava-Jatinho

Table 8: Examples of headlines annotated by humans (H) – annotators A, B, and C – and the SHAP scores for tokens that had a positive influence on the Satirical classification by Albertina's fine-tuned model (A). Tokens with a score of at least 0.06 (empirically defined value) are highlighted in bold.

corpus, such as behavior, entertainment, sports, world, and country.

Acknowledgements

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44. Additionally, this work was partially funded by national Portuguese funds through the FCT – Foundation for Science and Technology, I.P. (grant number UI/BD/153496/2022), within the scope of the project CISUC (UID/CEC/00326/2020) and by the European Social Fund, through the Regional Operational Program Centro 2020. The work presented in this paper also meets some goals of the FAPESP Grant #2022/03090-0. Finally, we also thank the Graduate Program in Computer Science (PPGCC) and Linguistics (PPGL) from UFSCar.

References

Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. 2022. A comparative study of machine learning and deep learning techniques for fake news detection. *Information*, 13(12).

Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164, Suntec, Singapore. Association for Computational Linguistics.

Paula Carvalho, Bruno Martins, Hugo Rosa, Silvio Amir, Jorge Baptista, and Mário J. Silva. 2020. Situational irony in farcical news headlines. In *Lecture Notes in Computer Science*, pages 65–75. Springer International Publishing.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Isabel Ermida. 2012. *News satire in the press: Linguistic construction of humour in spoof news articles*, pages 185–210. Cambridge Scholars Publishing, Newcastle.

Marcelo Fischer, Rejwanul Haque, Paul Stynes, and Pramod Pathak. 2022. Identifying fake news in brazilian portuguese. In *Natural Language Processing and Information Systems*, pages 111–118, Cham. Springer International Publishing.

Hugo Gonçalo Oliveira, André Clemêncio, and Ana Alves. 2020. Corpora and baselines for humour recognition in Portuguese. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1278–1285, Marseille, France. European Language Resources Association.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Zachary Horvitz, Nam Do, and Michael L. Littman. 2020. Context-driven satirical news generation. In *Proceedings of the Second Workshop on Figurative Language Processing*, page 40–50, Online. Association for Computational Linguistics.

Marcio Inácio, Gabriela Wick-Pedro, and Hugo Gonçalo Oliveira. 2023. What do humor classifiers learn? an attempt to explain humor recognition models. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 88–98, Dubrovnik, Croatia. Association for Computational Linguistics.

Radu Tudor Ionescu and Adrian Gabriel Chifu. 2021. Fresada: A french satire data set for cross-domain satire detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*, page 1–8, Shenzhen, China. IEEE.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Jwen Fai Low, Benjamin C.M. Fung, Farkhund Iqbal, and Shih-Chia Huang. 2022. Distinguishing between fake news and satire with transformers. *Expert Systems with Applications*, 187:115824.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Rafael A Monteiro, Roney LS Santos, Thiago AS Pardo, Tiago A De Almeida, Evandro ES Ruiz, and Oto A Vale. 2018. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 324–334. Springer.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco California USA. ACM.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt-*.

Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

Roney Lira de Sales Santos, Gabriela Wick-Pedro, Sidney Evaldo Leal, Oto Araujo Vale, Thiago Alexandre Salgueiro Pardo, Kalina Bontcheva, and Carolina Evaristo Scarton. 2020. Measuring the impact of readability features in fake news detection. In *Proceedings of the 12th language resources and evaluation conference*, pages 1404—-1413.

Renato M. Silva, Roney L.S. Santos, Tiago A. Almeida, and Thiago A.S. Pardo. 2020. Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:113199.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Berk Ustun and Cynthia Rudin. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2016. Monday mornings are my fave :) #not exploring the automatic recognition of irony in English tweets. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2730–2739, Osaka, Japan. The COLING 2016 Organizing Committee.

Aline A Vanin, Larissa A Freitas, Renata Vieira, and Marco Bochernitsan. 2013. Some clues on irony detection in tweets. In *Proceedings of the 22nd international conference on world wide web*, pages 635–636.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gabriela Wick-Pedro, Roney LS Santos, Oto A Vale, Thiago AS Pardo, Kalina Bontcheva, and Carolina Scarton. 2020. Linguistic analysis model for monitoring user reaction on satirical news for brazilian portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–320. Springer.

Gabriela Wick-Pedro and Oto Araújo Vale. 2020. Comentcorpus: descrição e análise de ironia em um corpus de opinião para o português do Brasil. *Cadernos de Linguística*, 1(2):01–15.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

A Appendix

In Table 9 we show these instances. In these cases we were not able to find a pattern that could explain the misleading of the classification model.

Satiric news classified as Real	63$_{0.005}$ **viagens**$_{0.03}$ de$_{0.007}$ Rodrigo Maia$_{0.017}$ pela$_{0.01}$ FAB$_{0.01}$ **desencadeiam**$_{0.037}$ a$_{0.011}$ **Operação_Lava-**$_{0.045}$ Jatinho$_{0.019}$.$_{0.025}$ **O_pré-**$_{0.031}$ candidato$_{0.009}$ do_DEM$_{0.018}$ à$_{0.007}$ **presidência_da_república**$_{0.031}$ será investigado na recém inaurugrada$_{0.012}$ Operação$_{0.006}$ Lava$_{0.008}$ -$_{0.011}$ Jatinho .$_{0.098}$ Levantamento do Estado de São$_{0.004}$ Paulo Rodrigo Maia viajou_63$_{0.005}$ vezes pela Força Aérea Brasileira$_{0.003}$ para compromissos pelo país, a_maioria$_{0.005}$ deles_no_Rio$_{0.005}$ de_Janeiro$_{0.009}$.$_{0.032}$ O$_{0.006}$ ministro$_{0.003}$ da$_{0.002}$ Fazenda$_{0.004}$,$_{0.003}$ Henrique Meireiles ,$_{0.002}$ ainda
Real news classified as Satiric	**Aécio**$_{0.048}$ **Neves:**$_{0.02}$ Sua$_{0.008}$ excelência$_{0.008}$, o fato . Fui_ingênuo,$_{0.019}$ cometi_erros$_{0.005}$ e_me_penitencio$_{0.018}$ por_eles$_{0.009}$,$_{0.004}$ **mas_não**$_{0.022}$ cometi$_{0.012}$ **nenhuma**$_{0.027}$ ilegalidade$_{0.013}$.$_{0.006}$ **A_narrativa_que_se_impõe**$_{0.031}$ como_um$_{0.01}$ tsunami_no_país_tende_a$_{0.019}$ considerar$_{0.013}$,$_{0.006}$ de$_{0.003}$ **antemão,**$_{0.021}$ todos_os$_{0.002}$ políticos culpados$_{0.01}$. **Fragmentos**$_{0.032}$ de$_{0.015}$ **imagens_e**$_{0.037}$ **manchetes**$_{0.026}$ **repetidos**$_{0.02}$ **à**$_{0.02}$ exaustão$_{0.011}$ definem **percepções**$_{0.042}$. Vivemos o tempo da opinião muitas vezes **desvinculada**$_{0.05}$ **da_informação**$_{0.043}$.$_{0.021}$ Sou **alvo**$_{0.021}$ de

Table 9: Instances incorrectly classified by the fine-tuned Albertina model. The SHAP scores for tokens that had a positive influence for the class are shown subscribed. Tokens with a score of at least 0.02 (empirically defined value) are highlighted in bold.