What do Humor Classifiers Learn? An Attempt to Explain Humor Recognition Models

Marcio Lima Inácio and **Hugo Gonçalo Oliveira**
CISUC - University of Coimbra
Department of Informatics Engineering
Polo II, Pinhal de Marrocos, 3030-290
Coimbra, Portugal
{mlinacio, hroliv}@dei.uc.pt

Gabriela Wick-Pedro
Federal University of São Carlos
Departamento de Letras
Rod. Washington Luís, 235, 13965-905
São Carlos, Brazil
gwpedro@estudante.ufscar.br

Abstract

Towards computational systems capable of dealing with complex and general linguistic phenomena, it is essential to understand figurative language, which verbal humor is an instance of. This paper reports state-of-the-art results for Humor Recognition in Portuguese, specifically, an F1-score of 99.6% with a BERT-based classifier. However, following the surprising high performance in such a challenging task, we further analyzed what was actually learned by the classifiers. Our main conclusions were that classifiers based on content-features achieve the best performance, but rely mostly on stylistic aspects of the text, not necessarily related to humor, such as punctuation and question words. On the other hand, for humor-related features, we identified some important aspects, such as the presence of named entities, ambiguity and incongruity.

1 Introduction

As part of usual human language, dealing with complex deep linguistic knowledge, such as figurative language, is an important element of research on Natural Language Processing (NLP). Verbal humor is a large instance of figurative language, whose understanding and generation are crucial for language fluency and the comprehension of deeper nuances of language (Tagnin, 2005).

Additionally, computational systems capable of processing humor might give other fields of research (e.g. Linguistics, Psychology, Philosophy, to name a few) insights about how this phenomenon works and how it is conceived through language.

Regarding such computational models, we must always call in question their trustworthiness before drawing any conclusion that they are suitable to solve any specific problem, especially in tasks deemed as extremely complex, such as Humor Recognition, Fake News Detection, Irony Recognition, and the like (Ribeiro et al., 2016; Monteiro et al., 2018). Thus, it is essential to question if it is possible to really understand what the machine is learning and if it is actually capturing information relevant to the phenomenon being studied. In this way, we can find flaws with the methods or resources used, which drives to further research on the subject to develop better models.

Within this context, we present a study on Humor Recognition with a special focus on identifying which features and pieces of information are mostly used by supervised Machine Learning (ML) models for this task, including classical ML classification algorithms and deep learning Large Language Models (LLMs). We further highlight that the entirety of this work was made for the Portuguese Language, much more underdeveloped on this task when compared to languages like English.

Towards our goal, we first replicated the current state-of-the-art methods for Humor Recognition in Portuguese (Gonçalo Oliveira et al., 2020). In addition, we fine-tuned a BERT model pretrained for Portuguese (Souza et al., 2020) for the same task. Results were further analyzed with SHAP (Lundberg and Lee, 2017), a tool for Machine Learning explainability. SHAP provided scores for each feature, word, or sub-word used by the respective models, which, together with careful manual analysis, helped in understanding what exactly the models had learned from the provided data. All experiments were carried out on the corpus created by Gonçalo Oliveira et al. (2020), which is, to the best of our knowledge, the only corpus in Portuguese created for the task of Humor Recognition.

Our results show that the BERT model outperformed all other ML methods in terms of F1-score, achieving a score of 99.6%. However, through careful analysis, we discovered that this model, alongside other methods based on content-features, based their decisions primarily on stylistic aspects of the texts, such as punctuation, and other phenomena not necessarily related to humor, for instance the presence of questions.

Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL2023), pages 88–98

We also noted aspects of the set of humor-related features proposed by Gonçalo Oliveira et al. (2020) that might not have been expected by their original authors, such as the relation between concreteness and humor, and the association of people named entities with humorous texts. However, some of their interpretations were reinforced by the ML models, for example, the connection of ambiguity and incongruity to humorousness.

The remainder of this paper is organized as follows: some relevant related work about Humor Recognition and ML Explainability is presented in section 2, followed by an overall description of our methodology, in section 3. Later, the results are presented and discussed in section 4, with the final remarks and future work mentioned in section 5. In the end of the paper, we note some limitations of the current work, as well as ethical aspects that should be considered in the future.

2 Related Work

This paper has relations with two main areas of research: general Humor Recognition, usually interpreted as a ML classification task, and ML Explainability, which aims at creating explanations for computational models, in order to inspect what information the model actually uses for inference.

2.1 Humor Recognition

Humor Recognition research dates back to the 2000s, when Mihalcea and Strapparava (2005) used a hand-crafted feature set (including features like alliteration, slang usage, and antonymy presence) to train supervised ML algorithms for classifying texts in two categories: humorous and non-humorous. Since then, Humor Recognition has been approached with this supervised ML point-of-view, varying with different sets of attributes, including:

- Stylistic, e.g., keywords and text similarity with other jokes (Sjöbergh and Araki, 2007);
- Semantic information, e.g., presence of vocabulary focused on professional communities, sentiment polarity, and words related to negative human traits (Mihalcea and Pulman, 2007);
- Surface-level characteristics, e.g., punctuation and word frequency (Barbieri and Saggion, 2014).

More recently, following the general trends on many different NLP tasks, the current state-of-the-art in this task is achieved by Deep Learning (Ren et al., 2021; Kumar et al., 2022) and LLMs (Devlin et al., 2019; Weller and Seppi, 2019).

For languages other than English, the HAHA series of shared-tasks (Castro et al., 2018; Chiruzzo et al., 2021) has encouraged much advance for research on recognizing verbal humor in Spanish. In their latest event, Grover and Goel (2021), the winners, used an ensemble of LLMs to outperform other contestants. For Portuguese, however, there is still few research on the matter; to the best of our knowledge, current systems are still based on classical ML algorithms with a specific set of hand-crafted features (Gonçalo Oliveira et al., 2020), in a similar fashion to those methods from the early 2000s. Hence, there is still much to advance for this specific language.

On the other hand, also for Portuguese, we acknowledge research on Irony Detection (Carvalho et al., 2009; de Freitas et al., 2014; Wick-Pedro and Vale, 2020; Corrêa et al., 2021), a task that is to some extent related to humor, especially when dealing with satirical content (Wick-Pedro and Santos, 2021; Carvalho et al., 2020).

2.2 Machine Learning Explainability

As most ML models, Humor Recognition systems lack a qualitative understanding about how their prediction is obtained, i.e., what exactly the machine has learned from the provided examples. This brings up concerns regarding how trustful and understandable such models are, as well as questions if they are indeed basing their decision on meaningful parts of the data (Ribeiro et al., 2016).

Traditionally, ML explainability has been tackled simply through the usage of models that are inherently interpretable, such as linear classifiers (Ustun and Rudin, 2016) or rule-based methods (Wang and Rudin, 2015). Additionally, modern Neural Network models still have some degree of interpretability, through close inspection of their parameters, e.g., attention weights, especially for Computer Vision (Xu et al., 2015). However, such approaches are still limited to specific models; furthermore, they can get too overwhelming as the number of parameters increases.

There is, however, research on creating model-agnostic ML explanations, for example with tools like LIME (Ribeiro et al., 2016) and SHAP (Lund-

berg and Lee, 2017), which focus on approximating a simpler interpretable model by perturbing the inputs and measuring how each attribute (or token, pixel, subword, etc.) contribute to the original more complex one. These methods target local explainability, i.e., approximating models that work well on a vicinity of a given input, which is possible to be generalized for the whole data space through careful analysis of different instances.

3 Methodology

Our work has two main fronts of research: first, the implementation of Humor Recognition systems for the Portuguese language; then, a deeper analysis of their performance, to assess their weaknesses and stimulate further research. This includes a discussion on how to overcome some challenges, followed by what could be developed towards improved systems.

3.1 Humor Recognition Methods for Portuguese

Our first step was to re-implement the methods described by Gonçalo Oliveira et al. (2020) and Clemêncio (2019), as their source code is not publicly available and it is the only previous work for Humor Recognition in Portuguese. Their approach consists of testing different ML algorithms (i.e., SVM, Random Forest, and Naïve Bayes) with different sets of attributes: content and humor-related features. As content features, the original authors used a bag-of-words with TF-IDF counts for 1,000 tokens (or n-grams) selected via a χ^2 test. For humor-related features, they used different kinds of information, namely: alliteration through character n-grams, out-of-vocabulary words, average word embedding similarity, Named Entity Recognition (NER) counts, count of antonymy pairs, sentiment polarity, slang usage, concreteness, imageability, and ambiguity.

As we will see in subsection 4.1, our re-implementation outperformed the original reported values, leading us to reconsider our code and find some minor details, which might explain this difference in the evaluation metrics. In our implementation, we did not use the χ^2 test for selecting which attributes would comprise the final 1,000 content features, instead we used the most frequent ones. Additionally, we used the NLPyPort toolkit (Ferreira et al., 2019) for the content-features and not only for the humor-related ones, as shown in the original paper. In fact, comparing our feature analysis in subsection 4.2 to the one by Gonçalo Oliveira et al. (2020), we have strong evidence that their tokenizer discards punctuation, which NLPyPort does not. Differences between versions of the tools and resources used might also be an option, but we find the tokenization difference to be the most plausible reason for this difference in the results.

During our work, we decided to keep these changes as they resulted in a clearly higher performance. In all other aspects, we followed the same methodology as Gonçalo Oliveira et al. (2020), testing the same ML algorithms on the same corpus, with the same feature sets obtained from the same resources.

In addition, we fine-tuned BERTimbau, a pre-trained BERT model for Portuguese (Souza et al., 2020)[1], for Humor Recognition during 3 epochs with a learning rate of 5×10^{-5}. This was motivated by the broad utilization of LLMs for performing this and other tasks, leading to the current state-of-the-art in other languages (e.g., English and Spanish), as mentioned in subsection 2.1.

3.2 Corpus

We used the data set provided by Gonçalo Oliveira et al. (2020)[2], with short humorous texts in two main formats: satirical news headlines and one-line jokes. The authors were careful when including negative examples (non-humorous texts) into the corpus, trying to add only instances with a similar format to the humorous examples collected. For example, they included real news headlines as a counterpart to the satirical ones. For one-liners, as most of the jokes have a question-answer pattern, they used texts with this same composition from a trivia website and from MultiEight-04 (Magnini et al., 2005), a corpus for Question Answering. They also included proverbs to account for those one-liners not written in a question-answer fashion. We present some examples of instances from the corpus in Table 1.

Since the original corpus has different configurations available, we used the balanced one with texts from all sources, with a total of 2,800 texts, 1,400 humorous and 1,400 non-humorous instances. We should also note that, since we do not have access to the original train-test split used

[1]Available at: https://huggingface.co/neuralmind/bert-base-portuguese-cased.

[2]Available at: https://github.com/NLP-CISUC/Recognizing-Humor-in-Portuguese.

Original text in Portuguese	Translation	Comments
	Humor examples	
O que é uma fofoca? É um animal mamarítimo.	*What is a gossip? It is a mamarine animal.*	The humorous effect comes from the fact that the word "fofoca" (*gossip*) sounds like "foca" (*seal*) with a doubled initial syllable, so the answer says it is a marine animal, but also with a doubled initial syllable.
Patrões exigem vacinação obrigatória contra o bicho do sindicalismo	*Employers demand mandatory vaccination against the trade unionism bug*	The humor in this satirical headline arises from a semantic shift, as the association of vaccination is typically with disease rather than unionism. Furthermore, since satire employs humor as a means of criticism, this example serves as a critique of the bosses' opposition to unionism.
	Non-humor examples	
Onde fica Hyde Park? nos Estados Unidos.	*Where is Hyde Park? In the United States.*	–
Presidente promulga dia de luto nacional pelas vítimas de violência doméstica.	*President proclaims national day of mourning for victims of domestic violence.*	–

Table 1: Examples of instances present in the corpus

by Gonçalo Oliveira et al. (2020), we made a new split with the same reported ratio (80% train and 20% test).

3.3 Feature and Model Analysis

After the implementation, training, and testing of the models, we first used the SHAP explainability tool (Lundberg and Lee, 2017) to calculate importance values for each of the features proposed by Gonçalo Oliveira et al. (2020), both content and humor-related. Since SHAP was originally developed for explaining single instances of the data set, in order to measure the overall importance of each feature, we use the absolute mean value (over all examples in the test corpus); this is complemented with visualization techniques, such as beeswarm plots, to better understand how each feature behaves in general.

We also carried out an analysis of the fine-tuned BERT model, identifying which pieces of information were actually used by the system to distinguish humorous from non-humorous texts. For this, we used SHAP once again. However, as this kind of model does not consist of a pre-defined set of features, we were not able to use the absolute mean value, as it would have to comprise every single sub-word in the model. Therefore, we decided to perform such analysis manually, by examining specific instances that are representative of the corpus; to select such examples, we followed a simpler approach, using a clustering algorithm, namely K-Means with $k = 56$ (2% of the data set), on sentence embeddings obtained from BERTimbau fine-tuned for Semantic Textual Similarity[3], and selected the centroid instances as a sample of the whole corpus. Then, we carefully analyzed those sentences and their SHAP values, to finally identify some clear patterns that BERT learned for classification.

It is important to mention that Gonçalo Oliveira et al. (2020) also did a feature analysis procedure using a χ^2 test. However, this approach is not related to specific models and how they interpret the input, but rather focuses on finding relations between the features and the true labels.

4 Results

This work has two main results. First, a new fine-tuned BERT model for Humor Recognition in Portuguese, which outperforms the current state-of-the-art for this task in terms of automatic evaluation metrics. Secondly, a deeper analysis of such models, identifying how well they are suited for the task in general.

4.1 Humor Recognition

The results for each of the implemented approaches, alongside those reported by Gonçalo Oliveira et al.

[3]Available at: https://huggingface.co/rufimelo/bert-large-portuguese-cased-sts.

(2020), are presented in Table 2. Due to space limitations, we show only the results obtained by the best model for each set of features.

Feature Set	Model	F1
Content features	SVM	96.4%
Humor features	Random Forest	78.8%
All features	Random Forest	97.1%
—	Fine-tuned BERT	99.6%
Gonçalo Oliveira et al. (2020)		
Content features	SVM	75%
Humor features	Random Forest	64%
All features	SVM	78%

Table 2: Best results for Humor Recognition with different models and different feature sets

We note that our re-implementation produced better results than those by Gonçalo Oliveira et al. (2020); for example, while their best model using all features (content and humor-related) reportedly and F1 of 78%, our models reached up to 97.1%. The probable reasons for such a large gap have been discussed in subsection 3.1.

In addition, we found it surprising that our models had such positive results – BERT had a nearly perfect score of 99.6% – for a task that is usually mentioned in the literature as extremely difficult and subjective (Veale, 2004; Hempelmann, 2008; Reyes Pérez, 2013; Kumar et al., 2022). From this observation, we decided to do an explainability analysis to identify exactly which features and pieces of information our trained models were leveraging on when classifying their input.

4.2 Explainability Analysis

In the first analysis, we used SHAP to calculate the importance values of each feature for the best methods reported in the previous subsection 4.1. For the SVM model using exclusively content features, Figure 1 presents the most important (i.e., larger average absolute SHAP value) features for the humor class. In the plot, each feature is represented in the Y axis, with each point representing an instance of classification; their color expresses the relative value of the feature in that specific instance, while their placement along the X axis indicates their importance. For example, we can see that the most important feature used by the model is the presence of a full stop (a period followed by an end-of-sentence special token), and that they are most important for the humor class (positive SHAP values, right of the central vertical bar) when their TF-IDF counts are low (blue). This same behavior can be seen for the second most important feature (period), indicating that the model is interpreting the mere presence of periods as an indicative of non-humorousness. This is probably a fault from the corpus, as will be further discussed in subsection 4.3.

Another interesting observation that can be drawn from this analysis is that the model leverages question-related features as indicatives of humor, for example the usage of question marks, and wh-question words ("qual é", "o que", "qual", and "que"[4]). One can note that the model considers them important to identify humor when their TF-IDF counts are higher (red points), meaning that it is associating questions to humor despite the presence of similar texts as negative examples of humor, as mentioned in subsection 3.2.

Due to space and resource limitations, we cannot extensively analyze all 1,000 content-features. However, we report that the next features in the list are still wh-question words, such as "porque", "qual é o", and "como" [5], or punctuation marks (colons, double quotes, and exclamation marks). We highlight, however, that the explainability results for all features will be made publicly available alongside the code and results obtained by the models, so that the research community can observe this data in its entirety.

The second analysis refers to only humor-related features, presented in Figure 2. The most important feature is the number of out-of-vocabulary words, which is seen as a strong indicative of humor. Then, the average level of concreteness follows with a not so clear disparity of how its values interact with its importance; however, there seems to be a preference of higher values to be positive contributions to the humor class, which is contradictory to the interpretation by Gonçalo Oliveira et al. (2020) that non-humorous texts are more concrete, while humor is more related to mental images.

Another remarkable note is that higher NER counts for people ("PESSOA") is usually taken as evidence to favor the humor class, which is again the opposite speculated by Gonçalo Oliveira et al. (2020). The authors mention that real headlines would contain more names of people, but we argue that they are also present in satirical headlines and

[4] "Which is", "what", "which", and "what". Translated by the authors.

[5] "Why", "which is the", and "how". Translated by the authors.

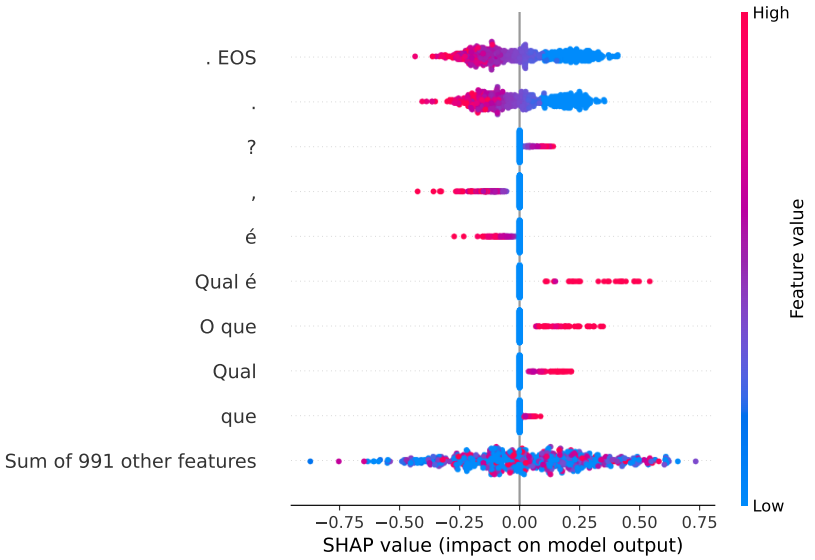

Figure 1: Beeswarm plot with the most important content features used by the SVM model

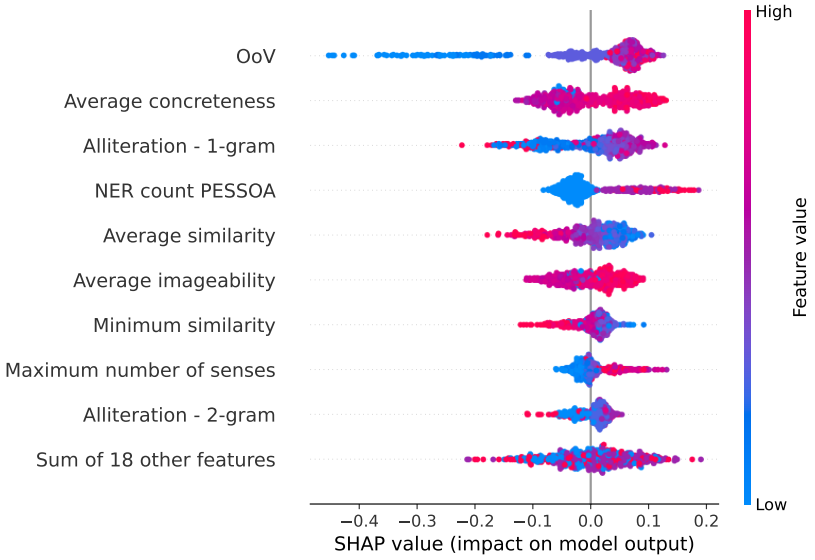

Figure 2: Beeswarm plot with the most important humor-related features used by the Random Forest model

one-liner jokes (e.g. "Por que a Angélica não mata baratas? Ela espera o Maurício Mattar."[6]), so that the model learned to link them to humor instead.

Finally, in accordance to the reasoning of Gonçalo Oliveira et al. (2020), the average and minimum similarity features are evidences of humor when they have lower values, which represents a higher incongruity among the words. Thus, it seems fruitful to model incongruity as word similarity. Also, the model favors high numbers of senses to classify an instance as humor, reaffirming the argument that humor resides in ambiguity.

[6]"Why doesn't Angélica kill cockroaches? She waits for Maurício Mattar." Translated by the authors.

When combining both kinds of features, the observations do not vary much: the Random Forest model relies mainly on punctuation (full stop, period, question mark), and wh-question words. It is, however, noticeable that humor-related features such as concreteness, imageability, person NER counts, and average similarity are considered more important than question words in this scenario. We highlight, once more, that an extensive display of these results will be made available.

4.3 Explainability Analysis of BERT

As mentioned in subsection 3.3, for the fine-tuned BERT model, we needed to do a careful manual

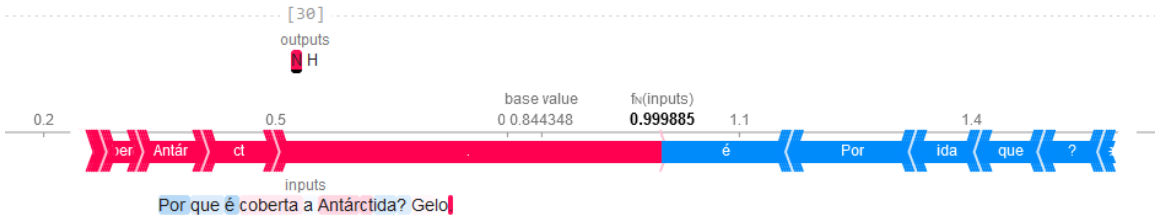

Figure 3: Example of BERT explanation obtained via SHAP

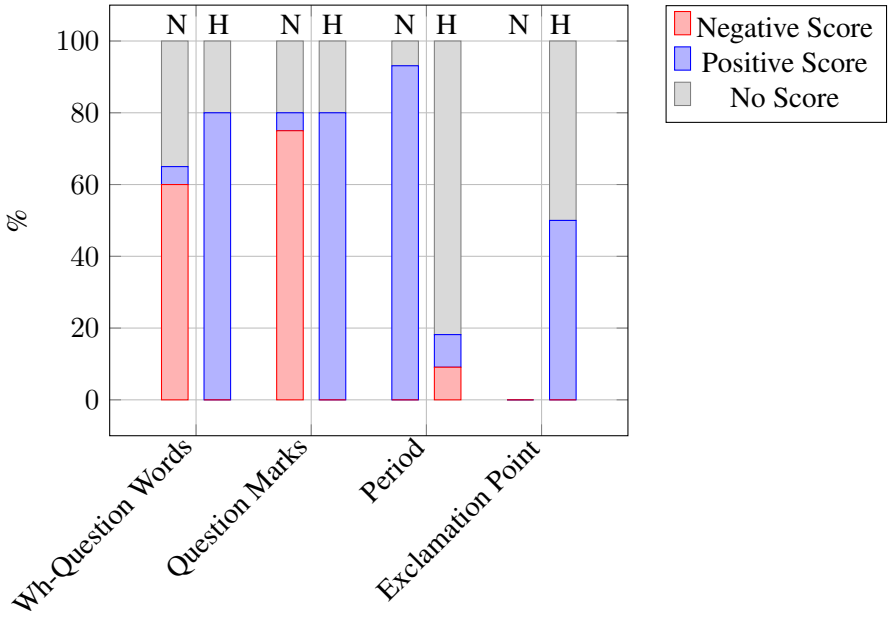

Figure 4: Results of the BERT explanation analysis

analysis of a subset of representative instances of the corpus obtained via clustering. All 56 examples were explained by SHAP and handed to a linguist with previous experience on analyzing humorous and figurative language. An example of a SHAP explanation can be seen in Figure 3.

In the figure, the text "Por que é coberta a Antárctida? Gelo."[7] was classified as non-humorous ("N" is highlighted) with a confidence score ($fx(\text{inputs})$) of 0.999885. For the analysis, the BERT model starts at a base value – for this specific case, 0.844348 – obtained by masking out all input subwords. Then, each subword contributes to this value positively (in red) or negatively (in blue) until it reaches the final confidence score; moreover, larger bars represent a larger absolute contribution.

From this analysis, we have drawn two main observations, which are related to the results reported in the previous subsection 4.2. First, the usage of wh-question words – such as "o que", "quais", and "como"[8] – are in general taken by the model as evidence to classify an instance as humor; from all 29 examples classified as non-humorous, 20 (69.97%) have wh-question words, from which they contributed negatively in 12 instances (60%), positively in only 1 instance (5%), and in the remaining 7 texts (35%) they did not get any scoring.

Another evidence for this association of wh-question words with humor in the model is that, in 25 instances classified as humor, from which 15 (60%) had such type of words, 12 (80%) occurrences contributed positively to the classification, while the remaining 3 (20%) did not contribute at all. There were no cases in which BERT considered the presence of wh-question words as negative evidences for this class.

The second result of this analysis is about punctuation, which was also observed in other models as being extremely important. For instances classi-

[7] "What is Antarctica covered by? Ice." Translated by the authors.

[8] "What", "which", and "how." Translated by the authors.

fied as non-humorous, 20 (69%) contained question marks, 15 (75%) of which were assigned with negative SHAP values, 4 (20%) were not scored at all, and only 1 (5%) received a positive score. Meanwhile, for examples classified as humor, 15 (60%) had question marks, from which 12 (80%) were considered as positive evidences for this class, and the remaining 3 (20%) had no score; similarly to wh-question words, no question mark was considered as negative for the humor class.

We argue that these observations contribute to the point-of-view that BERT, similarly to the models discussed in subsection 4.2, is focusing on the textual form rather than specific humor-related linguistic devices by connecting the mere existence of questions to humor. Nonetheless, as mentioned in subsection 3.2, the original authors of the corpus were careful to also include question-answer texts with no humor, and the model still reached more than 99% F1-Score (Table 2), meaning that it is very likely classifying such instances correctly. In this context, as illustrated by Figure 3, another punctuation mark comes into place: the period, specially a full stop, which was also highly scored in the other methods discussed before.

All 29 examples classified as non-humor end with a period, from which 27 (93.10%) received positive SHAP scores and the remaining 2 (6.90%) were not scored at all; no period was deemed as a negative evidence for the non-humor class. Meanwhile, for the sample of 25 instances classified as being humorous, only 9 (36%) had periods, sometimes with more than one resulting in 11 periods, from which 1 (9.09%) was positive, 1 (9.09%) was negative and the remaining 9 (81.82%) received no scoring, indicating that BERT tends to not even consider periods for the humor class.

We highlight that the difference in the occurrence of periods between the humor and non-humor classes in the analyzed sample may indicate that this discrepancy also exists in the corpus. Likely, this specific aspect of the text format was overlooked by the original authors, which may explain why the models primarily use this punctuation mark to distinguish humor from non-humor.

Additionally, exclamation points are present only in the examples classified as humorous, with 4 (50%) being positive and 4 (50%) not having attributed any value to this instance. All these results are summarized in Figure 4.

From all these observations, we point out how difficult it is to find negative examples when creating a corpus for Humor Classification – and arguably to any classification task. LLMs are so powerful in finding surface-level patterns that even slight details (such as punctuation) can and will be used in the task, even if they are not necessarily part of the linguistic mechanism that produces the humorous effect, such as ambiguity, incongruity, and surprise (Attardo and Raskin, 1991; Tagnin, 2005; Reyes Pérez, 2013; Kao et al., 2016; Wick-Pedro and Vale, 2020; Aleksandrova, 2022).

5 Conclusion

In this paper, we presented a re-implementation of the previous state-of-the-art method for Humor Recognition in the Portuguese language, alongside a novel fine-tuned BERT model for the same task, reaching a nearly-perfect F1 score of 99.64%.[9]

However, a deeper analysis of the models using a Machine Learning explainability method, SHAP, enabled us to understand which pieces of information the models were relying on to do such classification. We came into the conclusion that BERT and models based on TF-IDF counts did not learn specific mechanisms of humor, but were instead leveraging mainly stylistic characteristics of the texts, such as punctuation and the presence of wh-questions.

Furthermore, the analysis of how humor-related features were interpreted by the ML model led to interesting observations not considered by their proposers (Gonçalo Oliveira et al., 2020). For example, the association of humor with person named entities or higher levels of concreteness; however, some of their reasoning can also be reinforced by how the model used some of the knowledge provided, e.g. humor was considered related to higher levels of ambiguity and incongruity within the text, which is up to par with linguistic descriptions of verbal humor (Raskin and Attardo, 1994; Tagnin, 2005; Aleksandrova, 2022).

As a final conclusion, we emphasize how challenging it is to create a text classification corpus for supervised ML in such a way that the model actually learns about the linguistic phenomenon in question, rather than resorting to specific attributes and shortcuts not directly related to the problem being studied. We find that humor is a specially difficult task to create such a corpus, as it is a largely di-

[9] All the code, models, results, and analysis is available at: https://github.com/Superar/HumorRecognitionPT.

verse phenomenon (verbal humor can be conveyed in many different ways), with an equally large universe of negative examples (non-humorous texts also present themselves in various formats).

From these conclusions, we can draw some fruitful paths for future research. First, we mention the creation of a new corpus for Humor Recognition in Portuguese, taking into account some of the flaws found in the corpus by Gonçalo Oliveira et al. (2020). However, as it is – to the best of our knowledge – the only available corpus for this task, it can still be evaluated if it is fit after some process of normalization, starting with punctuation, e.g., by adding full stops to the humor examples, which would be a less expensive process; some early experiments in this sense show a decrease of 4 percentage points in F-Score obtained by the BERT model when discarding or normalizing punctuation. Another point to be considered for the creation of a new corpus is the responsibility of which texts to include; as we mention later in our Ethics Statement, the corpus used in this work contains texts annotated as jokes that contain rather problematic content, e.g. riddles that are openly racist.

Another possibility for future work is to change the models and how they work. One could use methods, such as the one proposed by Kao et al. (2016), that are not based on ML, but rather on formalizing linguistic theories of humor to a computational environment. We also find it appealing to explicitly include linguistic knowledge into the ML models, so that they are powered with some information beyond the textual surface, argued by other researchers as vital to deal with complex phenomena such as humor (Hempelmann, 2008; Amin and Burghardt, 2020). This goal could also be achieved by exploring further the humor-related features, which were proposed originally from a linguistic point-of-view; other extra-linguistic aspects of Humor could also be studied, for instance how different cultural backgrounds affect the perception and definition of humorousness.

Limitations

As main limitation of this work, we mention the lack of an extensive analysis of the explainability results, limiting our examination to the most highly-scored features; additionally, we not consider the interaction among the features themselves. We also think that the analysis of the BERT model could use a larger set of representative instances of the corpus; regarding this selection, we also mention that there are probably other methods rather than clustering to ensure that the analyzed subset is actually a good representation of the data set in its entirety. Finally, we agree that the classification models deserve a deeper analysis on their performance, for example, by carrying out K-fold cross validation tests.

Ethics Statement

We believe that humor is a positive and constructive form of human expression to unite and reduce tensions while respecting cultural differences, beliefs, and people's identities. However, we acknowledge that humor, when used in a Christian or offensive way to discriminate, ridicule, or disparage individuals or groups, especially those who have been historically marginalized or oppressed, can have negative consequences.

So if there are jokes that promote violence, hatred, or prejudice, including but not limited to racial, gender, and sexual stereotypes, xenophobia, and similar forms of discrimination, then they ought not to be deemed acceptable. In this context we find it crucial to report that the corpus used in this paper contains some texts (annotated as humor) that are openly racist, specially against black people. Other texts considered as jokes have different groups represented in a negative light, for example alentejanos (people from a region in Portugal), jeweish people, and blonde women. Some other sensitive subjects are also present in the corpus, for instance suicide, and pedophilia.

It is crucial to take into account the potential effects that computer models designed for mood detection could have on individuals and society, both during the development phase and when utilizing them. Ensuring that these models are impartial and free of undesired bias is of utmost importance to prevent the perpetuation of stereotypes that could ultimately result in negative outcomes.

In conclusion, we would like to emphasize that models used for the recognition of humor have inherent limitations due to their subjective nature, which may vary significantly depending on cultural, social, and individual contexts. Therefore, these models are constantly evolving and improving, and evaluating their efficacy is an ongoing process.

Acknowledgements

This work was funded by national funds through the FCT – Foundation for Sci-

ence 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.

References

Elena Aleksandrova. 2022. Pun-based jokes and linguistic creativity: Designing 3R-module. The European Journal of Humour Research, 10(1):88–107.

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In Proceedings of the the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 29–41, Online. International Committee on Computational Linguistics.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. Humor - International Journal of Humor Research, 4(3-4).

Francesco Barbieri and Horacio Saggion. 2014. Automatic Detection of Irony and Humour in Twitter. In International Conference on Computational Creativity, Ljubljana. Association for Computational Creativity (ACC).

Paula Carvalho, Bruno Martins, Hugo Rosa, Silvio Amir, Jorge Baptista, and Mário J. Silva. 2020. Situational Irony in Farcical News Headlines. In Paulo Quaresma, Renata Vieira, Sandra Aluísio, Helena Moniz, Fernando Batista, and Teresa Gonçalves, editors, Computational Processing of the Portuguese Language, volume 12037, pages 65–75. Springer International Publishing, Cham.

Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). In Proceeding of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion - TSA '09, page 53, Hong Kong, China. ACM Press.

Santiago Castro, Luis Chiruzzo, and Aiala Rosá. 2018. Overview of the HAHA Task: Humor Analysis based on Human Annotation at IberEval 2018. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), pages 187–194, Sevilla. CEUR-WS.org.

Luis Chiruzzo, Santiago Castro, Santiago Góngora, Aiala Rosá, J. A. Meaney, and Rada Mihalcea. 2021. Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish. Procesamiento del Lenguaje Natural, 67:257–268.

André Clemêncio. 2019. Reconhecimento Automático de Humor Verbal. MSc, Universidade de Coimbra, Coimbra.

Ulisses B Corrêa, Leonardo Coelho, Leonardo Santos, and Larissa A de Freitas. 2021. Overview of the IDPT task on irony detection in portuguese at IberLEF 2021. Procesamiento del Lenguaje Natural, 67:269–276.

Larissa A. de Freitas, Aline A. Vanin, Denise N. Hogetop, Marco N. Bochernitsan, and Renata Vieira. 2014. Pathways for irony detection in tweets. In Proceedings of the 29th Annual ACM Symposium on Applied Computing, pages 628–633, Gyeongju Republic of Korea. ACM.

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.

João Ferreira, Hugo Gonçalo Oliveira, and Ricardo Rodrigues. 2019. Improving NLTK for Processing Portuguese. page 9 pages.

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.

Karish Grover and Tanishq Goel. 2021. HAHA@IberLEF2021: Humor Analysis using Ensembles of Simple Transformers. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), pages 883–890, Málaga. CEUR-WS.org.

Christian F. Hempelmann. 2008. Computational humor: Beyond the pun? In The Primer of Humor Research, number 8 in Humor Research, pages 333–360. Victor Raskin, Berlin, New York.

Justine T. Kao, Roger Levy, and Noah D. Goodman. 2016. A Computational Model of Linguistic Humor in Puns. Cognitive Science, 40(5):1270–1285.

Vijay Kumar, Ranjeet Walia, and Shivam Sharma. 2022. DeepHumor: A novel deep learning framework for humor detection. Multimedia Tools and Applications, 81(12):16797–16812.

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.

Bernardo Magnini, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de

Rijke, Paulo Rocha, Kiril Simov, and Richard Sutcliffe. 2005. Overview of the CLEF 2004 Multilingual Question Answering Track. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, Multilingual Information Access for Text, Speech and Images, volume 3491, pages 371–391. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rada Mihalcea and Stephen Pulman. 2007. Characterizing Humour: An Exploration of Features in Humorous Texts. In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 4394, pages 337–347. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz, and Oto A. Vale. 2018. Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. In Aline Villavicencio, Viviane Moreira, Alberto Abad, Helena Caseli, Pablo Gamallo, Carlos Ramisch, Hugo Gonçalo Oliveira, and Gustavo Henrique Paetzold, editors, Computational Processing of the Portuguese Language, volume 11122, pages 324–334. Springer International Publishing, Cham.

Jonathan D. Raskin and Salvatore Attardo. 1994. Non-literalness and non-bona-fide in language: An approach to formal and computational treatments of humor. Pragmatics & Cognition, 2(1):31–69.

Lu Ren, Bo Xu, Hongfei Lin, and Liang Yang. 2021. ABML: Attention-based multi-task learning for jointly humor recognition and pun detection. Soft Computing, 25(22):14109–14118.

Antonio Reyes Pérez. 2013. Linguistic-based Patterns for Figurative Language Processing: The Case of Humor Recognition and Irony Detection. Procesamiento del Lenguaje Natural.

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.

Jonas Sjöbergh and Kenji Araki. 2007. Recognizing Humor Without Recognizing Meaning. In Francesco Masulli, Sushmita Mitra, and Gabriella Pasi, editors, Applications of Fuzzy Sets Theory, volume 4578, pages 469–476. Springer Berlin Heidelberg, Berlin, Heidelberg.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I, pages 403–417, Berlin, Heidelberg. Springer-Verlag.

Stella E. O. Tagnin. 2005. O humor como quebra da convencionalidade. Revista Brasileira de Linguística Aplicada, 5(1):247–257.

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

Tony Veale. 2004. Incongruity in humor: Root cause or epiphenomenon? Humor - International Journal of Humor Research, 17(4).

Fulton Wang and Cynthia Rudin. 2015. Falling Rule Lists. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, volume 38 of Proceedings of Machine Learning Research, pages 1013–1022, San Diego, California, USA. PMLR.

Orion Weller and Kevin Seppi. 2019. Humor Detection: A Transformer Gets the Last Laugh. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Gabriela Wick-Pedro and Roney L. S. Santos. 2021. Complexidade textual em notícias satíricas: Uma análise para o português do Brasil. In Anais Do XIII Simpósio Brasileiro de Tecnologia Da Informação e Da Linguagem Humana (STIL 2021), pages 409–415, Brasil. Sociedade Brasileira de Computação.

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.