PUNTUGUESE: A Corpus of Puns in Portuguese with Micro-edits

Marcio Lima Inácio[*†], **Gabriela Wick-Pedro**[‡], **Renata Ramisch**[§], **Luís Espírito Santo**[*†¶], **Xiomara S. Q. Chacon**[‖], **Roney Santos**[**], **Rogério F. de Sousa**[††], **Rafael T. Anchiêta**[††], **Hugo Gonçalo Oliveira**[*†]

[*]Centre for Informatics and Systems of the University of Coimbra (CISUC)
[†]Intelligent Systems Associate Laboratory (LASI)
{mlinacio, hroliv}@dei.uc.pt
[‡]Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
gabiwick@gmail.com
[§]Redação Nota 1000
renata.ramisch@gmail.com
[¶]AILab - Vrije Universiteit Brussel (VUB)
luis.espirito.santo@vub.be
[‖]Institute of Mathematics and Computer Science (ICMC/USP)
xiomara.qchacon@usp.br
[**]Federal University of Bahia (UFBA)
roneyleft@gmail.com
[††]Federal Institute of Piauí (IFPI)
{rogerio.sousa, rta}@ifpi.edu.br

Abstract

Humor is an intricate part of verbal communication and dealing with this kind of phenomenon is essential to building systems that can process language at large with all of its complexities. In this paper, we introduce PUNTUGUESE, a new corpus of punning humor in Portuguese, motivated by previous works showing that currently available corpora for this language are still unfit for Machine Learning due to data leakage. PUNTUGUESE comprises 4,903 manually-gathered punning one-liners in Brazilian and European Portuguese. To create negative examples that differ exclusively in terms of funniness, we carried out a micro-editing process, in which all jokes were edited by fluent Portuguese speakers to make the texts unfunny. Finally, we did some experiments on Humor Recognition, showing that PUNTUGUESE is considerably more difficult than the previous corpus, achieving an F1-Score of 68.9%. With this new dataset, we hope to enable research not only in NLP but also in other fields that are interested in studying humor; thus, the data is publicly available.

Keywords: Punning Humor, Corpus Creation, Humor Recognition

1. Introduction

With the most recent advances in Natural Language Processing (NLP) systems, research expands even more into complex and subtle language uses (Cambria and White, 2014). An example is the computational processing and generation of creative and humorous texts (Hempelmann, 2008; Amin and Burghardt, 2020). Dealing with this kind of linguistic phenomena is essential to build systems that can deal with language at large, as they are an important part of language fluency (Tagnin, 2005).

Moreover, it is widely known that NLP research is generally focused on a few languages, especially English, sometimes neglecting other tongues (Bender, 2019). For our language of interest, Gonçalo Oliveira et al. (2020) developed a corpus of European Portuguese jokes, shown to be flawed for Machine Learning after further studies, as it led the algorithms to associate mainly questions and punctuation with humor, which is not always the case in real scenarios (Inácio et al., 2023).

To this extent, this paper focuses on developing a new higher-quality corpus of punning humor in Brazilian and European Portuguese, named PUNTUGUESE, to foster further research about humorous and figurative language within the Portuguese-speaking community. Our new corpus — containing 4,903 jokes, 2,850 of which are publicly available at `https://github.com/Superar/Puntuguese` — is considerably larger than the previous one and has also the advantage of being manually curated. We include some additional annotation layers, such as categories from a taxonomy of punning humor, and highlighting of punning signs. Besides, PUNTUGUESE has examples of non-humorous texts created via micro-editing to minimize the influence of textual form during learning, hopefully enabling machine learning models to focus on learning puns with minimal noise related to writing style and other textual features.

The remainder of the paper is organized as follows. In section 2, we describe some previous related work, namely some corpora of verbal humor,

LREC-COLING 2024, pages 13332–13343

and similar phenomena. In section 3, we present the creation process of PUNTUGUESE, including the data gathering and micro-editing processes. Section 4 shows a clustering-based analysis of the editions. Afterward, in section 5, we do some experiments on Humor Recognition and considerations about their results. Finally, the conclusions and future work paths are shown in section 6, followed by some considerations about Ethics and limitations of our work in sections 7 and 8, respectively.

2. Related Work

This work is closely related to the works of Gonçalo Oliveira et al. (2020), the only known corpus of humorous texts in Portuguese; Inácio et al. (2023), which motivated the creation of a new corpus; and Hossain et al. (2019), an inspiration to our methodological approach.

Gonçalo Oliveira et al. (2020) created a corpus of Portuguese jokes from different sources, such as satiric newspapers and collections of one-line jokes. Since their objective was to create a dataset that could be used for supervised learning, they also gathered non-humorous texts as negative examples, being attentive to collect data with similar formats, e.g. question-answer pairs to account for riddle jokes, and proverbs as a counterpart to non-riddle ones and headlines. Their humor classification experiments, using different kinds of content and humor-based features, resulted in a maximum F1-Score of 87% on the one-liners set, 82% for headlines, and 75% when dealing with all different types of text in the corpus.

More recently, Inácio et al. (2023) used the same corpus for Humor Recognition and obtained impressive results, achieving an F-Score of 99% with a BERT model. From a Machine Learning Explainability analysis, they observed that the algorithm linked the presence of questions and some sorts of punctuation to the presence of humor in the text, unveiling some subtle differences, that were being used as shortcuts for the classification, resulting in such high results. This means that, even though Gonçalo Oliveira et al. (2020) were careful enough to search for similar-looking texts, the stylistic differences between the different sources were enough to create some kind of data leakage, which motivated the creation of a new corpus through a different methodology.

In this context, we decided to follow a similar approach as in Humicroedit, by Hossain et al. (2019). In their dataset, the authors collected news headlines from Reddit[1] that were later edited via crowdsourcing to turn each text into a funny one. The annotators were instructed to make the smallest change possible (micro-edit) so that the example pairs only differ in terms of humor-induction capabilities. Finally, each edited headline was judged in terms of funniness on a scale from 0 to 3. In our work, we explore this idea of micro-edits, turning humorous texts into their non-funny counterparts. This process is better detailed in subsection 3.2.

Specifically regarding the Portuguese language, it is also worth mentioning other related initiatives, such as corpora for irony detection in tweets (Carvalho et al., 2009; Wick-Pedro and Vale, 2020) and satirical news (Carvalho et al., 2020; Wick-Pedro and Santos, 2021). Specifically for puns, some corpora are also available for the English language (Miller et al., 2017; Simpson et al., 2019).

3. Corpus Creation

As previously mentioned, our corpus comprises punning jokes in two varieties of Portuguese: European and Brazilian Portuguese. After gathering the puns, we carried out a process of micro-editing, similar to that of Hossain et al. (2019), but the other way around, i.e. to eliminate the humor aspect of the provided jokes. This should ensure that the instances of each class (Humor and Non-humor) differ only in the funniness dimension. More details on the corpus creation process are given below.

3.1. Data Gathering

In PUNTUGUESE, we decided to focus on a specific type of humor: puns, which are considered to be a simpler kind of text capable of expressing funniness through word ambiguity (Kao et al., 2016). To this extent, we first created some gathering guidelines based on our chosen definition of punning humor:

> "A pun is a form of wordplay in which one sign (e.g., a word or phrase) suggests two or more meanings by exploiting polysemy, homonymy, or phonological similarity to another sign, for an intended humorous or rhetorical effect." (Miller et al., 2017)

We also provided some additional clarification in the guidelines to help identify the puns:

- A sign can be a single word (or token), a phrase (a sequence of tokens), or a part of a word (a subtoken);
- The humorous effect must rely on the ambiguity of said sign;
- The ambiguity must originate from the word's form (written or spoken);
- Every pun must have a "pun word" (the ambiguous sign that is in the text) and an "alternative word" (the sign's ambiguous interpretation) identified. If it is not possible to identify both, the text is not considered a pun and should not be included.

[1] www.reddit.com

We also imposed some limitations regarding the textual form. For instance, as our focus was on one-liners, the jokes had to be short (one or two sentences) and should not include long narrative arcs or dialogues between characters.
Finally, the gatherer would mark each pair of pun and alternative signs according to the nature of their ambiguous relation following the taxonomy provided by Hempelmann and Miller (2017): homophonic, homographic, both, or neither. This kind of annotation can help researchers to filter the data according to their study interests.
After defining the guidelines, we implemented a simple web interface used by two researchers to collect jokes from three different sources, which are presented in Table 1.

Source	Type	#Puns
Maiores e Melhores	Web blog	45
O Sagrado Caderno das Piadas Secas	Instagram page	752
UTC - Ultimate Trocadilho Challenge	YouTube channel	4,106
Total		4,903

Table 1: Number of puns collected from each source

"Maiores e Melhores[2]" is an entertainment Portuguese web blog with at least three articles on classic puns in the language. After that, we gathered jokes from "O Sagrado Caderno das Piadas Secas[3]," a famous Portuguese Instagram page with short funny jokes, most of which are puns; these had to be transcribed manually from posted images. Finally, for the Brazillian portion of the corpus, we reached out to the creators of "UTC[4]", a famous pun competition among professional comedians on Brazilian YouTube, and they willingly provided us with their repository, which included thousands of puns that we analyzed manually.
Comparing with the different datasets provided by the previous work of Gonçalo Oliveira et al. (2020), shown in Table 2, we can see that PUNTUGUESE is at least twice as large as their largest one (headlines) when comparing data with similar format (one-liners), the difference is even larger. We also highlight that our corpus has been manually curated according to the chosen definition of pun, whereas the previous corpus was created completely automatically from given sources. Moreover, our dataset has a better coverage of punning humor, as the authors of the previous corpus did not focus specifically on this sort of humor.

Corpus	#Humor	#Non-Humor
One-liners	700	700
Headlines	2,000	2,000
All	1,400	1,400

Table 2: Number of instances in the previous corpora for Humor Recognition in Portuguese. Source: Gonçalo Oliveira et al. (2020)

Furthermore, as mentioned previously, PUNTUGUESE has annotations for the pun and alternative signs, as well as which are their relation (according to the mentioned taxonomy by Hempelmann and Miller (2017)). We also did a basic annotation of possibly problematic jokes, i.e. texts that may perpetuate negative stereotypes or deal with sensitive content, which is better discussed in section 7. The number of signs that fall in each category of the taxonomy, along with the number of problematic jokes, are shown in Table 3. It is important to note that the quantities are related to punning signs; as there are puns with more than one punning sign, the total number is higher than the number of jokes gathered.

Type of pun	Quantity
Only homophonic	953
Only homographic	10
Both homophonic and homographic	672
Not homophonic nor homographic	3,352
Problematic jokes	106

Table 3: Quantity of signs of each category according to the pun taxonomy alongside the number of problematic puns from an ethical point of view.

We can see that homographic-only jokes are extremely rare, as in Portuguese there are few words that are written the same but pronounced differently. We can also see that jokes that are not homophonic nor homographic are, by far, the most common ones. These are jokes that use punning signs that sound or look similar but not the same as their alternative signs; this phenomenon can be due to a substantial part of the puns being built using neologisms. Some examples of each type of joke can be seen in Table 4.
Finally, we highlight that the data is publicly available[5]; however, we open up only a fraction (50% = 2,053) of the jokes from UTC, as requested by their authors. Nonetheless, we highlight that this fraction still has more examples than the previous available dataset of one-liners (Table 2).

[2] https://www.maioresemelhores.com/
[3] https://www.instagram.com/osagradocaderno/
[4] https://www.youtube.com/@castrobrothers/
[5] https://github.com/Superar/Puntuguese

Homophonic	Homographic	Pun	Comment
✓	✗	Porque é que os polícias não gostam de sabão? Porque preferem deter gente. (*Why do the policemen do not like soap? Because they prefer arresting people.*)	This pun is funny because the verbal phrase "deter gente" (*arrest people*) sounds exactly the same as "detergente" (*detergent*).
✗	✓	Para que é que se plantam garfos? Para depois colher. (*Why does one plant forks? To harvest later.*)	This pun takes advantage of the polysemy of the word "colher", meaning either *spoon* (pronounced [ku.ʎˈɛɾ]) or *to harvest* (pronounced [ku.ʎˈeɾ]).
✓	✓	Um homem ia-se mandar dum prédio, passa um físico lá em baixo: Não faça isso! Você tem muito potencial! (*A man was about to jump from a building, a physicist passed below: Don't do that! You have a lot of potential!*)	This joke uses the multiple meanings of the word "potencial" (*potential*), meaning either unrealized abilities or a specific kind of energy studied in the field of Physics.
✗	✗	A pessoa que inventou o autocorrect devia arder no inverno. (*The person who created the AutoCorrect should burn in the winter.*)	The funny effect is created through the word "inverno" (*winter*) which sounds similar to "inferno" (*hell*).

Note: The phonetic transcriptions were obtained from the dictionary by Ashby et al. (2012) using the standard Lisbon pronunciation (http://www.portaldalinguaportuguesa.org/index.php?action=fonetica).

Table 4: Examples of jokes of each type (homophonic, homographic, both, or none).

3.2. Corpus Micro-editing

After collecting examples of punning humor, we needed to gather — or create — negative instances for the corpus while avoiding the problems raised by Inácio et al. (2023). For this, we resorted to the Humicroedit methodology of making minimal editions to the texts (Hossain et al., 2019). However, instead of turning non-humorous headlines into funny ones, we edited our puns so that they lost their comic effect. This way, we believe that PUNTUGUESE should be considerably more difficult than the corpus by Gonçalo Oliveira et al. (2020), as the classes differ exclusively in terms of humor and not much in textual form. Still, we believe that models trained in this harder corpus will effectively be better at recognizing punning humor.

We split the texts evenly across 18 fluent speakers of Portuguese (eight from Portugal and ten from Brazil). All annotators graduated either in Computer Science or Linguistics, and most of them are researchers in NLP or work with related areas. Each editor received only jokes from their specific variety of Portuguese, so we are more confident that they would know specific expressions and cultural aspects (e.g. celebrity names) in the texts.

Similarly to the Humicroedit process, we merged some tokens, namely named entities, acquired through Spacy[6], and the punning signs obtained during the gathering process. Finally, the editors were provided with a web interface developed with Streamlit[7] — a simple Python library to create data visualization and manipulation web apps — and a small set of instructions:

- Make the minimum amount of editions (preferably one);
- The new text must make sense (to ensure grammaticality);
- The new text must not be funny;
- Focus on editing open-class words (nouns, adjectives, verbs, and adverbs);
- Edit other kinds of tokens to ensure grammaticality.

[6]We used the `pt_core_news_sm` model. https://spacy.io/models/pt

[7]https://streamlit.io/

Original	Edited	Comment
Qual é a sobremesa mais popular na Rússia? O Putin flan. (*What is the most popular dessert in Russia? The flan Putin.*)	Qual é a sobremesa mais popular na Rússia? O pudim flan. (*What is the most popular dessert in Russia? The flan pudding.*)	The joke takes advantage of the phonetic similarity between Russia's president's name "Putin" and the Portuguese word for pudding "pudim".
Um parto não costuma demorar muito tempo. Mas para as grávidas parece maternidade. (*A childbirth doesn't usually take long. But for pregnant women, it feels like motherhood.*)	Um parto não costuma demorar muito tempo. Mas para as grávidas parece uma eternidade. (*A childbirth doesn't usually take long. But for pregnant women, it feels like an eternity.*)	This pun uses the pronunciation similarity between "uma eternidade" (*an eternity*) and "maternidade" (*motherhood*) to create the humorous effect. Mentioning first the time duration of childbirth also creates an expectation for the word "uma eternidade", which is then broken by "maternidade".
Qual cantora superou seu deficit de atencao? Rita Li-na (*Which singer overcame her attention deficit? Rita Li-na.*)	Qual cantora superou seu deficit de atencao? Ana Carolina (*Which singer overcame her attention deficit? Ana Carolina.*)	The original joke uses the name of a famous Brazilian singer, Rita Lee, which sounds similar to a common attention deficit medication Ritalina.

Table 5: Examples of edited jokes in the corpus.

We also provided some examples for the editors to facilitate their understanding of the task. Such examples are shown in Table 5.

All 4,903 jokes have been edited following this methodology. The distribution of the editions is shown in a log-scale histogram (Figure 1). As we can see, the vast majority of jokes had exactly one edition (around 4,300), with an average edition rate of 1.184, these counts include both changing or deleting a word. When looking exclusively at deletions, only a small set of jokes had deleted tokens (230), also with a low average rate (1.5).

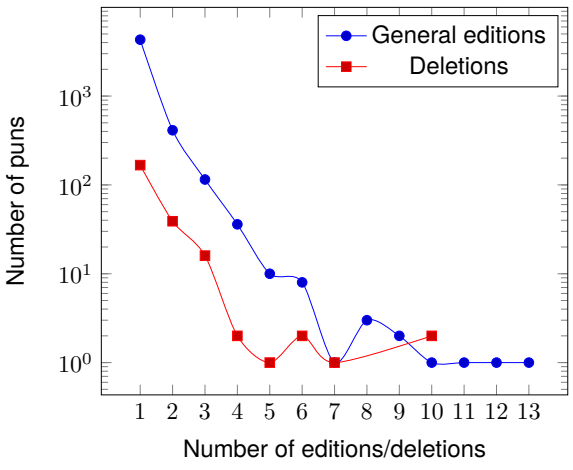

Figure 1: Log-scale histogram of the number of editions and deletions in puns

4. Editions Analysis

Since this work is inspired by Hossain et al. (2019), we attempted to replicate their clustering-based analysis to obtain a general overview of the editions and which words are used for the humorous effects. To this extent, we used BERTimbau (Souza et al., 2020), which creates embeddings for each subtoken in the texts, i.e. each token is comprised of one or more subtokens, each of which with its vectorial representation. These embeddings, obtained from the model's last layer, are then averaged into a vectorial representation for each token. It is worth mentioning that we decided to use a BERT model, instead of GloVe, to avoid out-of-vocabulary words — which are common in punning humor — due to its subword tokenization and positional encoding. The embeddings for each of the edited tokens (both in the original and edited versions of the texts) were then clustered using KMeans ($K = 20$ for each experiment). Finally, the clusters were manually analyzed and are described in Tables 6 and 7.

Differently to Hossain et al. (2019), we did not obtain many well-defined clusters according to the semantic field of the words, except for some specific groups, e.g. Profanity, Food, Car-related vocabulary, and Artists and football players.

From the rest of the cluster analysis, it is evident that these groupings appear to encompass a variety of semantic elements. The most prominent clusters, concerning the original tokens, include fictitious names derived from real proper nouns, such as "Britney Espinhas" (from *Britney Spears*),

Original tokens		
Cluster Name	Examples	Size
Neologisms with proper nouns	*Britney Espinhas*; *Semsunga*; *Slipknor*	474
Neologisms with proper nouns	*Unimedium*; *Tamarinheira*; *Orando Tango*	450
Neologisms with proper nouns	*Auxílio Moro Adia*; *Sérgio Molho*; *Gilberto Giz*	388
Neologisms	*Em Gana*; *Fricção Científica*; *Desinfestante*	304
Generic common nouns and neologisms	*carocho*; *muesli*; *d'abelha*	297
Neologisms, functional words, and common nouns	*massa corrida*; *Ed it mais cedo*; *vão*	283
Neologisms, functional words, common nouns, and MWEs	*caneta*; *acesa*; *dar uma voltinha*	269
Generic common nouns	*antônimo*; *assustador*; *pirão*	193
Functional words and MWEs	*os*; *um*; *dos*	186
Neologisms	*Y-amaha*; *for, miga*; *Ah, tá. Cama*	173
Generic proper nouns	*Fatkovic*; *Dolce & Cabana*; *Wolverine*	170
Generic common nouns	*bonito*; *cerveja*; *pés*	158
Generic common and proper nouns	*sentinela*; *pontinho*; *vinho*	146
Neologisms	*laptopspirose*; *sixpack*; *Orirental*	136
Words with capital letters	*BYEanas*; *VOLVOrine*; *CaranGUEIXA*	68
Profanity	*bunda*; *pornô*; *preservativo*	24
Generic proper nouns	*LG*; *Ásia*; *Cabo*	14
Positive sentiment	*preferida*; *favorito*; *preferido*	10
Generic verbs	*dá*; *tem*; *faz*	7

Table 6: Description of the clusters for the edited tokens in the original texts (puns).

"Semsunga" (*Samsung*), "Hallspadinha" (*Halls*), "Gilberto Giz" (*Gilberto Gil*), and "Canivete Sangalo" (*Ivete Sangalo*). Furthermore, clusters featuring generic words and neologisms formed by the amalgamation of other words, such as "massa corrida" (*plaster*), "vão" (*they go*), "inverso" (*inverse*), "passar elas" (*to pass them*, sounds like *runway*), and "fotografilha" (amalgamation of *photography* and *daughter*) stand out. There are also clusters comprising functional words, common nouns, and multi-word expressions (MWEs), such as "caneta" (*pen*), "acesa" (*lit*), "dar uma voltinha" (*enjoy a ride*), and "perdeu a linha" (*lose composure*), as well as clusters that group neologisms and common nouns, such as "carocho" (*Portuguese dogfish*), "muesli," "béé" (sound of a bleating sheep), "empato" (*I tie a game*), and "d'abelha" (*the bee's*).

As for the clusters of edited tokens, the most representative one consists of capitalized common nouns and proper nouns, including "Pantufas" (*slippers*), "Pochete" (*fanny pack*), "Latam" (name of a brazilian air company), and "Pokémon." Finally, we have identified clusters containing common nouns, such as "interessante" (*interesting*), "matemático" (*mathematician*), "treinos" (*trainings*), and "ouvidos" (*ears*), as well as clusters encompassing functional words, common nouns, and MWEs, such as "salvam" (*they save*) "lê" (*he/she reads*), "Eu te amo" (*I love you*), "um" (article *a*), and "não aprende" (*he/she does not learn*).

Despite the generic results, we have chosen to report this experiment to ensure that we replicate most of the work of the Humicroedit corpus.

5. Humor Recognition

As mentioned earlier, the main goal we expect to achieve with this corpus is to overcome the problems made explicit by Inácio et al. (2023), namely guaranteeing that the dataset is up to par with the expected complexity required to solve such a difficult task that is Humor Recognition, and not a corpus in which one can achieve 99% F-Score with a simple BERT model.

5.1. Classification with both varieties

First, we used the same classification methods provided by the previous authors[8] in our dataset to assess if it is indeed harder.

For the classification, we test two feature sets: content and humor-related features. Content features are TF-IDF counts of the 1,000 most common tokens, while humor-related features consider

[8] https://github.com/Superar/HumorRecognitionPT

Edited tokens		
Cluster Name	Examples	Size
Generic common nouns	*interessante*; *matemático*; *Caixão para homem.*	469
Generic common nouns	*de tendinite*; *ingênuo*; *beco*	429
Proper and capitalized common nouns	*Pantufas*; *Latam*; *Sacerdote*	424
Common nouns, functional words, and MWEs	*salvam*; *Eu te amo*; *um*	376
Proper and capitalized common nouns	*Unimed*; *era o Ibirapuera*; *Leptospirose*	365
Generic common nouns	*renal*; *mamífero*; *goleiro*	314
Generic proper nouns	*Alasca*; *Pepino di Capri*; *Everest*	232
Interjections and capitalized common nouns	*Em Portugal*; *Meu Deus*; *Sai de Baixo*	209
Generic proper nouns	*Alexandre Pato*; *Ciro Gomes*; *João Silva*	206
Common nouns, proper nouns, and MWEs	*Flor do Jardim*; *bolsa*; *mulher*	203
Proper and capitalized common nouns	*Água*; *John*; *Finlândia*	179
Food	*Chokito*; *Milk Shake*; *Hortifruti*	177
Proper female nouns	*Anitta*; *Andorinha*; *Barbie*	112
Artists and football players	*Luan*; *Alceu Valença*; *Zeca baleiro*	110
Car-related vocabulary	*BWM*; *Corsa*; *triciclo*	40
Functional words	*a*; *na*; *para toda a*	33
Personal relationship	*namorada*; *mestre*; *companheiros*	28
Negation and affirmation words	*Nada*; *sim*; *Nenhum*	27
Unrelated words	*vai*; *afilhado*; *Oi, gatinha*	4
Romeo	*Romeu*	2

Table 7: Description of the clusters for the edited tokens in the edited texts (non-puns).

many textual aspects, such as named entities, ambiguity, out-of-vocabulary words, imageability, concreteness, and others (Gonçalo Oliveira et al., 2020). The feature sets were tested using the Random Forest algorithm, reported by Inácio et al. (2023) as their best results. We also tested a fine-tuned BERTimbau model with a classification head (Souza et al., 2020). The average results across a 10-fold cross-validation are shown in Table 8.

Method	F1-Score
PUNTUGUESE	
Content features	17.9%
Humor features	35.1%
Content + humor features	27.0%
BERTimbau	68.9%
Inácio et al. (2023)	
Content features	96.4%
Humor features	78.8%
Content + humor features	97.1%
BERTimbau	99.6%

Table 8: Classification results.

From the table, we can observe that PUNTUGUESE is significantly more difficult than the previous corpus, especially for the content features (from 96.4% to 17.9%, a decrease of approximately 81%), which, according to Inácio et al. (2023), achieved high performance by resorting mainly on full stops, question marks, and question words. Since our dataset was constructed to preserve these general characteristics across classes, such features are arguably less relevant for classification; a future explainability analysis can attest to this claim more confidently. It is also possible to see that BERTimbau still achieved the best performance across all methods, but noticeably lower than with the previous dataset (an approximate decrease of 31%). This result indicates that transformers can be a promising path for research, but are still far from perfect for dealing with complex phenomena such as humor.

5.2. Classification by variety

In addition, we analyzed the same cross-validation results by grouping the test splits by variety to check if the fact that the Brazilian Portuguese data is dominating impacts the results. The only larger difference we observed was regarding the BERTimbau

model, which had an average F1-Score of 71.8% for Brazilian Portuguese and 53.6% for European Portuguese. As the other feature sets and methods were not so dissimilar, we attribute this difference to the model used: BERTimbau was pre-trained exclusively for Brazilian Portuguese.

In this sense, we decided to carry out a new 10-fold cross-validation evaluation, but this time by isolating the varieties. We also include newly released pre-trained models from the Albertina PT-* project (Rodrigues et al., 2023), which provides pre-trained DeBERTa (He et al., 2021) models for both varieties of Portuguese. This approach ensures that each variety is used to fine-tune a model optimized for its specific linguist nuances. These results can be seen in Table 9.

Method	F1-Score	
	PT-BR	PT-PT
Content features	19.0%	15.2%
Humor features	35.3%	26.9%
Content + humor features	27.0%	15.3%
BERTimbau	71.8%	53.6%
Albertina PT-*	72.1%	62.0%

Table 9: Classification results for each variety.

It is worth noting that we used Albertina PT-* base models as they are more comparable in size to BERTimbau base — all of them have around 100M parameters. Also, Albertina PT-* models did not provide satisfactory results with the same hyperparameters used for BERTimbau by Inácio et al. (2023) (3 epochs, with starting learning rate of 5×10^{-5}). Therefore, we fine-tuned it using the same parameters used by Rodrigues et al. (2023) for the ASSIN 2 entailment detection task (Real et al., 2020) and the PLUE benchmark (Gomes, 2020): 5 epochs with learning rate 1×10^{-6}.

In these experiments, we can see that the performance for European Portuguese (PT-PT) is consistently lower compared to Brazilian Portuguese (PT-BR), which is expected since it has less training data. By comparing the results inside a single variety, we can attest that transformer-based language models are surely more powerful, as largely discussed in recent years (Devlin et al., 2019; Siekiera et al., 2022; Zhang et al., 2023; Hessel et al., 2023). For PT-BR, we see that Albertina PT-BR is slightly better (72.1%) than using BERTimbau (71.8%); however, it required more epochs and some hyperparameter tuning. On the other hand, for PT-PT, Albertina PT-PT poses an advantage (62.0% compared to 53.6% for BERTimbau), because the model is pre-trained with texts from this specific variety of the language; the performance of Albertina PT-PT is still better than BERTimbau with a larger mixed training set, as mentioned earlier (53.55%).

Such results, although interesting, show that we still have a long way to develop new methods able to deal with such a complex and difficult phenomenon which is humor. We believe that this new corpus, from the way it was created, can pose a motivating challenge for understanding and creating systems that can deal with humor and figurative language.

5.3. Machine Learning Explainability

To attest if the new corpus is indeed better than the previous one in terms of Machine Learning, we conducted a preliminary study using SHAP (Lundberg and Lee, 2017), the same explainability method used in Inácio et al. (2023) to find the flaws of the previous dataset. Details of this experiment are reported in Inácio and Gonçalo Oliveira (2024). In sum, by looking at the 150 most influential tokens for classification with the BERTimbau model, depicted in Figure 2, we observed that no token dominates the classification process, especially question words and punctuation[9].

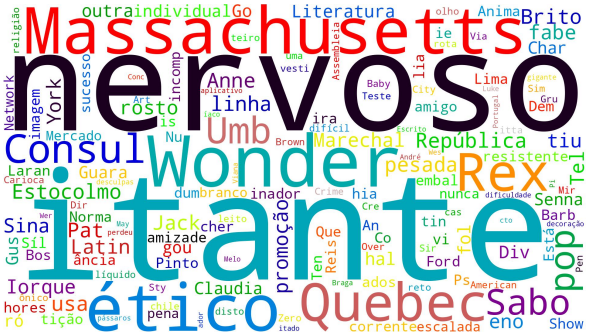

Figure 2: 150 most important tokens for humor recognition with BERTimbau. Source: Inácio and Gonçalo Oliveira (2024)

As mentioned by Inácio and Gonçalo Oliveira (2024), there is more variety of important tokens compared to the previous corpus, including some punning signs and punchline words, such as "O que escrevem no placar quando o Elvis joga fora de casa? Elvisitante" (*What do people write on the scoreboard when Elvis plays away from home? Elvisitor.*). The image also shows some words that are part of the edited passages, as in "Qual é o estado americano que não cai duas vezes no mesmo lugar? Massachusetts." (*Which American state does not fall twice in the same spot? Massachusetts.*).

These observations suggest that the creation methodology for PUNTUGUESE adresses some data leakage issues present in the previous corpus. However, we highlight the importance of conducting more comprehensive research in this regard.

[9]To provide a sense of scale, we mention that "nervoso" has a score of 0.485, while "Consul" scores 0.464.

6. Conclusion

This work aimed at creating PUNTUGUESE, a new corpus of humorous texts in Portuguese to overcome known limitations of the single current available dataset (Inácio et al., 2023).

The new corpus contains 4,903 puns in both European and Brazilian Portuguese, classified according to the taxonomy of Hempelmann and Miller (2017), and with pun and alternative signs highlighted, which can be used in future research.

To ensure that PUNTUGUESE is suitable for Machine Learning — guaranteeing that algorithms learn to identify humor — we carried out a Humicroedit-like methodology (Hossain et al., 2019) to create negative instances through micro-edits. In this sense, each pun has a non-humorous counterpart that is similar but differs only in terms of its humor-inducing effects. This approach resulted in a dataset that reached a maximum F1-Score of 68.9% by using a method that reached 99.6% on the previous corpus, indicating that it can be a strong benchmark to test Humor Recognition systems for Portuguese.

This new corpus opens up many paths for future research and the development of new resources. The puns could be ranked or graded according to their level of funniness, sentiment, or emotion. The editions can be further analyzed to understand what choices were made by the editors to make the texts unfunny, which can help better understand how verbal humor is constructed and perceived, including from a cultural perspective of comparing Brazilian and Portuguese puns. In this sense, after analyzing thousands of puns, we observed some interesting patterns used for the construction of the jokes (e.g. agglutination or juxtaposition of two existing words, neologisms that resemble the sound of a foreign language, and others); a deep investigation can motivate the creation of a new taxonomy for puns or the expansion of existing ones (Hempelmann and Miller, 2017; Aleksandrova, 2022).

Naturally, as we are working in the field of Natural Language Processing, the corpus spurs the development of better methods to deal with not only humor recognition but also generation. The fact that we made the pun and alternative signs explicit can also lead to the development of Pun Disambiguation models, which aim at deducing the alternative meaning for the punning sign present in the text (Miller and Gurevych, 2015).

With this work, we expect to motivate research not only for Computer Science and computational systems but also for the fields of Linguistics, Psychology, Sociology, and many others that can bring different points of view on such an interesting phenomenon that is humor.

7. Ethical Considerations

As discussed in the Humanities, the ambiguous nature of humor enables it to be used as a way to insult and disrespect people in subtle ways. It can also be used to spread and legitimize dynamics of prejudice, power, and oppression in society (Crawford, 2003; Bemiller and Schneider, 2010). Therefore, it is of utmost importance to raise concerns about such tendencies within our data.
During the data collection process (subsection 3.1), the gatherers were instructed to keep in a note which jokes they would personally consider "problematic", i.e. jokes that could harm minority groups, reinforce social prejudice, deal with delicate subjects (for instance suicide or genocide), and any other criteria to their choice.
In this sense, alongside the corpus, we publish a list of identifiers, indicating which jokes fall under this category, so that we make an effort to mitigate this aspect in PUNTUGUESE. On the other hand, as we did not carry out a full annotation task taking into consideration multiple points of view (Rosso, 2023), we still include all collected jokes within the general data, but the provided list makes it simple to filter out such texts.
We also highlight that, as the list was created by only two people with very specific demographics, it is possible (and probable) that we still missed jokes that could fall under this category.

8. Limitations

One of our main concerns about the corpus is its limitation regarding the representation of not only the European variety of the Portuguese language but also from other countries, such as Angola and Cape Verde. We further acknowledge that the lack of annotation regarding the level of funniness, which is common in verbal humor corpora, which can limit its usefulness in some scenarios.
Besides, the identification of punning and alternative signs, as well as the classification according to the pun taxonomy, was done independently by the two gatherers, without calculating agreement scores. This occurred first because these were simply hints to identify if a text was a pun; however, we consider this kind of information too valuable to keep it out of the corpus.
Finally, the most significant limitation of this work is that a joke was exclusively edited by a single person. Consequently, we can only ensure that the non-funny texts are considered as such by their respective editors, given the inherently subjective nature of humor. Ultimately, we can refer to these texts as non-puns, since we can follow a more objective definition, as described in subsection 3.1. We believe that additional efforts to collect funniness ratings can assure that negative instances are in fact less funny.

9. Acknowledgements

We would like to give thanks to Castro Brothers and their staff, who gladly shared their puns spreadsheet with us for research, with special thanks to Marcos Castro. We also thank all other researchers that took part in the task force for editing the jokes: Ana Alves, Ana Carolina Rodrigues, Ana Caroline M. Brito, André C. Santos, Bruno Ferreira, Henrico Brum, Isabel Carvalho, and Patrícia Ferreira.
This work is funded by national 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 supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI.

10. Bibliographical 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 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Simone Ashby, Sílvia Barbosa, Silvia Brandão, José Pedro Ferreira, Maarten Janssen, Catarina Silva, and Mário Eduardo Viaro. 2012. A rule based pronunciation generator and regional accent databank for portuguese. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pages 1886–1887. ISCA.

Michelle L. Bemiller and Rachel Zimmer Schneider. 2010. IT'S NOT JUST A JOKE. *Sociological Spectrum*, 30(4):459–479.

Emily M. Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters.

Erik Cambria and Bebo White. 2014. Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine*, 9(2):48–57.

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.

Mary Crawford. 2003. Gender and humor in social context. *Journal of Pragmatics*, 35(9):1413–1430.

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.

J. R. S. Gomes. 2020. PLUE: Portuguese Language Understanding Evaluation. https://github.com/ju-resplande/PLUE.

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. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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.

Christian F. Hempelmann and Tristan Miller. 2017. Puns: Taxonomy and Phonology. In *The Routledge Handbook of Language and Humor*, 1 edition, Routledge Handbooks in Linguistics, pages 95–108. Routledge.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest. In *Proceedings of the 61st Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "President Vows to Cut Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines. In *Proceedings of the 2019 Conference of the North*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcio Inácio and Hugo Gonçalo Oliveira. 2024. Exploring multimodal models for humor recognition in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 568–574, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

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.

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

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.

Tristan Miller and Iryna Gurevych. 2015. Automatic disambiguation of English puns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 719–729, Beijing, China. Association for Computational Linguistics.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 Task 7: Detection and Interpretation of English Puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The ASSIN 2 Shared Task: A Quick Overview. In Paulo Quaresma, Renata Vieira, Sandra Aluísio, Helena Moniz, Fernando Batista, and Teresa Gonçalves, editors, *Computational Processing of the Portuguese Language*, volume 12037, pages 406–412. Springer International Publishing, Cham.

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-*.

Roberto Labadie Tamayo y Berta Chulvi y Paolo Rosso. 2023. Everybody hurts, sometimes overview of HUrtful HUmour at IberLEF 2023: Detection of humour spreading prejudice in twitter. *Procesamiento del Lenguaje Natural*, 71(0):383–395.

Julia Siekiera, Marius Köppel, Edwin Simpson, Kevin Stowe, Iryna Gurevych, and Stefan Kramer. 2022. Ranking Creative Language Characteristics in Small Data Scenarios. In *International Conference on Computational Creativity*, Bolzano-Bozen. Association for Computational Creativity (ACC).

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting Humorousness and Metaphor Novelty with Gaussian Process Preference Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728, Florence, Italy. Association for Computational Linguistics.

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.

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.

Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Huadong Wang, Deming Ye, Chaojun Xiao, Xu Han, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2023. Plug-and-Play Knowledge Injection for Pre-trained Language Models.