Improving Pun Detection with an Ensemble of Traditional Machine Learning Methods

Jhúlia de Souza Leal[1], Marcio Lima Inácio[2], Hugo Gonçalo Oliveira[2], Rafael Torres Anchiêta[3]

[1]Universisty of São Paulo – (USP)
Institute of Mathematics and Computer Science (ICMC)
São Carlos – SP – Brazil

[2]University of Coimbra
CISUC/LASI – Centre for Informatics and Systems of the University of Coimbra
Department of Informatics Engineering
Polo II, Pinhal de Marrocos
3030-290 Coimbra, Portugal

[3]Federal Institute of Maranhão – (IFMA)
Caxias – MA – Brazil

`lealjhu@gmail.com`,{`mlinacio`,`hroliv`}`@dei.uc.pt`,`rafael.torres@ifma.edu.br`

Abstract. *Humor is a remarkably complex emotional process, defined as any object or event that causes laughter or amusement or is considered funny. Therefore, recognizing humor is considered one of the most challenging tasks in Natural Language Processing. In this paper, we approached the pun detection task for the Portuguese language. Puns are a form of wordplay that exploits multiple meanings of a term or similar-sounding words to create an intended humorous or rhetorical effect. Our strategy is straightforward: we trained and evaluated an ensemble learning approach of traditional machine learning models on* PUNTUGUESE, *a recent corpus of Portuguese puns. With this, we outperformed a BERT-based model by 11 p.p. in accuracy and achieved state-of-the-art results. More than that, we performed a detailed error analysis and found that our approach has limitations in identifying puns that contain neologisms.*

1. Introduction

Computational humor recognition is considered one of the most challenging tasks in Natural Language Processing (NLP) since humor is a remarkably complex emotion [Kalloniatis and Adamidis 2024]. Humor may be defined as "any object or event that causes laughter, amusement, or is considered funny" [Attardo 2020], or the effect that would make the audience laugh, or even "a non-serious social disagreement" [Banas et al. 2011].

Although the humor recognition task is complex, enabling systems to detect it can have genuinely impactful outcomes, as evidenced by a documented failure of stock market algorithms to interpret an April Fool's joke press release[1]. In this example, Tesla

[1]`https://www.wsj.com/articles/BL-MBB-35151`

announced it is creating a rival to the Apple Watch, dubbed the Model W. But... it was a joke. From that announcement, the stock jumped about \$2 on heavy volume when the news hit shortly before the market closed.

According to Kalloniatis and Adamidis [Kalloniatis and Adamidis 2024], humorous content may appear in different forms, such as one-liners, narrative jokes, dialogue, multimodal ways, puns, and others. Puns, in particular, are a common source of humor. They are a form of wordplay that exploits multiple meanings of a term or similar-sounding words to create an intended humorous or rhetorical effect [Attardo 2020]. This duality in meaning makes puns particularly challenging for NLP models to detect and interpret accurately, as it may require understanding context, phonetics, and semantics [Gameiro et al. 2024]. For example, the sentences below express different types of puns. The first is based on how words sound similar but have different meanings and spellings. In the second, the words are spelled similarly but have different meanings. The third example includes two punny words in one statement, relying on the sound of two words blended to make the joke.

1. A pessimist's blood type is always **B-negative**.
2. Every calendar's days are **numbered**.
3. Everyone thinks my runny nose is funny, but **it's snot**.

The above examples highlight the difficulty of working with this kind of phenomenon. Pun detection refers to classifying a sentence or a text based on whether it contains a pun or not [Miller et al. 2017]. Identifying puns is an important research question and has some real-world applications, such as machine translation and computational education. In the first one, recognizing puns is important, particularly for sitcoms and other comedic works. This sort of wordplay may be complicated for non-native speakers to detect, let alone translate. In the second, computer-assisted detection and classification of puns could help digital humanists produce similar surveys of other works, as wordplay is a timeless scholarship topic in literary criticism and analysis [Miller et al. 2017].

Given the relevance of identifying puns, several authors have addressed this task in different languages, such as English [Monika and Vij 2019, Jaiswal and Monika 2019, Yatsu and Araki 2018, Kao et al. 2016, Yang et al. 2015], Spanish [Castro et al. 2018, Labadie Tamayo et al. 2023], and French [Ermakova et al. 2022, Ermakova et al. 2024]. However, this research area is still under-explored for Portuguese. In this paper, we approach the pun detection task for the Portuguese language to mitigate this gap.

For that, we used the PUNTUGUESE corpus [Inacio et al. 2024], a very recent and one of the only datasets focused on the Portuguese pun detection task. Based on this, we developed a straightforward, faster, and lower-cost computational strategy. We combined three classifiers into an ensemble learning approach and fed it with vectorized bi-grams using the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme. This strategy outperformed a BERT-based model by 11p.p. in accuracy, achieving state-of-the-art results, showing that classical Machine Learning models can still be up to par for this task. Moreover, we performed a detailed error analysis and identified that our approach misclassified 120 puns, 77 of which were neither homophonic nor homographic. This result shows that our model has difficulty detecting neologisms, opening doors for future work.

The remainder of this paper is organized as follows: Section 2 briefly presents related work. In Section 3, we introduced the corpus used for training and evaluating our approach. Section 4 details our strategy to detect puns. In Section 5, we reported and analyzed the achieved results. Finally, Section 6 concludes the paper and indicates future directions.

2. Related work

Humor recognition for Portuguese started to be explored by Clemêncio [Clemêncio 2019] and Gonçalo Oliveira et al. [Gonçalo Oliveira et al. 2020]. They created a corpus of Portuguese jokes and developed a set of humor-related features based on relevant literature to classify humor. Their strategies achieved an F1-score of 80% for one-liners and 76% for satirical headlines with the Random Forest classifier. Inácio et al. [Lima Inácio et al. 2023] fine-tuned the BERTimbau model [Souza et al. 2020] on the same corpus, surpassing the results of Clemêncio and Gonçalo Oliveira, achieving a 99.6% F1-score. However, after some machine learning explainability experiments with SHAP [Lundberg and Lee 2017], they found that such positive results were due to data leakage in the dataset; namely, the model relied on punctuation and the presence of questions to do the classification.

In addition to the humor recognition task, some authors worked on related tasks, such as irony detection [Carvalho et al. 2020, Corrêa et al. 2021, Anchiêta et al. 2021, Luz et al. 2023]. The strategies ranged from superficial features, such as TF-IDF, to deep learning. Moreover, there are works on the analysis of satirical news [Wick-Pedro and Santos 2021, Wick-Pedro et al. 2024], in which the authors analyzed the textual complexity of satirical and true news in Brazilian Portuguese and found a greater complexity in the authentic texts.

Specifically for the pun detection task, besides creating the PUNTUGUESE corpus to mitigate the problems found in the previous dataset, Inácio et al. [Inacio et al. 2024] also assessed several strategies to detect puns. The best-performing strategy was a fine-tuned BERTimbau model [Souza et al. 2020], with an F1-score of 68.9% in 10-fold cross-validation. In a posterior work [Inácio and Gonçalo Oliveira 2024], three methods of multimodal transformers were explored to combine transformer-based representations with humor-related features from the literature [Clemêncio 2019]. Their results did not improve over their previous approach.

For the English language, the most researched one, recent works mainly focus on deep learning approaches [Diao et al. 2018, Diao et al. 2019, Ren et al. 2021], using Bidirectional Long Short-Term Memory (Bi-LSTM) networks with attention mechanisms and language models [Zou and Lu 2019, Zhou et al. 2020, Xu et al. 2024] for the pun detection task.

Our strategy is more straightforward and faster than these approaches. It does not require high computational resources and is based on traditional supervised learning algorithms, such as Random Forest, Logistic Regression, and Support Vector Machine. In what follows, we present the corpus used to train and evaluate our approach.

3. Corpus

PUNTUGUESE is a curated collection of punning texts in Brazilian and European Portuguese with its public portion containing 2,850 puns, 2,053 attributed to Brazilian Portuguese, and 797 to European Portuguese [Inacio et al. 2024]. Each pun was manually annotated with its punning mechanisms: the pun words (i.e., the triggers for the text to be considered a pun) and their alternative words (what other meanings these triggers have). Moreover, every pair of punning and alternative words in each joke is classified according to their lexical relationship, whether they are homographs or homophones. In addition to the punning texts, PUNTUGUESE includes a non-humorous counterpart for each entry, created through micro-editing, making the corpus parallel and balanced. Table 1 presents an example of puns in European and Brazilian Portuguese, alongside their corresponding non-punning text.

Table 1. Examples of puns and non-puns in the PUNTUGUESE corpus. Punning words are in bold. Edited words to create the non-punning texts are underlined.

Language variety	**Pun**	**Non-Pun**
Brazilian	*Qual o nome do filho do Mc **Kevinho**? MC **Kessuco**.* (What is the name of Mc **Kevinho's** son? MC **Kessuco**.)	*Qual o nome do filho do Mc Kevinho? Marcelo.* (What is the name of Mc Kevinho's son? Marcelo.)
European	*Hoje vi o Jorge Jesus num anúncio de detergentes em que ele dizia: Este é o melhor detergente que **tenho Tide**!* (Today, I saw Jorge Jesus in a detergent advert in which he said: This is the best detergent **I have had!**)	*Hoje vi o Jorge Jesus num anúncio de detergentes em que ele dizia: Este é o melhor detergente que tenho!* (Today, I saw Jorge Jesus in a detergent advert in which he said: This is the best detergent I have!)

From Table 1, the European Portuguese pun requires knowledge of both regional accents and cultural context, as "tenho Tide" ("I have Tide", a detergent brand) sounds like "tenho tido" (meaning "I have had"), and the pun involves Jorge Jesus, a well-known Portuguese football coach, whose accent mostly fits with the replacement "tido" $\rightarrow$ "tide". Similarly, the Brazilian Portuguese pun relies on cultural background and linguistic nuances. MC Kevinho is a Brazilian funk musician, and the pun plays on his name: "Kevinho" sounds like "Quer vinho" (meaning "Want wine"), and "Kessuco" sounds like "Quer suco" (meaning "Want juice"), when pronounced with a Brazilian accent.

The PUNTUGUESE dataset is split into training (70%), testing (20%), and validation (10%) subsets. It uses a stratified sampling approach to maintain an even distribution of language varieties, as presented in Table 2.

In the following section, we detail our strategy to detect puns.

4. Ensemble strategy

To deal with the pun detection task, we developed a pipeline based on three steps: preprocessing, modeling, and evaluation.

Table 2. Distribution of the PUNTUGUESE corpus.

Language variety	Train	Val	Test	Total
Brazilian	1,437	206	410	2,053
European	558	79	160	797
Total	1,995	285	570	2,850

For the preprocessing stage, we removed stopwords using the Portuguese list from the Natural Language ToolKit (NLTK) [Bird et al. 2009]. Next, we converted the text into bigrams and vectorized them, applying the TF-IDF weighting scheme [Sammut and Webb 2011] from the scikit-learn library [Pedregosa et al. 2011]. The vocabulary of the document-term matrix, which has a size of 23,769, was learned from the training set of the PUNTUGUESE corpus.

After preprocessing, we combined three classifiers, Random Forest, Logistic Regression, and Support Vector Machine, from the scikit-learn library [Pedregosa et al. 2011] into an ensemble learning strategy. We chose these classifiers empirically and adjusted their parameters based on the grid search algorithm. For Random Forest, we used the value for the `criterion` parameter equal to `entropy`. For Logistic Regression, we used the default parameters, and for Support Vector Machine, we used the `probability` parameter equal to `true`.

We combined these three classifiers through the voting classifier. This machine-learning model gains experience by training on several models and forecasts an output (class) based on the class with the highest likelihood of becoming the output. We adopted the soft voting classifier type to predict the output class. This classifier computes the average probabilities of the classes given by the base models to determine which one will be the final prediction.

Following the modeling of the pun detection task as an ensemble learning approach, we evaluate our strategy on the test set of the PUNTUGUESE corpus and compare it with BERT-based approaches. The results obtained and the analysis performed are detailed below.

5. Results and Analysis

To better understand the results of this work, we organized them into three subsections: results with single supervised learning approaches, results with the ensemble method, and a manual analysis of predictions.

5.1. Traditional Machine Learning Models

As shown in Table 3, we initially evaluated the Logistic Regression, Random Forest, and Support Vector Machine classifiers individually.

As we can see from this table, all the classifiers obtained a poor result for the pun detection task when analyzed separately, with a maximum of 40% accuracy using the Random Forest method. These results align with those obtained by the original PUNTUGUESE authors, who reported a maximum of 17.9% average F-Score using a Random Forest algorithm with TF-IDF features. Yet, we still achieved a better result for the same

Table 3. Results for each classifier separately.

Classifier	Class	Precision	Recall	F-score	Accuracy
Logistic Regression	Not a pun	0.23	0.25	0.24	0.21
	Pun	0.18	0.17	0.18	
Random Forest	Not a pun	0.40	0.41	0.41	0.40
	Pun	0.40	0.39	0.39	
Support Vector Machine	Not a pun	0.21	0.21	0.21	0.20
	Pun	0.20	0.20	0.20	

algorithm (40% average F-Score), which we attribute to having a more complete preprocessing pipeline than the previous work. This improvement was achieved despite Inácio et al. [Inacio et al. 2024] utilizing a larger version of PUNTUGUESE that includes a private dataset portion.

5.2. Ensemble Learning

After combining these classifiers into an ensemble learning approach, the results improved significantly. Table 4 shows the results obtained by the voting classifier compared to the BERTimbau model, as described by Inácio et al. [Inacio et al. 2024][2], trained and tested in the same standard PUNTUGUESE splits. We also compare the results with larger models (with 900M parameters) from the Albertina PT-* family [Rodrigues et al. 2023], reported by Inácio and Gonçalo Oliveira [Inácio and Gonçalo Oliveira 2024]. The authors evaluated the Albertina PT-* models via cross-validation on PUNTUGUESE and, although the models are not publicly available, their predictions are[3]; therefore, the table shows the average scores across all folds. Furthermore, we also compare the results with the multilingual versions of BERT [Devlin et al. 2019] and DeBERTa [He et al. 2023].

Table 4. Comparison of results between the voting classifier and BERT.

Approach	Class	Precision	Recall	F-score	Accuracy
Ours	Not a pun	0.79	0.81	0.80	0.80
	Pun	0.80	0.79	0.80	
BERTimbau [Inacio et al. 2024]	Not a pun	0.67	0.77	0.71	0.69
	Pun	0.73	0.61	0.67	
Albertina PT-BR [Inácio and Gonçalo Oliveira 2024]	Not a pun	0.50	0.51	0.50	0.50
	Pun	0.50	0.48	0.48	
Albertina PT-PT [Inácio and Gonçalo Oliveira 2024]	Not a pun	0.50	0.52	0.51	0.50
	Pun	0.50	0.48	0.48	
mBERT [Devlin et al. 2019]	Not a pun	0.64	0.69	0.67	0.65
	Pun	0.67	0.62	0.64	
mDeBERTa V3 [He et al. 2023]	Not a pun	0.74	0.83	0.78	0.77
	Pun	0.81	0.71	0.75	

[2] https://huggingface.co/Superar/pun-recognition-pt
[3] https://github.com/Superar/multimodal-humor-recognition

One can see that the ensemble strategy outperformed the BERTimbau model by 11 p.p. (17.65%) in accuracy, from 0.69 to 0.80, suggesting that classifications by the traditional algorithms are somehow complementary. Moreover, Inácio et al. [Inácio and Gonçalo Oliveira 2024] did not achieve solid results, not surpassing 50% accuracy. Furthermore, our approach also surpassed two fine-tuned multilingual models: BERT and DeBERTa V3 base models. DeBERTa enhances the BERT and RoBERTa [Liu et al. 2019] models by employing ELECTRA-Style pre-training [Clark et al. 2020] with Gradient Disentangled Embedding Sharing, achieving state-of-the-art results on most natural language understanding tasks. To fine-tune these models, we empirically defined the following hyperparameters, as shown in Table 5. For example, the batch size is 8, the loss function is cross entropy, the learning rate is 2×10^{-5}, and so on.

Table 5. Hyperparameters used to fine-tune mBERT and mDeBERTa V3.

Parameter	**Value**
Batch	8
Loss function	CrossEntropy
Learning rate	2×10^{-5}
Optimizer	AdamW
L2 regularization	0.01
Epoch	5

We also conducted a 5-fold cross-validation experiment to determine whether these results are an artifact of a specific train/val/test split. We performed ten executions and calculated the average of the metrics. The results obtained correspond with Table 4, indicating that they are not merely an artifact of a particular division.

It is important to highlight that our approach requires much less computational resources than language models. Furthermore, it is faster and simpler than Transformer-based models since it uses only traditional supervised machine learning algorithms and the TF-IDF weighting scheme.

5.3. Error Analysis

By analyzing the confusion matrix of the voting classifier (Table 6), we realize there is room for improvement since the model produced 111 false positives and 120 false negatives.

Table 6. Confusion matrix of the voting classifier.

		Actual	
		Pun	**Not pun**
Predicted	**Pun**	450	111
	Not pun	120	459

To understand the results obtained, we performed an error analysis on the 120 false negatives. We found that the types of misclassified puns were distributed as follows: 71 are neither homographs nor homophones, 24 are only homophones, 1 is only a homograph, and 26 are both homographs and homophones. It is important to say that these numbers are related to punning signs, and since there are puns with more than one punning sign, the total number exceeds the number of jokes identified as false negatives. For example, Table 7 presents a pun that is neither a homograph nor a homophone and, simultaneously, only a homophone.

Table 7. Examples of jokes with different punning signs.

Homographic	Homophonic	Pun	Comment
✗	✗	*Qual é o contrário de* ***menu****? Youvestido.* (What is the opposite of **menu**? Youdress.)	The funny effect is created through the word "menu" (menu), which sounds similar to "me nu" (me nude).
✗	✓	*Qual é o contrário de menu?* ***Youvestido****.* (What is the opposite of menu? **Youdress**.)	In this case, the funny is created through the word "Youvestido" (you dress) that sounds exactly like "you vestido" (you dressed).

Observing the results of the ensemble learning approach, we can see that it has more difficulty detecting puns that are neither homographs nor homophones. These jokes use punning signs that sound or look similar but differ from their alternative signs, i.e., neologisms. Other puns with many misclassifications were only homophones (24) and homographs and homophones (26). Table 8 presents an example of these types of puns.

More than identifying the type of puns, we used the LIME tool [Ribeiro et al. 2016] to understand the results a little bit better. From this analysis, we realize that punning signs have zero weight when the puns are neither homographs nor homophones. For instance, in the first example in Table 8, the punning sign "inverno" (winter) has zero weight. We believe this occurs because this word is not frequent in the corpus. When the puns are homographs and homophones, and only homophones, the pun signs have negative weight, indicating that they are not puns. In Table 8, the pun signs "deter gente" (arrest people) and "potencial" (potential) have negative weights. We believe this is because these punning signs appear more frequently in non-jokes. We expect this analysis to help develop more robust methods for detecting puns in Portuguese.

Finally, since the individual models complement each other within the ensemble, we decided to analyze whether their correct classifications are somehow related to the pun types. In other words, we wanted to check if each model specializes in classifying a specific kind of pun. To this extent, we depict the distributions of correct classifications (true positives) for each pun type in Figure 1.

In the figure, we can see that the distributions are similar across the models, regardless of the type of pun being classified: Random Forest (RF) consistently outperforms the Linear Regression (LR) and SVM models. This shows that the features that the models are learning are not necessarily aligned with the underlying punning mechanism of the

Table 8. Examples of misclassified puns.

Homographic	Homophonic	Pun	Comment
✗	✗	*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).
✗	✓	Porque é que os polícias não gostam de sabão? Porque preferem deter gente. (Why do the policemen 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).
✓	✓	*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 when 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.

joke. We also highlight that "homographic only" jokes are extremely rare in the dataset; in fact, there is only one joke of such kind in the test portion of PUNTUGUESE.

6. Conclusion

In this paper, we tackled the challenging task of pun detection in Portuguese using an ensemble learning approach. By combining three classifiers, Random Forest, Logistic Regression, and Support Vector Machine, we significantly outperformed a BERT-based model, increasing accuracy by 11 p.p., which is now the state-of-the-art of this task for Portuguese. Our results show that, when properly optimized, a traditional supervised learning approach can be both efficient and effective, outperforming deep learning models in terms of computational cost and simplicity. More than that, we performed a detailed error analysis and found that our approach has limitations in identifying puns that contain neologisms. These findings suggest that incorporating additional linguistic features or leveraging more contextual information could enhance pun recognition in Portuguese.

In future work, we intend to explore hybrid models that integrate ensemble learning with deep learning techniques to improve the detection of puns. The source code used in our experimentation is publicly available at: `https://github.com/`

liara-ifpi/soltando-puns.

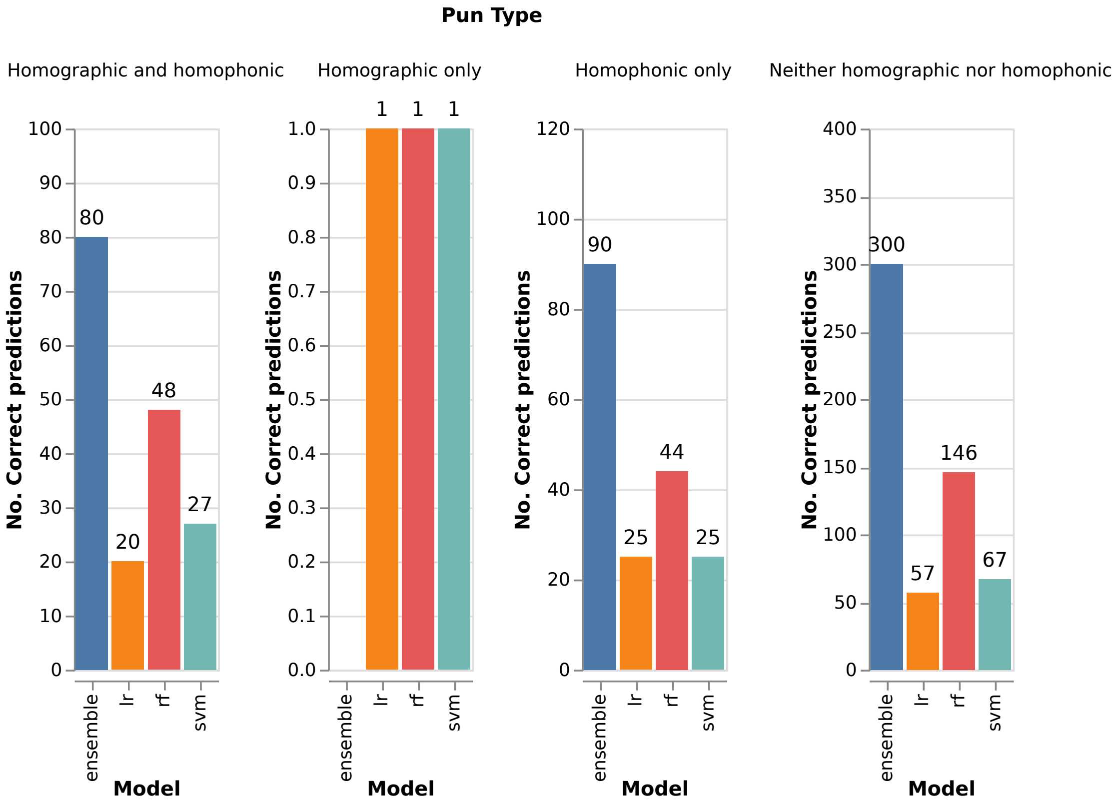

Figure 1. Distribution of true positive classifications for each model across pun types.

Acknowledgments

This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055, Center for Responsible AI; and by national funds through FCT – Foundation for Science and Technology I.P. (grant number UI/BD/153496/2022), in the framework of the Project UIDB/00326/2025 and UIDP/00326/2025.

References

Anchiêta, R. T., Neto, F. A. R., Marinho, J. C., do Nascimento, K. V., and Moura, R. S. (2021). Piln idpt 2021: Irony detection in portuguese texts with superficial features and embeddings. In *IberLEF@ SEPLN*, pages 917–924, Málaga, Spain. CEUR.

Attardo, S. (2020). *The linguistics of humor: An introduction*. Oxford University Press.

Banas, J. A., Dunbar, N., Rodriguez, D., and Liu, S.-J. (2011). A review of humor in educational settings: Four decades of research. *Communication Education*, 60(1):115–144.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Carvalho, P., Martins, B., Rosa, H., Amir, S., Baptista, J., and Silva, M. J. (2020). Situational irony in farcical news headlines. In *International Conference on Computational Processing of the Portuguese Language*, pages 65–75, Evora, Portugual. Springer.

Castro, S., Chiruzzo, L., and Rosá, A. (2018). Overview of the haha task: Humor analysis based on human annotation at ibereval 2018. In *IberEval SEPLN*, pages 187–194.

Clark, K., Luong, M., Le, Q. V., and Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations*, Addis Ababa, Ethiopia. Curran Associates, Inc.

Clemêncio, A. D. F. (2019). Reconhecimento automático de humor verbal. Master's thesis, Universidade de Coimbra.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Diao, Y., Lin, H., Wu, D., Yang, L., Xu, K., Yang, Z., Wang, J., Zhang, S., Xu, B., and Zhang, D. (2018). WECA: A WordNet-encoded collocation-attention network for homographic pun recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2507–2516, Brussels, Belgium. Association for Computational Linguistics.

Diao, Y., Lin, H., Yang, L., Fan, X., Wu, D., Zhang, D., and Xu, K. (2019). Heterographic pun recognition via pronunciation and spelling understanding gated attention network. In *The World Wide Web Conference*, pages 363–371, San Francisco, CA, USA. Association for Computing Machinery.

Ermakova, L., Bosser, A.-G., Miller, T., Thomas, T., Preciado, V. M. P., Sidorov, G., and Jatowt, A. (2024). Clef 2024 joker lab: Automatic humour analysis. In *European Conference on Information Retrieval*, pages 36–43, Glasgow, UK. Springer.

Ermakova, L., Miller, T., Regattin, F., Bosser, A.-G., Borg, C., Mathurin, É., Le Corre, G., Araújo, S., Hannachi, R., Boccou, J., et al. (2022). Overview of joker@ clef 2022: automatic wordplay and humour translation workshop. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 447–469, Bologna, Italy. Springer.

Gameiro, P., Inácio, M. L., Gonçalo Oliveira, H., and Alves, A. (2024). Sequence labeling for pun location and detection in portuguese. In *EPIA Conference on Artificial Intelligence*, pages 254–266, Viana do Castelo, Portugal. Springer.

Gonçalo Oliveira, H., Clemêncio, A., and Alves, A. (2020). Corpora and baselines for humour recognition in Portuguese. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1278–1285, Marseille, France. European Language Resources Association.

He, P., Gao, J., and Chen, W. (2023). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda. Curran Associates, Inc.

Inácio, M. L. and Gonçalo Oliveira, H. (2024). Exploring multimodal models for humor recognition in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 568–574, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Inacio, M. L., Wick-Pedro, G., Ramisch, R., Espírito Santo, L., Chacon, X. S. Q., Santos, R., Sousa, R., Anchiêta, R., and Goncalo Oliveira, H. (2024). Puntuguese: A corpus of puns in Portuguese with micro-edits. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13332–13343, Torino, Italia. ELRA and ICCL.

Jaiswal, A. and Monika (2019). Pun detection using soft computing techniques. In *International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 5–9, Faridabad, India. IEEE.

Kalloniatis, A. and Adamidis, P. (2024). Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58(2):43.

Kao, J. T., Levy, R., and Goodman, N. D. (2016). A computational model of linguistic humor in puns. *Cognitive science*, 40(5):1270–1285.

Labadie Tamayo, R., Chulvi-Ferriols, M. A., and Rosso, P. (2023). Everybody hurts, sometimes overview of hurtful humour at iberlef 2023: Detection of humour spreading prejudice in twitter. *Procesamiento del lenguaje natural*, 71:383–395.

Lima Inácio, M., Wick-pedro, G., and Goncalo Oliveira, H. (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.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, page 4768–4777, Long Beach, California, USA. Curran Associates, Inc.

Luz, A. I., Santos, H., Medeiros, M. M., and Anchiêta, R. T. (2023). Improving irony detection by balancing methods and feature selection. In *Anais do XII Brazilian Workshop on Social Network Analysis and Mining*, pages 216–221, João Pessoa, Brazil. SBC.

Miller, T., Hempelmann, C., and Gurevych, I. (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.

Monika and Vij, S. (2019). Humour agent detection in puns. In *4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*, pages 1–6, Ghaziabad, India. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ren, L., Xu, B., Lin, H., and Yang, L. (2021). Abml: attention-based multi-task learning for jointly humor recognition and pun detection. *Soft Computing*, 25(22):14109–14118.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, San Francisco, California, USA. Association for Computing Machinery.

Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing neural encoding of portuguese with transformer albertina pt-*. In *EPIA Conference on Artificial Intelligence*, pages 441–453, Faial Island, Azores. Springer.

Sammut, C. and Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417, Rio Grande do Sul, Brazil. Springer.

Wick-Pedro, G., da Silva, C. F., Inácio, M. L., Vale, O. A., and de Medeiros Caseli, H. (2024). Using large language models for identifying satirical news in Brazilian Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 156–167, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Wick-Pedro, G. and Santos, R. L. (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*, pages 409–415, Online. SBC.

Xu, Z., Yuan, S., Chen, L., and Yang, D. (2024). "a good pun is its own reword": Can large language models understand puns? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11766–11782, Miami, Florida, USA. Association for Computational Linguistics.

Yang, D., Lavie, A., Dyer, C., and Hovy, E. (2015). Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Yatsu, M. and Araki, K. (2018). Comparison of pun detection methods using Japanese pun corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3602 – 3605, Miyazaki, Japan. European Language Resources Association (ELRA).

Zhou, Y., Jiang, J.-Y., Zhao, J., Chang, K.-W., and Wang, W. (2020). “the boating store had its best sail ever”: Pronunciation-attentive contextualized pun recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 813–822, Online. Association for Computational Linguistics.

Zou, Y. and Lu, W. (2019). Joint detection and location of English puns. 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 2117–2123, Minneapolis, Minnesota. Association for Computational Linguistics.