Albertina in Action: An Investigation of its Abilities in Aspect Extraction, Hate Speech Detection, Irony Detection, and Question-Answering

Júlia da Rocha Junqueira[1], **Claudio Luis Junior**[1], **Félix Leonel V. Silva**[1]
Ulisses Brisolara Côrrea[1], **Larissa A. de Freitas**[1]

[1]Centro de Desenvolvimento Tecnológico – Universidade Federal de Pelotas (UFPel)
Pelotas – RS – Brazil

{julia.rjunqueira,clsmachado,flvdsilva,ulisses,larissa}@inf.ufpel.edu.br

Abstract. *The field of natural language processing has witnessed significant advances in recent decades, driven by the application of deep learning. Combined with using a neural architecture named Transformers, the advances are superior and outstanding. In this work, we used a BERT based model for the Brazilian Portuguese language, called Albertina, to tasks of Aspect Extraction, Hate Speech Detection, Irony Detection, and Question-Answering. Lastly, we compare the results in each task obtained with the BERTimbau and Albertina base and large models.*

Resumo. *O campo de processamento de linguagem natural testemunhou avanços significativos nas últimas décadas, impulsionados pela aplicação de aprendizado profundo. Combinando com o uso de uma arquitetura neural chamada Transformers, os avanços são ainda mais superiores e marcantes. Neste trabalho, usamos um modelo baseado em BERT para a língua portuguesa do Brasil, chamado Albertina, nas tarefas de Extração de Aspecto, Detecção de Discurso de Ódio, Detecção de Ironia e Perguntas-Respostas. Por fim, comparamos os resultados obtidos em cada tarefa com os modelos de base e grande de BERTimbau e Albertina.*

1. Introduction

In the last decade, the field of Natural Language Processing (NLP) has witnessed significant advances due to its application of Deep Learning (DL). The application of DL in NLP has produced more efficient and precise results for various tasks in Natural Language Understanding, particularly topic classification, sentiment analysis, question answering, and language translation [LeCun et al. 2015].

As an example of this improvement, using Transformers for NLP problems, such as BERT (Bidirectional Encoder Representations for Transformers), has become very common. Its popularity has increased because this model could represent the types of syntactic and semantic abstractions traditionally necessary for language processing. Moreover, they can model complex interactions between different levels of hierarchical information [Tenney et al. 2019].

Given this context, the development of AlbertinaPT-* [Rodrigues et al. 2023], a model based on DeBERTa, brings significant benefits for text processing in Portuguese,

allowing a more accurate and comprehensive analysis of emotions, opinions, aspects, negative speeches, and ironies present in texts written, enabling the advance of research and innovation in language technology for European Portuguese and Brazilian Portuguese.

Therefore, this work explores the tasks of Aspect Extraction (AE), Hate Speech Detection (HS), Irony Detection (ID), and Question-Answering (QA) using methods based on the BERT based model, Albertina PT-*. For this, we do our experiments in Brazilian Portuguese through the use of Albertina PT-BR, observing how the power of Albertina PT-* can be harnessed to improve the quality and accuracy of these tasks.

The paper is structured as follows: **Theoretical Background** covers the technical information relevant to understand the addressed tasks; **Related Works** reviews relevant works previously published in the literature, with a particular focus on studies covering NLP models concerning the Portuguese language; **Methodology** describes the steps taken to perform the experiments, including information about datasets, fine-tuning, and the data flux across tasks; **Experiments** shows the configuration and hyperparameters used to approach each task; **Final Remarks** summarizes the work and briefly discusses potential future studies.

2. Theoretical Background

Sentiment Analysis involves categorizing texts into positive, negative, or neutral polarities. In terms of data processing approaches, the literature describes different levels of granularity: the document-level analysis considers the entire text as a whole; the aspect-level requires additional steps, as it involves Aspect Extraction (AE) before classifying the sentiments associated with each aspect [Hoang et al. 2019]. The AE task focuses on identifying and extracting specific aspects or features discussed in a given text, usually a review. For instance, a hotel review could include specific aspects as room cleanliness, food quality, and staff friendliness [Liu 2015].

The HS task involves identifying whether various forms of communication, such as text, audio, and others, contain expressions of hatred or incite violence towards individuals or specific groups. A significant arena for spreading HS online is social media. The social media posts include paralinguistic signals (e.g., emoticons and hashtags), and their linguistic content contains plenty of poorly written text which are difficult to analyze. Another area for improvement is the lack of consensus on what constitutes HS, which makes the task difficult even for humans [Kovács et al. 2021].

The ID task corresponds to the ability to classify texts binarily, whether their respective content contains ironic behavior utilizing algorithms and models that can detect said behavior. It is difficult to determine what consists of irony and where the line is drawn when compared to sarcasm; for example, the base consensus is that irony revolves around the understanding that it is used to express the opposite of the literal meaning of what trying to be expressed by whoever is communicating, and while sarcasm can be understood the same way, it is used as a form of verbal irony, in which it carries a mocking or contemptuous tone, usually meant to mock, provoke or criticize something [Lee and Katz 1998]. Still, irony in itself can also depend on the level of aggressiveness shown and vocal clues [Van Hee et al. 2018].

The QA task combines several research fields, such as Information Retrieval, Information Extraction, and NLP. The methods used for the task aim to solve and

propose answers relevant to the question selected. The task can be divided into three modules: question classification, information retrieval, and answering extraction [Allam and Haggag 2012]. Question classification returns the type of answer that the question informed is expecting; for example, if you ask "What year the computer was invented?" the model is expected to return a year as an answer. Information retrieval returns search results based on the question submitted and its type; if it does not find information that matches, no further processing is carried out. Answering extraction returns the answer to the question asked.

3. Related Works

Below we describe some works in the literature about AE, HS, ID, and QA. Furthermore, we describe the work proposed by [Rodrigues et al. 2023], the Albertina model's.

In 2022, [da Silva et al. 2022] proposed the first shared task dedicated to identifying aspects and extracting the polarity in texts written in Portuguese, the Aspect-Based Sentiment Analysis in Portuguese (ABSAPT). ABSAPT comprised two sub-tasks: Aspect Extraction (AE) and Aspect Sentiment Classification (ASC). The results showed that with the ABSAPT 2022 hotel reviews dataset, the BERT methods, specifically "base-multilingual-cased" and "base-portuguese-cased", to AE, achieved an Accuracy of 0.67 [Gomes et al. 2022]. These studies have showed promising outcomes in enhancing the accuracy of Sentiment Analysis for Portuguese texts.

The solution to HS task proposed by [Leite et al. 2020] split ToLD-BR dataset into three parts: 80% for training, 10% for development, and 10% for testing. They utilized Bag-of-Words (BoW) to represent the examples and an AutoML model to establish the baseline model (BoW + AutoML). To accomplish this, they employed the auto-sklearn library for BERT-based models. The simple transformers library was utilized for convenient training and evaluation. Default arguments were used for parameter tuning, and a specific seed was defined to ensure reproducibility. Two versions of BERT, namely mBERT and BERTimbau base [Souza et al. 2020], were employed. The resulting F-Measure scores were 0.74 for BoW + AutoML, 0.75 for mBERT, and 0.76 for BERTimbau base.

In 2021, [Corrêa et al. 2021] proposed the first shared task dedicated to identifying the presence of irony in texts (tweets and news) written in Portuguese. The results showed that, with the IDPT 2021 tweets dataset, the classical feature-based models outperformed Deep Learning methods, achieving a BAcc of 0.52. [Jiang et al. 2021] introduced a solution to address the problem by utilizing BERTimbau, weight loss, and ensemble learning. The author claimed that the best-performing strategy involved leveraging two datasets used in IDPT 2021 for assisting in model classification and generalization, and this strategy achieved a BAcc of 0.48. Due to the relatively small size of the IDPT 2021 dataset [Subies 2021] they opted to employ Data Augmentation techniques. [Jiang et al. 2021] applied random masking to 15% of the tokens and utilized BERTimbau base and hyperparameter Grid Search to predict the masked tokens. The experiments with BERTimbau presented a BAcc of 0.49.

The QA task development was based on Guillou's work, using the BERTimbau model, fine-tuned on the SQUAD v1.1 in the Portuguese dataset [Guillou 2021]. [Spindola et al. 2021] also cites Guillou's work in their paper "Portuguese-Based Ques-

tion Answering System about the Blue Amazon". The authors combined BERTimbau model fine-tuned by Guilou with BertViz to visualize the attention weights and compared the results on their dataset Blue Amazon QA [Spindola et al. 2021]. Guillou, in his work, has split the dataset into two parts, training and validation. Since the dataset contains many words in each paragraph, the length was limited to 384 characters, allowing only one long example in the dataset, to give it different input features. Then, the Trainer API for feature-complete training in PyTorch was used to fine-tune and evaluate the model. The results obtained were 70.49% of Exact Match and 0.82 of F-Measure, using different hyperparameters on the BERTimbau base model.

According to [Rodrigues et al. 2023], the Albertina model's ability to achieve superior performance with less training time/computation likely results from resorting to all pre-trained layers, including the first layer, concerning word embeddings and the last layer, concerning masked token prediction, in contrast to the common practice in the literature of resetting these two layers to random weights to continue the pre-training. The tasks the author chose to demonstrate the model's functioning were: Remote Procedure Call, Semantic Textual Similarity, Recognizing Textual Entailment, and WNLI. In our experiments, we used the Albertina model's in AE, HS, ID, and QA task.

4. Methodology

Our work is composed of four main steps (Figure 1). Initially, the Albertina model is used. After, we applied fine-tuning in the AE, HS, ID, and QA tasks. And we test in datasets ABSAPT 2022 [da Silva et al. 2022], ToldBR [Leite et al. 2020], IDPT 2021 [Corrêa et al. 2021] , and SQUAD v1.1 [Rajpurkar et al. 2016]. Finally, we analyzed the results obtained in each task, comparing them to BERTimbau.

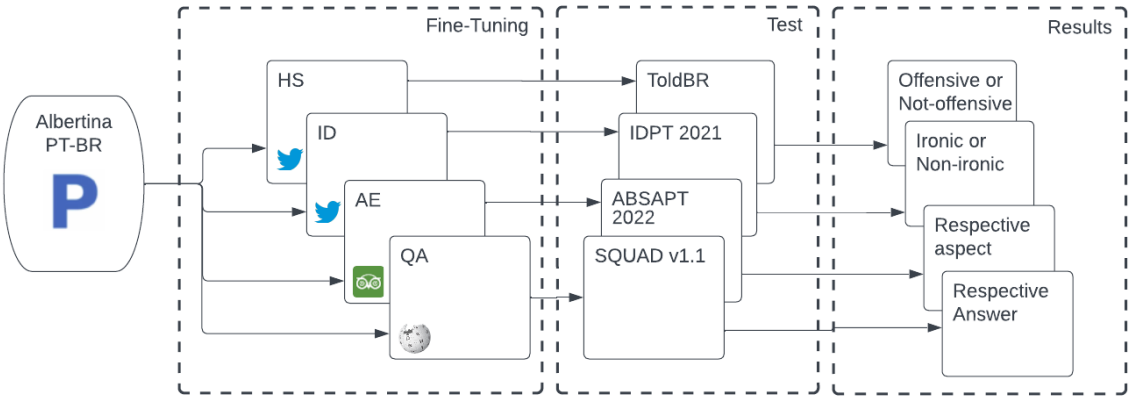

Figure 1. Methodology of this work.

For the AE task, we used the Trainer API[1] from Huggingface for the training and fine-tuning, and lastly, the Evaluate library for validating and evaluation of the results. The dataset utilized for AE task comprises reviews sourced from TripAdvisor, which were compiled by [da Silva et al. 2022] The training data consists of 847 reviews, divided in 77 aspects, and includes 3111 sentiment polarity annotations. Among these annotations are 2112 positive examples, 472 neutral examples, and 527 negative examples. On the

[1]https://huggingface.co/docs/transformers/main_classes/trainer

other hand, the test dataset comprises 184 reviews involving 70 aspects and 686 sentiment polarity annotations. Of these annotations, 450 are positive, 105 are neutral, and 131 are negative.

For the HS task, the ToLD-BR dataset was used. It comprises tweets gathered between July and August 2019 utilizing GATE Cloud's Twitter Collector tool[2]. Two distinct strategies were employed for tweet collection. The first strategy involved searching for specific keywords and predefined hashtags like *"gay"*, *"little woman"*, and *"northeasterner"*. The second strategy involved gathering tweets that mentioned influential figures such as Brazil's former President Jair Bolsonaro, and soccer player Neymar Jr. This method imposed no restrictions on keywords or hashtags, resulting in the collection of over 10 million unique tweets, out of which 21000 were randomly selected to compose the dataset. It should be noted that the first strategy accounted for 60% of the collected data. To annotate the dataset, 42 annotators were involved in classifying 1500 tweets as LGBTQ+phobia, obscene, insult, racism, misogyny, or xenophobia. Ultimately, the dataset consisted of 9245 offensive tweets and 11693 non-offensive tweets, each classified by three annotators.

For the ID task, the dataset utilized was the IDPT 2021 tweets [Corrêa et al. 2021], and was manually classified by linguistics and computer science students. The dataset is divided into two columns, one composed of text, the actual tweets, and another that determines whether or not what is written in the said tweet is a form of irony/sarcasm. We an differentiate what composed ironic and non-ironic as follows: (1) Ironic are sentences contradict the meaning between what is intended and what is written, e.g.: "*Que time horrível esse do Vasco, quase fez gol!!!*" ["*What a horrible team Vasco is, (they) almost scored!!!*"]; (2) Non-ironic are sentences that do not contain linguistic mechanisms that alternate their meaning, e.g.: "*Frustração hoje tem nome: Economia! — se sentindo triste*" ["*Frustration today has a name: Economy! — feeling sad*"]. Tweets were composed of 12736 ironic, and 2476 non-ironic. The testing dataset has a similar constitution but is a lot smaller, consisting of only 300 tweets, 177 tweets ironic and 123 tweets non-ironic.

For the QA task, the Trainer API for feature-complete training in PyTorch was used to fine-tune and evaluate the results, such as [Guillou 2021]. The dataset used was the SQUAD v1.1-PT, created by automatically translating the content of SQUAD using the Google Cloud API. The dataset was split into 87599 rows of paragraphs for train and 10570 for validation. Its data is composed of a title, a context, a question, and an answer, based on Wikipédia articles, where the answer to every question is a segment of text from the corresponding reading context [Rajpurkar et al. 2016].

5. Experiments

For most experiments, we used a batch size of 8, 3 epochs, a learning rate of $1 * 10^{-5}$, loss function CrossEntropy and optimizer AdamW (Table 1). The model training process is tailored to balance computational limitations with the need to achieve reasonable model performance. It is important to note that these hyperparameter choices were made based on constraints and may not necessarily represent the optimal configuration for the problem.

[2]https://cloud.gate.ac.uk/info/help/twitter-collector.html

Table 1. Hyperparameters Across Experiments.

Hyperparameters	Base	Large
Attention Heads	12	16
Batch Size (*)	8	2
Epochs	3	3
Hidden Size	768	1536
Hidden Layers	12	24
Learning Rate	1e-5	1e-5
Loss Function	CrossEntropy	CrossEntropy
Optimizer	AdamW	AdamW
Parameters	100 M	900 M

*: The QA experiment is an exception regarding the displayed batch size. We use a batch size of 16 and 8 for the base and large models, respectively.

The model Albertina (PT-BR) was tested on four tasks AE, ID, HS, and QA. Each test dataset was evaluated on several metrics, such as Accuracy (Acc), Precision, Recall, and F-Measure [Brownlee 2016], except for the QA task, which was evaluated based on Exact Match (EM) and F-Measure only.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Instances} \tag{1}$$

$$Precision = \frac{True\ Positives}{True\ Positives + True\ Negatives} \tag{2}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3}$$

$$F - Measure = 2\frac{Precision.Recall}{Precision + Recall} \tag{4}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \tag{5}$$

$$ExactMatch = \frac{TruePositives}{TotalNumberofInstances} * 100 \tag{6}$$

In Table 2, we show the results of our experiments for each task with the models BERTimbau base and Albertina PTBR base. In Table 3, we present the results of our experiments for each task with the models BERTimbau large and Albertina PTBR large.

To ensure the selection of meaningful aspects, we employ a filtering process in the AE task, excluding aspects with fewer than 20 occurrences, only the remaining 25 aspects for the experiments. It is possible to observe that the results from both models were low compared to the other tasks and the work presented in the Related Works session, which resulted in a 67% Accuracy by [Gomes et al. 2022]. This results from our technical limitations since this task demand a higher processing power and requires a larger memory

Table 2. Results Obtained Using BERTimbau and Albertina PT-BR Base Models.

	Task	Dataset	Acc	Precision	Recall	F-Measure	EM%
BERTim.	AE	ABSAPT 2022	0.26	0.21	0.26	0.19	-
	HS	ToLD-BR	0.88	0.89	0.88	0.88	-
	ID	IDPT 2021	0.41	0.36	0.41	0.25	-
	QA	SQUAD v1-PT	-	-	-	0.56	43.29
Albertina	AE	ABSAPT 2022	0.22	0.12	0.22	0.13	-
	HS	ToLD-BR	0.78	0.72	0.77	0.74	-
	ID	IDPT 2021	0.40	**0.40**	**0.99**	**0.57**	-
	QA	SQUAD v1-PT	-	-	-	**0.57**	**45.12**

Table 3. Results Obtained Using BERTimbau and Albertina PT-BR Large Models.

	Task	Dataset	Acc	Precision	Recall	F-Measure	EM%
BERTim.	AE	ABSAPT 2022	0.27	0.23	0.27	0.22	-
	HS	ToLD-BR	0.89	0.90	0.89	0.89	-
	ID	IDPT 2021	0.40	0.16	0.40	0.22	-
	QA	SQUAD v1-PT	-	-	-	0.62	47.15
Albertina	AE	ABSAPT 2022	0.21	0.04	0.20	0.07	-
	HS	ToLD-BR	0.58	0.34	0.58	0.43	-
	ID	IDPT 2021	**0.41**	**0.41**	**1.0**	**0.58**	-
	QA	SQUAD v1-PT	-	-	-	0.32	**47.30**

capacity. Furthermore, BERTimbau returned better results than Albertina, in both models, base and large.

For HS, the results indicated that the Albertina PT-BR model was inferior in every aspect compared to the results using BERTimbau. The model achieved an Accuracy score of 78% (base) and 58% (large), while the BERTimbau model performed significantly better with an Accuracy of 88% (base) and 89% (large). Similarly, Recall, was significantly higher for the BERTimbau model with a score of 88% (base) and 89% (large). On the other hand, Albertina PT-BR achieved a Recall score of 77% (base) and 58% (large). Lastly, the F-Measure, further confirms the superiority of the BERTimbau model. BERTimbau achieved an F-Measure of 88% (base) and 89% (large), while Albertina PT-BR only managed a score of 74% (base) and 43% (large). The higher F-Measure of BERTimbau indicates a better balance between Precision and Recall, reflecting its overall better performance in identifying HS instances accurately. A significant difficulty in detecting hate speech is identifying the context of what is or is not hate speech, many of the texts are also poorly written, such as "*Ui Nooooooofa que lindo fofa Nosso galao e mara ne amiga rajkazblanks*" ["*Wow how beautiful cute Our gallon and mara right friend rajkazblanks*"], another factor that affected the results achieved were the hyperparameters used, which were reduced by the limitation of machine resources during the experiments performed in this work.

The model was loaded and fine-tuned for ID, and with the Trainer API, the metric

was evaluated. The metrics used for the evaluation were Accuracy, Precision, Recall, and F-Measure, and that, along with the Actual Values and Predicted Values, formed True Positives, True Negatives, False Positives, and False Negatives. This was possible because the task was a binary evaluation with only two feasible outcomes. In the base model, compared to the BERTimbau utilized in [Corrêa et al. 2021], the result for Accuracy in the Albertina PT-BR was 40.67%, compared to 41%. The Precision results marked 40.8% compared to the 36%. The Recall results marked 99.18% compared to the 41%. The F-Measure results were 57.81% compared to 25%. The large models showed the following results: Accuracy 41%, compared to 40%. The Precision results marked 41% compared to the 16%. The Recall results marked 100%, compared to the 40%. The F-Measure results were 58.15% compared to 22%. These general results show that the Albertina PT-BR model does show an increase in all aspects, which can be compared to previous results with BERTimbau. ID can be a challenging task to obtain due to the complex and context-dependent nature of the language, and without context being provided to the model, it is likely to miss some tweets such as "*O problema do Brasil não é a violência, são as vítimas.*" ["*The problem with Brazil is not violence, but the victims.*"], and that, combined with the fact that lower parameters were used in order to achieve results because of hardware limitations, result percentages were worsened.

For the QA task, after the fine-tuning of the model, we used the Trainer API to evaluate the metrics. In this task, only two metrics were used for the evaluation, F-Measure, and Exact Match, since the task's resolution is composed of answers, not a binary representation. We obtained 45.12% of EM on Albertina PT-BR, compared to 43.29% to BERTimbau, both on the base versions. In the large versions, we obtained 47.30% of EM on Albertina PT-BR, in comparison 47.15% to BERTimbau. It is observable that our results cannot be compared by EM to the other tasks since it evolves on another kind of representation, but, compared to the work of [Guillou 2021], it was considerably lower side-by-side to his results, 70.49%. This is due to our hyperparameters choice of configuration; we precisely choose to lower our hyperparameters so that all the tasks could run without achieving technical limitations. In addition, the EM metric tends to be very limited since it needs to account for minor variations, such as differences in punctuation, capitalization, or word order. Therefore, if a predicted answer is semantically correct but slightly different than the correct match, it is considered wrong.

6. Final Remarks

With this work, we can conclude that the Arbertina PT-BR model can be better in some tasks when compared to BERTimbau model, such as in ID and QA. In other tasks, such as HS, the results can be inferior in every aspect, including Accuracy, Precision, Recall, and F-Measure, significantly reducing the score percentage.

Regarding our future research, there are several aspects that can be considered for further investigation. Firstly, the scope can be expanded by integrating another datasets, enabling a more comprehensive analysis. Additionally, it would be beneficial to evaluate the models in relation to different tasks and implement preprocessing techniques to address any data imbalance issues, thus enhancing the accuracy of the datasets. Moreover, it is worthwhile to explore alternative hyperparameters for tasks that exhibit lower performance, such as AE and HS, in order to potentially achieve improved results.

References

Allam, A. M. N. and Haggag, M. H. (2012). The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3).

Brownlee, J. (2016). *Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End*. Machine Learning Mastery.

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.

da Silva, F. L. V., da S. Xavier, G., Mensenburg, H. M., Rodrigues, R. F., dos Santos, L. P., Araújo, R. M., Corrêa, U. B., and de Freitas, L. A. (2022). Absapt 2022 at iberlef: Overview of the task on aspect-based sentiment analysis in portuguese. *Procesamiento del Lenguaje Natural*, 69.

Gomes, J. R. S., Garcia, E. A. S., Junior, A. F. B., Rodrigues, R. C., Silva, D. F. C., Maia, D. F., da Silva, N. F. F., Filho, A. R. G., and da Silva Soares, A. (2022). Deep learning brasil at ABSAPT 2022: Portuguese transformer ensemble approaches. In *Proceedings of the Iberian Languages Evaluation Fórum (IberLEF 2022), co-located with the 38th Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), Online. CEUR. org*, Online. CEUR. org.

Guillou, P. (2021). Portuguese bert base cased qa (question answering), finetuned on squad v1.1.

Hoang, M., Bihorac, O. A., and Rouces, J. (2019). Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*.

Jiang, S., Chen, C., Lin, N., Chen, Z., and Chen, J. (2021). Irony detection in the portuguese language using bert. *Proceedings http://ceur-ws. org ISSN*, 1613.

Kovács, G., Alonso, P., and Saini, R. (2021). Challenges of hate speech detection in social media. *SN Computer Science*, 2.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, C. J. and Katz, A. N. (1998). The differential role of ridicule in sarcasm and irony. *Metaphor and symbol*, 13(1):1–15.

Leite, J. A., Silva, D. F., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

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. *arXiv preprint arXiv:2305.06721*.

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*.

Spindola, S., José, M. M., Oliveira, A. S., Cação, F. N., and Cozman, F. G. (2021). Interpretability of attention mechanisms in a portuguese-based question answering system about the blue amazon. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 775–786. SBC.

Subies, G. G. (2021). Guillemgsubies at idpt2021: Identifying irony in portuguese with bert. In *Proceedings of the Iberian Languages Evaluation Fórum (IberLEF 2021), co-located with the 37th Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), Online. CEUR. org*, pages 910–916, Online. CEUR. org.

Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Van Hee, C., Lefever, E., and Hoste, V. (2018). SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.