Sabiá in Action: An Investigation of its Abilities in Aspect-Based Sentiment Analysis, Hate Speech Detection, Irony Detection, and Question-Answering

Júlia da Rocha Junqueira *, Émerson P. Lopes *, Claudio Luis da S. M. Junior *, Félix Leonel V. Silva *
Eduarda Abreu Carvalho * Larissa Freitas * and Ulisses Brisolara*
* Technological Development Center (CDTec), Federal University of Pelotas (UFPel), Pelotas, Brazil
Email: {julia.rjunqueira, eplopes, clsmachado, flvdsilva, eduarda.carvalho, larissa, ub.correa}@inf.ufpel.edu.br

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10650878

***Abstract*—This research investigates the efficacy of the versatile Sabiá-7B model, employed to decipher the complexities of Portuguese language across various tasks. Leveraging state-of-the-art architecture, the model extends LLaMA-7B pre-training to represent the Portuguese language better. This study focuses on evaluating Sabiá-7B's performance in Aspect-Based Sentiment Analysis (ABSA), Hate Speech Detection (HS), Irony Detection (ID), and Question-Answering (QA) tasks using a few-shot approach. Employing the few-shot method and prompt engineering throughout task executions, our research revealed that Sabiá-7B exhibits notable proficiency, mainly when provided with ample examples during few-shot extraction. However, particular challenges emerged, especially in QA tasks, where the model displayed limitations in generating precise answers compared to expected exact responses. This limitation resulted in the inclusion of extraneous words, potentially classified as irrelevant, impeding the accurate identification of an exact match. Our investigation sheds light on the strengths and potential limitations of Sabiá-7B in various NLP domains. As AI capabilities continue to advance, understanding these intricacies becomes essential for practical applications and the field's ongoing development.**

I. Introduction

As society becomes increasingly interconnected through digital communication, understanding the nuances of human expression in written language is paramount. Sabiá [1], with its state-of-the-art architecture, has positioned itself as a versatile tool for deciphering the intricacies of text. This investigation delves into the model's effectiveness across various tasks, shedding light on its adaptability and potential impact in real-world applications.

The Sabiá-7B model initiates with the LLaMA-7B weights and is further pretrained for the Brazilian Portuguese language, using ClueWeb 2022 dataset [2], with hyperparameters closely aligned with PALM's settings. The training process involves the AdaFactor optimizer, auxiliary loss for stability, and a similar infrastructure, including TPU v2-512 and gradient checkpointing. The throughput and efficiency metrics for the tasks suggest that the primary advantages of the models, particularly the Portuguese variant, arise from domain-specific knowledge acquired through monolingual pretraining.

Therefore, our experiments focuses on evaluating the outcome of the tasks Aspect-Based Sentiment Analysis (ABSA), Hate Speech Detection (HS), Irony Detection (ID), and Question-Answering (QA) for Sabiá-7B using few-shot approach. The objective is to assess Sabiá-7B's performance in these tasks with a few-shot approach (without any fine-tuning of the pre-trained model), aiming to explore the potential of improving the quality and precision of the model's outputs in various natural language understanding tasks.

The paper is structured in the following sections: **Theoretical Background** covers important concepts regarding domain knowledge about the strategies used, as well as the technical information relevant to understand the addressed tasks, and finally, a brief background in Large Language Models; **Related Works** reviews relevant works previously published in the literature, with a particular focus on studies covering Natural Language Processing (NLP) models concerning the Portuguese language or other lower-resourced languages; **Methodology** describes the steps taken to perform the experiments, including information about datasets, few-shot strategies, and the data flux across tasks; **Experiments** shows the metrics and comparison of the results; **Final Remarks** summarizes the work and briefly discusses potential future studies.

II. Theoretical Background

The interconnectivity facilitated by digital means, entangled with the explosive growth of web and social media, significantly influences the NLP community and, more specifically, studies in Sentiment Analysis area. This can be attributed to the prevalent and continuous stream of digitally documented data, which presents sentiments and opinions towards a topic, a product, or even a service [3], [4], and a significant portion of this data can be easily accessed on the web.

Sentiment Analysis is a subfield of NLP and implies in the classification of people's sentiments and emotions presented in written text. The literature describes different levels of granularity for SA: document-level analysis, which evaluates the sentiment that it is being expressed over an entire document; sentence-level analysis, which evaluates the sentiment for each sentence in a document, allowing for the analysis of more complex texts; and aspect-level, also know as ABSA, in which the sentiment is evaluated for each individual aspect of the entity that is the target of the opinion text [3].

Social media also plays a significant role for the two following tasks: HS and ID, in which both can be utilized. For the HS task, it is necessary to discern whether various forms of communication, such as text, audio, and others, encompass expressions of hatred or encourage violence towards individuals or specific groups, and social media serves as a significant platform for disseminating Hate Speech online. The posts on these channels incorporate paralinguistic signals (e.g., emoticons and hashtags), and their linguistic content comprises a plethora of poorly written text that is challenging to analyze. Another area for improvement is the need for more consensus on what constitutes HS, rendering the task challenging even for humans.

Irony can also be found in social media, and the ID task involves analyzing text from a communication medium that underlines an ironic or non-ironic tone and accurately identifying its nature. Irony can be identified in many ways, specially when focusing on the human aspect of things. Some examples of these indicators includes a specific tone of voice being used or finding oneself in a particular situation where it's evident that the statement can only be interpreted as irony. Nuances are easily noticeable when the one analyzing such behavior is ourselves. However, for a machine, considering its ironic tendencies, becomes a highly difficult task, as they are incapable of having the context, of understanding, and appropriately classifying the polarity in a binary manner. In an early ID study, Kreuz [5] proposed that lexical cues found in ironic mediums could be utilized for automated detection. While pattern-based recognition is a useful initial approach, it is merely the first step toward developing a more sophisticated feature-based classification system [6].

On the other hand, the goal of QA methodologies is to effectively respond to user queries by suggesting contextually relevant answers. This intricate task is typically divided into three integral modules: question classification, information retrieval, and answer extraction [7]. Question classification involves anticipating the expected type of answer based on the nature of the posed question, while information retrieval generates search results aligned with the identified question type. Finally, answer extraction involves formulating a coherent response to the user's inquiry. For example, the query "When is Ada Lovelace's date of birth?" undergoes a three-step process. Firstly, in the question classification module, the system identifies the expected type of answer, recognizing that the user seeks a specific date. Moving to information retrieval, the system scans relevant sources to gather data in relation to the woman. If successful, the answer extraction phase produces a user-friendly response, such as "Ada Lovelace was born in December 10th, 1815". Question-Answering plays a pivotal role in enhancing user interactions with information systems, enabling precise and efficient retrieval of pertinent information from vast datasets or knowledge bases.

The Large Language Models (LLM) are a proposal of language models that achieve promising results on general purpose language generation by using a large amount of data during the training process. The surge of interest in LLMs underscores the transformative impact these models are having on various domains [8]. Unlike earlier models designed for specific tasks, LLMs exhibit a remarkable capacity to handle diverse challenges. This versatility positions them as powerful tools with the potential to transcend narrow applications [9], contributing to the growing belief that they could play a significant role in shaping the era of Artificial General Intelligence (AGI) [10].

Sabiá-7B is a specialized model for the Portuguese language developed by Pires, Abonizio, Almeida, and Nogueira [1]. It is designed to enhance the capabilities on the understanding of the language structures, cultural nuances, and knowledge of the target language, which may not be correctly obtained with a multilingual training approach. It was trained based on LLaMA-1-7B [11], using an additional of 7 billion tokens from a Portuguese subset of ClueWeb22 [12], for approximately of 1.4 epochs (for a total of 10 billion tokens).

Few Shot Learning (FSL) is a method within machine learning that addresses the challenge of training models with limited labeled examples. As demonstrated by Yang in the following passage: "Suppose a P is used to evaluate the performance of a computer program on a task class T. If a program improves performance on T task by using experience E, then we say that the program learns about the T and the program." [13]. The goal of this method is to enable models to make accurate predictions with only a small number of training examples. This approach is particularly valuable in scenarios where pre-training is expensive, time-consuming, or impractical.

III. Related Works

Below we describe some relevant works in the literature about ABSA, HS, ID, and QA.

In 2022, Silva *et. al.* (2022) [14] proposed the Aspect-Based Sentiment Analysis in Portuguese (ABSAPT), the first shared task dedicated ABSA in the Portuguese Language. ABSAPT comprised two sub-tasks: Aspect Extraction (AE) and Aspect Sentiment Classification (ASC). For the first task, methods based on Transformers encoder-only models achieved the best results, with an Accuracy of up to 0.67 [15]. On the second task, focused on the classification of the sentiment for those aspects, the best approach consists in the use of encoder-decoder Transformers models, with a ensemble of four fine-tuned PTT5 [16] models, that obtained an Balanced Accuracy score of 0.82. These studies have showed promising outcomes in enhancing the accuracy of Sentiment Analysis for Portuguese texts.

While remaining focused on Sentiment Analysis task, the use of LLMs is evaluated for several of its sub-tasks by Wenxuan, Yue, Liu, Sinno, and Lidong [17]. That work shows that for simpler document and sentence-level SA, the use of LLMs with only few-shot learning can obtain better results than the usage of smaller fine-tuned encoder-decoder models. Their results also show that, for ABSA, when considering only the ASC task, the results are almost the same, but when

both tasks (AE and ASC) are unified, the results of fine-tuned models still is far better.

Analogously, for the HS task, Leite *et al.* (2020) [18] tackled it by employing a data-driven approach to the ToLD-BR (Toxic Language Dataset for Brazilian Portuguese) dataset. The authors split it into standard training, development, and testing portions and opted for Bag-of-Words representation and AutoML for the initial model (BoW + AutoML). Leveraging the auto-sklearn and simple transformers libraries, they streamlined the process with default parameter tuning and ensured repeatability through a fixed seed. Their exploration of two BERT models, mBERT and BERTimbau base [19], yielded promising results, surpassing the BoW + AutoML baseline with F1-Score of 0.75 and 0.76.

Regarding the ID task, a shared task was introduced by Correa (2021) [20], which was solely focused on detecting irony within texts, including tweets and news articles in Portuguese. His work found that feature-based models performed better than Deep Learning methodologies when using the IDPT (Irony Detection in Portuguese) 2021 tweet dataset. In another study, Jiang [21] proposed an effective solution by utilizing weight loss techniques and ensemble learning with BERTimbau. To aid in model classification and generalization, the author used the two datasets used in IDPT 2021. However, these datasets have a limited size [22], so researchers opted to use Data Augmentation techniques. Jiang applied random masking to 15% of the tokens and used hyperparameter Grid Search along with BERTimbau base to predict the masked tokens. A work by Aytekin and Erdem [23] in 2023 explored the use of Generative Pre-trained Transformer (GPT) Models in ID in the English language through the implementation of zero-shot and few-shot examples. In the study, specifically the models text-davinci-003 and gpt-3.5-turbo used a few-shot learning method, and showed themselves to be a valid form of irony detection as the text-davinci-003 model scored the highest F1 Score when compared to the other models used in the same study and the highest recall overall in a binary classification task when compared to the other models used in the official competition.

And, for the fourth task, the domain of question answering using FSL and LLMs has witnessed noteworthy progress in recent years, marked by significant contributions in these realms. Some notable advancements in these areas are listed below.

Ram, Kirstain, Berant, Globerson, Levy [24] wrote a study that delves into the challenges posed by the few-shot setting in question answering benchmarks, where only a limited number of training examples, in the hundreds, are available. The authors observed that standard models struggle in this realistic scenario, shedding light on the existing disparity between prevalent pretraining objectives and the requirements of question answering. Then, they proposed a new form of approach of returning answers. In this approach, given a passage with multiple sets of recurring spans, the authors mask all recurring spans except one in each set. The outcomes are promising, with the resulting model achieving surprisingly strong results across multiple benchmarks, including a notable 72.7% of F1-Score on SQuAD (the English version) with just 128 training examples.

A very significant challenge in QA revolves around the limited access to high-quality QA datasets, especially in languages other than English. Languages with fewer resources, such as Brazillian Portuguese, frequently face a scarcity of extensive QA datasets, posing challenges for researchers to investigate and test the latest neural techniques in QA.

The work proposed by Nunes, Primi, Pires, Lotufo, and Nogueira [25], explores the capabilities of LLMs in handling high-stakes multiple-choice tests in Brazilian Portuguese, and has a direct and positive impact on the challenges mentioned. The study contributes to overcoming the scarcity of extensive QA datasets for languages like Brazilian Portuguese by leveraging advanced language models, specifically GPT-4 with Chain-of-Thought prompts. The success and accuracy achieved by GPT-4 in handling questions from the Exame Nacional do Ensino Médio (ENEM), a multidisciplinary entrance examination used by Brazilian universities, demonstrate the potential of LLMs in addressing complex QA tasks in Portuguese.

Another work that we used as baseline to our results was Bahak, Taheri, Zojaji, and Kazemi's study [26], where they examine ChatGPT's role as a Question Answering System (QAS), undertaking a comparative analysis with other existing QASs. The primary focus of the work lies in evaluating ChatGPT's proficiency in extracting responses from given paragraphs, a fundamental QAS capability. Additionally, the study delves into performance comparisons in scenarios devoid of a contextual passage. Multiple experiments, exploring response hallucination and considering question complexity, were conducted on ChatGPT. Evaluation leveraged established Question Answering (QA) datasets, encompassing SQuAD, NewsQA, and Persian-QuAD, spanning English and Persian languages.

The study reveals that ChatGPT falls behind task-specific models in question answering efficacy. They showed that context provision and prompt engineering enhance its performance, especially for questions lacking explicit answers in paragraphs. However, the model struggles with more complex "how" and "why" queries. The evaluation highlights instances of hallucinations, it can be observed clearly through the results comparing effectiveness between various Language Models on SQuAD 1.1, where ChatGPT returned the worst Exact Match (EM), 44.4, between LUKE [27], 90.2, XLNet [28], 89.9, and SpanBERT [29], 88.8.

In the work of da Rocha Junqueira *et al.* [30], the BERTimbau Base and Large models were employed accross varios NPL tasks, including Sentiment Analysis, Aspect Extraction, Hate Speech Detection and Irony Detection tasks. The research consists of four main steps: an initial the BERTimbau Base and Large models are used, followed by fine-tuning being applied to the four tasks used, after that they were tested on the TweetSentBR [31], ABSAPT 2022 [14], ToLD-BR [18] and IDPT2021 [20] datasets; The evaluation culmimated in

the analysis and comparison of results. The following hyperparameters were utilized: 32 batch size, 4 epochs, AdamW as the optimizer, CrossEntropy as the Loss Function, and 2e-5 for the learning rate. Furthermore, the study used six metrics for the experiments: Accuracy (Acc), Precision (P), Recall (R), F1-Score (F1), Specificity (S) and Balanced Accuracy (BAcc), the results achieved are also displayed in Table II.

Furthermore, da Rocha Junqueira [32] presented results for Aspect Extraction (AE), Sentiment Analysis (SA), Hate Speech Detection (HS), Irony Detection (ID) and Question Answering (QA) tasks at a later date using the Albertina PT-BR Large and Base models. After a sequence of fine-tuning and testing on the same four datasets, the models were configured with the following hyperparameters for Base and Large versions, respectively: 12 and 16 attention heads, 8 and 2 as batch size (except for QA, which had 16 and 8, respectively), 3 epochs for both, 768 and 1536 hidden size, 12 and 24 hidden layers, CrossEntropy as loss function, 100 M and 900 M parameters, and a learning rate of 1e-5 for both, with AdamW as the optimizer. The outcome was then compared to the results of the previous work. This new model demonstrated promising, with results varying depending on the implemented task. ID showed the most gains in terms of performance, with Albertina PT-BR presenting slighty lower numbers only in Accuracy in the Base version of the models at 41% to 40%. QA also showed improvements, and since it had a different kind of evaluation, the metrics used were F1-Score and EM% (Exact Match %). The results in this work contributed to the practical application and testing of the Albertina PT-BR model, assessing its capabilities in the Brazilian Portuguese language. These findings will be used to determine if the Sabiá model achieved acceptable results.

IV. Methodology

Our work is composed of three main steps. Firstly, we pre-process the few-shot examples by selecting and separating the samples of the datasets (ToLD-BR [18], IDPT 2021 [20], and SQUAD v1.1 [33]) assigned to each task, generating a "train set". Next, we incorporate the few-shot examples as prompts for each inference, serving as input to the model, which is the previously introduced Sabiá-7B [1]. Finally, we analyzed the results obtained in each task.

TABLE I
THE DIVISION USED FOR EACH DATASET.

Sets	ABSA	HS	ID	QA
Original Train	3111	16750	15211	87599
Original Test	686	2094	300	10570
Few-Shot Examples	11	10	20	4
Test Set	686	150	300	4139

On the ABSA task, eleven examples were selected as few-shots. The few-shot examples had nine different aspects, and four of them had Positive polarity, four had Negative polarity, and the remaining three had Neutral polarity. These few-shots were added to the start of the prompt, before the actual example to be predicted. All examples were formatted as "Text: REVIEW TEXT Aspect: ASPECT Sentiment: POLARITY", changing all three upper cased parts on the few-shot examples, and for the review that should be predicted the 'POLARITY' was left empty, as that is what the model should generate. The evaluation on this task used all 686 examples from the test set of the ABSAPT shared task.

On the other hand, ten texts were selected from the ToLD-BR dataset to serve as few-shot examples to the HS task: five containing hate speech and five without. Some of the examples used in this assignment were: *"@user mas de fato existe o hacker então, não é mesmo? ou só porque não é russo tá liberado hacker celulares e vazar mensagens?"*, (from Brazilian Portuguese, *"@user, but there really is such a thing as a hacker, is not there? or is it just because you're not Russian that you're allowed to hack cell phones and leak messages?"*) which is not hate speech, and *"carol não quer deixa eu trabalhar mane , toda manda msg essa ****"*, (from Brazilian Portuguese, *"carol doesn't want to let me work dude, she sends me all these messages, that *****"*), which contains hate speech. Additionally, we utilized 150 examples from a test dataset, distinct from the few-shot examples, to serve as a reference for the model. This set was a combination of an equal distribution of examples, half containing hate speech and half without.

A double quantity of examples was chosen for the ID task. The twenty examples, composed of an equal distribution of ironic and non-ironic contexts, were extracted from the IDPT 2021 dataset and carefully curated into a new dataset for few-shot examples used in the model for Irony Detection. This new dataset contained phrases like *"Que pena que eu me esqueci de trazer as folhas de biologia! Agora não posso estudar."* (from Brazilian Portuguese, *"What a shame that I forgot my biology papers! Now I can't study."*) with a polarity of 1, indicating irony. A different dataset was used as a blank reference for result percentages to avoid any biases related to the original training dataset and the few-shot examples extracted from it. This second dataset was also present in the IDPT 2021 competition and contains 300 entries. It was originally used solely for testing purposes and has no association with the dataset composed of the few-shot examples.

To tackle the QA task, we used a combination of prompt engineering plus four few-shot examples. For the prompting, it was passed *"The answer to each question is a segment of text from the corresponding reading passage. The answer should be extension based, objective answer only. Answer the question accurately and succinctly, containing only your main answer, as short as possible, as in the examples below:"*, in Brazillian Portuguese, as input sentences. Moreover, for the examples, a meticulous selection of samples was undertaken to exemplify four distinct question types in Brazilian Portuguese: "What?", "Where?", "Who?" and "When?". This deliberate choice aimed to ensure a balanced representation of the response quality across various inquiry categories. The decision to utilize only four samples was dictated by the substantial

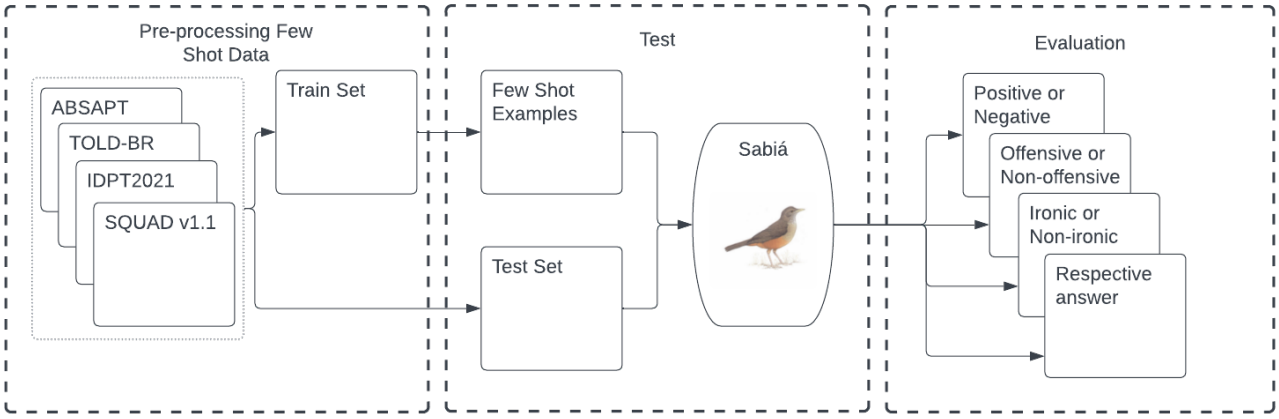

Fig. 1. Methodology of this work.

token count of the dataset context, thereby necessitating a strategic and focused sampling approach. For the experiments, we executed over 4,000+ questions of the validation subset of SQUAD v1.1. The model returned (for each inference) a combination of the prompt + few-shot examples + context + question and then the respective answer; which were post-processed using regular expressions, so we could evaluate the results more easily.

Regarding the generality and reproduction of our work in other languages, the model used in our approach is exclusively for Portuguese, and does not offer the specific advantages of the Brazilian Portuguese training for any other languages. However, the methodology presented can be reused if there is a dataset and model available and trained in the intended language.

V. Experiments

The Sabiá model was tested on four tasks ABSA, ID, HS, and QA. Each test dataset was evaluated on several metrics, such as Accuracy, Precision, Recall, and F1-Score [34], except for the QA task, which was evaluated based on Exact Match (EM) and F1-Score only, and the ABSA task, which was evaluated also on the Balanced Accuracy.

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

$$Precision = \frac{True\ Positives}{True\ Positives + True\ Negatives} \quad (2)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

$$F1 - Score = 2 * \frac{Precision.Recall}{Precision + Recall} \quad (4)$$

$$BAcc = \frac{(Recall_{Pos} + Recall_{Neu} + Recall_{Neg})}{3} \quad (5)$$

$$ExactMatch = \frac{TruePositives}{TotalNumberofInstances} * 100 \quad (6)$$

In Table II, we show the results of previous works that use the same datasets using other transformers models trained for the Portuguese language (BERTimbau and Albertina PT-BR), for the tasks of HS, ID and QA.

On the ABSA task, the BAcc score – the main metric for the ABSAPT shared task – is worse than almost all teams that participated in ABSAPT, that had BAcc of up to 0.82. Comparing the F1-Score, our results are worse than all participants. Upon closer examination of the model's output, we find that the main source of errors are the example labeled as "Neutral", that are the hardest for this task, as the difference from "Negative" to "Positive" is more noticeable than the difference of one of those to "Neutral" polarity. The model that we used was unable to make this separation, and didn't generated any "Neutral" prediction. To obtain a better understanding of the model capabilities, we also calculated the metrics for all examples that have only "Positive" or "Negative" polarities, as an evaluation of Sabiá's capabilities for those two categories. In this scenario, the F1-Score goes up from 0.53 to 0.79 – which would be the second result by F1-score in ABSAPT, if the neutral examples didn't also affect those methods –, and the BAcc goes up from 0.61 to 0.91.

For the HS task, the results can be further evaluated through the detailed analysis shown in the Table II. The F1-Score was used as a criterion for comparing the models investigated. The Sabiá model stood out by achieving superior results, scoring 93%. In contrast, the BERTimbau base and large models performed competitively, recording scores of 89% and 88%, respectively. It is worth noting that the Albertina base and large models performed significantly less well, scoring 74% and 43%, respectively. Considering the limited data in the Sabiá model's test set is crucial compared to the other models. Despite the disparity in the size of the data set, Sabiá showed superior efficiency, suggesting a remarkable capacity for the HS task.

Despite the high efficiency of the Sabiá model, there is a difficulty in getting poorly formulated or abbreviated sentences right, exemplified by expressions such as: *"olha quem fala, maravilha p c..."* (from Brazilian Portuguese, *"look who's talking, it's so f... wonderful"*) and *"Eu fico tão orgulhosa*

TABLE II
RESULTS OBTAINED USING BERTIMBAU, ALBERTINA PT-BR AND SABIÁ MODELS.

Model	Task	Dataset	Acc	Precision	Recall	F-Measure	BAcc	EM%
BERTimbau Base	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
BERTimbau Large	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 PT-BR Base	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
Albertina PT-BR Large	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**
Sabiá-7B	ABSA	ABSAPT 2022	0.77	0.64	0.61	0.53	**0.61**	-
	HS	ToLD-BR	**0.94**	**0.92**	**0.94**	**0.93**	**0.94**	-
	ID	IDPT 2021	**0.46**	**0.50**	0.46	0.44	-	-
	QA	SQUAD v1-PT	-	-	-	0.54	-	39.17

de vc, sabe? Eu vi seu amadurecimento, vc sempre foi um destaque para mim, vc sempre foi tão vc mesma, lutou muito e agora está lutando com a transição que eu sei que não é algo fácil. Eu tenho orgulho de vc" (from Brazilian Portuguese, *"I'm so proud of u, you know? I've seen you mature, u've always been a highlight for me, u've always been so urself, you've struggled a lot, and now you're struggling with the transition, which I know isn't easy. I'm proud of u."*). Interestingly, sentences wrongly marked as hate speech include more careful and positive expressions, while others with potentially aggressive content were classified as not containing hate speech. This scenario highlights the complexity of the HS task. This difficulty can be attributed, in part, to the absence of text context during the detection process. In addition, the presence of texts with inferior writing quality contributes to additional challenges in correct interpretation.

Also presented in Table II, the results of the ID task reveal that BERTimbau Base outperforms BERTimbau Large in all evaluated metrics, demonstrating superior overall metric results. The Base and Large versions of Albertina PT-BR exhibit comparable performance across all metrics. As the models differ, a direct comparison can't be made between their inner workings and the results. However, considering only the numerical results, Sabiá emerges as a standout performer, showcasing the highest accuracy, reaching five-percentual-point advantage over the results of BERTimbau Base and Albertina PT-BR Large at 46%. In terms of F1-Score, Sabiá shows a nine-percentual-point advantage over the results of Albertina PT-BR Large at 50.3% and both Recall and F1-Score showed a similar pattern between themselves, surpassing BERTimbau Base and Large results but not the Albertina PT-BR ones in both its sizes. This underscores the significance of model selection and methodology in Irony Detection applications, with Sabiá emerging as a notably effective choice.

As kind of expected, the model seemed to have a hard time with specific phrases that lacked context and by themselves would normally pass as non-ironic if we do not consider outside factors like an individual's financial situation, taking a text like *"Só mais de R$.: 3.000,00 gastos... só... :v :v"* (from Brazilian Portuguese, *"Just more than R$ 3.000 spent... Just that... :v :v"*) and considering not ironic, missing completely the contrast between a magnitude of the value described and the absurdity of considering it a small amount, along with emoticons to facilitate comprehension. The case of false positives are also present, with the model identifying even small texts like "pls" as ironic.

An example of a different methodology in this study with Sabiá results not shown in Table II is how the "prediction" column in the datasets was organized. Originally, the dataset was composed of ones and zeroes, but as a test it was re-organized with "POSITIVE" and "NEGATIVE" as indicators of irony to determine if that would get better results. What ended up happening was worsened results with the incorrect identification of polarity of phrases like *"Que imagem linda!!! Vou guardar para sempre me lembrar de não confiar em nada que este presidente do Senado fizer ou falar. #RodrigoPacheco #FlavioBolsonaro #Brasil"* (from Brazilian Portuguese, *"What a beautiful image!!! I'm gonna save it forever as a reminder to not trust in anything this Senate's president does or says. #RodrigoPacheco #FlavioBolsonaro #Brasil"*) as a NEGATIVE, composed with interesting reactions to text with emoticons, like the evaluation of a phrase containing " 👍 " with the polarity of " 😂 😂 ", and suffice to say that is not the expected return of the polarity of a phrase. What could be happening is the identification of a pattern in which the emoticons appear, as in, maybe in other phrases with the thumbs-up emoticon, the tears of joy emoticon immediately follows it, and that ends up ingrained into the model's algorithm as a plausible polarity.

In our evaluation of the Sabiá-7B model in the QA task, we obtained insightful results that shed light on the model's

performance. As per Table II, the EM rate was found to be 39.17%. This indicates that the Sabiá model correctly answered approximately 39.17% of the questions posed in our evaluation. Considering that this metric evaluates only the predicted answer that is identical to the ground truth answer, this result shows that Sabiá struggles with providing precise answers that exactly match the reference answers, when comparing to the other various language models that are shown in Table II. However, if we look at the results returned from the model prediction, it can be seen that most of the results are not completely incorrect, but rather, they are not identical. For example, given the answer *"Em que cidade o Super Bowl 50 aconteceu?"* (from Brazilian Portuguese, *"In which city did Super Bowl 50 take place?"*), the expected answer was only *"Santa Clara"* while the model predicted *"Santa Clara, Califórnia"*.

The F1-Score is a combination of Precision and Recall. For QA, it considers both false positives (model claims something is a correct answer when it's not) and false negatives (model misses correct answers). Sabiá returned a result of of 54.86% and is a decent performance, but it also suggests room for improvement. This means that the model is providing correct answers more frequently than not, but there are still instances where it either misses correct answers or provides incorrect ones. In addition, we tested the model without using a few-shot and prompt, and the result was much worse, EM of 8.33 and F1-Score of 33.04%, and most of the time it caused the model to return very large generated answers or hallucinate.

As cited in Pires, Abonizio, Almeida, and Nogueira [1] work, the model shows higher and improved results in native Portuguese datasets when comparing to automatically translated datasets. The SQUAD v1-PT dataset [33] is automatically translated using Google Translate, and made available by the Deep Learning Brasil group. This dataset occasionally contains English words and terms that may hold cultural relevance specific to the English language. As an example, in the question "Quantos *turnovers* a Cam Newton tem?" ("How many turnovers does Cam Newton have?"), *turnovers* is a word belonging to the English language, and has no definition for it in Portuguese. The training on a specific language may make the model lose some of the knowledge for the English language, but in turn greatly increasing the performance when terms that only belong to the target language are in the text.

Furthermore, the results obtained by Sabiá are very promising for Brazilian Portuguese, since the task was worked on an automatically translated dataset, and even so, the model returned a result similar to that of Bahak, Taheri, Zojaji, and Kazemi's work [26], using ChatGPT, with a EM of 44.4%.

VI. Final Remarks

In this study, we conducted an extensive analysis of the Sabiá-7B model's performance accross various NLP tasks, including Aspect-Based Sentiment Analysis, Hate Speech Detection, Irony Detection, and Question Answer. Throughout the task executions, we employed the few-shot method and prompt engineering to evaluate the model.

Our findings suggest that Sabiá-7B demonstrates impressive performance, especially when given sufficient examples during few-shot extraction. However, the model appears to encounter challenges with certain types of prompts, more specifically the QA task in the answer generation aspect where it exhibits limitations in generating precise answers compared to the expected exact responses. This limitations may lead to the inclusion of additional words that might be classified as irrelevant, not identifying it as an exact match when it should be.

In conclusion, our investigation into the performance of the Sabiá-7B model across diverse tasks has provided valuable insights into its capabilities. Leveraging the few-shot method and prompt engineering allowed us to thoroughly assess the model's adaptability and proficiency in handling nuanced tasks.

Furthermore, the outcomes underscore both the strengths and potential limitations of Sabiá-7B in various natural language processing domains. As we continue to push the boundaries of AI capabilities, understanding the intricacies of model behavior becomes crucial for applications and future advancements in the field.

Regarding our future work, this research lays a solid foundation for further investigations and enhancements in the development of other applications. Taking into consideration the characteristics of the results between the tasks, future studies can focus on optimizing the model to better cater to the specific requirements of each domain.

Exploring enhanced prompt engineering approaches and fine-tuning methods may yield significant gains in the model's ability to comprehend complex contexts. Other additions that could be made: broadening the scope of the inquiry to incorporate supplementary datasets; applying transfer learning techniques; testing other hyper-parameters and strategies.

Additionally, given the dynamism of the natural language processing field, future efforts may be directed towards exploring more advanced model architectures and methodological innovations to further enhance the effectiveness and versatility of these models in an ever-evolving landscape.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We would like to thank the FAPERGS - Brasil for Financial Support, Award Agreement 22/2551-0000598-5. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

[1] R. Pires, H. Abonizio, T. S. Almeida, and R. Nogueira, "Sabiá: Portuguese large language models," in *Intelligent Systems*, M. C. Naldi and R. A. C. Bianchi, Eds. Cham: Springer Nature Switzerland, 2023, pp. 226–240.

[2] A. Overwijk, C. Xiong, and J. Callan, "Clueweb22: 10 billion web documents with rich information," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 3360–3362.

[3] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015.

[4] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using bert," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 2019.

[5] R. Kreuz and G. Caucci, "Lexical influences on the perception of sarcasm," in *Proceedings of the Workshop on Computational Approaches to Figurative Language*, A. Feldman and X. Lu, Eds. Rochester, New York: Association for Computational Linguistics, Apr. 2007, pp. 1–4. [Online]. Available: https://aclanthology.org/W07-0101

[6] A. Maladry, E. Lefever, C. Van Hee, and V. Hoste, "The limitations of irony detection in dutch social media," *Language Resources and Evaluation*, pp. 1–32, 2023.

[7] A. M. N. Allam and M. H. Haggag, "The question answering systems: A survey," *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, vol. 2, no. 3, 2012.

[8] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[9] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *arXiv preprint arXiv:2307.03109*, 2023.

[10] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[12] A. Overwijk, C. Xiong, X. Liu, C. VandenBerg, and J. Callan, "Clueweb22: 10 billion web documents with visual and semantic information," 2022.

[13] M. Yang, "A survey on few-shot learning in natural language processing," in *2021 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*. IEEE, 2021, pp. 294–297.

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

[15] J. R. S. Gomes, E. A. S. Garcia, A. F. B. Junior, R. C. Rodrigues, D. F. C. Silva, D. F. Maia, N. F. F. da Silva, A. R. G. Filho, and A. da Silva Soares, "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, 2022.

[16] D. Carmo, M. Piau, I. Campiotti, R. Nogueira, and R. Lotufo, "Ptt5: Pretraining and validating the t5 model on brazilian portuguese data," *arXiv preprint arXiv:2008.09144*, 2020.

[17] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," 2023.

[18] J. A. Leite, D. F. Silva, K. Bontcheva, and C. Scarton, "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*, 2020.

[19] F. Souza, R. Nogueira, and R. Lotufo, "Bertimbau: pretrained bert models for brazilian portuguese," in *Proceedings of the 9th Brazilian Conference on Intelligent Systems*, 2020.

[20] U. B. Corrêa, L. Coelho, L. Santos, and L. A. de Freitas, "Overview of the idpt task on irony detection in portuguese at iberlef 2021," *Procesamiento del Lenguaje Natural*, vol. 67, 2021.

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

[22] G. G. Subies, "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*, Online. CEUR. org, 2021, pp. 910–916.

[23] M. U. Aytekin and O. A. Erdem, "Generative pre-trained transformer (gpt) models for irony detection and classification," in *2023 4th International Informatics and Software Engineering Conference (IISEC)*. IEEE, 2023, pp. 1–8.

[24] O. Ram, Y. Kirstain, J. Berant, A. Globerson, and O. Levy, "Few-shot question answering by pretraining span selection," *arXiv preprint arXiv:2101.00438*, 2021.

[25] D. Nunes, R. Primi, R. Pires, R. Lotufo, and R. Nogueira, "Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams," *arXiv preprint arXiv:2303.17003*, 2023.

[26] H. Bahak, F. Taheri, Z. Zojaji, and A. Kazemi, "Evaluating chatgpt as a question answering system: A comprehensive analysis and comparison with existing models," *arXiv preprint arXiv:2312.07592*, 2023.

[27] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "Luke: deep contextualized entity representations with entity-aware self-attention," *arXiv preprint arXiv:2010.01057*, 2020.

[28] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[29] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Transactions of the association for computational linguistics*, vol. 8, pp. 64–77, 2020.

[30] J. da Rocha Junqueira, F. da Silva, W. Costa, R. Carvalho, A. Bender, U. Correa, and L. Freitas, "Bertimbau in action: An investigation of its abilities in sentiment analysis, aspect extraction, hate speech detection, and irony detection." *The International FLAIRS Conference Proceedings*, vol. 36, no. 1, May 2023. [Online]. Available: https://journals.flvc.org/FLAIRS/article/view/133186

[31] H. B. Brum and M. das Graças Volpe Nunes, "Building a sentiment corpus of tweets in brazilian portuguese," in *Proceedings of the 11th International Conference on Language Resources and Evaluation*, 2018.

[32] J. da Rocha Junqueira, C. L. Junior, F. L. V. Silva, U. B. Côrrea, and L. A. de Freitas, "Albertina in action: An investigation of its abilities in aspect extraction, hate speech detection, irony detection, and question-answering," in *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. SBC, 2023, pp. 146–155.

[33] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *arXiv e-prints*, p. arXiv:1606.05250, 2016.

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