Journal on Interactive Systems, 2025, 16:1, doi: 10.5753/jis.2025.5564

Sentiment Analysis of Shared Content in Brazilian Reddit Communities

Giovana Piorino [**Federal University of Minas Gerais** | ***giovana.piorino@dcc.ufmg.br***]
Vitor Moreira [**Federal University of Minas Gerais** | ***vitormoreira@dcc.ufmg.br***]
Luiz Henrique Quevedo Lima [**Federal University of Minas Gerais** | ***luiz.quevedo@dcc.ufmg.br***]
Ana Clara Souza Pagano [**Federal University of Minas Gerais** | ***anapagano@ufmg.br***]
Adriana Silvina Pagano [**Federal University of Minas Gerais** | ***apagano@ufmg.br***]
Ana Paula Couto da Silva [**Federal University of Minas Gerais** | ***ana.coutosilva@dcc.ufmg.br***]

Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

Received: 15 February 2025 • **Accepted:** 05 August 2025 • **Published:** 23 August 2025

Abstract: The growth of social media in the present decade is one of the main drivers of studies on user-generated content. *Reddit*, a social network that has been gaining popularity among Brazilians, has become a source for sentiment analysis studies aimed at evaluating automated models for this task. This article reports a study on the development and evaluation of a dataset of human-annotated Reddit comments and its comparison with sentiment classification models. Comments retrieved from Brazilian Reddit communities were labeled by annotators and submitted to automated classification using 10 models with different architectures. Human labeling showed moderate agreement coefficients and reasonable disagreement, highlighting the subjectivity of the task. Models based on LLMs and BERT performed well with Brazilian Portuguese texts. The comparison revealed similarities in the challenges faced by humans and models, suggesting opportunities to improve automated language understanding. Both humans and models face similar difficulties in sentiment assignment, language characteristics of the texts being a major challenge for model classification, which points to the need for further advancement in this respect.

Keywords: Sentiment Analysis, Classification Models, Human Labeling, Linguistic Analysis

1 Introduction

Social networks have broken down communication barriers, allowing people to interact with friends and family around the world, participate in debates, and learn about a wide variety of subjects [Amedie, 2015]. In addition, they have enabled the rapid dissemination of up-to-date news [Siddiqui and Singh, 2016]. Every year, the number of users increases, with growth rates of more than 5% per year, having reached 5.07 billion users at the beginning of April 2024 [Kemp, 2024]. However, this expansion has also meant a growing number of vulnerable people who see their emotions negatively affected by interactions on these platforms [Kramer *et al.*, 2014]. In this respect, a survey aimed at understanding the effect of cyber aggression on adults in Italy, carried out at the University of Turin, found that of the 341 respondents, 43% reported having been victims of cyber aggression. According to 95.1% of the respondents, these incidents took place on social networks and were considered potentially harmful to mental health [Martella *et al.*, 2021].

Given the volume of user-generated content on a daily basis (more than 4.4 billion posts in the second half of 2023 in the case of the Reddit platform [Reddit, 2023]), moderating it has become a challenging and costly problem. To deal with this, large companies such as X, formerly Twitter, adopted machine learning and human review models [X, 2024]. These models assist platforms in taking the measures they deem necessary regarding content identified as violating their guidelines. However, even though there are models capable of carrying out these tasks, their performance is limited in languages with less available data, such as Brazilian Portuguese.

Some research initiatives have produced Brazilian Portuguese datasets with human-labeled sentiment annotations for social media texts. Most of the work has focused on texts retrieved from Twitter, with texts having specific language patterns, which is why models trained with them show limitations when applied to texts from other social networks. In the case of Reddit, there is a shortage of datasets of texts originally written in Brazilian Portuguese, extracted from Reddit and annotated with sentiment labels, as well as models that automatically analyze the sentiments in texts extracted from this social network.

This paper seeks to expand NLP (Natural Language Processing) resources for Brazilian Portuguese by exploring sentiment in texts extracted from Reddit [1]. Reddit is a platform that allows users to interact through anonymous posts (submissions) and comments. Users are organized into communities (subreddits) and subscribe to the communities most aligned with their topics of interest.

The main contributions of our work are twofold. The first one, previously published in [Piorino *et al.*, 2024], is the creation of an annotated Reddit dataset for sentiment analysis. The data annotated with one of three sentiments (*Positive*, *Negative* and *Neutral*), along with the characterization of the language in the texts labelled for each sentiment class can be used to support new sentiment classification models and improve existing ones so that they are better suited to the specific characteristics of the Portuguese language. We also

[1] http://reddit.com/

clarify that the present work is an extended and revised version of the work explained above.

Our second contribution, further exploring the results presented in [Piorino *et al.*, 2024], is a thorough evaluation of the sentiment analysis task in the domain under study, considering different types of classification models found in the literature: (i) open-source models – LeiA (Lexicon for Adapted Inference) [Almeida, 2018], Pysentimiento [Pérez *et al.*, 2024], XLM-RoBERTa [2](*Cross Lingual Language Model - Robustly Optimised BERT-Pretraining Approach*)[Barbieri *et al.*, 2022], BERTimbau [Souza *et al.*, 2020] and BERTabaporu [Costa *et al.*, 2023]; and (ii) large language models (closed-source models) – Sabia-3 [Abonizio *et al.*, 2024] and GPT-4 [OpenAI, 2024].

Additionally, we compare the performance of automated classification with human annotations, identifying key linguistic features that can aid in understanding cases of disagreement. This analysis helps in identifying the main problems faced by models during the classification task in the context of the Brazilian Portuguese language. As a means to allow reproducibility and foster follow-up studies, we have released our collected dataset for public use. Due to user privacy protection, the data we provide contains only the body of the comments and the labels from both the manual and automatic annotation processes.[3]

The remainder of this paper is organized as follows. Section 2 provides a brief account of related work. Section 3 describes our methodology, while Section 4 presents our main results. Finally, Section 5 concludes the paper and offers possible directions for future work.

2 Related work

There is a diverse and extensive body of literature on sentiment analysis of user-generated content on social media [Dang *et al.*, 2020], [Pereira, 2021], the vast majority of the works focusing on content published in English [Zhang *et al.*, 2025], [Melton *et al.*, 2021], [Nandurkar *et al.*, 2023]. Beyond the scope of studies focused on the English language, the authors in [Bibi *et al.*, 2024] survey 40 different research works on sentiment analysis using machine learning in various languages, including Italian, German, Urdu, Arabic, Spanish, and French. The study by [Corso *et al.*, 2024] analyzes Reddit comments from communities related to France and Italy to analyze discussions about Russia's invasion of Ukraine. Given our focus on Portuguese language, we present below a non-exhaustive list of studies that have explored sentiment analysis in social network texts in Brazilian Portuguese, as well as more recent studies that focus on analyzing the performance of learning models based on *transformer architecture* for the task of classifying sentiment in texts.

One of the first works focusing on developing learning models and/or analyzing their performance in sentiment analysis for Portuguese texts is [Hutto and Gilbert, 2014]. The authors proposed an extension of VADER (*Valence Aware Dictionary for Sentiment Reasoning*) for Portuguese, called LeIA [Almeida, 2018] an acronym for (*Léxico para Inferência Adaptada*), which labels texts with a *Positive, Negative and Neutral* class and can be adapted to different contexts, without being restricted to the scope of texts from a specific social network.

Focusing on analyzing the sentiment expressed in Twitter content, the work in [Garcia and Berton, 2021] explored topic detection and sentiment analysis on Brazilian Twitter texts on topics related to COVID-19. The topic extraction approach used was LDA (*Latent Dirichlet Allocation*) and sentiment analysis for Portuguese texts drew on mUSE(*Multilingual Universal Sentence Encoder for Semantic Retrieval*) and SemEval 2018. The authors in [Yang *et al.*, 2019] also collected tweets in Brazilian Portuguese in order to compile a dataset [4] of 15,000 tweets extracted between January and July 2017. The tweets were classified with the labels *Positive, Negative and Neutral* by annotators whose annotation obtained *Krippendorf's Alpha* metrics of 0.529, considered moderate agreement.

Proposals for improvements in task pipelines related to sentiment analysis are presented [Oliveira *et al.*, 2023]. The work aims to analyze some aspects and difficulties of the task in general and how to improve it. Similarly, our work aims to identify potential models that are more suitable for performing the task, as well as to analyze linguistic aspects of their classification errors as a way to assess the challenges of the task.

[Souza *et al.*, 2024] explore Brazilian content, more specifically user opinions about Brazilian public spaces, but their data is restricted to *Twitter*. They collected approximately 100,000 *tweets* and performed sentiment classification of the texts based on the BERTimbau model. They examined some issues pertaining to labeling were, such as the differences between adopting three labels (*Positive*, *Negative* and *Neutral*) compared to a polarization gradient (a scale from 1 to 5 from *Positive* to *Negative*, for example). In addition, they extracted the main topics using *BERTopic* in order to comprehensively characterize the content in their dataset. Authors in [da Silva Oliveira *et al.*, 2024] report on a comparison in the performance of two LLMs which we also addressed in our research (though in different versions): ChatGPT, in their study GPT-3.5-turbo-0613, and Maritalk, Sabia-65B in their study. The also compared performance between zero-shot and few-shot strategies. Despite being related to a slightly different task (toxicity detection in a *Twitter*), the work explores interesting strategies in the use of LLMs for labeling and various related performance analyses.

Turning our attention to the Reddit, the authors [Demszky *et al.*, 2020] used the *GoEmotions* model based on a dataset with approximately 58,000 comments manually labeled with emotion categories, written in English and translated into Portuguese. The study also explored language patterns for the identified emotions in the labeled comments and obtained evaluation metrics for the annotations and the model. However, due to the large number of emotion categories present in the labeling, the metrics yielded moderate or weak values for inter-annotator agreement.

[Koncar *et al.*, 2021] compared a wide range of text and

[2] https://huggingface. co/docs/transformers/model_doc/xlm-roberta

[3] https://github.com/SentPortugueseDataset/SentimentAnalysisReddit

[4] https://bitbucket.org/HBrum/tweetsentbr/src/master/

user characteristics to analyze and predict controversy in a multilingual environment on Reddit, Portuguese being one of the languages included. One of their findings regarding Portuguese is that a large number of the *Negative* comments collected on the network discussed Brazilian politics. In [Júnior *et al.*, 2022], the authors used a dataset of texts extracted from Twitter and Reddit to evaluate different configurations of preprocessing pipelines for texts in Brazilian Portuguese, which could be implemented before applying topic modeling methods. The adaptations performed showed improvements in all metrics.

Finally, the work presented in [Pereira *et al.*, 2023] implement a labeling pipeline using the GPT-3 model to perform sentiment classification on tweets related to the 2022 Brazilian presidential elections. This automatic labeling stage, based on GPT-3 prompts, served as an intermediate process for applying other analytical techniques, with the model's classifications subsequently reviewed manually. The validation indicated a satisfactory performance of GPT-3 in carrying out this task. [Herculano *et al.*, 2024] explore the analysis of different Brazilian Reddit communities related to mental health, employing BERT models for text classification tasks aimed at investigating potential depressive disorders among users. The classification is based on three distinct levels of an index representing the severity of depressive disorder, derived from the textual content of user comments. The study also includes training performance comparisons between the BERTimbau and BERTabaporu models, with BERTabaporu showing a slight advantage in F1-Score and Accuracy metrics.

Despite the recent growth of the Reddit social network, there are still few references in the literature on textual analysis and sentiment annotation tasks in texts from this network, especially in Brazilian Portuguese. Thus, our study seeks to expand the resources of NLP in Brazilian Portuguese, providing a dataset annotated with sentiments, along with the results of the central metrics of human annotation evaluation and a characterization of the language of the texts in the dataset. In addition, we carried out a performance analysis of a diverse set of learning models for sentiment classification in Portuguese. Most of the models analyzed are based on the *transformers* architecture [Vaswani *et al.*, 2017]. Our analyses seek to understand how language patterns specific to Brazilian Portuguese and the Brazilian community on Reddit impact the accuracy of the models studied.

3 Methodology

This section first presents the methodology used to collect the original dataset. Figure 1 provides an overview of the methodology proposed in this work. For data collection, we first selected the communities of interest. The community selection aims to identify the main Brazilian communities in terms of number of subscribers on Reddit, thereby choosing the most relevant communities in terms of content generation within the national context (Step 1: Data Collection). The originally collected data is then filtered, in order to generate a representative sample of the selected communities and time period (Step 2: Data Sampling). This sample is essential for conducting the manual annotation process (Step 3: Manual Annotation). We then analyze the linguistic characteristics of our human-labeled dataset (Step 4.a: Linguistic Characterization) and propose and evaluate a set of models to automatically identify sentiment expressed in Portuguese Reddit comments (Step 4.b: Classification Models and Evaluation).

Table 1. Selected subreddits and their total number of posts and comments (2022).

Subreddit	Posts	Comments
r/brasil	115,876	2,382,928
r/desabafos	115,876	1,487,076
r/futebol	35,826	1,272,009
r/saopaulo	7,308	88,894
r/eu_nvr	12,631	221,348
r/botecodoreddit	7,059	62,999
r/conversas	21,967	355,761
r/investimentos	9,756	156,695
r/tiodopave	2,371	12,106
r/brasilivre	67,301	1,308,441
Total	390,924	7,348,257

3.1 Dataset

Reddit is an online social media organized into subcommunities by areas of interest or subreddits, in which users discuss different topics through post-comment interactions, called *threads*. Our original database consists of user activity (posts and comments[5]) between January and December 2022 in the top 10 Brazilian communities with the highest number of active users. Table 1 shows the main statistics for the 10 selected communities. The data was collected from the Pushshift platform, which has been collecting, analyzing and archiving Reddit content since 2015 [Baumgartner *et al.*, 2020]. This data was previously presented in [Lima *et al.*, 2024] and used for the task of classifying the toxicity of the comments shared in these subreddits.

Ethical issues: We carefully addressed ethical considerations during the comment collection process, particularly with respect to user privacy and the types of data shared in the dataset. Since Reddit interactions are inherently anonymous, no personally identifiable information is included or disclosed without consent. The public dataset contains only the text of the posts along with the sentiment labels assigned by human annotators and the models applied. As a result, it contains no sensitive user data.

3.2 Data annotation

To perform the human annotation of the sentiment associated with each comment, 2,000 comments were selected from the original collected base, following a stratified sample of the total number of comments in each community analyzed. These comments were divided into 4 groups, with 500 comments each, called *Group1, Group2, Group3, Group4*. Each group was annotated by 3 different annotators.

[5] In this work, the term 'comments' will be used to cover both comments and posts.

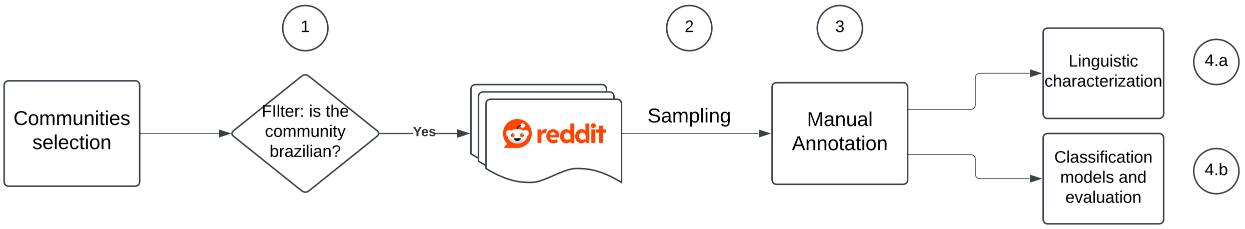

1
2
3
Communities selection
FIlter: is the community brazilian?
Yes
reddit
Sampling
Manual Annotation
Linguistic characterization
4.a
Classification models and evaluation
4.b

Figure 1. Methodology overview.

The annotators were university students who were invited to participate anonymously and instructed to read and classify each comment as *Positive, Negative, Neutral or I don't know*, taking into account the predominant feeling in each text. Whenever they felt it was impossible to determine a sentiment, the option to be chosen should be *I don't know*. To help identify the predominant feeling, the annotators were asked to pay special attention to two points: (i) *Negative* comments generally express emotions of fear, guilt, hurt, sadness, anger, anguish, anxiety and depression; and (ii) *Neutral* comments do not have any characteristic that could account for classifying them as *Negative* or *Positive*.

It is worth noting that, when providing instructions to annotators, we aimed to maintain a standard of direct guidelines, an approach also adopted by [Brum and das Graças Volpe Nunes, 2018] when working with Brazilian Portuguese datasets derived from social media comments. Additionally, following the insights of [Parmar *et al.*, 2023], a study that analyzed bias in annotator instructions, we avoided including specific examples, such as sample comments, to prevent the formation of strong associations between labels and specific content or phrasal constructions.

At the end of the annotation process, each comment was labeled with the sentiment assigned by the majority of the annotators and the agreement between them was measured by three commonly used metrics: *Fleiss' Kappa* [Fleiss, 1971], *Krippendorf's Alpha* [Krippendorff, 2019] and *Observed Agreement* [Fleiss, 1975].

3.3 Text Analysis

Before starting our text analysis, the 2,000 annotated comments were filtered using regular expressions in order to detect the content of the comments to be excluded from the analysis: website addresses, mentions of other users, hashtags, quoted texts, dates and emojis. Giggly texting acronyms to express laughter were removed; so were comments containing grammatical words that occurred in isolation and had no information value for our analysis; this was based on the list of (*stopwords*) in the NLTK library [NLTK, 2023b] and a model from [spaCy, 2023]. Thus, 19 comments were removed from the analysis upon filtering.

3.3.1 Word clouds

To have an overview of the collected and human annotated comments, we used wordclouds [Mueller, 2024]. This approach makes it possible to identify the most frequent words in the interactions between users of the Reddit communities under analysis. In our study, we explored some characteristics present in each of the sentiments analyzed, for example, by analyzing unique words frequently found in each of the classes (*Positive*, *Negative*, *Neutral*).

3.3.2 *Type-Token Ratio (TTR)*

With the tokenization of comments done by the NLTK library [NLTK, 2023a], we determined lexical diversity using the TTR measure. The TTR result is obtained by dividing the number of distinct tokens by the total number of tokens in the comment. We complemented the analysis by evaluating the size (in number of tokens) of the comments in each group.

The TTR metric provides an estimate of how varied the vocabulary is in a given text. Higher TTR values often indicate more sophisticated or complex language. In our context, it is interesting to gain some insight into how each sentiment is expressed by users. Moreover, the TTR metric is easy to compute and is widely used in linguistic analysis. As stated by the authors in [Rosillo-Rodes and Sánchez, 2025], this metric is used and emphasized for large text corpora, being considered a classic indicator of lexical diversity. Its use is justified as a well-established and stable measure of lexical diversity, widely applied in linguistic studies, language acquisition, stylometry, among other fields.

3.3.3 Part-of-speech tagging (*POS Tagging*)

To examine the predominant part-of-speech classes in labeled comments, we performed POS tagging [Petrov *et al.*, 2011] with a pre-trained [spaCy, 2023] model[6], a large-scale model specifically trained for the Portuguese language, based on a treebank annotated according to the Universal Dependencies pattern [Freitas *et al.*, 2008]. This treebank is mainly based on the work of [Rademaker *et al.*, 2017]. The spaCy model comprises several linguistic analysis components, such as tokenization, lemmatization, dependency parsing and named entity recognition. For the purposes of this section, we specifically employ its morphologizer component, related to the part-of-speech tagging task.

3.3.4 Named Entity Recognition (NER)

In this section, we leveraged the NER pipeline component of the aforementioned pre-trained spaCy model to perform the Named Entity Recognition task. To adapt the model to

[6] pt_core_news_lg

our data, we used our POS Tagging model and the dataset WikiNER [Nothman *et al.*, 2013]. This technique classifies entities into 3 categories: PERSON (PER), LOCATION (LOC) and ORGANIZATION (ORG). Entities that do not fall into these categories are classified as MISCellaneous (MISC).

3.3.5 N-gram analysis

To complement our language analysis, we carried out an n-gram analysis to explore contiguous sequences of n items.

3.3.6 Topic classification (BERTopic)

In order to characterize the most frequent themes in the texts and how they relate to the sentiments labeled by the annotators, we used the BERTopic model [Grootendorst, 2022] to extract topics from the comments. The comments were converted into vectors with the help of the BERTimbau model [Souza *et al.*, 2020], in which there is an additional fine-tuning stage aimed at the similarity of textual semantics [Fonseca *et al.*, 2016; May, 2021; Real *et al.*, 2020].

To ensure a more consistent modeling of the topics, the dimensionality of the vectors was reduced using UMAP (*Uniform Manifold Approximation and Projection for Dimension Reduction*), a technique that improves subsequent groupings. The HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) algorithm was also used to group the vectors based on semantic similarities. Finally, c-TF-IDF (*Class-based Term Frequency-Inverse Document Frequency*) and MMR (*Maximal Marginal Relevance*) were applied and adjusted to improve the definition of keywords for topics and to diversify their semantic content, respectively. Recommendations from the model's documentation [7] were used to adjust these parameters. For UMAP, the number of neighbors was adjusted to 10 and the number of components to 8. For HDBSCAN, the minimum number of clusters is 10 and the minimum number of samples is 8. The MMR parameter was updated to a rate of 0.8.

3.3.7 Semantic labeling (PyMUSAS)

For the semantic analysis of the comments, the pyMUSAS tool was used, based on the USAS7[8] (*UCREL Semantic Analysis System*) framework adapted to the Python language.

In short, it presents a structure organized into codes, which represent distinct semantic categories [Piao *et al.*, 2015]. Each comment can be framed in one or more semantic categories, providing a broad and abstract view of the contents present in the comments and how they are related to the labeled sentiments.

3.4 Sentiment Classification Models

To measure the correlation between human and automatic sentiment classification in the sampled Reddit comments, we applied a set of models with a diversity of architectures (classical and deep neural network models), methodologies (pre-trained and untrained models) as well as prompt-based models [Akbik *et al.*, 2018]. It should be noted that for our model application we used different filters than those used for text analysis in Section 3.3. Hence, for exploring sentiment classification models, we only removed URLs from the comments, retaining most of their content, including emojis and texting acronyms.

3.4.1 Open-Source Models

In this section, we present the main open source classification models found in the literature that were applied to the Reddit data sample [Wu and Wan, 2025] to classify the sentiment expressed by users. We considered VADER/LeiA model **VADER/LeiA** [Almeida, 2018] as our baseline and the other models analyzed are variations of the BERT (Bidirectional Encoder Representations from Transformers) model [Devlin *et al.*, 2019]. In the case of the Pysentimiento [Pérez *et al.*, 2024] and XLM-RoBERTa [Barbieri *et al.*, 2022] models, we used their pre-trained embeddings on other bases. The BERTimbau and BERTabaporu models were trained using our annotated dataset. The models were adapted to the classification task using the *Hugging Face* platform.[9] 512 tokens were used for text input, along with AdamW [Kingma and Ba, 2017] and a *learning rate* equal to $1.0e^{-5}$. In the training phase, we applied k-fold cross validation, with the parameter $k = 10$.

VADER/LeiA [Almeida, 2018]. Leia (Lexicon for Adapted Inference) is an adapted version of VADER (Valence Aware Dictionary and Sentiment Reasoner) model [Gilbert, 2014] for the Portuguese language, differing in the lexicon (in Portuguese) applied to define the sentiment value assigned to each comment analyzed.

Pysentimiento [Pérez *et al.*, 2024]. This model has pre-trained versions with texts in Spanish and Portuguese [Pérez *et al.*, 2024]. The Portuguese version is based on fine-tuning using the BERTabaporu model [Pablo Botton da Costa, 2022] on a training corpus of 15,000 tweets in Brazilian Portuguese [Brum and das Graças Volpe Nunes, 2018], manually annotated for the sentiment analysis task.

XLM-RoBERTa [Barbieri *et al.*, 2022]. Based on the BERT model, the version of XLM-RoBERTa (*Cross Lingual Language Model - Robustly Optimized BERT-Pretraining Approach*) used in this work was previously trained on approximately 10 million *tweets* in the Portuguese language, specifically for the task of sentiment classification [Barbieri *et al.*, 2022].

BERTimbau [Souza *et al.*, 2020]. Model trained on the *brWaC (Brazilian Web as Corpus)* web text corpus [Wagner Filho *et al.*, 2018]. From 3.53 million documents, 2.68 billion tokens were extracted, enabling high performance in classification tasks with texts written in the Portuguese

[7] https://maartengr.github.io/BERTopic/index.html

[8] https://ucrel.lancs.ac.uk/usas/

[9] https://huggingface.co/

language.

BERTabaporu [Costa *et al.*, 2023]. This model was trained on a set of 238 million tweets in Portuguese, written by 100,000 different users, resulting in more than 2.9 billion tokens. The fact that it was trained with texts from a social network makes it attractive for use in other online social media platforms. In the results presented in [Costa *et al.*, 2023], *BERTabaporu* outperforms the *BERTimbau* model in three tasks: *stance detection, mental health status and political alignment.*

3.4.2 Large Language Models

We also analyzed two large proprietary language models (LLMs): GPT-4 [OpenAI, 2024] and Sabia-3 [Pires *et al.*, 2023]. The first model is trained with data from several languages (including Brazilian Portuguese), while the second is a (*fine-tuned*) version for Portuguese data. For both models, we defined a *prompt* which indicates, in a well-structured and clear way, the main instructions to perform the sentiment classification task. Furthermore, since our sentiment classification task is conducted in Portuguese, we used a prompt in that language.

Box 1 shows the *prompt* used for the classification task, using the *zero-shot* approach. For the *few-shot* approach, we used the text of the zero-shot prompt *prompt*, adding 28 examples of each class from our training set [Brown *et al.*, 2020]. Finally, for both models, we considered the maximum size of input tokens for each model to be equal to their respective default settings, which respect the maximum limit set by the model, and temperature [10] to be equal to zero.

> Você é um assistente que classifica comentários do Reddit em Português do Brasil (PT-BR) como *Positivo*, *Negativo* ou *Neutro*. Você receberá o texto de um comentário e a sua tarefa é classificar o sentimento do texto fornecido.
>
> Use somente as informações abaixo para fazer a predição:
>
> 1. Para cada comentário se limite a escolher apenas uma dessas três opções, sem acrescentar texto explicativo e sem marcar outras opções que não sejam uma dessas três: *Positivo*, *Negativo* ou *Neutro*;
> 2. Marque somente como *Positivo* os comentários que tiver certeza, alta confiança de que tenham o sentimento positivo;
> 3. Marque somente como *Negativo* os comentários que tiver certeza, alta confiança de que tenham o sentimento negativo;
> 4. Marque somente como *Neutro* os comentários que tiver certeza, alta confiança de que tenham o sentimento neutro.
>
> Para cada comentário abaixo marque uma das opções: *Positivo* ou *Negativo* ou *Neutro*.

Box 1. Zero-shot prompt used for LLMs.[11]

Table 2. Annotator agreement.

Metric	All annotations	Sentiment only
Fleiss' Kappa	0.40	0.51
Krippendorf's alpha	0.47	0.53
Observed agreement	0.60	0.70

Table 3. Examples of comments that achieved total agreement among annotators.

Sentiment	Sample comment
Positive	Ahh para, eu curto cidadezinha, às vezes eu vou pra uns lugares desses, fico uns 2 ou 4 dias, acho super legal. English gloss: *"Ahh stop it, I like small towns, sometimes I go to places like that, I stay for 2 or 4 days, I think it's super cool."*
Negative	Intervencionismo externo visando ganho próprio e sem estudar a situação complexa e possíveis consequências. Um clássico dos Estados Unidos de m*rda. English gloss: *"External interventionism aimed at self-gain and without studying the complex situation and possible consequences. Classic United States cr*p."*
Neutral	Subsídio para quem vender preferencialmente para o mercado interno ou, ao contrário, cobrar mais imposto sobre o produto exportado. English gloss: *"Subsidies for those who sell preferentially to the domestic market or, on the contrary, charge more tax on the exported product."*

4 Results

In this section, we present the main results obtained in the evaluation, modeling and characterization of the annotated dataset.

4.1 Agreement between annotators

The metrics for analyzing the results of agreement between annotators were applied to the subsets *All annotations*, covering all comments labeled with the four available categories, and *Sentiments only*, covering comments labeled without the *I don't know* label, in order to verify the impact of this uncertainty label on the results. Table 2 shows the results. *Krippendorf's Alpha* and *Fleiss' Kappa* showed values that can be interpreted as moderate agreement between the annotators. Observed agreement, on the other hand, shows considerably higher values than the other metrics, but is not as robust, as it does not consider that agreement between annotators may have occurred by chance. In general, it can be seen that the quality of the metrics improves considerably by including only the comments labeled with sentiments and disregarding the *I don't know* category.

With regard to total agreement between annotators, i.e. all annotators assigning the same label, the percentage of

[10] Temperature is a parameter that allows you to modify the output of a language model, making it more predictable or creative.

[11] See English gloss in Appendix.

Table 4. Percentage of comments by manual label and for total disagreement.

Classification	Percentage
Negative	48.05%
Neutral	20.95%
Positive	16.30%
Total disagreement	10.55%
I don't know	4.15%

comments that achieved total agreement was 44.65% in the subset that considers all annotation labels. In the subset of comments with only sentiment labels, total agreement increased to 57%. Some examples of these comments can be found in Table 3.

In order to establish sentiment classifications for later analysis of comparison with automatic models and text characterization, we assigned the occurrences of partial agreement to the predominant annotated sentiment, i.e. the agreement of two or more annotators on the same label. Table 4 shows that almost half of the comments were mostly labeled as *Negative*, indicating a considerable class imbalance. *Positive* and *Neutral* labels had similar proportions. 10.55% of the comments obtained total disagreement, i.e. each annotator gave a different label. This amount of disagreement may be the result of different perspectives that each annotator may have on what is Positive or Negative [Mokhberian *et al.*, 2023] or the presence of sarcastic content, or lack of additional context to facilitate the assignment of a sentiment.

Table 5 shows some examples of comments in which there was total disagreement between the annotators. Comments #2 and #4 exemplify the scenario where the lack of contextualization of the discussion *thread* related to politics may have contributed to the disagreement found. Comment #1 presents an example of the presence of irony. The lack of context can lead to one of the annotators identifying such an ironic comment, therefore labeling it as *Negative*, while another annotator did not, labeling it as *Positive* or *Neutral*. Finally, some comments present conflicting ideas, with positive and negative aspects in the same sentence, such as comment #5, which expresses the idea that Brazil is a very good place, but criticizes its politicians with curses. This conflict increases the complexity of the annotator's analysis of the comment, a situation ripe for disagreement. In short, some language resources such as irony, as well as the possible lack of general context and the divergence and weighting between polarized ideas in the same comment are points that can generate greater difficulty for human annotation, a source of uncertainty and divergence. Due to the very structure of the Reddit social network, in which numerous discussions can develop around the same initial post, this difficulty is aggravated, especially with regard to the lack of context provided by a comment. One option that can be used to alleviate this difficulty is to provide the entire discussion *thread* around the comment, but this technique can also make labeling more laborious for annotators.

In addition to the cases of total disagreement, an interesting analysis to be carried out is related to the degree of uncertainty present in the labeling task. For the total of 2,000 labeled comments, the option *I don't know* was selected by at least one of the three annotators in 23.05% of the total set. However, in only 4.1% of the comments did two or more annotators label the same comment with *I don't know*, a significant drop which may indicate that it is more difficult for two or more annotators to characterize sentiment uncertainty for the same text. An example of a text in which 2 or more annotators showed uncertainty in labeling is: *"Curti muito sua dupla personalidade, hehe."*[12], which seems to be a sarcastic comment, making the labeling task more difficult even for humans.

Table 6 shows in detail the performance of each trio of annotators and their respective metrics. We can see that Group 4 had the best performance in comment annotation while group 3 had the worst performance. However, in general, the annotations obtained reasonably close agreement metrics, pointing to average and moderate agreement between their respective annotators. In addition, Table 7 shows the labeling of feelings by each annotator within each group of comments.

We also found great variation in the labels chosen by the annotators. Table 8 shows the results of the metrics for annotator agreement in each group of comments. We can see that for the same group of texts, annotator 1 labeled 13.60% of the comments as *I don't know*, while annotator 3 labeled only 0.40% of the comments. These results reaffirm what has been said in the literature about the subjectivity of sentiment evaluation and the difficulty of this task. Additionally, analyzing the data in Table 7, it can be seen that group 1 generally had a higher labeling of *Positive* comments, and annotator 3 in this group was the one who labeled the most Positively, by a large margin of difference compared to the others. On the other hand, annotator 2 in group 4 was the one who labeled the most negatively, even though, in general, this annotator obtained a considerably constant proportion of annotations compared to the other annotators in the group, which is the one with the best agreement metrics. Group 3, the group with the lowest agreement metrics, showed considerable disparities between labeling proportions, with annotator 1 in this group showing a reasonable discrepancy in labeling proportions in relation to annotators 2 and 3.

4.2 Language Characterization

Language patterns were compared by grouping comments based on the predominant label of the 3 labels given to each comment by the annotators. A p-value < 0.5 was used in all analyses to ensure statistical significance and we present the results with their average the 95% confidence interval. Table 9 shows the data after filtering out texts and, consequently, comments that did not contain relevant content. This data was used in the analysis.

Word Cloud: Figure 2.a shows the most frequent words in comments labeled *Positive*. Some of these words (for example, *leio [I read], escreva [write], gramática [grammar]* are linked to educational themes, suggesting that our corpus tends to refer to education as a Positive sentiment. For comments labeled *Negative*, the most frequent words shown in Figure 2.b include *Brasileiros [Brazilians], debate*,

[12] English gloss: *"I really liked your dual personality, hehe."*

Table 5. Comparison between labels assigned by each annotator in cases of total disagreement.

#	Comment	Annotator 1	Annotator 2	Annotator 3
1	vc é um poeta amigo, faz letras ? English gloss: *”you’re a poet my friend, are you a literature student?”*	Negative	Neutral	Positive
2	Vc está certo, só com lula 2022 essa bandidagem fascista vaza English gloss: *”You’re right, only with Lula 2022 will this fascist thuggery end”*	I don’t know	Negative	Neutral
3	Café é igual carne. Tem café de terceira e tem café de primeira. Não adianta comprar café pilão achando que vai ser bom. English gloss: *”Coffee is like meat. There’s third rate coffee and there’s first rate coffee. It is no use buying third rate coffee like Pilão expecting it to be good.”*	Positive	Negative	Neutral
4	Sim, rola grupo de oração e leitura de versículos bíblicos para reflexão além de cantar hinos da harpa cristã English gloss: *”Yes, there is a group for prayer and reading biblical verses for reflection as well as singing hymns from the Christian harp”*	I don’t know	Neutral	Positive
5	eu ñ odeio o brasil, acho um lugar muito bom o f*da mesmo são os políticos English gloss: *”i don’t hate Brazil, i think it’s a great place, it’s the politicians that are the fucking problem*	Negative	I don’t know	Positive
6	Se continuar do jeito que tá, em poucas décadas uma revolução popular vai começar a decapitar empresário em praça pública English gloss: *”If it stays the way it is, in a few decades a popular revolution will start beheading businessmen in public squares”*	I don’t know	Negative	Neutral

Table 6. Evaluation metrics for annotator agreement for each group disregarding label *I don’t know*.

Metric	Group 1	Group 2	Group 3	Group 4
Kappa de Fleiss	0.41	0.39	0.34	0.44
Alfa de Krippendorf	0.48	0.48	0.40	0.50
Observed agreement	0.60	0.58	0.56	0.64

referring to the political debate during the data collection period. The occurrence of the word *pronto [that’s it]* is worth noting, which can be related to scenarios of irony, such as the following example: *“...Proibe o livro do Hittler, pronto acabou o nazismo…”* [13] . Exploring word clouds based on bigrams for the *Positive* comments (Figure 2.c), we find temporal references in bigrams such as, *hoje dia [today] [day]*, *anos influência [years][influence]* and *ano passado [year][past]*, as well as a strong emphasis on the word *vale pena [is][worth]* which has a connotation of advice, and can be exemplified in: *“......Vale muito mais a pena você ser uma pessoa acertiva (o que é totalmente diferente de ser arrogante ou deselegante), leal, e justa…”* [14] . Finally, when we analyze the word clouds based on bigrams for the comments labeled *Negative* (Figure2.d), the emphasis is on bigrams such as *“lula bolsonaro” [Lula] [Bolsonaro]*, *“oriente médio” [middle east]”*, *“salário mínimo”[minimum wage]”* and *“estado unidos” [united states]*. These bigrams reflect the influence of political polarization in Brazil and geopolitical issues in the analyzed corpus.

Type-Token Ratio (TTR): When analyzing TTR, there were differences between the average number of characters per comment per label, especially between *I don’t know* and the other labels. Comments labeled *I don’t know* had the lowest average, with 43.13 [29.08, 60.15]. *Neutral* comments had an average of 81.41 [69.65, 94.45], while the *Negative* and *Positive* comments had averages of 98.46 [90.57, 106.66] and 99.13 [80.11, 122.28]. These results may indicate that comments with *Negative* and *Positive* sentiment tend to be longer than *Neutral* comments and those that need more context to be interpreted, labeled *I don’t know*. However, when applying the Mann-Whitney test[15][Tallarida and Murray, 1987], no difference was found between *Neutral* and *Positive* comments. On the other hand, when we carried out the statistical test both to compare *Negative* and *Neutral* comments and to compare *Negative* and *Positive* comments, differences between them were significant.

With regard to the mean and the TTR confidence interval,

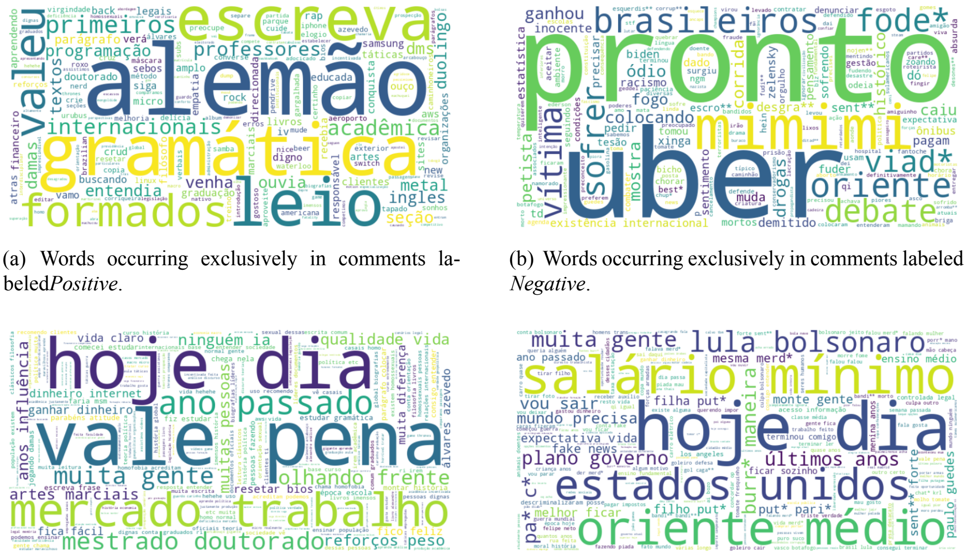
[13] English gloss:*“...Ban Hittler’s book, that’s it, Nazism is over…”*

[14] English gloss:*“......It’s much more worth it being someone who is confident (which is totally different from being arrogant or impolite), loyal, and fair...”*

[15] The Mann-Whitney test is a non-parametric test used to determine whether or not two groups of independent samples belong to the same population.

Table 7. Distribution of sentiment label per annotator.

	Group1			Group2			Group3			Group4		
	Annotator 1	**Annotator 2**	**Annotator 3**	**Annotator 1**	**Annotator 2**	**Annotator 3**	**Annotator 1**	**Annotator 2**	**Annotator 3**	**Annotator 1**	**Annotator 2**	**Annotator 3**
Positive	22.8%	16.6%	35.4%	19.6%	18.8%	19.8%	17.6%	24.0%	10.0%	15.2%	15.4%	14.8%
Negative	47.6%	46.8%	38.4%	45.4%	50.2%	44.0%	55.6%	43.6%	48.0%	42.0%	57.2%	46.6%
Neutral	16.0%	28.0%	25.8%	24.4%	21.4%	14.0%	19.4%	27.2%	25.2%	25.8%	27.2%	38.4%
I don't know	13.6%	0.86%	0.4%	10.6%	9.6%	22.2%	7.4%	5.2%	16.8%	17.0%	0.2%	0.2%



(a) Words occurring exclusively in comments labeled*Positive*.

(b) Words occurring exclusively in comments labeled *Negative*.

(c) Bigrams in comments labeled *Positive*.

(d) Bigrams in comments labeled *Negative*.

Figure 2. Word clouds

Table 8. Percentage of comments labeled *I don't know* per annotator and group.

Annotators	Group1	Group2	Group3	Group4
Annotator 1	13.60%	10.60%	7.40%	17.00%
Annotator 2	8.60%	9.60%	5.20%	0.20%
Annotator 3	0.40%	22.20%	16.80%	0.20%

Table 9. Number of comments by manual label and for total disagreement.

Classification	Number of Comments
Negative	960
Neutral	413
Positive	319
Total disagreement	210
I don't know	79

Table 10. Percentage of Part-of-Speech tags.

Classification	NOUN	VERB	ADJ	PROPN	ADV
Negative	35.51%	30.54%	18.28%	6.28%	4.17%
Neutral	34.97%	27.83%	18.08%	10.07%	3.92%
Positive	34.49%	32.19%	17.82%	6.31%	3.66%
I don't know	29.84%	26.16%	14.15%	18.41%	2.71%

the labels had the following values: *I don't know* had 0.98 [0.96, 0.99], *Neutral* 0.97 [0.96, 0.98], *Negative* 0.97 [0.96, 0.97] and *Positive* 0.97 [0.97, 0.98]. An analysis using the Mann-Whitney test showed that the only label with a significant difference compared to the others was *I don't know*. For the remainder of our results we will then report results using the Mann-Whitney test.

Part-of-Speech tagging (POS tagging): The average and confidence interval for diversity of POS tags for each label are as follows: *I don't know* shows 0.73 [0.66, 0.79], *Neutral*, 0.57 [0.55, 0.60], *Negative*, 0.50 [0.48, 0.51] and *Positive*, 0.56 [0.52, 0.59]. These results corroborate those obtained for TTR, especially concerning the difference between the *I don't know* label and the others in terms of diversity. It is worth noting that the *Negative* label had the lowest average.

Table 10 shows the representativeness of the main POS tags in relation to the total number of words tagged in each category of comments. For a more in-depth look, we analyzed the average number of words classified with specific tags per comment, starting with adjectives (ADJ). The average and confidence interval for each label are as follows: *I don't know* cluster had 0.93 [0.62, 1.29], *Neutral*, 1.94 [1.58, 2.36], *Negative*, 2.41 [2.21, 2.62] and *Positive*, 2.38 [1.90, 2.98]. The *I don't know* label had the lowest average. When we apply the statistical test to the other labels, we see that there are significant differences between all of them, based on comparisons between *Negative* and *Neutral*, *Negative* and *Positive*, and *Positive* and *Neutral*.

For nouns (NOUN), the average and confidence interval for each category are as follows: *I don't know* shows 1.96 [1.38, 2.65], *Neutral* 3.76 [3.24, 4.33], *Negative* 4.681 [4.31, 5.06] and *Positive* 4.61 [3.74, 5.66]. As in the case of adjectives, the category *I don't know* has the lowest average. The statistical test revealed significant differences when comparing *Negative* with *Positive* and *Negative* with *Neutral*, but this is not the case when comparing the *Neutral* with *Positive*.

The mean and confidence interval of label categories for verbs (VERB) are as follows: the label *I don't know* shows 1.71 [1.09, 2.52], the label *Neutral*, 2.99 [2.58, 3.43], the label *Negative*, 4.03 [3.68, 4.38] and the label *Positive*, 4.30 [3.45, 5.33]. The same pattern is observed for *I don't know* in the 3 labels. The statistical test indicated significant differences when comparing *Negative* with *Positive* comments and *Negative* with *Neutral* comments, but showed no significant differences when comparing *Neutral* with *Positive* comments.

Finally, *I don't know* is the only label category with more POS tags for proper nouns (PROPN) than adjectives (ADJ), as can be seen in Table 10. This also shows that this label contains the highest number of proper nouns.

Named Entity Recognition ('NER'): Comments classified as *I don't know* show a predominance of entities of the PER type, representing 51% of the entities identified, followed by 19% of entities of the LOC type, 16% of ORG and 14% of MISC. *Neutral* comments show a distribution of 43% PER entities, 25% LOC, 16% ORG and 16% MISC. *Positive* comments show 40% PER entities, 24% LOC, 11% ORG and 24% MISC. Finally, *Negative* comments show 44% PER entities, 35% LOC, 12% ORG and 10% MISC. These results highlight the predominance of PER entities in the *I don't know* comments, the number of LOC entities in the *Negative* comments and the significant presence of MISC entities in the *Positive* comments.

In addition, considering the 2,000 comments, our analyses showed an increase in the number of entities mentioned from January to February and from February to March, possibly due to the war between Russia and Ukraine. Moreover, there are peaks around October, coinciding with the election period in Brazil, with the exception of the group *I don't know*, probably because it has few comments, as shown in Table 8.

N-grams: Our analysis of n-grams shows that the bigram results in comments classified as *Positive* evidence life-related topics. As for *Negative* comments, we find the bigram *lula, bolsonaro [Lula] [Bolsonaro]*. In trigrams of comments labeled with a *Positive* sentiment, there are words related to counseling on relationships (e.g. *sociedade [society], vê [see], casais [couples]*). In trigrams of *Negative* comments, we find the combination *bandido [crook], bandido [crook], morto [dead]*, possibly related to political debates and ideological stance.

Topic extraction (BERTopic): We performed topic extraction, obtaining 15 topics, ordered by their frequency of occurrence in the comments, as shown in Table 11. Analysing comments in which there was total disagreement between annotators, the most related topics proportionally are, in descending order: 14, 3, 1, 9 and 13. While topics 3 and 1 are more generic, related to routine, family and everyday situations, topics 14, 9 and 13 are related to politics in different spheres: topic 14 is more related to political ideologies, especially Nazism; topic 9 is related to the concept of *fake news*, the result of election polls and political parties, and topic 13 deals with issues and themes concerning the government during Jair Bolsonaro's presidential term.

As for comments in which there was total agreement between annotators, topics 11, 12, 7 and 2 stand out. Considering the sets of words, these are generally polarized topics that convey Positive (topic 11) or Negative ideas (topics 7, 2), in addition to topic 12, which criticizes Brazilian governments.

Table 11. Topics and most frequent words.[16]

Topic	Most frequent words
0	pessoa, pessoas, ficar, nada, fazer, aí, coisa, ainda, vida, porque
1	carro, acho, nunca, vou, uso, desse, lembro, sei, ver, achei
2	burro, bozo, ai, and, of, p*ca, vem, comida, pode, comentário
3	nome, filho, criança, banho, banheiro, tomar, p*ta, durante, lembro, deve
4	dinheiro, pagar, salário, fazer, trabalho, mercado, todos, ganhar, história, sobre
5	brasil, país, estado, eua, direita, países, rússia, china, nuclear, esquerda
6	time, goleiro, jogo, gol, futebol, palmeiras, jogador, vasco, paulo, passado
7	f*da, odeio, mano, tô, p*rra, pqp, tomara, gosto, pena, horrível
8	palavras, entender, dia, 11, pois, pessoas, falando, palavra, países, comecei
9	falou, entendi, falei, resultado, fake, hoje, disse, pt, pesquisa, dia
10	bandido, quer, bunda, p*u, pq, mãos, matar, passou, cima, bola
11	obrigado, sorte, entendi, comentários, man, respeito, espero, deus, feliz, boa
12	lula, bolsonaro, bolsonarista, governo, auxílio, gastos, mal, época, presidente, contra
13	população, política, direito, popular, governo, político, saúde, economia, bolsonaro, passar
14	socialismo, amp, x200b, hitler, nacional, comunismo, alemães, contrário, dizem, igreja

With regard to comments that at least one annotator labeled as *I don't know*, topics 10, 14 and 2 stand out. Topic 10 has 32.6% of its comments labeled *I don't know* by at least one annotator, and features content related to crime, cursing and sexual content. Topic 14, in which 30.4% of comments have at least one label for *I don't know*, relates to ideological and political issues. Finally, topic 2, which has 28% of its comments with at least one annotator assigning the label *I don't know*, has generic content related to colloquial language and slang and everyday stories. With regard to comments which all the annotators labeled *I don't know*, 3 comments stand out in topics 2 and 3, generally related to everyday stories and facts and colloquial language and slang. Possibly because they draw on very specific contexts within a post, they are considered more difficult to label.

Semantic labeling (PyMUSAS): The results obtained from the semantic categorisation of the comments yielded a total of 163,704 labels for general semantic levels (main semantic categories that exclude punctuation, for example), bearing in mind that each word in each comment has one or more possible labels within the domain of the USAS[17] categories.

Considering all the comments, the categories that occur the most are *proper nouns, colloquial language and slang and swear words*, which make up 29.64% of the total occurrences, *abstract terms, which cover general actions, affection, classification, evaluation, comparison, possession, importance, ease/difficulty, degree, exclusivity and security*, which accounted for 17.6%, and *social terms, covering actions, states and processes, reciprocity, participation, merit, personality traits, people, relationships, family, groups, obligation, power*, which accounted for 9.17% of all categorised occurrences.

Considering only sentiment annotations, the categories *numerical terms* and *judgements of appearance and physical attributes, such as appearance, colour, shape, texture and temperature* are highly relevant in the comments labeled as *Positive* by annotators. In this case, the second category makes up 6.7% of all *Positive* comments, compared to 1.9% of *Negative*, and 2.8% of *Neutral* comments.

In contrast, for comments labeled as *Negative*, the categories *concepts of movement, location, travel and transport*

[16] See English gloss in Appendix

[17] https://ucrel.lancs.ac.uk/usas/Lancaster_visual/Frames_Lancaster.htm

Table 12. Performance metrics for models.

Metric	VADER/Leia	Pysentimiento	XLM-RoBERTa	BERTimbau	BERTabaporu	Sabia-3 Zero-Shot	Sabia-3 Few-Shot	GPT Zero Shot	GPT Few Shot
Accuracy	0.51	0.67	0.61	0.73	0.76	0.80	**0.82**	0.76	**0.82**
Precision	0.55	0.72	0.62	0.74	0.77	0.82	0.82	0.81	**0.84**
Recall	0.51	0.67	0.61	0.73	0.76	0.80	**0.82**	0.76	**0.82**
F1-Score	0.52	0.68	0.62	0.73	0.76	0.80	0.82	0.76	**0.83**

and *concepts of climate and environmental issues* stand out. The first category constitutes 9.5% of all occurrences categorised as *Negative* comments, compared to 3.3% for *Positive* and 5.0% for *Neutral*. For comments labeled as *Neutral*, the categories *concepts of science and technology*, *concepts of money, business, work and industry*, as well as *abstract terms covering general actions, affection, classification, evaluation, comparison, possession, importance, ease/difficulty, degree, exclusivity and security* stand out in relation to the proportions of *Negative* and *Positive* comments.

In addition, the category comprising *proper nouns, colloquial language and slang and swear words* makes up a considerable part of both *Positive* comments (28.65% of all Positive comments) and *Negative* comments (29.9%). Thus, colloquial language and slang, swear words and proper names may not be considered predominant characteristics in determining the sentiment of a comment, since for both sentiments, such language uses have a similar occurrence. This will be taken up when comparing human annotation with the best models, which often classify topics comprising a few swear words, more negatively compared with the average human labeling for negatives. Therefore, such semantic patterns can lead the model to label comments as *Negative* excessively, due to the difficulty in dealing with such language patterns in the texts.

For comments in which there was total disagreement between annotators, we find the categories *architecture, types of buildings and houses, constructions, residence, furniture and household accessories*, *concepts of money, business, work and industry* and *entertainment in general, music, theater, sports and games*. Considering the total number of occurrences of the category *architecture, types of buildings and houses, constructions, residence, furniture and household accessories* for all comments, 12.38% of them fall within comments showing total disagreement. For the categories *concepts of money, business, work and industry* and *entertainment in general, music, theater, sports and games*, the percentages are 10.37% and 10.10% respectively. These categories make up the highest proportions of total disagreement among all categories. These results indicate a certain difficulty in agreeing with annotations on specific subjects that involve the annotator's world knowledge, such as architecture, entertainment and the financial market, for example.

Finally, the analysis of predominant categories in the comments that the annotators labeled *I don't know* shows a predominance of the categories *artistic concepts, arts, crafts*, *food, drinks, tobacco and drugs, agriculture and horticulture* and *education and studies*.

4.3 Evaluation of automatic classification models

To evaluate the models described in Section 3.4, the annotated dataset was divided as follows. Of the 2,000 comments initially annotated, 1,706 were assigned one of the labels for the sentiment classes. Of these 1,706, we selected 10% of the sample, i.e. 171 comments, to be used as the test set for the models. The remaining comments were used to train and validate the BERTimbau and BERTabaporu models.

The models were compared using the Accuracy, Precision, Recall and F1-Score metrics [Sokolova and Lapalme, 2009]. Due to the unbalanced nature of the data analysed (the class of *Negatives* is 2.5 times bigger than the class of *Positives* and almost 2 times bigger than the class of *Neutrals*), we used the weighted average per class of these metrics [Hinojosa Lee *et al.*, 2024]. Table 12 shows the values for the set of metrics analyzed, with the values in bold for the models with the best performance in each of the metrics.

As expected, the baseline (VADER/Leia), which uses a lexical dictionary, is the model with the worst performance, close to a random choice for one of the classes to label a particular comment. Models that are BERT-based, but have been trained with data from other domains, perform between 1.2 and 1.3 times better on average when compared to the baseline.

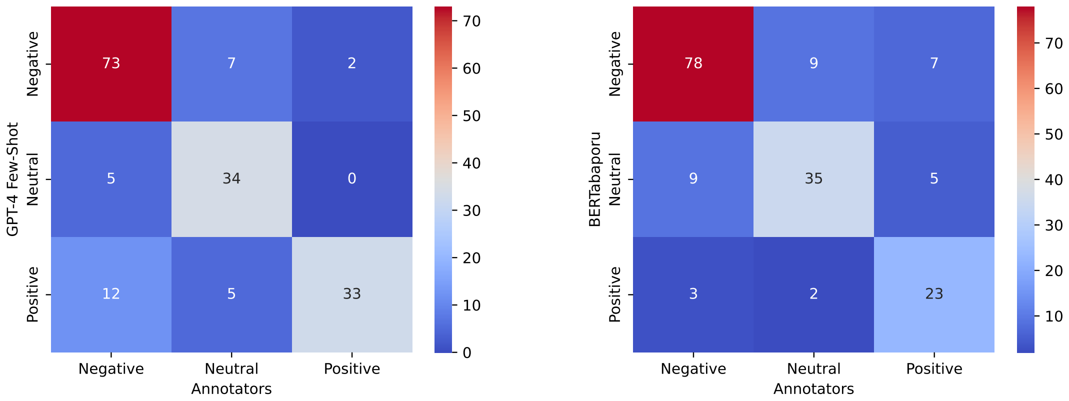
The best performance is reached by models trained with data from the original dataset (BERTimbau and BERTabaporu) or when large language models are used (Sabia-3 and GPT). The BERTabaporu model yields similar results to those obtained by Large Language Models, especially in cases where the zero-shot approach is used.

While our results corroborate the accuracy of different versions of Large Language Models for different machine learning tasks, including sentiment labeling [Zhang *et al.*, 2023; Mughal *et al.*, 2024], open-source models trained with our labeled data performed very similarly to Large Language Models. These results are extremely important, since proprietary models are expensive to use and do not allow their results to be interpreted.

4.4 Human and Automatic Labeling

This section presents a detailed comparison between human labeling and automatic labeling by the two best-performing models: BERTAbaporu (open-source) and GPT-4 Few-Shot (LLM), referred to as GPT-4. For this analysis, we considered the comments belonging to the test set made up of 171 comments.

Percentage of comments per sentiment. Table 13 shows the percentage of comments labelled as *Positive, Negative, Neutral* by the annotators and the models. We can see that the



(a) Confusion matrix between annotators and GPT-4 Few-Shot.

(b) Confusion matrix between annotators and BERTabaporu.

Figure 3. Analysis of labeled categories by best models and annotators using confusion matrices

GPT-4 Few-Shot model tends to classify fewer comments as *Negative* ($\approx 48\%$) when compared to the annotators. On the other hand, the BERTabaporu model tends to be more conservative in labelling sentiment *Negative* ($\approx 55\%$) when compared to the annotators. Despite being more conservative, the BERTabaporu model is closer to the behaviour of the annotators, who classified approximately 53% of the comments analysed as *Negative*.

It is interesting to note that GPT-4 tends to classify a higher percentage of comments as *Positive* ($\approx 29\%$), compared to BERTabaporu ($\approx 16\%$) and the annotators ($\approx 20\%$). We present two comments that GPT-4 classified as *Positive*, in contrast to the annotators and BERTabaporu, who classified them as *Negative*: (i) *" Percebo que as belíssimaS e moralmente corretas não são da cidade de Taubaté."*[18] and; (ii) *"Sei não em aposto que o Cristiano Ronaldo é mais cheiroso e bonito que essa mina, e menos chato CERTEZA."*[19] . These comments illustrate how specific adjectives, which generally have a positive connotation (*belíssimas, moralmente corretas, cheiroso, bonito*)[20], can be used in to construe pejorative meanings, with the function of diminishing some other entity, such as the city of Taubaté and a woman. Considering this dataset, GPT-4 seems to be more sensitive to the use of irony and other language resources used by Brazilian users in online discussions.

Finally, considering the *Neutral* sentiment, BERTabaporu tends to classify a higher percentage of comments in this category ($\approx 29\%$). For example, for the comment *"Caso ela esteja flertando com o cara. kkkkk"*, [21] which annotators labelled as *Negative*, BERTabaporu labelled it as *Neutral*, most probably because it failed to identify sarcasm, whereas GPT-4 labelled it as *Positive*, possibly because of the giggly texting acronym *kkkkk* considered as non-aggressive.

Correlation between labelings. When comparing annotators' labeling with the models', the following correlations are observed between the classes generated by the models and annotators. The Pearson correlation between annotators and BERTabaporu is 0.68, while GPT-4 had a correlation of 0.70. The correlation between BERTabaporu and GPT-4 is approximately 0.56. This moderate level of correlation shows a certain tendency among the models to assign the same labels to a given comment.

Table 13. Percentage of comments labeled for each sentiment.

Sentiment	Annotators	GPT-4 Few-Shot	BERTabaporu
Negative	52.63%	47.95%	54.97%
Positive	20.46%	29.23%	16.37%
Neutral	26.90%	22.80%	28.65%

Table 14. Observed agreement between best models and annotators for each sentiment

Sentiment	GPT-4 Few-Shot	BERTabaporu
Negative	81.11%	86.67%
Positive	94.29%	65.71%
Neutral	73.91%	76.09%

An analysis of the *Kappa de Cohen* metric shows a high agreement between the different types of annotation carried out. The value of this metric for the labels assigned by annotators and GPT-4 was 0.71, while for annotators and BERTabaporu it was 0.66. For the two models, the value was 0.6. In addition, Table 14 shows the observed agreement for each of the annotated classes.

Figure 3 corroborates some of the insights presented earlier when analysing the proportions of labels for each category between the models and the annotators, together with the observed agreement between the models and the annotators, described in Table 14. One of them is GPT's preference for *Positive* labels: since its labelling rate for this category is reasonably higher than that of the annotators, its rate of success in this category is expected to be higher. In addition, the lowest concordance rate is for the label *Neutral*, possibly pointing to a greater tendency towards polarisation.

On the other hand, for the BERTabaporu model, as shown in Table 14, *Positive* labels had the lowest rate of agreement, while *Negative* labels had the highest one. This is also related to the proportion of the model's labelling

[18] English gloss: *" I understand that the most beautiful and morally correct women are not from the city of Taubaté."*

[19] English gloss: *"I don't know, I bet Cristiano Ronaldo is more perfumed and handsome than that girl, and less boring FOR SURE."*

[20] English gloss: *beautiful, morally correct, perfumed, handsome*

[21] English gloss: *"In case she's flirting with the guy. kkkkk"*

described above, in which the rate of *Positive* labels was the lowest of all the comparisons made. Therefore, the model has a tendency to assign *Negative* labels to comments.

Exploring disagreement. We now present a more in-depth analysis of the cases where there was disagreement in the choice of labels. Overall, the GPT-4 model had 31 comments with labelling in disagreement with annotators; BERTabaporu had 35 comments in this regard. An intersection of this set, in which both models disagreed with annotators, results in 15 comments. Finally, cases in which GPT-4 disagreed with annotators but BERTabaporu agreed with them, ammout to 16 comments; and cases in which BERTabaporu disagreed with annotators but GPT-4 agreed with them yield 20 comments.

First, considering GPT-4's labels, Figure 3 shows that, for comments that annotators labelled as *Negative*, the model tends to concentrate its errors on *Positive* labels. This corroborates the findings previously described in Table 13, showing that the model had a significantly higher rate of Positive labelling than the annotators. As for *Neutral* labelling, the model showed a slight tendency to label them as *Negative*. Finally, the model missed only 2 of the comments considered *Positive* by the annotators, a fact also related to its high rate of Positive labelling. Considering a reasonable frequency of polarising errors (annotators labelling a comment as *Negative* and the model as *Positive*, and vice versa), GPT4 has greater difficulty in interpreting uses of irony and sarcasm, as well as interpreting colloquial language and slang and swear words in non-pejorative comments.

For the BERTabaporu model, the confusion matrix shows that, for all classes of annotations, there is a tendency for the model to erroneously label comments as *Neutral*, while comments classified as *Neutral* by the annotators are considered *Negative* in higher numbers by the model. These trends are related to a certain difficulty of the model in identifying comments of the class *Positive*, and a certain difficulty in analyzing the importance of some words for the context, such as comparisons between adjectives, colloquial language and slang and swear words.

Regarding misclassifications by GPT-4, but correctly labeled by BERTabaporu, we find 16 comments, of which 11 were labeled as *Negative* by annotators, while GPT-4 classified 10 of them as *Positive* and 2 as *Neutral*. The remaining 5 comments were labeled as *Neutral* by annotators, while the model evaluated them as *Negative*. Examples of comments with labeling disagreements include: *"Mas quem desfez a baixa do ipi foi o ministro Alexandre o Glande rs"*[22] e *"Stranger Things é genérico pra car*lh*, totalmente compreensível kkkkkkkk."*[23].

In general, it is observed that the disagreement between *Negative* labels assigned by annotators, and *Positive* by GPT-4 occurs due to the model's apparent difficulty in associating laughter with something pejorative. In all these cases, BERTabaporu followed the same labeling assigned by annotators, which can be explained by the model's refinement using a training set extracted from our dataset.

There are also cases in which annotators labeled a comment as *Neutral*, while the model classified it as *Negative*. Within this scope, a certain content pattern emerges: the comment expresses an opinion that includes both negative and positive points, leading to a neutral overall assessment of the comment. However, the model tends to highlight the negative aspects of the text, erroneously labeling it as *Negative*.

We then analysed misclassifications by BERTabaporu compared to annotators but which GPT was able to label correctly. In this case, we find 20 comments, half of which have a *Positive* label, 6 have a *Negative* label, the remaining 4 being *Neutral* (as labeled by annotators). Emphasizing the greater presence of *Positive* comments that the model misclassified, BERTabaporu, in turn, labeled 6 of these 10 comments as *Negative*, and the remaining 4 as *Neutral*. Here, we can observe that while the GPT model tends to classify *Negative* comments as *Positive*, the BERTabaporu model more frequently classifies *Positive* comments as *Negative*. These cases reveal difficulty in distinguishing different uses of colloquial language and slang and swear words, as the model frequently associates these expressions with a negative context, which is not always accurate.

Finally, Table 15 shows 5 of the 15 comments where both models disagreed with the annotators' labeling. These examples provide some indications of possible difficulties that automatic labeling may have in performing this classification task. The critical tone of comment #1 about the relationship between men and women emerges subtly, with the ironic use of the giggling texting acronym *kkkkkkkk*. As discussed, common words in positive contexts tend to lead GPT-4 to generalize to *Positive* labels, while BERTabaporu had difficulty capturing the overall context, assigning the label *Neutral*.

Comment #2 includes many negative words and colloquial language and slang, which may justify its classification as *Negative* by the models. However, the expression *"'..hoje só quero saber do meu e estou me dando muito melhor..."'*[24] suggests a positive message, leading the annotators to assign *Positive*. Comment #3, on the other hand, uses irony and a subtle tone, resulting in the annotators classifying it as *Negative*, while the models label it as *Neutral*. In comments #4 and #5, labeled *Neutral* by the annotators, there is general advice. However, due to the intensity of the descriptions, the models take polarized stances, as they consider isolated words with positive or negative connotation instead of the overall context.

Based on the examples described here, the scenario where the models disagree with the annotators the most can be found when irony is used. For these cases, a possible mitigation is to make the entire *thread* of comments available for automatic labeling. However, this type of approach can limit the use of proprietary models such as GPT-4, since a larger number of tokens must be analyzed. This is a relevant point for the use of open-source models such as BERTabaporu, which performed satisfactorily on our dataset.

However, it is worth noting that this classification task is

[22] English gloss: *"But the one who undid the ipi reduction was minister Alexander the Glands rs"*

[23] English gloss: *"Stranger Things is so f*cking generic,that's totally understandable kkkkkkkk."*

[24] English gloss: *"'..today I only care about what is mine and I'm doing much better..."*

Table 15. Comparison between labels assigned by annotators and best models

#	Comment	Annotators	GPT-4 Few Shot	BERTabaporu
1	Porque na cabeça dela o simples fato dela ser mulher faz com que automaticamente um homem perceba e se interesse por ela, caso ela esteja flertando com o cara. kkkkkkkk English gloss: *”Because in her mind, the mere fact that she's a woman automatically makes a man notice her and take an interest in her, if she's flirting with the guy. kkkkkkkk”*	Negative	Positive	Neutral
2	Já fui bonzinho só me f*di, e hoje só quero saber do meu e estou me dando muito melhor. O mundo é uma competição. Quem é mais fraco é esmagado. English gloss: *”I used to be a goody-goody, I just got screwed, and today I only care about what's mine and I'm doing much better. The world is a competition. Whoever is weaker is crushed.”*	Positive	Negative	Negative
3	Esse cara fez alguma coisa como deputado? English gloss: *”Has this guy done anything as an assembly member?”*	Negative	Neutral	Neutral
4	Tomar Roacutan. Mesmo assim continuo não sendo grandes coisas, mas comparado com antigamente... Jesus Cristo! English gloss: *”To take Roacutan. I'm still not great, but compared to the old days... Jesus Christ!”*	Neutral	Positive	Negative
5	Vai de Chrome mesmo, não cai na furada do Brave. É cheio de falhas. English gloss: *”Go for Chrome anyway, don't fall for Brave. It's full of bugs.”*	Neutral	Negative	Negative

also challenging for human annotation, mainly due to its subjective nature, especially in cases where there is a contrast within the same sentence, of the kind: *x is good, but y is bad*. Thus, a more thorough and complex analysis is required in order to identify which of the two clauses in the sentence has more weight in the categorization of sentiments, or even if the clauses have equal weight, motivating a Neutral labeling.

In sum, our analyses make it possible to identify language patterns which impact model performance due to the need to interpret a sentence as a whole; it also shows which patterns are difficult for both automatic and human labeling. This fact corroborates the overall difficulty of the task and reveals how the language characteristics of the collected comments can significantly impact the performance of both human and automatic labeling.

Explainability of models. One of the main advantages of using open-source models is the possibility of applying explainability models such as SHAP (SHapley Additive exPlanations) [Lundberg and Lee, 2017], in order to understand which tokens had the greatest relevance in classifying a comment in a given sentiment class. Figure 4 presents some examples as well as the results obtained from applying the SHAP model to comments classified by BERT-based models, such as BERTabaporu, BERTimabu, XLM-RoBERTa.

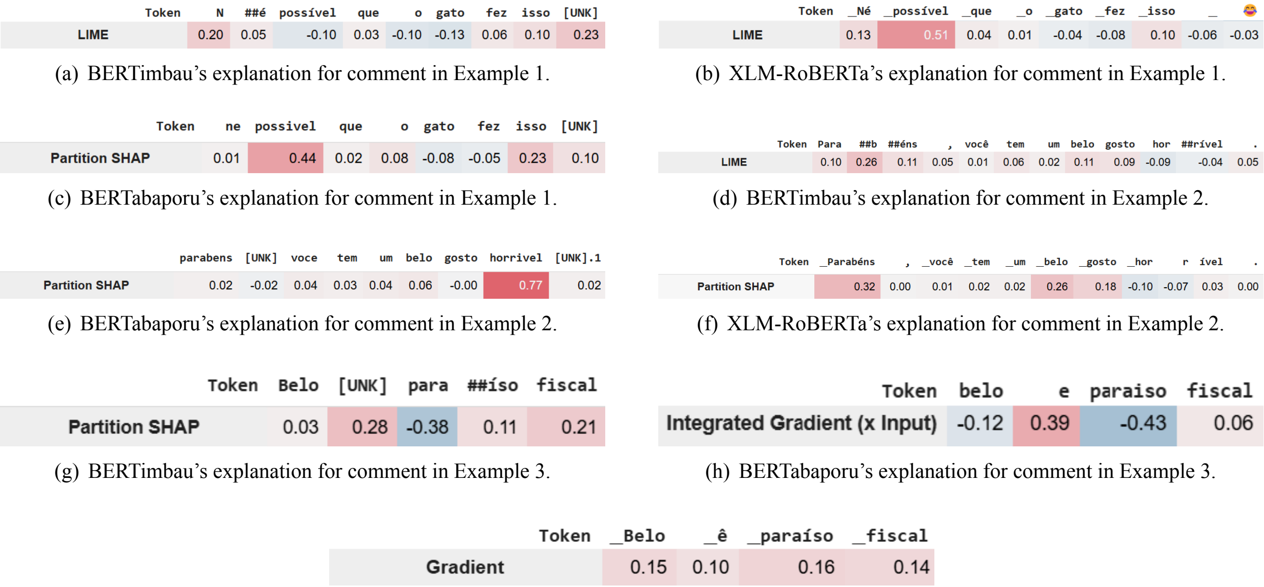
In Example 1, the comment *”Né possível que o gato fez isso 😂”* [25] is classified as *Positive* by the annotators and *Negative* by the BERTabaporu model. The SHAP results showed that the word *possible*, with a weight of 0.44 was the determining factor for the Negative labeling, followed by *this* with 0.23 and the emoji with 0.10, not recognized as a valid token by the model. In Example 2, the comment *“Parabéns, você tem um belo gosto horrível.”* [26] was classified as *Negative* by the annotators and by BERTabaporu, in contrast to the other BERT-based models which classified it as *Positive*. The word *horrible*, with a weight equal to 0.77, was the most relevant to the result, with the other words also showing favorable weights for labeling, but to a lesser extent.

Finally, in Example 3, the comment *“Belo ê paraíso fiscal”* [27], interpreted as sarcastic by the annotators and classified as *Negative*, was also classified as *Negative* by BERTimbau, but incorrectly labeled by BERTabaporu and XLM-RoBERTa, which classified it as *Neutral* and a*Positive*, re-

[25] English gloss: *“It's not possible that the cat has done this 😂”*

[26] English gloss: *“Congratulations, you have a beautiful horrible taste.”*

[27] English gloss: *“Belo is a fiscal paradise”*



Token	N	##é	possível	que	o	gato	fez	isso	[UNK]
LIME	0.20	0.05	-0.10	0.03	-0.10	-0.13	0.06	0.10	0.23

(a) BERTimbau’s explanation for comment in Example 1.

Token	_Né	_possível	_que	_o	_gato	_fez	_isso	_	😂
LIME	0.13	0.51	0.04	0.01	-0.04	-0.08	0.10	-0.06	-0.03

(b) XLM-RoBERTa’s explanation for comment in Example 1.

Token	ne	possivel	que	o	gato	fez	isso	[UNK]
Partition SHAP	0.01	0.44	0.02	0.08	-0.08	-0.05	0.23	0.10

(c) BERTabaporu’s explanation for comment in Example 1.

Token	Para	##b	##éns	,	você	tem	um	belo	gosto	hor	##rível	.
LIME	0.10	0.26	0.11	0.05	0.01	0.06	0.02	0.11	0.09	-0.09	-0.04	0.05

(d) BERTimbau’s explanation for comment in Example 2.

	parabens	[UNK]	voce	tem	um	belo	gosto	horrivel	[UNK].1
Partition SHAP	0.02	-0.02	0.04	0.03	0.04	0.06	-0.00	0.77	0.02

(e) BERTabaporu’s explanation for comment in Example 2.

Token	_Parabéns	,	_você	_tem	_um	_belo	_gosto	_hor	r	ível	.
Partition SHAP	0.32	0.00	0.01	0.02	0.02	0.26	0.18	-0.10	-0.07	0.03	0.00

(f) XLM-RoBERTa’s explanation for comment in Example 2.

Token	Belo	[UNK]	para	##íso	fiscal
Partition SHAP	0.03	0.28	-0.38	0.11	0.21

(g) BERTimbau’s explanation for comment in Example 3.

Token	belo	e	paraiso	fiscal
Integrated Gradient (x Input)	-0.12	0.39	-0.43	0.06

(h) BERTabaporu’s explanation for comment in Example 3.

Token	_Belo	_ê	_paraíso	_fiscal
Gradient	0.15	0.10	0.16	0.14

(i) XLM-RoBERTa’s explanation for comment in Example 3.

Figure 4. BERTimbau’s, XLM-RoBERTa’s and BERTabaporu’s explanation for selected comments for analysis.

spectively. BERTabaporu interprets the words “ê” and “fiscal” as determining the label, placing special emphasis on the word “ê”.

5 Conclusions, limitations and further research

The findings of our research corroborate the literature on the development of datasets by means of human annotation in tasks that involve a great deal of subjectivity, such as sentiment analysis. One of the findings concerns agreement between annotators, which in our study was moderate according to Krippendorf’s Alpha and Fleiss’s Kappa.

Regarding the content of our dataset, our results show that almost half of the comments were labeled as *Negative* by the majority, indicating a considerable class imbalance and possibly showing a potentially more harmful interaction environment. For instance, the largest subreddit in our dataset, r/brasil, featured many discussions related to the Brazilian elections and the Russia–Ukraine war. The second-largest subreddit analyzed was r/desabafos, where users typically share emotional or personal struggles. These topics are often associated with controversy and a generally pessimistic tone. We are aware that when using these data to train classification models, this negative bias can potentially be *learned* by the model. To mitigate these limitations, one should adopt strategies such as data balancing, data augmentation, among others, as well as use evaluation metrics that are more robust in imbalanced scenarios. Although we have not applied these strategies in the current work, our best open model still achieved good performance.

With regard to the results of the agreement metrics, the annotations obtained similar values, indicating medium and moderate agreement between the annotators. With regard to uncertainty, only in 4.1% of the comments did two or more annotators label the same comment with *I don't know*, which revealed greater difficulty for two or more annotators to characterize sentiment for the same text.

The language characterization of the comments revealed that comments labeled *Negative* and *Positive* tended to be longer than comments labeled *Neutral* and those labeled *I don't know*. The length of the comment can have an impact on labeling, since the larger the context, the greater the chance that the annotators will be able to make an interpretation and assign a sentiment.

With regard to the most frequent part-of-speech tags for each type of sentiment, the comments classified as *I don't know* stand out, as they had a predominance of entities of the PER type, as well as a greater number of tags of the proper noun class (PROPN), which may suggest that these comments require recognizing these entities and, consequently, world knowledge, in order to be able to assign a sentiment, a problem that seems to have been faced by the annotators.

Topic analysis revealed that for comments in which there was total disagreement between annotators, topic 14, which is more concerned with political ideologies, ranked first. This result can be related to the findings on the performance of the model, which labeled comments on political issues as *Negative* in greater number than human annotators. With regard to comments that at least one annotator labeled as *I don't know*, topics related to crime, swearing and sexual content and ideological and political issues were the most prominent.

In terms of methodology, our study showed that the quality of the metrics improved considerably when we separated the dataset into two subsets and included only the comments labeled with sentiment, disregarding the *I don't know* category. The same was the case when calculating the percentage of total agreement between annotators on the same label, which was higher when the *I don't know* category was disregarded.

Our comparison between the models yielded good metrics combining LLMs with Few-Shot techniques, with GPT-4

Few-Shot performing best, with an F1-Score of 0.83 and Cohen's Kappa of 0.71 compared to annotators' labeling. Considering Open-Source models, BERTabaporu performed best, with an F1-Score of 0.76 and a Cohen's Kappa of 0.66 compared to annotators' labeling. Other models we examined had similar and satisfactory metrics, such as Sabia-3 and Zero-Shot versions for LLMs (both GPT-4 and Sabia-3), as well as BERTimbau, which performed considerably better than our *baseline*, VADER/Leia, which had an F1-Score of 0.52.

In addition, we analyzed some labeling difficulties among the best models (GPT-4 Few-Shot and BERTabaporu), as well as difficulties faced by the annotators themselves. In this respect, we found that models tend to generalize colloquial language and slang as well as swear words to offensive contexts and tend to label comments as *Negative* when this kind of language occurs in non-offensive contexts; likewise, they fail to detect irony and sarcasm. On the other hand, like annotators, models also have difficulty in interpreting comments in which opinions express positive and negative aspects concomitantly, such as in criticism or more elaborate and developed pieces of advice. These insights contribute to a deeper understanding of the difficulty of the annotation task as a whole, in order to identify key problems for models and differentiate them from cases in which there is also disagreement in human annotation.

We would also like to point out some limitations of our study, such as the number of comments, especially when it comes to using models. Due to the testing and training stages for BERT-based models, a small set of comments was analyzed, limiting possible analyses that could be carried out on a larger sample. In addition, we acknowledge the subjectivity of the annotation task, which presented a proportion of disagreement between annotators. Finally, we should also point out limitations regarding collection and presentation of comments for annotation which could be improved to add more robustness to the annotation task, such as making complete comment threads available to provide more context and possibly reduce disagreement between annotators.

In line with the literature, our study corroborates the complexity of the task of creating a dataset, given the challenge of dealing with moderate levels of agreement between annotators. In order to compile the dataset, the majority vote, or the aggregation of the different answers, is decisive for the single reference label that will be assigned. In tasks that involve a high degree of subjectivity, such as sentiment analysis, the majority decision reduces the representativeness of the various opinions that may exist in an even larger population. In this sense, recent studies [Fornaciari *et al.*, 2021; Frenda *et al.*, 2023] propose a shift towards a more inclusive approach of all annotators' perspectives as an alternative to majority as reference or *ground truth*. In future work, we intend to explore perspectivization in order to mitigate the problem of the level of agreement between annotators.

In summary, despite the limitations discussed in this section, we believe our work represents an important first step toward providing sentiment analysis models tailored to Portuguese-language Reddit data.

Declarations

Acknowledgements

We acknowledge the valuable contribution of students from UFMG who participated in the annotation task.

Funding

This research was funded by grants from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

Authors' Contributions

All authors equally contributed to the conception of this study. GP, VM, and LL performed the experiments. GP, VM, and LL wrote the article. AP and AS contributed to the methodology, reviewed the article's content, assisted in writing, and provided guidance. ACSP contributed to writing and proofreading the English version. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The materials used in this study are available at the link: https://github.com/SentPortugueseDataset/SentimentAnalysisReddit. This is meant to enhance consultation, reproducibility, and, consequently, support the principles of open science and knowledge sharing. The materials available are under the MIT license.

References

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 technical report. DOI: `https://doi.org/10.48550/arXiv.2410.12049`.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics. `https://aclanthology.org/C18-1139/`. Access on 12 August 2025.

Almeida, R. J. A. (2018). Leia - léxico para inferência adaptada. `https://github.com/rafjaa/LeIA`. Access on 12 August 2025.

Amedie, J. (2015). The impact of social media on society. *Advanced Writing: Pop Culture Intersections*. `https://scholarcommons.scu.edu/engl_176/2/`. Access on 12 August 2025.

Barbieri, F., Espinosa Anke, L., and Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources* *and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association. DOI: https://doi.org/10.48550/arXiv.2104.12250.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839. DOI: https://doi.org/10.48550/arXiv.2001.08435.

Bibi, A., Ihsan, U., Ashraf, H., and Jhanjhi, N. (2024). Multilingual sentiment analysis using deep learning: Survey. *Preprints*. DOI: https://doi.org/10.1109/ICSSIT55814.2023.10060993.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. DOI: `https://doi.org/10.48550/arXiv.2005.14165`.

Brum, H. and das Graças Volpe Nunes, M. (2018). Building a Sentiment Corpus of Tweets in Brazilian Portuguese. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). DOI: https://doi.org/10.48550/arXiv.1712.08917.

Corso, F., Russo, G., and Pierri, F. (2024). A longitudinal study of italian and french reddit conversations around the russian invasion of ukraine. In *ACM Web Science Conference*, Websci '24, page 22–30. ACM. DOI: https://doi.org/10.48550/arXiv.2402.04999.

Costa, P. B., Pavan, M. C., Santos, W. R., Silva, S. C., and Paraboni, I. (2023). BERTabaporu: Assessing a genre-specific language model for Portuguese NLP. In Mitkov, R. and Angelova, G., editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 217–223, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria. DOI: https://doi.org/10.26615/978-954-452-092-2_024.

da Silva Oliveira, A., de Carvalho Cecote, T., Alvarenga, J. P. R., de Souza Freitas, V. L., and da Silva Luz, E. J. (2024). Toxic speech detection in Portuguese: A comparative study of large language models. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 108–116, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics. `https://aclanthology.org/2024.propor-1.11/`. Access on 12 August 2025.

Dang, N. C., Moreno-García, M. N., and De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3). DOI: https://doi.org/10.3390/electronics9030483.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *arXiv preprint*. DOI: https://doi.org/10.48550/arXiv.2005.00547.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *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. DOI: https://doi.org/10.18653/v1/N19-1423.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378. DOI: https://doi.org/10.1037/h0031619.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 31(3):651–659. DOI: https://doi.org/10.2307/2529549.

Fonseca, E., Santos, L., Criscuolo, M., and Aluisio, S. (2016). Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15. `https://www.linguamatica.com/index.php/linguamatica/article/view/v8n2-1`. Access on 12 August 2025.

Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., and Poesio, M. (2021). Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics. DOI: http://dx.doi.org/10.18653/v1/2021.naacl-main.204.

Freitas, C., Rocha, P., and Bick, E. (2008). A new world in floresta sintá(c)tica – the portuguese treebank. *Calidoscópio*, 6(3):142–148. DOI: https://doi.org/10.4013/cld.20083.03.

Frenda, S., Pedrani, A., Basile, V., Lo, S. M., Cignarella, A. T., Panizzon, R., Marco, C., Scarlini, B., Patti, V., Bosco, C., and Bernardi, D. (2023). EPIC: Multi-perspective annotation of a corpus of irony. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2023.acl-long.774.

Garcia, K. and Berton, L. (2021). Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. *Applied Soft Computing*, 101:107057. DOI: https://doi.org/10.1016/j.asoc.2020.107057.

Gilbert, C. H. E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. `https://ojs.aaai.org/index.`

php/ICWSM/article/view/14550. Access on 12 August 2025.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*. DOI: https://doi.org/10.48550/arXiv.2203.05794.

Herculano, A., de Paula, T.-H., Fernandes, D., and Rego, A. (2024). Depreredditbr: Um conjunto de dados textuais com postagens depressivas no idioma português brasileiro. In *Anais do VI Dataset Showcase Workshop*, pages 77–90, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/dsw.2024.243994.

Hinojosa Lee, M. C., Braet, J., and Springael, J. (2024). Performance metrics for multilabel emotion classification: Comparing micro, macro, and weighted f1-scores. *Applied Sciences*, 14(21). DOI: https://doi.org/10.3390/app14219863.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225. DOI: https://doi.org/10.1609/icwsm.v8i1.14550.

Júnior, A. P. D. S., Cecilio, P., Viegas, F., Cunha, W., Albergaria, E. T. D., and Rocha, L. C. D. D. (2022). Evaluating topic modeling pre-processing pipelines for portuguese texts. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '22, page 191–201, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3539637.3557052.

Kemp, S. (2024). Digital 2024 april global statshot report. https://datareportal.com/reports/digital-2024-april-global-statshot. Access on 12 August 2025.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization. DOI: https://doi.org/10.48550/arXiv.1412.6980.

Koncar, P., Walk, S., and Helic, D. (2021). Analysis and prediction of multilingual controversy on reddit. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, page 215–224, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3447535.3462481.

Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24). DOI: https://doi.org/10.1073/pnas.1320040111.

Krippendorff, K. (2019). *Content Analysis: An Introduction to Its Methodology*. Sage Publications. DOI: https://doi.org/10.4135/9781071878781.

Lima, L. H. Q., Pagano, A. S., and da Silva, A. P. C. (2024). Toxic content detection in online social networks: a new dataset from Brazilian Reddit communities. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 472–482, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics. https://aclanthology.org/2024.propor-1.48/. Access on 12 August 2025.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.. DOI: https://doi.org/10.48550/arXiv.1705.07874.

Martella, M., Bert, F., Colli, G., Lo Moro, G., Pagani, A., Tatti, R., Scaioli, G., and Siliquini, R. (2021). Consequences of cyberaggression on social network on mental health of italian adults. *European Journal of Public Health*, 31. DOI: https://doi.org/10.1093/eurpub/ckab165.589.

May, P. (2021). Machine translated multilingual sts benchmark dataset. https://github.com/PhilipMay/stsb-multi-mt. Access on 12 August 2025.

Melton, C. A., Olusanya, O. A., Ammar, N., and Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding covid-19 vaccines on the reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, 14(10):1505–1512. Special Issue on COVID-19 – Vaccine, Variants and New Waves. DOI: https://doi.org/10.1016/j.jiph.2021.08.010.

Mokhberian, N., Marmarelis, M. G., Hopp, F. R., Basile, V., Morstatter, F., and Lerman, K. (2023). Capturing perspectives of crowdsourced annotators in subjective learning tasks. *arXiv preprint arXiv:2311.09743*. DOI: https://doi.org/10.48550/arXiv.2311.09743.

Mueller, A. (2024). wordcloud. https://pypi.org/project/wordcloud/. Access on 12 August 2025.

Mughal, N., Mujtaba, G., Shaikh, S., Kumar, A., and Daudpota, S. M. (2024). Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis. *IEEE Access*, 12:60943–60959. DOI: https://doi.org/10.1109/ACCESS.2024.3386969.

Nandurkar, T., Nagare, S., Hake, S., and Chinnaiah, K. (2023). Sentiment analysis towards russia - ukrainian conflict: Analysis of comments on reddit. In *2023 11th International Conference on Emerging Trends in Engineering Technology - Signal and Information Processing (ICETET - SIP)*, pages 1–6. DOI: https://doi.org/10.1109/ICETET-SIP58143.2023.10151571.

NLTK (2023a). Nltk - sample usage for tokenize. https://www.nltk.org/howto/tokenize.html. Access on 12 August 2025.

NLTK (2023b). Nltk - stopwords. https://www.nltk.org/search.html?q=stopwords. Access on 12 August 2025.

Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175. DOI: https://doi.org/10.1016/j.artint.2012.03.006.

Oliveira, D. N. d., Utsch, M. N. R., Machado, D. V. P. d. A., Pena, N. G., Oliveira, R. G. D. d., Carvalho, A. I. R., and Merschmann, L. H. d. C. (2023). Evaluating a new automl approach for sentiment analysis and intent recognition tasks. *Journal on Interactive Systems*, 14(1):92–105. DOI:

https://doi.org/10.5753/jis.2023.3161.

OpenAI (2024). Gpt-4 technical report. DOI: `https://doi.org/10.48550/arXiv.2303.08774`.

Pablo Botton da Costa (2022). bertabaporu-base-uncased (revision 1982d0f). DOI:`https://doi.org/10.57967/hf/0019`.

Parmar, M., Mishra, S., Geva, M., and Baral, C. (2023). Don't blame the annotator: Bias already starts in the annotation instructions. In Vlachos, A. and Augenstein, I., editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789, Dubrovnik, Croatia. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2023.eacl-main.130.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artif. Intell. Rev.*, 54(2):1087–1115. DOI: https://doi.org/10.1007/s10462-020-09870-1.

Pereira, R., Alves, A., Vidal, D., Moura, F., Cabral, L., Paulino, R., Serrufo, M., and Figueiredo, K. (2023). Análise de sentimento de postagens de usuários no twitter combinando gpt-3 e aprendizado de máquina: Um estudo de caso sobre o 2º turno das eleições presidências brasileiras. In *Anais do XIV Workshop sobre Aspectos da Interação Humano-Computador para a Web Social*, pages 20–27, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/waihcws.2023.233507.

Petrov, S., Das, D., and McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*. DOI: https://doi.org/10.48550/arXiv.1104.2086.

Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A., and Rayson, P. (2015). Development of the multilingual semantic annotation system. In Mihalcea, R., Chai, J., and Sarkar, A., editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1268–1274, Denver, Colorado. Association for Computational Linguistics. DOI: https://doi.org/10.3115/v1/N15-1137.

Piorino, G., Moreira, V., Lima, L., Pagano, A., and Silva, A. (2024). Análise de sentimentos de conteúdo compartilhado em comunidades brasileiras do reddit: Avaliação de um conjunto de dados rotulados por humanos. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web*, pages 54–62, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/webmedia.2024.242020.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-031-45392-2_15.

Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alemany, L. A., and Martínez, M. V. (2024). pysentimiento: A python toolkit for opinion mining and social nlp tasks. DOI: `https://doi.org/10.48550/arXiv.2106.09462`.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy. `http://aclweb.org/anthology/W17-6523`. Access on 12 August 2025.

Real, L., Fonseca, E., and Oliveira, H. G. (2020). The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer. DOI: https://doi.org/10.1007/978-3-030-41505-1_39.

Reddit (2023). Transparency report: July to december 2023. `https://www.redditinc.com/policies/transparency-report-july-to-december-2023`. Access on 12 August 2025.

Rosillo-Rodes, Pablo, M. M. S. and Sánchez, D. (2025). Entropy and type-token ratio in gigaword corpora. *Phys. Rev. Res.*, pages –. DOI: https://doi.org/10.1103/rxxz-lk3n.

Siddiqui, S. and Singh, T. (2016). Social media its impact with positive and negative aspects. *International Journal of Computer Applications Technology and Research*, 5:71–75. DOI: http://dx.doi.org/10.7753/IJCATR0502.1006.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45(4):427–437. DOI: https://doi.org/10.1016/j.ipm.2009.03.002.

Souza, C. N., Martínez-Arribas, J., Correia, R. A., Almeida, J. A., Ladle, R., Vaz, A. S., and Malhado, A. C. (2024). Using social media and machine learning to understand sentiments towards brazilian national parks. *Biological Conservation*, 293:110557. DOI: https://doi.org/10.1016/j.biocon.2024.110557.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23*. DOI: http://dx.doi.org/10.1007/978-3-030-61377-8_28.

spaCy (2023). Portuguese models. `https://spacy.io/models/pt`. Access on 12 August 2025.

Tallarida, R. J. and Murray, R. B. (1987). *Mann-Whitney Test*, pages 149–153. Springer New York, New York, NY. DOI: https://doi.org/10.1007/978-1-4612-4974-$0_4$6.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.. DOI: https://doi.org/10.48550/arXiv.1706.03762.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). `https://aclanthology.org/L18-1686`. Access on 12 August 2025.

Wu, Y. and Wan, J. (2025). A survey of text classification based on pre-trained language model. *Neurocomputing*, 616:128921. DOI: https://doi.org/10.1016/j.neucom.2024.128921.

X (2024). Dsa transparency report - april 2024. `https:`

//transparency.x.com/dsa-transparency-report.html. Access on 12 August 2025.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2019). Multilingual universal sentence encoder for semantic retrieval. DOI: https://doi.org/10.48550/arXiv.1907.04307.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., and Bing, L. (2023). Sentiment analysis in the era of large language models: A reality check. DOI: https://doi.org/10.48550/arXiv.2305.15005.

Zhang, X., Qi, X., and Teng, Z. (2025). Performance evaluation of reddit comments using machine learning and natural language processing methods in sentiment analysis. In Zhou, K., editor, *Computational and Experimental Simulations in Engineering*, pages 14–24, Cham. Springer Nature Switzerland. DOI: https://doi.org/10.48550/arXiv.2405.16810.

6 Appendix

English translation for *prompt* instructions in Box 2 and topics and most frequent words in Table 16.

You are an assistant who classifies Reddit comments in Brazilian Portuguese (PT-BR) as *Positive*, *Negative* or *Neutral*. You will receive the text of a comment and your task is to classify the sentiment of the text provided.

Use only the information below to make the prediction:

1. For each comment, limit yourself to choosing just one of these three options, without adding explanatory text or assigning any label other than one of these three: *Positive*, *Negative* or *Neutral*;
2. Only assign *Positive* to those comments that you are confident that they have a positive sentiment;
3. Only assign *Negative* to those comments that you are confident that they have a negative sentiment;
4. Only assign *Neutral* to those comments that you are confident that they have a neutral sentiment.

For each comment below, assign one of these labels: *Positive*, *Negative*, or *Neutral*.

Box 2. Instructions provided to LLM models

Table 16. Topics and most frequent words.

Topic	Most frequent words
0	person, people, stay, nothing, do, there, thing, still, life, because
1	car, I think, never, will, use, that, remember, know, see, found
2	ass, bozo, then, and, of, dick, comes, food, can, comment
3	name, son, child, bath, bathroom, take, whore, during, remember, must
4	money, pay, salary, do, work, market, all, earn, story, about
5	brazil, country, state, usa, right, countries, russia, china, nuclear, left
6	team, goalkeeper, game, goal, soccer, palmeiras, player, vasco, paulo, past
7	fuck, hate, bro, I'm, fuck, fuck, hope, like, pity, horrible
8	words, understand, day, 11, because, people, speaking, word, countries, started
9	spoke, understood, spoke, result, fake, today, said, pt, poll, day
10	bandit, wants, ass, cock, because, hands, kill, went, over, ball
11	thank you, luck, I understood, comments, man, respect, i, hope, god, happy, good
12	lula, bolsonaro, bolsonarista, government, aid, spending, bad, time, president, against
13	population, politics, right, popular, government, political, health, economy, bolsonaro, pass
14	socialism, amp, x200b, hitler, national, communism, germans, against, say, church