

Article

Benchmarking Psychological Lexicons and Large Language Models for Emotion Detection in Brazilian Portuguese

Thales David Domingues Aparecido ^{1,2,3,†} , Alexis Carrillo ^{2,†} , Chico Q. Camargo ^{3,4,5}  and Massimo Stella ^{2,*} 

¹ Curso de Ciências Moleculares (Pró-Reitoria de Graduação), Universidade de São Paulo, Av. Prof. Lúcio Martins Rodrigues, 370-Butantã, São Paulo, SP 05508-020, Brazil; thalesdavidom@usp.br

² CogNosco Lab, Department of Psychology and Cognitive Science, University of Trento, 38122 Trento, Italy; ae.carrilloramirez@unitn.it

³ Department of Computer Science, University of Exeter, Exeter EX4 4QE, UK; f.camargo@exeter.ac.uk

⁴ Alan Turing Institute, London NW1 2DB, UK

⁵ Department of English Language and Literature, Ewha Womans University, Seoul 03760, Republic of Korea

* Correspondence: massimo.stella-1@unitn.it

† These authors contributed equally to this work.

Abstract

Emotion detection in Brazilian Portuguese is less studied than in English. We benchmarked a large language model (Mistral 24B), a language-specific transformer model (BERTimbau), and the lexicon-based EmoAtlas for classifying emotions in Brazilian Portuguese text, with a focus on eight emotions derived from Plutchik's model. Evaluation covered four corpora: 4000 stock-market tweets, 1000 news headlines, 5000 GoEmotions Reddit comments translated by LLMs, and 2000 DeepSeek-generated headlines. While BERTimbau achieved the highest average scores (accuracy 0.876, precision 0.529, and recall 0.423), an overlap with Mistral (accuracy 0.831, precision 0.522, and recall 0.539) and notable performance variability suggest there is no single top performer; however, both transformer-based models outperformed the lexicon-based EmoAtlas (accuracy 0.797) but required up to 40 times more computational resources. We also introduce a novel “emotional fingerprinting” methodology using a synthetically generated dataset to probe emotional alignment, which revealed an imperfect overlap in the emotional representations of the models. While LLMs deliver higher overall scores, EmoAtlas offers superior interpretability and efficiency, making it a cost-effective alternative. This work delivers the first quantitative benchmark for interpretable emotion detection in Brazilian Portuguese, with open datasets and code to foster research in multilingual natural language processing.

Keywords: large language models; emotional profiling; text analysis; EmoAtlas; Brazilian Portuguese



Academic Editor: Arslan Munir

Received: 17 July 2025

Revised: 9 August 2025

Accepted: 16 September 2025

Published: 1 October 2025

Citation: Domingues Aparecido, T.D.; Carrillo, A.; Camargo, C.Q.; Stella, M. Benchmarking Psychological Lexicons and Large Language Models for Emotion Detection in Brazilian Portuguese. *AI* **2025**, *6*, 249. <https://doi.org/10.3390/ai6100249>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The emotional profile of a text refers to the set of emotions it elicits and to how intense they are [1]. Detecting the emotional profile of a text is recognized as an important yet challenging task, with applications ranging from predicting emotional content in discussions about stock markets [2] to understanding political discourse [3]. A wide variety of tools have been developed to evaluate the emotional content of texts. These tools range from simple lexicon-based scoring methods to complex transformer-based large language models (LLMs). Within the computer science field of natural language processing, the analysis of emotional profiles of text—or, more simply, emotional analysis—is often used

to refer to studies and tools focused on detecting the overall feeling of pleasantness (often called “positive”) or unpleasantness (often called “negative”) [4]. That said, the field of emotional analysis is broader, as it looks at more than just positive/negative sentiment [5].

In this work, we choose to address the problem of detecting these specific emotions instead of the simpler “sentiment” category, as extensive studies have shown that the latter is not sufficient to fully capture the complexity of human emotional experience [6]. Here, we focus on benchmarking techniques for the detection of eight specific emotions defined by Plutchik as basic emotions [7]. These emotions are anger, fear, disgust, trust, surprise, anticipation, joy, and sadness. We relate these emotions to their respective translations in Brazilian Portuguese: raiva, medo, nojo, confiança, surpresa, antecipação, alegria, and tristeza.

Emotional analysis in text is supported by a broad spectrum of computational tools, among which EmoAtlas stands out as an advanced and interpretable framework [1]. Implemented in Python 3.11 EmoAtlas integrates psychologically validated emotion lexicons with artificial intelligence, leading to a set of functions performing syntactic, semantic, and emotional analyses of texts. The core of EmoAtlas’ functionalities includes (i) emotion analysis across text and (ii) the extraction of textual forma mentis networks (TFMNs), i.e., specialized cognitive networks that reconstruct the syntactic and semantic relationships between words as expressed by authors [8]. Textual forma mentis networks go beyond simple word co-occurrence [9]: they capture how concepts are associated structurally and contextually within sentences, revealing the underlying knowledge and emotional framing embedded in narratives. TFMNs are built to represent the connotation of specific words in a given corpus by examining their semantic frame, i.e., the set of words syntactically or semantically associated with a given concept within a text [10]. For instance, TFMNs can be used to investigate how different groups of social media users framed trauma-related jargon in online social media posts about sexual assault [9]. The emotional profiling is performed through statistical testing: EmoAtlas quantifies how much emotional words concentrate around a target word in a given text compared to random sampling from a reference corpus [1]. Once combined, the emotional profiling of specific semantic frames can provide relevant insights about how specific concepts were described and framed in a text.

EmoAtlas detects eight fundamental emotions from Plutchik’s model [7] and supports analysis in 18 languages. By leveraging TFMNs and psychological lexica, EmoAtlas can quantify how emotions are conveyed and framed through specific word associations across linguistic contexts. For instance, Semeraro and colleagues have shown that the word “bed” was framed along negative associations in low hotel reviews [1]. The framework produces interpretable and reproducible outputs and achieves performance comparable to state-of-the-art transformer models (such as BERT) but with markedly greater computational efficiency and no requirement for large-scale retraining. EmoAtlas thus enables both text-level and word-level emotional profiling, making it a powerful tool for multilingual, explainable emotion detection and cognitive network analysis across diverse types of textual data.

In this work, we propose to validate EmoAtlas as an innovative and cost-effective alternative for emotion profiling in Brazilian Portuguese while also assessing the performance of widely used state-of-the-art tools for emotion detection in the same language.

1.1. Related Literature

Research in emotional profiling can be broadly divided into two categories: approaches that rely on features endogenous to text and pipelines that rely on features exogenous to text, such as lexicons, to predict emotions. In methods that use only endogenous features,

the features are translated into numerical vectors, transforming emotion evaluation into a classification or regression problem [11,12]. These features can be either explicit and interpretable—such as word frequency counts, bag-of-words models, n-gram analysis, and co-occurrence networks [4,13,14]—or latent and less directly interpretable, including word embeddings generated by neural networks [15,16]. Typically, these approaches rely on supervised learning from texts pre-labeled with emotion profiles.

Large language models (LLMs) represent an evolution of methods employing latent features. Utilizing transformer architectures [17], LLMs can generate sequences of words that resemble human text. Furthermore, unlike traditional models, LLMs can be repurposed for various NLP tasks without requiring additional training or access to pre-labeled datasets [18]. Indeed, emotional profiling is only one of many tasks, as LLMs have been used widely across multiple contexts [19–21].

Despite their power, a major limitation of embedding-based models is their lack of interpretability: embeddings generated through attention mechanisms are opaque to the experimenter and difficult to analyze or explain [22]. This issue is especially pronounced in the case of LLMs, where interpretability decreases as a consequence of the model's complexity. Even the smallest and simplest LLMs often contain billions of parameters. Due to this complexity, implementing these models can be computationally expensive and resource-intensive, even for relatively straightforward tasks such as emotion profiling on large corpora.

An alternative to heavy embedding-based methods is to use text-exogenous features, such as a set of predefined words or expressions (called a lexicon) to identify emotions in text. In these lexicons, words or expressions are typically associated with sentiment or specific emotions [23,24]. Popular lexicons constructed for sentiment analysis include VADER (Valence Aware Dictionary for sEntiment Reasoning) [23] and the Hedonometer [25], both of which classify texts based on sentiment (e.g., positive/negative or happy/sad). For emotional analysis covering all eight emotions from Plutchik's theory [7], a widely used resource is EmoLex (also known as the National Research Council–NRC–Emotion Lexicon). This lexicon was developed through a large-scale study in which over 14,000 English words were rated by human participants [24].

One advantage of lexicon-based methods is that the lexicons can be translated into any language where the words have equivalents, without the need for additional training [1]. This can be particularly useful considering that while emotional profiling tools have been applied and validated across several datasets in different languages, most studies in natural language processing focus on English-language content, limiting the adoption and evaluation of new techniques in other linguistic contexts. Moreover, lexicon-based methods can be adapted or tuned for specific domains or topics of interest [22]. The disadvantage of lexicon-based methods is that most lexicons do not account for the context in which words are used—even though the emotional or sentimental meaning of a word may vary depending on the situation. Another limitation is that, when using lexicons for emotion profiling, the overall performance of the algorithm is inherently constrained by the quality and coverage of the lexicon employed [24].

As already mentioned above, EmoAtlas [1] is a natural language processing algorithm that performs emotional profiling through a combination of text-exogenous lexicons and text-endogenous algorithms. The tool stands as a low-computational cost alternative to heavy emotional profiling tools [1].

EmoAtlas operates in two distinct modes. In the first mode, EmoAtlas conducts text-level emotional analysis using a bag-of-words approach, relying on EmoLex to detect eight basic emotions in the text. In the second mode, however, EmoAtlas leverages computational syntactic parsing to capture both syntactic and semantic relationships between words,

constructing textual forma mentis networks (TFMNs) [8] as seen in Figure 1. Thanks to its hybrid structure, EmoAtlas does not require additional training on large text corpora, making it easier to adapt to other languages, such as Portuguese. This characteristic also makes EmoAtlas considerably less complex and potentially less costly to implement and run compared to methods involving the fine-tuning of large models such as BERTimbau on translated datasets or utilizing LLMs via API (for reference see [1]). The combination of low computational cost and easy adaptability to other languages is the main motivation for validating EmoAtlas in Brazilian Portuguese.

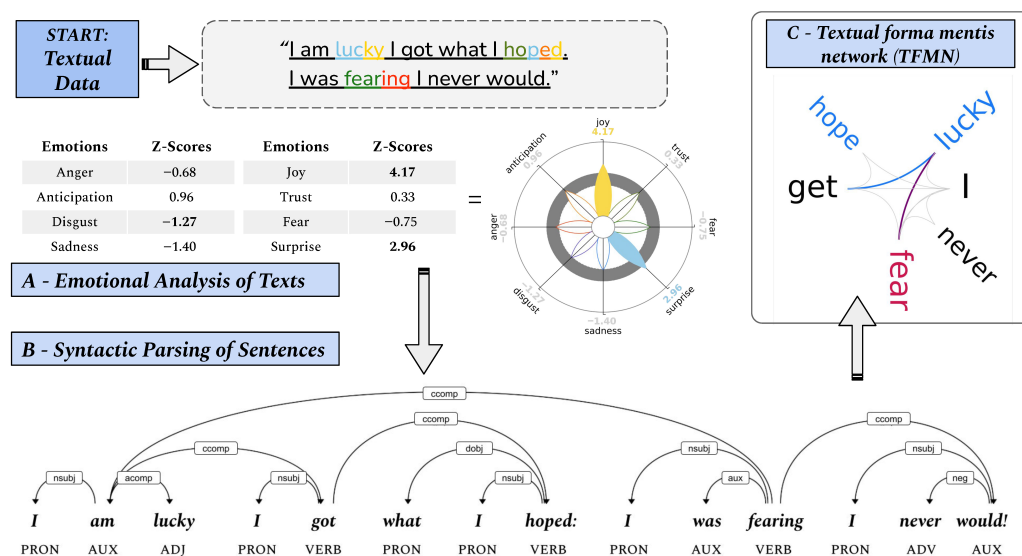


Figure 1. Top left: EmoAtlas reads a text sentence by sentence, mapping words into emotions. The tool can work as in “bag-of-words” mode, just counting words, or as a network builder. In the “bag-of-words” mode, words eliciting up to eight emotions are counted and compared against random sampling. Emotional flowers can represent word counts or z-scores. Individual petals indicate scores/z-scores. When using z-scores, colored petals fall outside the rejection region $|z| < 1.96$ ($\alpha = 0.05$), which is indicated by a gray ring. Bottom: In network-building mode, each sentence goes through syntactic parsing via spaCy. On the resulting syntactic tree, the distances t_{ij} separating non-stopwords i and j are measured. Every sampled i and j are linked if $t_{ij} < K$, giving rise to a textual forma mentis networks where (i) stopwords do not appear, and (ii) other concepts are linked despite stopwords not being included. By default $K = 3$, measuring local syntactic relationships on the syntactic tree. Top right: Textual forma mentis network of the sentence underlined in (top, left). Cyan (red) indicates positive (negative) words as outlined in EmoLex. Cyan (red) links connect positive words. Purple (gray) links connect positive and negative (neutral) words.

1.2. Emotional Profiling in Brazilian Portuguese: Challenges and Limitations

Although emotional profiling has historically focused on the English language, some pioneering works have already addressed emotion profiling in Brazilian Portuguese [26–28]. Notably, however, most works referring to emotional profiling deal primarily with sentiment—i.e., assigning polarity (positive/negative) [3]—and studies involving other emotions are often constrained to specific contexts [2].

Regarding methodology, researchers have implemented solutions using both supervised learning over text-endogenous features with Latent Semantic Analysis (LSA) [29], and more recently, BERT and BERTimbau [2,28,30]. These approaches are pioneering and foundational, given that most natural language processing is often dominated by analyses in English [4]. However, both LSA and transformer networks often require large annotated corpora for training, most of which are available only in English. To overcome this, many researchers need to build their own corpora and recruit annotators to label large sets of texts [2,3] or employ alternative data generation strategies such as translating

annotated datasets from other languages [28] or using data augmentation techniques [31]. LLMs could be used via zero- or few-shot learning, but they still entail high computational costs. There have also been initiatives involving text-exogenous methods, but these remain limited [32,33].

Given the scarcity of large, annotated corpora for Brazilian Portuguese, there is a clear need for a comprehensive benchmark of emotion detection methodologies. This study addresses this gap by proposing a mixed-methods approach to evaluate three distinct tools: a low-resource lexicon-based model (EmoAtlas), a language-specific transformer (BERTimbau), and a large language model (mistral-small-24B, released in March 2025). We employ a mixed-methods approach using both natively human-annotated datasets (stock market tweets and news headlines) and synthetic data generated or translated by LLMs (see Section 2.3). This dual approach allows us to not only evaluate model performance on authentic human language but also to introduce a novel benchmarking strategy we term “emotional fingerprinting”. This method uses LLM-generated data to provide insights into how different language models process and represent emotions.

2. Materials and Methods

2.1. Emotional Profiling Task

This study evaluated model performance in an emotional profiling task. This task involved assessing each model’s ability to predict emotion presence or absence within texts. For each emotion, ground truth labels underwent binarization. An emotion received a *True* label if present in a text; it received a *False* label otherwise. Each text was labeled with one or more emotions from a predefined set in Portuguese: raiva (anger), alegria (joy), antecipação (anticipation), tristeza (sadness), medo (fear), nojo (disgust), surpresa (surprise), and confiança (trust). Texts could be associated with any number of these emotions or with a neutral category.

Each model produced binary predictions for every emotion. This process yielded prediction sequences from EmoAtlas, Mistral, plus a stochastic baseline based on the inherent class imbalance in the datasets. Accuracy, precision, and recall constituted the evaluation metrics. This procedure was executed across four datasets. These datasets contained human-generated and LLM-generated texts.

2.2. Confidence Interval Calculation

To quantify the uncertainty in our performance scores (accuracy, precision, and recall), we calculated 95% confidence intervals using a bootstrap method [34]. This procedure involved resampling the dataset with replacement for 2000 iterations. In each iteration, the model’s performance was re-evaluated on the resampled data. The mean of the 2000 bootstrap scores served as the point estimate for each metric. The 95% confidence interval was then determined by taking the 2.5th percentile as the lower limit and the 97.5th percentile as the upper limit of the distribution of bootstrap scores.

2.3. Datasets

Four datasets served for emotional profiling model evaluation. An additional dataset supported a specific application in Portuguese. Two datasets comprised texts manually annotated by native Portuguese speakers. The third dataset was originally in English and translated to Portuguese using two different LLMs, thus creating two translated versions. The fourth dataset consisted of texts generated by an LLM. This corpus was created to specifically elicit the predefined set of target emotions, addressing the limitation that no single preceding dataset contained all target emotions across diverse contexts.

2.3.1. Human-Annotated Portuguese Datasets

The stock market tweets dataset [2] consists of 4277 tweets in Brazilian Portuguese. Native Portuguese speakers, at least three volunteers per tweet, manually annotated these tweets for emotions. The corpus focuses on the Brazilian stock market domain, containing user comments on individual stocks. Emotions annotated are based on Plutchik's theory [7]. These include trust, disgust, fear, sadness, joy, surprise, anticipation, and anger. A neutral category is also present. Annotation reliability is indicated by a high percentage of tweets with a majority label for each emotion pair. Consensus levels among annotators appeared lower for "surprise" and "anticipation" compared to "joy"/"sadness" and "trust"/"disgust". The dataset structure includes columns for eight emotions (TRU, DIS, JOY, SAD, ANT, SUR, ANG, and FEAR) and a NEUTRAL column. Values range from -2 to 1 . Negative values indicate annotation disagreement. A value of 0 signifies emotion absence. A value of 1 denotes emotion presence. Over 52 percent of tweets contained at least one emotion label. For this work, tweets with fewer than 15 words were excluded, since EmoAtlas cannot work well with extremely small texts; cf. [1].

The news headlines dataset [29] constitutes a corpus of 1002 short news items in Brazilian Portuguese, extracted from the *globo.com* website. The dataset consists of news headlines and accompanying brief descriptions. Each short news item received evaluation from three participants, from a group of 13 volunteers who manually performed the emotional annotation. The corpus was annotated for six specific emotion categories: desgosto (disgust), medo (fear), tristeza (sadness), alegria (joy), surpresa (surprise), and raiva (anger). A neutral state was identified when no dominant emotion was present. For the purpose of this study, texts classified under the "desgosto" category were discarded. This action aligns the dataset with the "nojo" (disgust) category used in this study.

We also included a corpus of posts from Brazilian political elites on X (formerly Twitter) [3]. Researchers annotated these posts for relevance, stance, and sentiment regarding COVID-19 vaccines and vaccination. Data collection occurred between 2020 and 2022. The initial collection contained 9045 posts. Filtering for topic relevance identified 5937 posts (65.64%) as relevant. These relevant posts subsequently underwent annotation for stance and sentiment. Nine annotators, organized into three groups, performed the annotations. This annotation process involved extensive training and weekly consistency meetings. Sentiment classification used three categories: positive, negative, and unclear. Positive sentiment attributed to posts manifested emotions like hope, admiration, or gratitude or an overall positive emotional state. Negative sentiment messages expressed pessimism, fear, or an overall negative emotional state. Posts with unclear emotional states, distinct from "neutral" to account for ambiguity, received an unclear classification. The authors explicitly distinguished sentiment (overall emotional tone) from stance (position toward vaccines) during annotation.

2.3.2. Large Language Models for Dataset Translation

The GoEmotions dataset [35] was selected to provide a comprehensive corpus of human-annotated text. This dataset comprises 58,009 Reddit comments manually annotated for 27 specific emotion categories, alongside a neutral label. It has previously served to benchmark emotion detection in Portuguese [26,28].

To align with the study's target emotions, six specific emotions: anger, disgust, fear, joy, sadness, and surprise were filtered. Anticipation and trust were not among the original 27 annotated emotions. To prepare the dataset for Portuguese analysis, English emotion labels were mapped to their corresponding Brazilian Portuguese terms.

The original English texts underwent translation into Brazilian Portuguese using two open-source LLMs: Gemma 3 27B [36] and Mistral Small 24B (cf. <https://huggingface.co>

[/mistralai/Mistral-Small-24B-Instruct-2501](#) (Mistral-Small-24B-Instruct-2501), accessed on 8 July 2025). This dual-translation approach allowed assessment of potential noise introduced by translation processes in emotional profiling. It also enabled evaluation of emotional profiling consistency across different LLM-based translations. LLMs have demonstrated superior performance over traditional translation systems for informal text translation [37]. Translation parameters included a temperature of 0.3 to mitigate commentary and hallucination. The prompt employed was

Translate the following text from {source_lang} to {target_lang}: '{text}'.
Format your translation as a Python list and write nothing else than the Python list.
Translate accurately.

Resulting texts underwent pre-processing to remove unwanted text generated by the model, considered as AI-generated noise. The regular expression `[\\'“\\”(.*)\\'“\\”]` captured only the translated text. This process yielded 5242 texts annotated with Portuguese emotion labels.

2.3.3. LLM-Generated Texts

To evaluate model performance across every emotion in varied contexts, this study used LLMs for data generation. The DeepSeek R1 8B model (cf. <https://huggingface.co/deepseek-ai/DeepSeek-R1-0528-Qwen3-8B> (DeepSeek-R1-0528-Qwen3-8B), accessed on 8 July 2025), was employed to produce 4800 news headlines in Brazilian Portuguese. These headlines were divided evenly across six themes: Brazilian Politics, Global Economy, Technology and Innovation, Public Health, National Sports, and Culture and Entertainment. Each text aimed to elicit one of eight emotions: trust, disgust, fear, sadness, joy, surprise, anticipation, and anger.

The generation prompt underwent careful design to elicit strong emotional tones throughout the text. The prompt, written in Portuguese, was as follows:

*“Você é um jornalista profissional. Escreva uma notícia jornalística entre 70 e 75 palavras sobre o tema '{topic_pt}', com um tom emocional MUITO forte de '{emotion_pt}', evidente ao longo de todo o texto. Comece com um parágrafo de reflexão marcado como '**Parágrafo de reflexão:**', seguido pela notícia final marcada como '**Notícia Final:**'. Mantenha o estilo jornalístico, mas influenciado pela emoção de '{emotion_pt}'.”*

which can be translated in English as

*“You are a professional journalist. Write a news story of between 70 and 75 words on the topic '{topic_pt}', with a VERY strong emotional tone of '{emotion_pt}', evident throughout the text. Start with a reflection paragraph marked as '**Reflection paragraph:**', followed by the final news story marked as '**Final News Story:**'. Keep the journalistic style, but influenced by the emotion of '{emotion_pt}'.”*

Texts containing AI-generated meta-comments or interruptions were filtered out. Any entry containing the character sequence “\n” was removed as an indicator of stream segmentation in LLM outputs. After filtering, the final dataset consisted of 2057 texts.

2.4. Emotional Profiling Models

2.4.1. Large Language Models for Emotional Annotation

To control output variability and minimize potential LLM hallucination or deviations from the intended response format, the model’s inference temperature was set to 0.3. A custom prompting system guided Mistral for the emotional classification task. This prompt structure was specifically engineered to constrain model output to only emotion names, thereby simplifying subsequent automated parsing of predictions. The primary prompt, utilized for datasets incorporating a neutral emotion category, was

Analyze the following text and identify the elicited emotions from these options: anger, disgust, fear, joy, sadness, surprise, trust, anticipation or neutral. Neutral is for when the text elicits no other emotion according to you. Return ONLY the emotions name in lowercase, nothing else. Text: {text}

The placeholder {text} was dynamically substituted with the textual input from the various datasets (see Section 2.3). For datasets not containing a neutral emotion category, the prompt was adjusted to exclude that option:

Analyze the following text and identify the elicited emotions from these options: anger, disgust, fear, joy, sadness, surprise, trust or anticipation. Return ONLY the emotions name in lowercase, nothing else. Text: {text}

2.4.2. EmoAtlas

Unlike “black-box” embedding-based models, EmoAtlas operates via a structured pipeline integrating psychological lexicons, syntactic parsing, and network science tools [1]. Its core emotion detection capability does not rely on prior training on specific annotated datasets.

EmoAtlas offers two primary processing modes: bag-of-words and structured. In bag-of-words mode, EmoAtlas processes input text at the sentence level. It normalizes text using lemmatization, which is the default setting in the EmoAtlas library, to reduce words to their base forms. This mode primarily employs lexical analysis, counting the frequency of words and emojis associated with specific emotions based on a predefined lexicon [24]. These counts are then aggregated to generate an overall emotional profile for the entire text.

The structured mode utilizes AI for syntactic parsing, implemented via the spaCy library (specifically, the “pt_core_news_lg” package for Portuguese). This parsing constructs a dependency tree for each sentence, identifying grammatical relationships and part-of-speech tags for words. EmoAtlas identifies non-stopwords (e.g., nouns, verbs, adverbs, pronouns) and detects negation through examination of auxiliary verbs. It measures the syntactic distance between related non-stopwords within the tree; words within a threshold distance of 3 (the default value) establish links. This process constructs a TFMN for each sentence, which is subsequently merged into a document-level TFMN. Emotional analysis in this mode occurs at the word level. The system determines the emotional profile (emotion scores) of a specific target word or concept based on the emotional words linked to it within the text’s network structure, considering negation. This capability enables the analysis of how specific concepts receive emotional framing within a text.

EmoAtlas quantifies and classifies emotions using raw counts, scores, or z-scores. For this study, EmoAtlas computed emotion presence by analyzing z-scores for each text in the emotion-annotated datasets. An emotion is considered present if its z-score surpasses a statistical threshold of 1.96, indicating a significant deviation from a null model including words uniformly at random from the psychological lexicon; for more details see [1]. For each text, a list of these predicted emotions was compiled. For binary classification, a “True” label is assigned if the target emotion is present in the list of predicted emotions.

For the Portuguese COVID-19 tweets dataset, this study applied two distinct analyses. First, all tweets classified as “positive” and “negative” were concatenated into two separate strings. The `.draw_statistically_significant_emotions` method then generated two emotional flowers. This visualization allows for the examination of the relationship between specific emotions and the overall valence of the dataset.

In the structured mode, EmoAtlas evaluated the emotional framing of the term “vacina” (vaccine) within the same dataset. Initially, textual forma mentis networks (TFMNs) were generated for each individual tweet classified as “positive” or “negative.” These individual

TFMNs were subsequently merged into two weighted graphs, corresponding to posts with positive and negative sentiments, utilizing the *igraph* library. Weights indicate the frequency of links in a given collection of texts. The semantic frame of “vacina” (vaccine) was extracted from each of these merged networks, comprising all nodes connected to “vacina” and their interconnections. Links within these semantic frames underwent a filtering process based on weight. We used the same statistical filtering technique introduced in [9]. We retained links with weights in top quartile that also exhibited an edge weight increase of at least 100% compared to links obtained from randomly reshuffling words.

2.4.3. BERTimbau

To provide a task-specific fine-tuned transformer model for comparison, we employed the BERTimbau-Large model for emotional classification. BERTimbau is a BERT-based model pretrained on a large-scale Brazilian Portuguese corpus [30]. The BERTimbau Large model was fine-tuned to classify emotions based on the GoEmotions emotional model using a translated version of the GoEmotions annotated dataset [26]. It is important to note that the GoEmotions dataset is one of the datasets used for evaluation (see Section 2.3.2). Consequently, BERTimbau’s performance on the translated GoEmotions dataset should be viewed as a measure of its capacity to learn its training data rather than its generalizability to unseen text. For this study, we focused on the six emotions that align with Plutchik’s model of basic emotions: anger (raiva), disgust (nojo), fear (medo), joy (alegria), sadness (tristeza), and surprise (surpresa). The model outputs a probability score for each emotion class. To convert these scores into a binary classification (present or not present), a threshold of 0.03 was applied. This value was determined by performing a systematic search across a range of thresholds from 0.0 to 1.0, in increments of 0.001. For each threshold, we binarized the model’s predictions and calculated the corresponding precision and recall scores. The value of 0.03 was selected because it minimized the absolute difference between precision and recall.

2.4.4. Baseline Reference Model

To provide a comparative reference for Mistral, EmoAtlas, and BERTimbau, this study employed a stochastic baseline model. This baseline addresses the inherent class imbalance prevalent in the datasets, where emotion frequencies do not follow a uniform distribution.

This biased reference model predicts the presence of a given emotion e at random with a probability $p(e)$. This probability $p(e)$ is determined by the empirical ratio of “True” instances of emotion e to the total number of texts within the dataset. For instance, if a dataset contains 1000 texts, and 200 of these texts are labeled with “medo” (fear), then the baseline predicts the existence of fear in any given text with a 20% probability, irrespective of the text’s content.

The expected accuracy for such a stochastic model, for a specific emotion e , is calculated as

$$\text{Accuracy}(e) = p_{e_T} \cdot p_e + (1 - p_{e_T}) \cdot (1 - p_e)$$

where p_{e_T} represents the true proportion of texts eliciting emotion e in the dataset. For example, if $p_{e_T} = 0.2$ (as in the “medo” example), the expected accuracy would be $0.2 \times 0.2 + 0.8 \times 0.8 = 0.68$.

As this model operates stochastically, performance metrics for the baseline were estimated through 2000 Monte Carlo simulations on each dataset. The average scores obtained from these simulations for accuracy, precision, and recall represent the minimal performance level expected from a biased random guessing model, accounting for dataset imbalances. This provides a crucial benchmark against which the performance of more sophisticated models can be meaningfully interpreted.

3. Results

This section presents a performance comparison of EmoAtlas, Mistral, and BERTimbau for emotion profiling on Brazilian Portuguese text. We report accuracy, precision, and recall for each emotion, alongside a stochastic baseline, to provide a comprehensive evaluation of model effectiveness.

Across the human-annotated datasets (Portuguese News Headlines and Stock Market Tweets) and translated data (GoEmotions Gemma and GoEmotions Mistral), the Mistral model consistently achieved the highest accuracy. BERTimbau and EmoAtlas often followed with overlapping confidence intervals, indicating comparable performance for the three models. Precision and recall scores revealed differential performance across emotions. For emotions like anger/*raiva* and fear/*medo*, models showed similar levels of precision and recall. Conversely, for emotions such as joy/*alegria*, sadness/*tristeza*, and surprise/*surpresa*, models often exhibited an inverse extreme values between precision and recall, impacting their performance ranking on precision and recall.

3.1. News Headlines

On the news headlines dataset shown in Table 1, Mistral and BERTimbau generally outperformed both EmoAtlas and the Baseline model across most emotions and metrics. Mistral generally demonstrated superior performance, with notable differences in several emotions. BERTimbau demonstrated a precision of 0.50 (0.33, 0.67) for joy/*alegria*, which had a wide confidence interval. For sadness/*tristeza*, Mistral exhibited an inverse performance between precision (0.24) and recall (0.94). Despite its high recall for this emotion, Mistral's accuracy for sadness/*tristeza* was 0.56 (0.54, 0.59), a value below the baseline 0.75 (0.74, 0.77). For surprise/*surpresa*, both BERTimbau (0.13) and Mistral (0.20) had low recall scores, with their confidence intervals overlapping with the baseline 0.14 (0.10, 0.19).

Table 1. News headlines.

Emotion	Score	BERTimbau	Baseline	EmoAtlas	Mistral
<i>raiva</i> (anger)	Accuracy	0.75 (0.73, 0.77)	0.76 (0.75, 0.78)	0.79 (0.77, 0.81)	0.87 (0.85, 0.89)
	Precision	0.22 (0.18, 0.27)	0.14 (0.10, 0.18)	0.23 (0.18, 0.29)	0.53 (0.46, 0.60)
	Recall	0.32 (0.26, 0.38)	0.14 (0.10, 0.18)	0.23 (0.18, 0.29)	0.49 (0.42, 0.55)
<i>medo</i> (fear)	Accuracy	0.84 (0.82, 0.85)	0.76 (0.74, 0.77)	0.72 (0.69, 0.74)	0.80 (0.79, 0.82)
	Precision	0.44 (0.38, 0.50)	0.14 (0.10, 0.18)	0.26 (0.22, 0.30)	0.39 (0.35, 0.44)
	Recall	0.49 (0.43, 0.55)	0.14 (0.10, 0.19)	0.54 (0.48, 0.60)	0.68 (0.62, 0.74)
<i>alegria</i> (joy)	Accuracy	0.86 (0.84, 0.87)	0.75 (0.74, 0.77)	0.81 (0.80, 0.83)	0.86 (0.84, 0.87)
	Precision	0.50 (0.33, 0.67)	0.14 (0.11, 0.19)	0.30 (0.24, 0.37)	0.51 (0.41, 0.60)
	Recall	0.07 (0.04, 0.11)	0.14 (0.10, 0.19)	0.21 (0.16, 0.26)	0.23 (0.17, 0.28)
<i>tristeza</i> (sadness)	Accuracy	0.87 (0.86, 0.89)	0.75 (0.74, 0.77)	0.81 (0.79, 0.82)	0.56 (0.54, 0.59)
	Precision	0.56 (0.50, 0.62)	0.14 (0.10, 0.19)	0.32 (0.26, 0.38)	0.24 (0.21, 0.27)
	Recall	0.54 (0.48, 0.60)	0.14 (0.10, 0.19)	0.32 (0.26, 0.37)	0.94 (0.91, 0.97)
<i>surpresa</i> (surprise)	Accuracy	0.86 (0.85, 0.88)	0.75 (0.74, 0.77)	0.80 (0.78, 0.82)	0.87 (0.86, 0.89)
	Precision	0.64 (0.51, 0.77)	0.14 (0.11, 0.19)	0.17 (0.11, 0.23)	0.70 (0.59, 0.81)
	Recall	0.13 (0.09, 0.17)	0.14 (0.10, 0.19)	0.11 (0.07, 0.14)	0.20 (0.15, 0.25)

3.2. Stock Market Tweets

On the stock market tweet dataset, Mistral and BERTimbau appear to be more robust models for emotional profiling compared to EmoAtlas, as seen in Table 2. For disgust/*nojo*, Mistral, BERTimbau, and EmoAtlas showed wide overlapping confidence intervals on precision. For recall on disgust/*nojo*, all models performed below the 0.19 (0.16, 0.23) baseline level, while BERTimbau, EmoAtlas, and Mistral achieved 0.07 (0.05, 0.09), 0.11 (0.08, 0.13), and 0.12 (0.09, 0.15), respectively. The widest precision intervals for fear/*medo* were from BERTimbau (0.20, 0.73) and Mistral (0.11, 0.33). EmoAtlas achieved the highest recall for

fear/medo, with a value of 0.25 (0.18, 0.32). Finally, Mistral demonstrated the lowest recall for trust/confiança, with a score of 0.01 (0.00, 0.02).

The results for the stock market tweets dataset are shown below:

Table 2. Stock market tweets.

Emotion	Score	BERTimbau	Baseline	EmoAtlas	Mistral
<i>raiva</i> (anger)	Accuracy	0.92 (0.91, 0.93)	0.87 (0.86, 0.88)	0.90 (0.88, 0.91)	0.92 (0.91, 0.93)
	Precision	0.42 (0.33, 0.52)	0.07 (0.04, 0.11)	0.22 (0.16, 0.30)	0.42 (0.34, 0.50)
	Recall	0.24 (0.18, 0.31)	0.07 (0.04, 0.11)	0.17 (0.12, 0.23)	0.35 (0.28, 0.43)
<i>antecipação</i> (anticipation)	Accuracy	–	0.72 (0.71, 0.74)	0.71 (0.69, 0.73)	0.76 (0.74, 0.78)
	Precision	–	0.17 (0.13, 0.20)	0.16 (0.13, 0.19)	0.28 (0.23, 0.32)
	Recall	–	0.16 (0.13, 0.20)	0.18 (0.14, 0.22)	0.28 (0.24, 0.32)
<i>nojo</i> (disgust)	Accuracy	0.81 (0.80, 0.83)	0.69 (0.67, 0.71)	0.80 (0.79, 0.82)	0.82 (0.81, 0.84)
	Precision	0.65 (0.51, 0.78)	0.19 (0.16, 0.23)	0.44 (0.34, 0.54)	0.73 (0.62, 0.83)
	Recall	0.07 (0.05, 0.09)	0.19 (0.16, 0.23)	0.11 (0.08, 0.13)	0.12 (0.09, 0.15)
<i>medo</i> (fear)	Accuracy	0.94 (0.93, 0.95)	0.89 (0.88, 0.90)	0.81 (0.79, 0.82)	0.93 (0.92, 0.94)
	Precision	0.47 (0.20, 0.73)	0.06 (0.02, 0.10)	0.09 (0.06, 0.12)	0.21 (0.11, 0.33)
	Recall	0.05 (0.02, 0.09)	0.06 (0.02, 0.10)	0.25 (0.18, 0.32)	0.08 (0.04, 0.13)
<i>alegria</i> (joy)	Accuracy	0.89 (0.87, 0.90)	0.80 (0.79, 0.81)	0.85 (0.84, 0.87)	0.88 (0.87, 0.89)
	Precision	0.46 (0.36, 0.57)	0.11 (0.08, 0.15)	0.22 (0.16, 0.29)	0.45 (0.37, 0.53)
	Recall	0.14 (0.10, 0.18)	0.11 (0.08, 0.15)	0.13 (0.09, 0.17)	0.23 (0.18, 0.28)
<i>tristeza</i> (sadness)	Accuracy	0.88 (0.87, 0.90)	0.79 (0.78, 0.80)	0.82 (0.81, 0.84)	0.78 (0.77, 0.80)
	Precision	0.61 (0.45, 0.77)	0.12 (0.08, 0.16)	0.22 (0.16, 0.27)	0.28 (0.24, 0.32)
	Recall	0.08 (0.05, 0.12)	0.12 (0.08, 0.16)	0.18 (0.14, 0.22)	0.50 (0.45, 0.56)
<i>surpresa</i> (surprise)	Accuracy	0.87 (0.86, 0.89)	0.78 (0.77, 0.80)	0.85 (0.83, 0.86)	0.87 (0.86, 0.89)
	Precision	0.43 (0.31, 0.56)	0.12 (0.09, 0.16)	0.20 (0.14, 0.27)	0.46 (0.33, 0.59)
	Recall	0.09 (0.06, 0.12)	0.12 (0.09, 0.16)	0.08 (0.05, 0.12)	0.09 (0.06, 0.12)
<i>confiança</i> (trust)	Accuracy	–	0.67 (0.66, 0.69)	0.69 (0.67, 0.71)	0.79 (0.78, 0.81)
	Precision	–	0.20 (0.17, 0.24)	0.26 (0.22, 0.30)	0.43 (0.17, 0.70)
	Recall	–	0.20 (0.17, 0.24)	0.29 (0.24, 0.33)	0.01 (0.00, 0.02)

3.3. Indirect Annotated Data

On the translated GoEmotions datasets, performance metrics were similar regardless of whether Gemma or Mistral was used for translation. BERTimbau generally outperformed Mistral and EmoAtlas across most metrics, showing notable differences in precision and recall. This pattern was consistent for most emotions, with the exception of disgust/nojo, where EmoAtlas achieved higher accuracy than Mistral across both translated datasets.

It is important to interpret BERTimbau's performance with the understanding that they represent an upper performance bound. These results reflect the model's capacity to learn the specific nuances of its own training data and, therefore, should not be considered a benchmark for its generalizability to unseen or independent text. Instead, they demonstrate the performance ceiling achievable when a model is fine-tuned on the target corpus.

Performance on the GoEmotions dataset translated with Gemma is presented in Table 3. For anger/raiva, Mistral and BERTimbau had similar precision scores; Mistral's 0.68 (0.64, 0.71) was only marginally higher than BERTimbau's 0.63 (0.60, 0.66), and their confidence intervals overlapped. For fear/medo, the Baseline model's accuracy of 0.83 (0.82, 0.84) was higher than EmoAtlas's score of 0.77 (0.76, 0.78). In contrast, EmoAtlas demonstrated higher accuracy for disgust, with a score of 0.82 (0.81, 0.83), which was superior to Mistral's 0.78 (0.77, 0.79).

Performance metrics on the GoEmotions dataset translated with Mistral are presented in Table 4 with similar results. For anger/raiva, Mistral and BERTimbau showed comparable precision, with Mistral's score of 0.70 (0.67, 0.73) slightly exceeding BERTimbau's 0.66 (0.63, 0.68). In the case of anger/raiva, the Baseline model's accuracy of 0.83 (0.82, 0.84) was higher than EmoAtlas's score of 0.78 (0.76, 0.79). Conversely, EmoAtlas demonstrated

higher accuracy for disgust, scoring 0.82 (0.81, 0.83), which was superior to Mistral’s 0.77 (0.76, 0.78).

Table 3. GoEmotions translation using Gemma.

Emotion	Score	BERTimbau	Baseline	EmoAtlas	Mistral
<i>raiva</i> (anger)	Accuracy	0.85 (0.84, 0.86)	0.66 (0.65, 0.68)	0.74 (0.72, 0.75)	0.83 (0.82, 0.84)
	Precision	0.63 (0.60, 0.66)	0.21 (0.19, 0.23)	0.34 (0.31, 0.37)	0.68 (0.64, 0.71)
	Recall	0.69 (0.66, 0.72)	0.21 (0.19, 0.24)	0.26 (0.24, 0.29)	0.40 (0.37, 0.43)
<i>nojo</i> (disgust)	Accuracy	0.86 (0.85, 0.87)	0.77 (0.76, 0.78)	0.82 (0.81, 0.83)	0.78 (0.77, 0.79)
	Precision	0.49 (0.45, 0.52)	0.13 (0.11, 0.15)	0.29 (0.26, 0.33)	0.32 (0.29, 0.34)
	Recall	0.64 (0.60, 0.67)	0.13 (0.11, 0.16)	0.28 (0.25, 0.31)	0.56 (0.52, 0.59)
<i>medo</i> (fear)	Accuracy	0.94 (0.93, 0.94)	0.83 (0.82, 0.84)	0.77 (0.76, 0.78)	0.93 (0.92, 0.94)
	Precision	0.64 (0.60, 0.68)	0.10 (0.07, 0.12)	0.20 (0.18, 0.23)	0.65 (0.60, 0.69)
	Recall	0.80 (0.76, 0.83)	0.10 (0.07, 0.12)	0.48 (0.44, 0.53)	0.56 (0.51, 0.60)
<i>alegria</i> (joy)	Accuracy	0.92 (0.92, 0.93)	0.65 (0.64, 0.67)	0.76 (0.75, 0.77)	0.87 (0.86, 0.88)
	Precision	0.92 (0.90, 0.93)	0.22 (0.20, 0.24)	0.47 (0.44, 0.50)	0.73 (0.70, 0.75)
	Recall	0.72 (0.70, 0.75)	0.22 (0.20, 0.25)	0.51 (0.48, 0.54)	0.66 (0.64, 0.69)
<i>tristeza</i> (sadness)	Accuracy	0.92 (0.91, 0.93)	0.67 (0.66, 0.69)	0.75 (0.74, 0.76)	0.84 (0.83, 0.85)
	Precision	0.85 (0.82, 0.87)	0.20 (0.18, 0.23)	0.36 (0.33, 0.40)	0.59 (0.57, 0.62)
	Recall	0.74 (0.71, 0.76)	0.20 (0.18, 0.23)	0.30 (0.28, 0.33)	0.67 (0.65, 0.70)
<i>surpresa</i> (surprise)	Accuracy	0.93 (0.92, 0.93)	0.73 (0.72, 0.74)	0.81 (0.80, 0.82)	0.89 (0.88, 0.90)
	Precision	0.75 (0.72, 0.78)	0.16 (0.14, 0.18)	0.34 (0.30, 0.39)	0.77 (0.73, 0.80)
	Recall	0.82 (0.79, 0.84)	0.16 (0.14, 0.19)	0.18 (0.15, 0.20)	0.44 (0.41, 0.48)

Table 4. GoEmotions translation using Mistral.

Emotion	Score	BERTimbau	Baseline	EmoAtlas	Mistral
<i>raiva</i> (anger)	Accuracy	0.86 (0.85, 0.87)	0.66 (0.65, 0.67)	0.73 (0.72, 0.74)	0.84 (0.83, 0.85)
	Precision	0.66 (0.63, 0.68)	0.21 (0.19, 0.23)	0.33 (0.30, 0.36)	0.70 (0.67, 0.73)
	Recall	0.74 (0.72, 0.77)	0.21 (0.19, 0.24)	0.25 (0.23, 0.28)	0.44 (0.41, 0.47)
<i>nojo</i> (disgust)	Accuracy	0.86 (0.85, 0.87)	0.77 (0.76, 0.78)	0.82 (0.81, 0.83)	0.77 (0.76, 0.78)
	Precision	0.48 (0.45, 0.51)	0.13 (0.11, 0.16)	0.29 (0.26, 0.33)	0.31 (0.29, 0.34)
	Recall	0.75 (0.72, 0.78)	0.13 (0.11, 0.16)	0.26 (0.23, 0.29)	0.63 (0.59, 0.67)
<i>medo</i> (fear)	Accuracy	0.94 (0.93, 0.94)	0.83 (0.82, 0.84)	0.78 (0.76, 0.79)	0.93 (0.93, 0.94)
	Precision	0.62 (0.58, 0.66)	0.10 (0.07, 0.12)	0.20 (0.17, 0.22)	0.67 (0.62, 0.71)
	Recall	0.84 (0.80, 0.87)	0.10 (0.07, 0.12)	0.44 (0.39, 0.48)	0.58 (0.53, 0.62)
<i>alegria</i> (joy)	Accuracy	0.93 (0.92, 0.94)	0.65 (0.64, 0.67)	0.78 (0.77, 0.79)	0.88 (0.88, 0.89)
	Precision	0.91 (0.89, 0.93)	0.22 (0.20, 0.25)	0.50 (0.47, 0.53)	0.75 (0.73, 0.78)
	Recall	0.76 (0.74, 0.79)	0.22 (0.20, 0.25)	0.51 (0.48, 0.54)	0.72 (0.69, 0.74)
<i>tristeza</i> (sadness)	Accuracy	0.92 (0.92, 0.93)	0.67 (0.66, 0.69)	0.76 (0.75, 0.77)	0.85 (0.84, 0.86)
	Precision	0.80 (0.77, 0.82)	0.20 (0.18, 0.23)	0.38 (0.35, 0.42)	0.63 (0.60, 0.65)
	Recall	0.84 (0.82, 0.86)	0.21 (0.18, 0.23)	0.31 (0.28, 0.33)	0.72 (0.69, 0.75)
<i>surpresa</i> (surprise)	Accuracy	0.94 (0.93, 0.95)	0.73 (0.72, 0.74)	0.81 (0.80, 0.82)	0.89 (0.88, 0.90)
	Precision	0.80 (0.78, 0.83)	0.16 (0.14, 0.19)	0.34 (0.29, 0.38)	0.78 (0.75, 0.82)
	Recall	0.84 (0.82, 0.86)	0.16 (0.14, 0.19)	0.17 (0.14, 0.19)	0.46 (0.43, 0.49)

3.4. Fingerprinting LLMs for Emotional Detection

Performance on the DeepSeek-generated headlines dataset is presented in Table 5. For anticipation/antecipação, Mistral’s accuracy of 0.47 (0.45, 0.49) was below the baseline’s 0.77 (0.76, 0.79), despite achieving a recall of 0.99 (0.98, 1.00). For disgust/nojo, EmoAtlas’s recall of 0.02 (0.00, 0.04) was below the baseline’s 0.11 (0.06, 0.15), but its precision was 0.56 (0.20, 0.89) with a wide confidence interval. For fear/medo, both BERTimbau and Mistral demonstrated high recall, at 0.99 (0.97, 1.00) and 0.97 (0.94, 0.99), respectively. For joy/alegria and sadness/tristeza, Mistral’s accuracy was below the baseline, with scores of 0.73 (0.71, 0.75) and 0.64 (0.62, 0.66), respectively, while its recall for both emotions was high,

at 0.96 (0.93, 0.98) and 0.98 (0.97, 1.00). The lowest recall for sadness was from EmoAtlas, with a value of 0.06 (0.04, 0.09).

Table 5. DeepSeek-generated headlines.

Emotion	Score	BERTimbau	Baseline	EmoAtlas	Mistral
<i>raiva</i> (anger)	Accuracy	0.93 (0.92, 0.94)	0.77 (0.75, 0.78)	0.86 (0.85, 0.88)	0.91 (0.89, 0.92)
	Precision	0.72 (0.67, 0.77)	0.13 (0.10, 0.17)	0.43 (0.31, 0.55)	0.61 (0.56, 0.65)
	Recall	0.79 (0.73, 0.83)	0.13 (0.10, 0.17)	0.10 (0.07, 0.14)	0.85 (0.80, 0.89)
<i>antecipação</i> (anticipation)	Accuracy	–	0.77 (0.76, 0.79)	0.71 (0.69, 0.73)	0.47 (0.45, 0.49)
	Precision	–	0.13 (0.09, 0.17)	0.26 (0.22, 0.29)	0.20 (0.18, 0.22)
	Recall	–	0.13 (0.09, 0.17)	0.65 (0.59, 0.70)	0.99 (0.98, 1.00)
<i>nojo</i> (disgust)	Accuracy	0.87 (0.85, 0.88)	0.81 (0.80, 0.82)	0.89 (0.88, 0.91)	0.94 (0.93, 0.95)
	Precision	0.43 (0.38, 0.48)	0.11 (0.07, 0.15)	0.56 (0.20, 0.89)	0.77 (0.71, 0.82)
	Recall	0.76 (0.71, 0.82)	0.11 (0.06, 0.15)	0.02 (0.00, 0.04)	0.68 (0.61, 0.74)
<i>medo</i> (fear)	Accuracy	0.84 (0.82, 0.85)	0.78 (0.76, 0.79)	0.86 (0.85, 0.88)	0.87 (0.85, 0.88)
	Precision	0.44 (0.40, 0.48)	0.13 (0.09, 0.16)	0.40 (0.31, 0.48)	0.49 (0.44, 0.53)
	Recall	0.99 (0.97, 1.00)	0.13 (0.09, 0.17)	0.20 (0.16, 0.25)	0.97 (0.94, 0.99)
<i>alegria</i> (joy)	Accuracy	0.94 (0.93, 0.95)	0.78 (0.77, 0.80)	0.78 (0.76, 0.80)	0.73 (0.71, 0.75)
	Precision	0.74 (0.69, 0.79)	0.12 (0.08, 0.16)	0.31 (0.27, 0.35)	0.31 (0.28, 0.34)
	Recall	0.80 (0.74, 0.84)	0.12 (0.08, 0.16)	0.66 (0.60, 0.72)	0.96 (0.93, 0.98)
<i>tristeza</i> (sadness)	Accuracy	0.88 (0.87, 0.90)	0.75 (0.73, 0.77)	0.86 (0.84, 0.87)	0.64 (0.62, 0.66)
	Precision	0.57 (0.52, 0.61)	0.15 (0.11, 0.18)	0.54 (0.38, 0.71)	0.29 (0.26, 0.32)
	Recall	0.91 (0.87, 0.94)	0.15 (0.11, 0.19)	0.06 (0.04, 0.09)	0.98 (0.97, 1.00)
<i>surpresa</i> (surprise)	Accuracy	0.94 (0.92, 0.94)	0.80 (0.78, 0.81)	0.84 (0.82, 0.85)	0.94 (0.93, 0.95)
	Precision	0.71 (0.65, 0.76)	0.11 (0.08, 0.15)	0.30 (0.24, 0.36)	0.83 (0.77, 0.88)
	Recall	0.73 (0.67, 0.78)	0.11 (0.07, 0.16)	0.32 (0.26, 0.38)	0.63 (0.57, 0.70)
<i>confiança</i> (trust)	Accuracy	–	0.79 (0.78, 0.80)	0.86 (0.84, 0.87)	0.89 (0.88, 0.91)
	Precision	–	0.12 (0.08, 0.16)	0.42 (0.37, 0.47)	0.54 (0.48, 0.59)
	Recall	–	0.12 (0.08, 0.16)	0.56 (0.50, 0.63)	0.70 (0.64, 0.75)

3.5. Semantic Frame Analysis with EmoAtlas in Portuguese

In addition to emotion detection in texts, EmoAtlas can be used to identify whether a given concept was framed positively, negatively, or neutrally along a given set of documents. To illustrate this, we used EmoAtlas on a corpus of tweets about COVID-19 vaccines [3]. Figure 2 shows two Plutchik flowers displaying the emotions of negative and positive tweets. We can see that the flower created from positive tweets expresses fewer negative emotions such as fear, anger, and disgust while also eliciting more joy compared to the flower created from negative tweets. The figure also displays TFMNs generated from the semantic frame (i.e., the associates) of the word “vacina” for positive and negative tweets, as explained in the Materials and Methods section. It is also clear that the TFMN of positive tweets expresses fewer negative connections (edges colored red) than the TFMN of negative tweets.

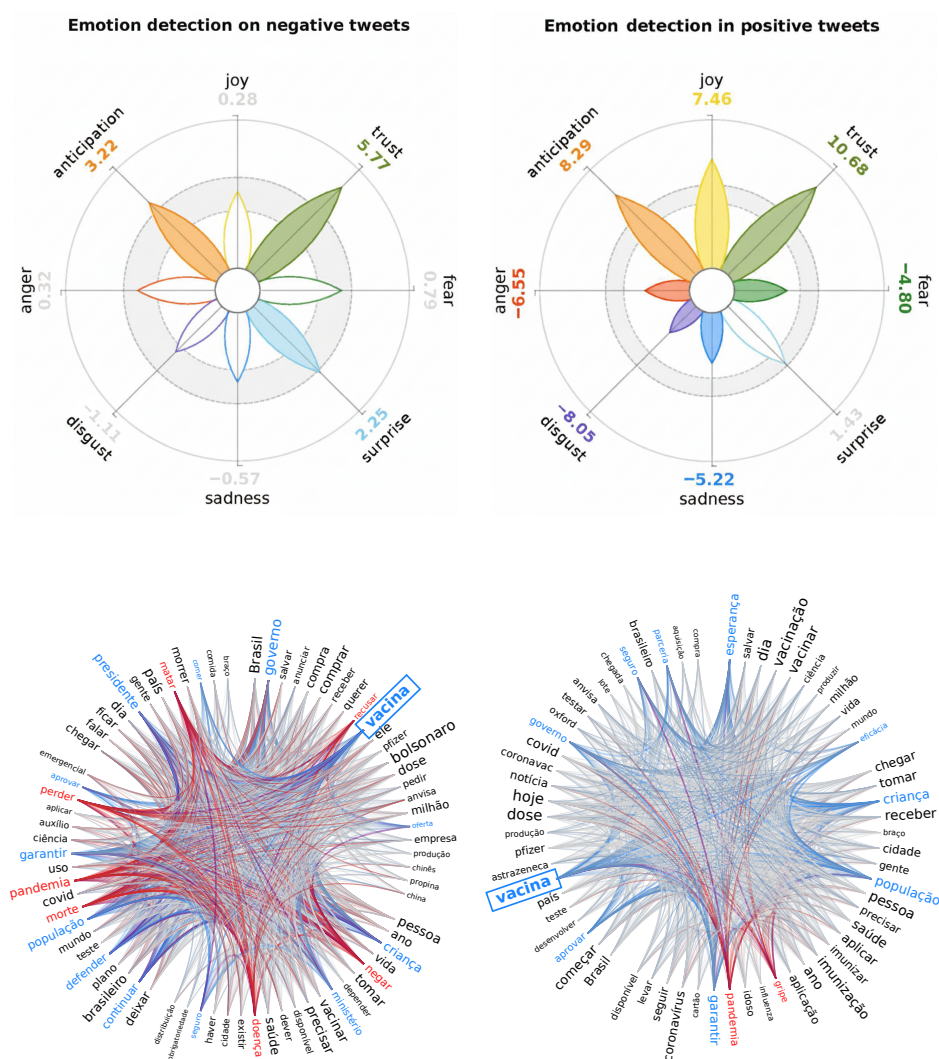


Figure 2. This figure presents the emotional flowers generated from Portuguese tweets about COVID-19 vaccination, classified as negative and positive by human annotators in the upper layer. The layer below presents TFMNs for the egonetworks around “vacina,” extracted from the same groups of tweets.

4. Discussion

This study benchmarked three distinct approaches for emotion detection in Brazilian Portuguese: a state-of-the-art large language model (Mistral), a language-specific transformer model (BERTimbau), and EmoAtlas, a lexicon-based method. Our analysis on natively human-annotated datasets from news headlines and stock market tweets showed a more balanced performance across all models, with results often close to a random baseline. In these real-world scenarios, we observed instances of accuracy below the baseline, such as Mistral for sadness/tristeza in news headlines and EmoAtlas for fear/medo in stock market tweets. All three models also struggled with recall for disgust/nojo in stock market tweets. In contrast, the translated GoEmotions dataset showed consistently low performance differences across translations, and similar scores compared to Portuguese human-annotated data. Our novel emotional fingerprinting approach revealed that transformer-based models generally outperformed EmoAtlas, with EmoAtlas showing wider confidence intervals for precision and a notable drop in recall. This complementary analysis highlights the distinct strengths and weaknesses of each model type when evaluated on different data sources.

On human-annotated datasets, performance across all models is not always superior to a stochastic baseline and with trade-off between precision and recall. Mistral's accuracy for sadness (0.56) and EmoAtlas's accuracy for fear (0.81) both fell below the baseline on news headlines and stock market tweets, respectively. Mistral's high recall of 0.94 for sadness in news headlines was accompanied with a low precision of 0.24. The wide confidence intervals for BERTimbau and Mistral on stock market tweets suggest instability in their predictions. This aligns with previous research by Hammes and de Freitas [26] where fine-tuning BERTimbau on translated corpora improved performance, but the quality of automatic translation and prevalence of idiomatic expressions could still degrade accuracy.

An analysis of misclassifications reveals that EmoAtlas, Mistral, and BERTimbau models produced high rate of false negatives. For example, Mistral failed to identify trust/*confiança* in stock market tweets with a rate of 0.987, while BERTimbau showed a high false negative rate for fear/*medo* (0.949) and disgust/*nojo* (0.931) on the same dataset. EmoAtlas similarly struggled with disgust/*nojo* and sadness/*tristeza* in machine-generated text, with rates of 0.977 and 0.936, respectively.

These high false negative rates are consistent with the documented challenges inherent to emotional classification of tweet data. The brevity of tweet content limits the contextual data available for analysis [1], and tweets often incorporate non-standard linguistic features such as slang, abbreviations, and hashtags [8]. This is compounded by the subjective nature of human language, which introduces ambiguity and context dependency [29]. Data imbalance in emotional datasets is another factor, with underrepresented classes leading to model bias [26]. Conversely, false positive rates indicate that values tend to be lower. Mistral had a false positive rate of 0.607 for anticipation on the machine-generated data and a rate of 0.499 for sadness in news headlines.

A qualitative analysis of misclassified examples provides further insight. For the *surprise* emotion in the stock market tweets dataset, models failed to detect its presence in examples such as @Paty_Meirelles na PETR4... *Só não pensei ver isso hoje, rs*. While the phrase "Só não pensei ver isso hoje" directly expresses surprise, the subtle phrasing and non-standard "rs" likely hindered detection. In other cases, such as "Algoritmo performando 12% acima do IBOV desde o início," surprise is inferred from market performance rather than explicit keywords. The use of technical jargon and stock tickers further complicates the task.

Conversely, Mistral had a false positive rate of 0.607 for *anticipation/antecipação* on machine-generated data and 0.499 for *sadness/tristeza* in news headlines. For fear/*medo* in the news headlines dataset, models misclassified examples characterized by violent themes, such as "Um morre e dois ficam feridos em assalto no Paraná" and "Padrasto mata enteado e simula suicídio do menino." These texts are negative, but the models appear to conflate general negative valence with the specific emotion of fear/*medo*.

An important difference between models is in their interpretability. EmoAtlas has an explicit pipeline which involves performing syntactic parsing, statistical analysis, and constructing TFMNs [1,8]. In contrast, LLMs like Mistral and BERTimbau function primarily as "black-box" models [38], offering high performance but limited transparency regarding their decision-making processes. This interpretability is especially valuable in contexts requiring justification or explanation of model predictions. Furthermore, EmoAtlas processes texts significantly faster, being approximately 40 times quicker on a CPU compared to the LLMs. This computational efficiency, coupled with its lower cost of implementation compared to fine-tuning large transformer models or using LLMs via API, positions EmoAtlas as a valuable and cost-effective alternative, especially in resource-constrained environments.

The application of EmoAtlas to COVID-19 vaccine tweets demonstrates its capacity for discerning emotional nuances and delivering interpretable outcomes through emo-

tional profiles and TFMNs for varying sentiments toward “vacina” [3]. This semantic frame analysis extends to investigations of various biases, including those related to gender, race, or political discourse. For example, prior research utilized TFMNs to reveal stereotypical perceptions of the gender gap in social discourse, showing how “woman” or “man” is associated with specific jargon [8]. This capability supports systematic examination of terms, such as “mulher” or “negro”, to ascertain if their semantic frames consistently elicit emotions. EmoAtlas provides a method for identifying and measuring biases with direct quantitative evidence, by revealing the linguistic features contributing to an emotional frame.

A significant challenge in Brazilian Portuguese emotion detection is the scarcity of large, high-quality annotated corpora [2,39]. To address this, our study employed datasets translated from English (GoEmotions) using different LLMs (Gemma and Mistral). The observation that both Mistral and EmoAtlas performed similarly on the GoEmotions data translated by either Gemma or Mistral suggests that potential noise introduced by LLM translation may not drastically compromise the utility of such datasets for benchmarking purposes.

Additionally, employing LLM-generated datasets (DeepSeek-generated headlines) introduced a novel benchmarking approach we termed “emotional fingerprinting”. Given LLMs’ advanced abilities for establishing linguistic conventions [40] while also displaying human-like biases when interacting together [41], one can only expect for LLMs to display also advanced emotional complexity, i.e., producing texts with non-trivial levels of emotional intensities. Our emotional fingerprinting revolved on the idea of having one LLM to generate such non-trivial emotional texts, across different topics, while having another LLM to reconstruct the original emotions. The variations in performance observed on this dataset might reveal model-specific “fingerprints”, like Mistral’s high recall/low precision vs. EmoAtlas’s converse for disgust. In other words, when analyzing performance on the DeepSeek Headlines dataset, we must consider its generative nature. An LLM (DeepSeek) produced both the text and its emotional labels. The alignment in performance between Mistral (the labeling LLM) and DeepSeek (the data-generating LLM) suggests overlapping feature space representations: since LLMs train on vast text corpora, this might lead them to develop similar internal representations of emotion-associated linguistic patterns. In other words, Mistral might be good at identifying what DeepSeek considers as joy, fear, or anticipation literally because both learned the same linguistic patterns for each one of those emotions, and therefore both would express them in the same way.

Our novel “emotional fingerprinting” approach suggests a new avenue for evaluating and understanding the emotional capabilities and biases of different language models by testing their ability to generate and recognize emotions associated with specific keywords or themes. This aligns with recent approaches in the literature using LLMs, like Centaur [42], as cognitive benchmarks for understanding psychological phenomena, including emotions.

EmoAtlas operates on an interpretable framework for emotional profiling [1,9]. As a lexicon-based, rule-driven model, EmoAtlas explicitly encodes human-defined emotional categories. It also incorporates their associated linguistic features. This makes its decision-making process transparent and directly verifiable. Differences in emotional labeling are quite crucial here, as discussed also in past works [35]. EmoAtlas relies on predefined, human-engineered emotional knowledge. LLMs, however, learn to recognize and reproduce emotional patterns implicitly from massive training data. This often occurs without an explicit, transparent conceptualization of emotions [24]. However, as mentioned above, Mistral’s higher scores on the DeepSeek synthetic dataset might simply reflect an ability to align with another LLM’s emotion generation mechanism. Interestingly, it does not prove a universally superior understanding or detection of genuine human emotion.

Despite its key contributions, this study has a few limitations. The human-annotated datasets used are specific to domains like stock market tweets and news headlines. The stock market dataset, in particular, showed lower annotation consensus for some emotions, and EmoAtlas, while powerful, is not ideally suited for very short texts (under 15 words). While LLM-translated data helps mitigate resource scarcity, it might not fully capture the cultural and linguistic specificities of Brazilian Portuguese emotion expression [2,29]. This study focused on a specific set of eight emotions and primarily benchmarked Mistral against EmoAtlas, leaving room for evaluation of other state-of-the-art models (like BERTimbau, cf. [39]) and a broader range of emotions. Future research could expand the benchmarking to include a wider array of state-of-the-art models and datasets, particularly larger, more diverse human-annotated corpora for Brazilian Portuguese.

Further investigation into the use of LLM-generated data for model “fingerprinting” could reveal more about how different AI architectures handle and represent emotional information. Exploring hybrid approaches that combine the global capabilities of LLMs [40] with the efficiency and interpretability of methods like EmoAtlas [1], potentially using LLMs for data annotation or lexicon refinement, could lead to more effective and accessible emotion detection tools for Brazilian Portuguese. Applying these tools to other complex Brazilian Portuguese NLP challenges, such as sarcasm detection or dialectal variations in emotion expression, would also be valuable.

5. Conclusions

This study provided a comprehensive benchmark of emotion detection models in Brazilian Portuguese, comparing a large language model (Mistral), a language-specific transformer model (BERTimbau), and a lexicon-based method (EmoAtlas). Our findings demonstrate that BERTimbau and Mistral outperformed the lexicon-based EmoAtlas, though they required more computational resources. The research also introduced a novel “emotional fingerprinting” methodology using LLM-generated datasets to probe the internal representations of these models. This approach revealed a key finding: a model’s high performance on synthetically generated data may not reflect a superior understanding of genuine human emotion but rather an alignment with the internal feature space of the generative model. By providing a quantitative benchmark, highlighting the trade-offs between performance, interpretability, and efficiency, and introducing a novel methodology for model evaluation, this work contributes valuable insights to the field of emotion analysis. The open datasets and code provided with this study are intended to foster future research in multilingual natural language processing, particularly in low-resource languages like Brazilian Portuguese.

Author Contributions: Conceptualization, T.D.D.A., A.C., and M.S.; methodology, T.D.D.A. and A.C.; software, T.D.D.A.; validation, T.D.D.A., A.C., and M.S.; formal analysis, T.D.D.A.; investigation, T.D.D.A., A.C., C.Q.C., and M.S.; data curation, T.D.D.A. and M.S.; writing original draft, T.D.D.A., A.C., C.Q.C., and M.S.; review and editing, T.D.D.A., A.C., C.Q.C., and M.S.; supervision, A.C. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: TDDA would like to thank the University of São Paulo for the International Exchange Scholarship. MS acknowledges the CALCOLO project for technical support in simulating LLMs locally. CALCOLO was funded by Fondazione VRT and Fondazione Caritro. AC and MS acknowledge support from the COGNOSCO grant (PS_22_27) funded by University of Trento.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: The data we generated for this paper is available on this Open Science Foundation Repository: <https://osf.io/qvu7m/> (accessed on 11 July 2025).

Acknowledgments: During the preparation of this manuscript/study, the authors used EmoAtlas for the purposes of emotional analysis. EmoAtlas is available online: <https://github.com/MassimoStella/emoatlas>, accessed on 15 September 2025.

Conflicts of Interest: Professor Massimo Stella and Dr. Alexis Carrillo Ramirez, co-authors of this manuscript, are also Guest Editors for the Special Issue “Understanding Transformers and large language models (LLMs) with Natural Language Processing (NLP)” in the *AI* journal. To ensure impartiality and avoid any perceived conflict of interest, this manuscript is being submitted to the general editorial board for independent peer review, separate from the Special Issue’s editorial process.

References

1. Semeraro, A.; Vilella, S.; Improta, R.; De Duro, E.S.; Mohammad, S.M.; Ruffo, G.; Stella, M. EmoAtlas: An emotional network analyzer of texts that merges psychological lexicons, artificial intelligence, and network science. *Behav. Res. Methods* **2025**, *57*, 77. [\[CrossRef\]](#)
2. Vieira da Silva, F.J.; Roman, N.T.; Carvalho, A.M. Stock market tweets annotated with emotions. *Corpora* **2020**, *15*, 343–354. [\[CrossRef\]](#)
3. Barberia, L.; Schmalz, P.; Trevisan Roman, N.; Lombard, B.; Moraes de Sousa, T. It’s about What and How you say it: A Corpus with Stance and Sentiment Annotation for COVID-19 Vaccines Posts on X/Twitter by Brazilian Political Elites. In Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, Albuquerque, NM, USA, 3–4 May 2025; Härmäläinen, M., Öhman, E., Bizzoni, Y., Miyagawa, S., Alnajjar, K., Eds.; The Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 365–376.
4. Hassani, H.; Beneki, C.; Unger, S.; Mazinani, M.T.; Yeganegi, M.R. Text Mining in Big Data Analytics. *Big Data Cogn. Comput.* **2020**, *4*, 1. [\[CrossRef\]](#)
5. Chen, Z.; Ye, J.; Tsai, B.; Ferrara, E.; Luceri, L. Synthetic politics: Prevalence, spreaders, and emotional reception of AI-generated political images on X. *arXiv* **2025**, arXiv:2502.11248.
6. Ekman, P.; Davidson, R.J. *The Nature of Emotion: Fundamental Questions*; Oxford University Press: Oxford, UK, 1994.
7. Plutchik, R. A psychoevolutionary theory of emotions. *Soc. Sci. Inf./Sur Les Sci. Soc.* **1982**, *21*, 529–553. [\[CrossRef\]](#)
8. Stella, M. Cognitive network science for understanding online social cognitions: A brief review. *Top. Cogn. Sci.* **2022**, *14*, 143–162. [\[CrossRef\]](#)
9. Abramski, K.; Ciringione, L.; Rossetti, G.; Stella, M. Voices of rape: Cognitive networks link passive voice usage to psychological distress in online narratives. *Comput. Hum. Behav.* **2024**, *158*, 108266. [\[CrossRef\]](#)
10. Fillmore, C.J.; Baker, C.F. Frame semantics for text understanding. In Proceedings of the WordNet and Other Lexical Resources Workshop, NAACL, Pittsburgh, PA, USA, 3–4 June 2001; Volume 6.
11. Bianchi, F.; Nozza, D.; Hovy, D. FEEL-IT: Emotion and Sentiment Classification for the Italian Language. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Online, 19–23 April 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2766–2774.
12. Sprugnoli, R. MultiEmotions-IT: A New Dataset for Opinion Polarity and Emotion Analysis for Italian. In Proceedings of the 7th Italian Conference on Computational Linguistics (CLiC-it 2020), Bologna, Italy, 1–3 March 2020; Accademia University Press: Torino, Italy, 2020; pp. 402–408.
13. Quispe, L.V.; Tohalino, J.A.; Amancio, D.R. Using virtual edges to improve the discriminability of co-occurrence text networks. *Phys. A Stat. Mech. Its Appl.* **2021**, *562*, 125344. [\[CrossRef\]](#)
14. Silva, F.N.; Amancio, D.R.; Bardosova, M.; Costa, L.d.F.; Oliveira, O.N., Jr. Using network science and text analytics to produce surveys in a scientific topic. *J. Inf.* **2016**, *10*, 487–502. [\[CrossRef\]](#)
15. Abdillahi, J.; Asror, I.; Wibowo, Y.F.A. Emotion Classification of Song Lyrics Using Bidirectional LSTM Method with GloVe Word Representation Weighting. *J. RESTI (Rekayasa Sist. Dan Teknol. Inf.)* **2020**, *4*, 723–729. [\[CrossRef\]](#)
16. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1532–1543. [\[CrossRef\]](#)
17. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805. [\[CrossRef\]](#)
18. Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A Survey on Large Language Model Based Autonomous Agents. *Front. Comput. Sci.* **2024**, *18*, 1–26. [\[CrossRef\]](#)
19. Zhai, X.; Chu, X.; Chai, C.S.; Jong, M.S.Y.; Istenic, A.; Spector, M.; Liu, J.B.; Yuan, J.; Li, Y. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity* **2021**, *2021*, 8812542. [\[CrossRef\]](#)

20. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [\[CrossRef\]](#)
21. Raza, M.; Jahangir, Z.; Riaz, M.B.; Saeed, M.J.; Sattar, M.A. Industrial applications of large language models. *Sci. Rep.* **2025**, *15*, 13755. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Rogers, A.; Kovaleva, O.; Rumshisky, A. A Primer in BERTology: What We Know About How BERT Works. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 842–866. [\[CrossRef\]](#)
23. Hutto, C.; Gilbert, E. Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8, pp. 216–225.
24. Mohammad, S.M.; Turney, P.D. Crowdsourcing a Word-Emotion Association Lexicon. *Comput. Intell.* **2013**, *29*, 436–465. [\[CrossRef\]](#)
25. Dodds, P.S.; Harris, K.D.; Kloumann, I.M.; Bliss, C.A.; Danforth, C.M. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE* **2011**, *6*, e26752. [\[CrossRef\]](#)
26. Hammes, L.O.A.; de Freitas, L.A. Utilizando bertimbau para a classificação de emoções em português. In Proceedings of the Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), Online, 29 November–3 December 2021; SBC: Tokyo, Japan, 2021; pp. 56–63.
27. Kansaon, D.P.; Brandão, M.A.; de Paula Pinto, S.A. Análise de Algoritmos para Detecção de Emoções em Tweets em Português Brasileiro. *Rev. Sist. Comput. (ISYS)* **2018**, *11*, 1–15. [\[CrossRef\]](#)
28. Oliveira, F.B.; Sichman, J.S. Portuguese Emotion Detection Model Using BERTimbau Applied to COVID-19 News and Replies. In Proceedings of the Intelligent Systems, Amsterdam, The Netherlands, 28–29 August 2025; Paes, A., Verri, F.A.N., Eds.; Springer: Cham, Switzerland, 2025; pp. 265–280. [\[CrossRef\]](#)
29. Martinazzo, B.; Dosciatti, M.M.; Paraíso, E.C. Identifying emotions in short texts for brazilian portuguese. In Proceedings of the IV International Workshop on Web and Text Intelligence (WTI 2012), Curitiba, Brazil, 20–25 October 2012; p. 16.
30. Souza, F.; Nogueira, R.; Lotufo, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proceedings of the Intelligent Systems*; Cerri, R., Prati, R.C., Eds.; Springer: Cham, Switzerland, 2020; pp. 403–417.
31. Veríssimo, V.; Costa, R. Using Data Augmentation and Neural Networks to Improve the Emotion Analysis of Brazilian Portuguese Texts. In Proceedings of the 26th Brazilian Symposium on Multimedia and the Web (WebMedia), Ribeirão Preto, Brazil, 23–27 October 2020; ACM: New York, NY, USA, 2020; pp. 9–16. [\[CrossRef\]](#)
32. de Souza, K.F.; Pereira, M.H.R.; Dalip, D.H. UniLex: Método Léxico para Análise de Sentimentos Textuais sobre Conteúdo de Tweets em Português Brasileiro. *Abakós* **2017**, *5*, 79–96. [\[CrossRef\]](#)
33. Carvalho, F.; Santos, G.; Guedes, G.P. AffectPT-br: An Affective Lexicon Based on LIWC 2015. In Proceedings of the 37th International Conference of the Chilean Computer Science Society (SCCC), Santiago, Chile, 5–9 November 2018; IEEE: New York, NY, USA, 2018; pp. 1–5. [\[CrossRef\]](#)
34. James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. Resampling Methods. In *An Introduction to Statistical Learning: With Applications in Python*; Springer International Publishing: Cham, Switzerland, 2023; pp. 201–228. [\[CrossRef\]](#)
35. Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; Ravi, S. GoEmotions: A Dataset of Fine-Grained Emotions. *arXiv* **2020**, arXiv:2005.00547.
36. Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. Gemma 3 technical report. *arXiv* **2025**, arXiv:2503.19786. [\[CrossRef\]](#)
37. Aldawsari, A.H.; Hamad. Evaluating Translation Tools: Google Translate, Bing Translator, and Bing AI on Arabic Colloquialisms. *Arab World English Journal (AWEJ) Special Issue on ChatGPT* **2024**. Available online: <https://ssrn.com/abstract=4814733> (accessed on 10 June 2025).
38. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Vianna, D.; Carneiro, F.; Carvalho, J.; Plastino, A.; Paes, A. Sentiment analysis in Portuguese tweets: An evaluation of diverse word representation models. *Lang. Resour. Eval.* **2024**, *58*, 223–272. [\[CrossRef\]](#)
40. Ashery, A.F.; Aiello, L.M.; Baronchelli, A. Emergent social conventions and collective bias in LLM populations. *Sci. Adv.* **2025**, *11*, eadu9368. [\[CrossRef\]](#)
41. Cau, E.; Failla, A.; Rossetti, G. Bots of a Feather: Mixing Biases in LLMs’ Opinion Dynamics. In *Proceedings of the International Conference on Complex Networks and Their Applications*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 166–176.
42. Binz, M.; Akata, E.; Bethge, M.; Brändle, F.; Callaway, F.; Coda-Forno, J.; Dayan, P.; Demircan, C.; Eckstein, M.K.; Éltető, N.; et al. A foundation model to predict and capture human cognition. *Nature* **2025**, *644*, 1002–1009. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.