Sentiment Analysis in Psychotherapy in Brazilian Portuguese: The Role of Domain-Specific Fine-Tuning

Lucas M. Vasconcelos[1], Felipe Corchs[2], Fátima L. S. Nunes[1], Ariane Machado-Lima[1]

[1]School of Arts, Sciences and Humanities (EACH), University of São Paulo (USP)
Rua Arlindo Béttio, 1000 – Ermelino Matarazzo, 03828-000, São Paulo-SP, Brazil

[2]Department of Psychiatry - University of Sao Paulo's School of Medicine (FMUSP)
Rua Doutor Ovídio Pires de Campos, 785 - 05403-010, São Paulo-SP, Brazil

{lucas.m.vasconcelos, felipe.corchs, fatima.nunes, ariane.machado}@usp.br

Abstract. *Text analysis has become an important tool in mental healthcare. However, the development of Natural Language Processing (NLP) applications for Brazilian Portuguese remains challenging, as most advanced pre-trained models are primarily designed for English and come from different domains than mental health. This study explores sentiment analysis in psychotherapy session transcripts using traditional NLP techniques, large language models, including the contextual model BERTimbau. We compare their performance on a pilot Brazilian labeled dataset of mental health dialogs. Results indicate that fine-tuning BERTimbau on domain-specific data, even when the available dataset is relatively small, significantly improves classification accuracy.*

1. Introduction

Text analysis has increasingly become a valuable tool in mental healthcare, enabling the automation and support of clinical tasks such as patient monitoring, risk assessment, and emotional state tracking [Calvo et al. 2017]. The modern healthcare system is a knowledge-driven industry characterized by the exponential growth of narrative information, such as physician case notes, discharge summaries, and radiology reports. The ability to extract meaningful insights for decision-making remains a significant challenge for healthcare systems [Panchbhai and Pathak 2022]. This challenge is particularly evident in the field of mental health, where the core of the clinical process resides in the patient-therapist dialog and unstructured clinical transcripts. The field of mental health presents a significant opportunity for the development of tools capable of assessing the complex sentimental dynamics of the patient-therapist interaction.

Despite its importance, the field remains constrained by traditional assessment methods. Manual review and coding of psychotherapy sessions are time-consuming and resource-intensive. Consequently, research often leans heavily on self-report measures. The integration of Machine Learning offers a transformative solution, providing the technological foundation to automate and analyze large-scale therapeutic data with enhanced precision and reproducibility [Goldberg et al. 2020].

One possible approach to achieving this integration is through the automatic recognition and interpretation of human emotions using Affective Computing [Picard 2000]. By observing affective cues through Natural Language Processing (NLP),

computers can identify meaningful patterns in therapeutic discourse. Within NLP, Sentiment Analysis (SA) has evolved from social media monitoring to the complex domain of mental health, where deep learning now enables the detection of subtle textual patterns previously deemed inaccessible [Tao and Tan 2005]. With the advent of new architectures in NLP, particularly those based on the Transformer framework, research in this field has increasingly incorporated models grounded in this paradigm, especially language models [Bengio et al. 2021]. The adoption of methods specifically tailored to Brazilian Portuguese may improve the performance of SA.

Therefore, the development of sentiment analysis tools specifically tailored to the linguistic and contextual nuances of psychotherapy should not be viewed as an end in itself, but rather as a foundational necessity. Creating such domain-specific models is a critical prerequisite for enabling more sophisticated downstream research and building advanced applications, such as clinical decision support systems, which depend entirely on a deep and accurate understanding of therapeutic discourse.

This paper performs a preliminary investigation into the application of sentiment analysis to Portuguese-language psychotherapy session transcripts. We conducted a comparative study of various natural language processing approaches, including traditional machine learning algorithms, deep learning architectures, Large Language Models (LLMs), and the Portuguese-specific BERTimbau model. A central aim of this research is to evaluate the impact of fine-tuning using a small, domain-specific dataset, given that limited labeled data is common in many fields due to the difficulty of manual text annotation. We analyzed the performance of models trained on out-of-context datasets versus those fine-tuned with data from the psychotherapy domain, thereby quantifying the substantial effect of contextual training on key performance metrics and demonstrating the importance of tailoring models to the specific nuances of therapeutic interactions.

The remainder of this article is organized as follows: Section 2 presents the theoretical background for this study; Section 3 reviews related studies and existing approaches to sentiment analysis in therapeutic and psychological contexts, highlighting current challenges and gaps in the literature; Section 4 details the materials and methods used in this study; Section 5 presents and analyzes the results obtained from the experiments; finally, Section 6 offers concluding remarks.

2. Background

Machine learning (ML) has been used in the SA area to identify sentiments in texts. This method leverages features and linguistic characteristics extracted from textual corpora using specific algorithms, allowing the text to be adequately represented for the application of ML models. In particular, supervised methods require a labeled dataset to effectively train the algorithms [Liu 2022]. This approach offers several advantages over lexicon-based methods, including the ability to capture context and adapt to new data [Van Der Veen and Bleich 2025].

ML algorithms require textual data to be represented in fixed-size numerical vectors, as raw text cannot be directly processed. Examples are Logistic Regression (RegLog), a traditional ML algorithm, and Long Short-Term Memory (LSTM), a deep learning algorithm. LSTM was designed to overcome the vanishing gradient problem common in traditional Recurrent Neural Networks [Hochreiter and Schmidhuber 1997]. It

achieves this through a gated architecture that regulates information flow, retaining or discarding information as needed, allowing efficient pattern capture in sequential data.

Traditional text techniques include Bag-of-Words (BoW), which uses word counts or binary indicators, and TF-IDF, which weighs terms based on frequency and rarity across documents. These approaches often produce sparse representations and ignore word order and syntax [Chollet 2021]. To overcome these limitations, word embeddings such as Word2Vec were introduced. They map words into dense vectors in a continuous space, preserving semantic and syntactic relationships. These vectors support arithmetic operations for analogies, e.g., king - man + woman $\simeq$ queen [Levy and Goldberg 2014].

More advanced models such as BERT (Bidirectional Encoder Representations from Transformers) use a transformer architecture to capture contextual relationships in both directions (left and right of a word) [Devlin et al. 2018]. Pretrained BERT can be fine-tuned for various NLP tasks. A Brazilian Portuguese version, BERTimbau, was trained on the brWaC corpus [Souza et al. 2020].

Large Language Models (LLMs) such as Bode [Garcia et al. 2024] and Sabiá [Pires et al. 2023], based on LLaMA2, are deep learning models trained on massive text datasets, enabling them to understand complex language patterns. Bode is optimized for Portuguese instruction-following tasks, while Sabiá excels in a wide range of NLP benchmarks in Portuguese [Brown et al. 2020, Radford et al. 2019].

To evaluate the ML models, several metrics can be employed. In this work, we used macro-averaged precision, recall, and the F1-score. These values are calculated according to Equations 1 to 3, where each of the l classes is treated as the positive class in turn, and the others as negative. The following values are considered for each class i: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

$$\text{Precision}M = \frac{1}{l}\sum_{i=1}^{l}\frac{TP_i}{TP_i + FP_i} \tag{1}$$

$$\text{Recall}M = \frac{1}{l}\sum_{i=1}^{l}\frac{TP_i}{TP_i + FN_i} \tag{2}$$

$$\text{F1-score}_M = \frac{2 \cdot \text{Precision}_M \cdot \text{Recall}_M}{\text{Precision}_M + \text{Recall}_M} \tag{3}$$

3. Related work

Recent studies have extensively explored the efficacy of specialized adaptation in NLP to bridge the performance gap in clinical and linguistic domains. One significant study investigated the safety performance of LLMs, concluding that fine-tuning specifically for mental health contexts significantly enhances response reliability and safety, proving more effective than simply increasing model size [Kalinich et al. 2026]. Within the specific scope of Brazilian Portuguese, comparative research between multilingual state-of-the-art models and fine-tuned LLMs demonstrated that linguistic-specific adaptation is essential for accurate sentiment analysis. This study revealed that fine-tuning not only reduces hallucinations but also allows specialized models to achieve higher precision in the nuances

of the Portuguese language compared to larger, general-purpose multilingual counterparts [da Fonseca Schuck et al. 2025]. Complementing these findings, researchers assessed the clinical impact of model adaptation in automated therapy, finding that fine-tuning models on datasets centered on therapeutic techniques significantly improved the quality of dialog and patient outcomes as evaluated through digital patient simulations [Yosef et al. 2025]. Collectively, these works underscore that domain-specific fine-tuning is a critical requirement for achieving high-accuracy, safe, and contextually aware applications in mental health and non-English linguistic environments.

This study is the first to systematically evaluate multiple Brazilian Portuguese language models, including different versions of BERTimbau and LLMs. While earlier research predominantly addressed English-language corpora and used binary or simplified sentiment classifications, this work explores a three-class structure (positive, neutral, negative) applied to real psychotherapy sessions in Brazilian Portuguese. Additionally, it integrates automatic transcription and sentiment classification in a complete processing pipeline. The inclusion of both traditional classifiers and deep learning models, along with the evaluation of underexplored language-specific LLMs, represents a novel and comprehensive approach to affective computing in mental health applications.

4. Material and methods

This section describes the experimental pipeline developed to evaluate sentiment analysis in psychotherapeutic contexts. The methodology encompasses the selection of datasets, the audio processing and transcription workflow, the manual annotation protocol, and the comparative evaluation of multiple machine learning architectures.

4.1. Datasets

Two distinct datasets were utilized: a general-purpose social media corpus (TweetSentBR) and a domain-specific clinical corpus, the Labeled COMVC-19 (LC).

4.1.1. TweetSentBR dataset

TweetSentBr (TSB) [Brum and Nunes 2017] is a *corpus* created from Twitter in Brazilian Portuguese, comprising 15,000 texts related to comments on Brazilian television programs. Sentiment annotation was manually performed by three annotators. This dataset has an imbalanced three-class distribution, consisting of 44% positive, 30% negative, and 26% neutral instances.

4.1.2. LC dataset

A domain-specific *corpus* was created using records of psychotherapy sessions of the COMVC-19[1] project, a program developed in 2020 to offer preventive mental health support to workers at the Hospital of Clinical Hospital of the School of Medicine of the University of São Paulo (HCFMUSP). During a period of 152 days, 16 therapists and 24 patients participated in 102 sessions, with one video recorded per session. All participants

[1] *"Psicoterapia no COMVC-19: um estudo computacional naturalístico"*

provided informed consent for the recordings to be used for research purposes. Due to the difficulty of expert labeling, two of these 102 sessions were randomly selected for manual annotation of the sentiment of each spoken phrase. The whole process, consisting of transcription and sentiment annotation was performed as follows. This project was approved by the Brazilian National Research Ethics Committee (CONEP) under the protocol number 30710820.8.0000.0068.

Transcription

The transcription of the audio sessions was automatically performed using Automatic Speech Recognition (ASR) technology, which allows for the automated conversion of human speech into text.

Four transcription tools were considered: IBM Watson, Google Speech, Vosk and Whisper. The first three tools had been previously evaluated in [Luna et al. 2022]. In this work, we extended this evaluation to include *Whisper* [Radford et al. 2023], using the same dataset described in [Luna et al. 2022]. *Whisper* is an artificial neural network used for ASR tasks. Its architecture comprises five models: *Tiny, Base, Small, Medium, and Large*. Each model has distinct configurations that affect the performance of the generated transcriptions. We evaluated the *Large* model [Radford et al. 2023].

The Word Error Rate (WER) metric [Ali and Renals 2018] was calculated for each transcription tool. This metric is calculated by summing the number of inserted (I), deleted (D), and substituted (S) words found in the generated transcription, and dividing this total by the number of words in the original reference sentence (N). This measure has a minimum value of zero but does not have a maximum value.

The tool with the highest performance (lowest WER) was also evaluated on a sample of the psychotherapy audios in order to validate the tool in the domain specific audios. For this, the first minute of 30 sessions of the COMVC-19 project was manually transcribed. The 30 sessions were randomly selected with the following distribution:

- 10 sessions between two men (HH);
- 10 sessions between a man and a woman (HM);
- 10 sessions between two women (MM).

Sentiment annotation

In a single utterance during a psychotherapy session, it is common for different sentiments to be expressed across multiple sentences. For this reason, the utterances were segmented by full stops, leading to 2,333 sentences (1,284 texts from one session and 1,049 from the other). The texts of the two sessions were concatenated, and the sentences were shuffled and randomly split into subsets of 100 sentences. Therefore, each subset contains isolated sentences without the full context of the interactions between patients and therapists. Each sentence subset was annotated by one of the nine licensed psychiatrists and psychologists with clinical experience, all affiliated with HCFMUSP. Consequently, no inter-rater reliability analysis or concordance check was performed across the classifications. We chose to annotate each sentence in isolation to protect the privacy of the psychotherapy sessions, since sharing full sessions with third parties would raise important ethical concerns. In addition, each sentence subset was annotated by only one expert due to limited expert availability. The annotation consisted of classifying the sen-

tence sentiment as positive, negative, or neutral. These annotated 2,333 sentences compose the Labeled COMVC-19 (LC) dataset. This dataset is characterized by a significant class imbalance: 64% neutral, 23% negative, and 13% positive instances.

4.2. Evaluation of sentiment analysis algorithms

Two large language models (Bode and Sabiá), LSTM, BERTimbau, TF-IDF and logistic regression were used to compose six different sentiment analysis approaches that were evaluated:

1. Bode (LLM) as an oracle;
2. Sabiá (LLM) as an oracle;
3. *BERTimbau* used as a pre-trained language model to generate text embeddings, without fine-tuning, followed by an LSTM classifier trained on the *TweetSentBR* dataset;
4. *BERTimbau* used as a pre-trained language model to generate text embeddings, without fine-tuning, with an LSTM classifier trained on the LC dataset;
5. *BERTimbau* as both a language model and classifier (fine-tuned using the LC dataset);
6. *TF-IDF* with a Logistic Regression classifier.

The inclusion of approaches 3 and 4 is primarily intended to evaluate the impact of domain-specific training on model performance. Specifically, approach 3 employs a classifier trained on the TweetSentBR dataset, a general-domain corpus consisting of social media texts unrelated to the psychotherapeutic context. By directly comparing its results with those of approach 4, which is trained on the domain-specific dataset (LC), it will be possible to quantify the significance of corpus specialization for the task of sentiment analysis in psychotherapy sessions.

To mitigate the risk of data leakage, the models were trained and evaluated using a 2-fold, session-based cross-validation. Specifically, in each cross-validation iteration, the training phase used data from one session, whereas testing was conducted on the other, entirely unseen session. The non-trainable models (Bode and Sabiá) and the LSTM pre-trained on *TweetSentBR* were evaluated exclusively on the test sessions, bypassing the training phase. This procedure ensured that all algorithms yielded performance metrics based on the same test sessions, resulting in paired samples suitable for subsequent statistical analysis.

For each test session, the macro F1-score ($F1_{\text{macro}}$) was calculated according to Equation 1. Consequently, each algorithm produced two independent measures that were used to compute the average $F1_{\text{macro}}$. Further details regarding the evaluation protocol for each algorithm are provided below.

4.2.1. Bode and Sabiá

For the evaluation of sentiment analysis using the LLMs Bode and Sabiá, each sentence from the LC dataset was input with a prompt asking the model to classify its sentiment as Positive, Neutral, or Negative, using a few-shot format with examples. To ensure deterministic and reproducible outputs, the generation *temperature* was set to 0:

“
You are a question-and-answer assistant. Each provided context will have a corresponding sentiment. There are three types of sentiment: Positive, Neutral, and Negative. Your task is to identify the corresponding sentiment given a sentence. Below are some examples:

Sentence: “Today I’m feeling good.” Answer: Positive
Sentence: “I had a fight with my family.” Answer: Negative
Sentence: “I graduated in medicine three years ago.” Answer: Neutral
”

The prompt above represents the English translation, whereas the original version in Portuguese used during the experiments is presented below:

“
Você é um assistente de perguntas e respostas. Cada contexto fornecido terá um sentimento correspondente. Existem três tipos de sentimento: Positivo, Neutro e Negativo. Sua tarefa é identificar o sentimento correspondente dada uma frase. Abaixo estão alguns exemplos:

Frase: “Hoje estou me sentindo bem.” Resposta: Positivo
Frase: “Tive uma briga com a minha família.” Resposta: Negativo
Frase: “Eu me formei em medicina há três anos.” Resposta: Neutro
”

In each iteration of the validation, the test fold was used for evaluation, and the training fold was discarded.

4.2.2. LSTM Trained on TSB or LC datasets

Two approaches using LSTM were compared using *BERTimbau*$_{base}$ embeddings: one trained on the TSB dataset and the other on the LC dataset. The rationale was to compare the classifier performance when trained using a larger, less domain-specific dataset (TSB, Twitter domain with 15,000 texts) versus a smaller, domain-specific dataset (LC, psychotherapy domain with an average of 1,166 sentences in the training fold).

The hyperparameters used for training were: *batch size*: 16; *epochs*: 5; *number of hidden units*: 128; *learning rate*: 0.001. Only these parameters were tested for the LSTM models, as the objective was to maintain a fixed configuration to focus the analysis on the impact of training data domain and size. When using the TSB dataset for training, just one LSTM was trained on this dataset, and during each cross-validation iteration, the test fold of the LC dataset was used for evaluation, disregarding the training fold. When using the LC dataset for training, two LSTMs were trained and tested using this dataset through the cross-validation process.

4.2.3. Fine-tuned *BERTimbau*

BERTimbau$_{base}$ and *BERTimbau*$_{large}$ were fine-tuned using the LC dataset. The cross-validation was performed using the LC training folds for fine-tuning and LC test folds for evaluation. Learning rates of 2e-5, 3e-5, and 5e-5 were tested with 4 and 10 training

epochs. Due to class imbalance, a weighted cross-entropy loss function was employed during training.

4.2.4. TF-IDF and Logistic Regression Classifier

Two Logistic Regression models were trained using TF-IDF representations extracted from the LC training folds during the cross-validation process. Dimensionality was limited to the top 5,000 terms ranked by weighted frequency and discriminative power. This approach aimed to balance model expressiveness and generalization capability while maintaining computational efficiency.

5. Results and discussion

5.1. Analysis of transcription tools

Table 1 compares the Word Error Rates (WER) obtained by *Whisper* and three other transcription tools when tested on the dataset described in another study [Luna et al. 2022]. The results show that *Whisper* stands out due to its low WER and stability (low standard deviation). *Whisper* achieved equally good performance when evaluated on a sample of sentences used to compose the LC dataset (Table 2). Therefore, this tool was selected to transcribe both sessions that were used to create the LC dataset.

Tabela 1. Word Error Rate (WER) comparison among ASR tools. Standard deviation is shown in parentheses. *: values obtained in [Luna et al. 2022].

Tool	IBM*	Google Speech*	Vosk*	Whisper
WER	17.65% (5.78)	13.42% (3.72)	18.47% (4.08)	**1.15% (0.13)**

Tabela 2. Word error rate (WER) achieved using *Whisper* on the sample of psychotherapy session audios. HM: sessions involving a man and a woman; MM: sessions involving two women; HH: sessions involving two men.

Category	Duration	Number of sessions	Whisper WER
HM	1 min	10	1.20%
MM	1 min	10	0.67%
HH	1 min	10	0.95%

5.2. Performance of sentiment analysis algorithms

Table 3 shows the mean, followed by its standard deviation of $f1_score_M$, $precision_M$ and $recall_M$ across the two sessions evaluated. The results evidence the superiority of the fine-tuned $BERTimbau_{base}$ classifier in achieving the best performance measures. These findings indicate that employing a language model capable of capturing contextual information significantly improves classification performance. The inferior performance of the LSTM using $BERTimbau_{base}$ embeddings, when compared to the fine-tuned $BERTimbau_{base}$, highlights the importance of adapting the model to the specific domain of the data. This suggests that fine-tuning enables $BERTimbau_{base}$ to learn contextual textual representations better aligned with the particularities of the psychotherapy domain, thereby improving the identification of patterns relevant to the classification task.

Tabela 3. Performance metrics obtained by the evaluated models. Standard deviation is shown in parentheses. The highest values are in bold.

			F1-score$_M$				
	$BERTimbau_{\text{base}}$	$BERTimbau_{\text{large}}$	LSTM LC	LSTM TSB	Bode	RegLog	Sabiá
Mean	**0.8312**	0.8126	0.6640	0.6558	0.6340	0.5717	0.3268
	(0.0491)	(0.0502)	(0.0293)	(0.0527)	(0.0282)	(0.0218)	(0.0176)
			Precision$_M$				
	$BERTimbau_{\text{base}}$	$BERTimbau_{\text{large}}$	LSTM LC	LSTM TSB	Bode	RegLog	Sabiá
Mean	**0.8313**	0.8182	0.6676	0.6514	0.6896	0.6069	0.7375
	(0.0491)	(0.0511)	(0.0380)	(0.0473)	(0.0281)	(0.0230)	(0.0442)
			Recall$_M$				
	$BERTimbau_{\text{base}}$	$BERTimbau_{\text{large}}$	LSTM LC	LSTM TSB	Bode	RegLog	Sabiá
Mean	**0.8312**	0.8172	0.6602	0.65136	0.6267	0.6067	0.3558
	(0.0145)	(0.0181)	(0.0283)	(0.0319)	(0.0264)	(0.0201)	(0.0176)

It was observed that the $BERT_{base}$ model outperformed $BERT_{large}$ in terms of performance, achieving higher *f1-score*M values. This result was expected because $BERT_{large}$ has a more complex architecture with a greater number of parameters, so it benefits less from a small fine-tuning dataset. This analysis highlights that, in scenarios where data is limited or exhibits specific characteristics, the $BERT_{base}$ model may offer a more appropriate balance between performance and computational efficiency, making it a preferable choice. The optimized hyperparameters were defined as follows: number of epochs set to 10 and learning rate set to $5E$–05.

After analyzing the performance metrics, it was observed that the LSTM trained on the LC dataset achieved better results. Specifically, the LSTM trained on TSB obtained an *f1-score*$_M$ of 0.6558, whereas the LSTM trained on LC dataset obtained an *f1-score*$_M$ of 0.6640.

The RegLog model establishes a linear relationship between the independent variables and the dependent variable. This can be a limitation in cases involving more complex interactions, further aggravated by the fact that TF-IDF generates a sparse matrix, which may lead to challenges for RegLog due to high dimensionality. Using TF-IDF as a language model also presents limitations. This representation does not consider text semantics and ignores word order, which impairs the understanding and modeling of textual information.

The use of Large Language Models (LLMs), such as Sabiá and Bode, is easy to apply since they do not require prior training. However, their performance in sentiment analysis specific to the psychotherapy context was suboptimal, mainly due to the lack of task-specific fine-tuning. Comparing the two LLM's, the Sabiá model showed lower performance compared to Bode. This result may be related to a domain mismatch in Sabiá's pre-training. It is plausible that its training corpus lacked sufficient exposure to the subtle sentimental lexicon of therapeutic interactions, thereby hindering its effectiveness in this context.

5.3. Limitations

The application of sentiment analysis in psychotherapy is hindered by significant technical challenges, primarily centered on data scarcity and the risks of error propagation.

High-quality, large-scale datasets are rare due to stringent ethical mandates and privacy regulations (such as HIPAA[2] and LGPD[3]), a problem particularly acute for under-resourced languages, such as Brazilian Portuguese. Furthermore, the reliance on clinical experts for the time-intensive annotation of nuanced emotional expressions introduces potential inter-annotator variability. These data-level hurdles are compounded by technical limitations within the processing pipeline; residual inaccuracies in foundational stages, such as Automatic Speech Recognition (ASR) or speaker diarization, can propagate through the analysis. Even minor transcription errors, such as the omission of a negative particle, can invert semantic meaning and lead to a cumulative skewing of quantitative results, highlighting the fragility of automated sentiment classification in clinical contexts.

The sentiment annotation process was conducted by a single annotator, and no inter-annotator agreement analysis was performed. As a result, the reliability and reproducibility of the labeling process cannot be formally assessed, and potential subjective bias may have influenced the ground truth used for model training and evaluation. The absence of agreement metrics, such as Cohen's Kappa or Krippendorff's Alpha, limits the interpretability of the reported performance results. Moreover, the dataset employed in this work is relatively small, which restricts the statistical power of the experiments and may hinder the generalizability of the findings. Future work should therefore include multi-annotator validation procedures and larger, more diverse datasets to strengthen empirical robustness.

5.4. Implications and impact

Despite the inherent limitations, this study offers significant clinical and methodological contributions to the intersection of mental health and applied NLP. From a clinical perspective, the findings demonstrate the feasibility of integrating automated tools into therapeutic practice, establishing a foundation for Clinical Decision Support Systems that augment clinical judgment through objective, longitudinal insights into session dynamics. Such systems could facilitate the identification of therapeutic turning points and provide a quantitative framework for supervision and training. Methodologically, this research underscores the necessity of domain adaptation, serving as a blueprint for applying sentiment analysis to specialized contexts characterized by unique linguistic conventions and data scarcity. Ultimately, the study suggests that specialized discourse, such as therapeutic dialog, demands context-aware modeling to successfully bridge the gap between computational analysis and real-world clinical utility

6. Conclusion

This study represents an important contribution to the evidence supporting domain-specific models for high-fidelity sentiment detection in psychotherapy. In particular, it shows that even a relatively small domain-specific dataset can substantially improve model performance. The use of a small dataset reflects practical constraints in real-world clinical settings: ethical issues, time demands for labeling, and the difficulty of recruiting specialists to perform manual annotation. By leveraging a hybrid strategy that utilizes models trained on specialized therapeutic corpora, the study provides a contextually aware framework for interpreting affective dynamics.

[2]Health Insurance Portability and Accountability Act.

[3]General Data Protection Act, from Portuguese *Lei Geral de Proteção de Dados*.

Referências

Ali, A. and Renals, S. (2018). Word error rate estimation for speech recognition: e-WER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24.

Bengio, Y., Lecun, Y., and Hinton, G. (2021). Deep learning for AI. *Communications of the ACM*, 64(7):58–65.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Brum, H. B. and Nunes, M. d. G. V. (2017). Building a sentiment corpus of tweets in Brazilian Portuguese. *arXiv e-prints*, pages arXiv–1712.

Calvo, R. A., Milne, D. N., Hussain, M. S., and Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

da Fonseca Schuck, A., Garcia, G. L., Manesco, J. R. R., Paiola, P. H., and Papa, J. P. (2025). Evaluating large language models for brazilian portuguese sentiment analysis: A comparative study of multilingual state-of-the-art vs. brazilian portuguese fine-tuned llms. *Journal of the Brazilian Computer Society*, 31(1):884–916.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Garcia, G. L., Paiola, P. H., Morelli, L. H., Candido, G., Júnior, A. C., Jodas, D. S., Afonso, L. C. S., Guilherme, I. R., Penteado, B. E., and Papa, J. P. (2024). Introducing bode: A fine-tuned large language model for Portuguese Prompt-Based Task.

Goldberg, S. B., Flemotomos, N., Martinez, V. R., Tanana, M. J., Kuo, P. B., Pace, B. T., Villatte, J. L., Georgiou, P. G., Van Epps, J., Imel, Z. E., et al. (2020). Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of counseling psychology*, 67(4):438.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kalinich, M., Luccarelli, J., Maria Jr, J. S., Williams, G., Moss, F., and Torous, J. (2026). Evaluating the effect of mental health fine-tuning relative to other model characteristics on LLM safety performance. *medRxiv*, pages 2026–01.

Levy, O. and Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Liu, B. (2022). *Sentiment analysis and opinion mining*. Springer Nature.

Luna, A. S., Machado-Lima, A., and Nunes, F. L. (2022). Analysis of transcription tools for Brazilian Portuguese with focus on disfluency detection. In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems*, pages 1–10.

Panchbhai, B. S. and Pathak, V. M. (2022). A Systematic Review of Natural Language Processing in Healthcare. *Journal of Algebraic Statistics*, 13(1).

Picard, R. W. (2000). *Affective computing*. MIT press.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese Large Language Models. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Tao, J. and Tan, T. (2005). Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer.

Van Der Veen, A. M. and Bleich, E. (2025). The advantages of lexicon-based sentiment analysis in an age of machine learning. *PloS one*, 20(1):e0313092.

Yosef, S., Zisquit, M., Cohen, B., Klomek, A. B., Bar, K., and Friedman, D. (2025). The impact of fine-tuning LLMs on the quality of automated therapy assessed by digital patients. *npj Mental Health Research*, 4(1):43.