Support for a Chat-Based Intervention to Reduce Alcohol Consumption via Emotion Detection

Luan Henrique da S. Barbosa[1], **Heder S. Bernardino**[1]

[1]Departamento de Ciência da Computação – Instituto de Ciências Exatas (ICE)
Universidade Federal de Juiz de Fora (UFJF)
Caixa Postal 20010 – 36036-900 – Juiz de Fora – MG – Brasil

luan.barbosa@estudante.ufjf.br, heder.bernardino@ufjf.br

Abstract. *Harmful alcohol consumption constitutes a public health problem associated with social and health impacts. Chat-based interventions conducted by psychologists emerge as an alternative to reduce the harm caused by alcohol, and text emotion detection can support professionals with additional information. It is expected that identifying emotions will allow consultants to monitor the patient's emotional state, facilitating more sensitive interventions. Thus, this work evaluates three datasets and five emotion classification models, aiming to provide a basis for the implementation of the service. The TweetEmotions, PlayStoreReviews, and GoEmotions datasets were analyzed, along with Logistic Regression, LinearSVC, Random Forest, XGBoost, and an adjusted BERTimbau-base model, as well as different preprocessing, balancing, and feature engineering strategies. The GoEmotions dataset proved to be the most suitable for the proposed scenario due to its multi-label characteristic. The results indicate consistent performance, with a macro F1-score of 0.40, making it possible to offer professionals an additional resource to support decision-making and contribute to better-targeted care and improved quality of support.*

1. Introduction

Alcohol, when consumed in excess, is linked to significant social and economic impacts, as well as direct harm to health. The term "harmful use of alcohol" is widely used to describe this pattern of consumption and is an aggravating factor in non-communicable diseases, injuries, and premature deaths [OECD 2021]. In 2019, alcohol was responsible for approximately 2.6 million deaths worldwide [World Health Organization 2024]. In Brazil, its use is also associated with the worsening of chronic diseases, such as cancer and depression [Ministério da Saúde 2023]. Despite a slight reduction in global consumption since 2010, there are still few interventions and public policies aimed at reducing the harm caused by alcohol [World Health Organization 2024].

In this context, the Álcool & Saúde project emerges, aiming to assist users in reducing alcohol consumption through digital tools, such as a brief self-intervention via a web-based system[1] and a mobile app[2]. These tools offer supportive features, including dose logs, tips, *gamification*, and control plans. The app Álcool & Saúde (A&S[2]) also provides a chat feature designed for conducting guided interventions led by psychologists. The intervention takes place through the exchange of messages between the user and the

[1]http://www.alcoolesaude.com.br/

[2]https://play.google.com/store/apps/details?id=br.com.alcoolesaude

consultant, consisting of a structured, patient-centered therapy process. Figure 1 shows an example of a conversation in the chat of A&S from the consultant's perspective.

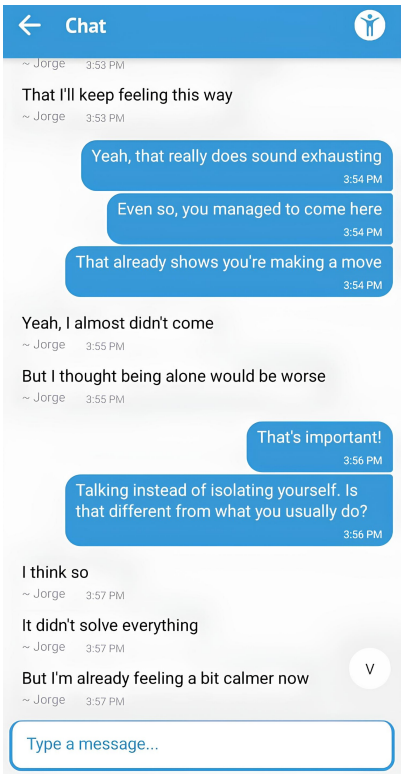

Figure 1. Example of a chat conversation from the consultant's perspective.

Considering that difficulties in emotional regulation are associated with harmful alcohol consumption, especially when it is used as a confrontive coping strategy for negative emotional states [Matei-Mitacu et al. 2024], these factors may interfere with the continuity of the intervention. Thus, this study proposes the inclusion of an emotion detection service as a tool to support decision-making for psychologists in the chat application. This proposal involves the task of classifying emotions in Portuguese texts using machine learning techniques. Therefore, this research presents an initial analysis focused on the pre-training and evaluation of emotion classification models that could be integrated into the chat application of A&S. Pre-training is necessary due to the limited amount of labeled data available within the application's context.

This resource is expected to allow consultants to monitor the user's emotional state in real time, facilitating interventions that are more sensitive to the emotional context. This can contribute to faster response times, better-targeted care, and improved quality of support. The work also allows for the evaluation of the use of emotion detection systems in an applied context with minimal risk to participants. Furthermore, it contributes to the investigation of emotion classification models applied to texts in Portuguese, considering that much of the literature focuses on languages such as English [Duarte 2019, Cortiz et al. 2021, Siqueira et al. 2024].

The remainder of this article is organized as follows. Section 2 describes the related work. Section 3 presents the methodology adopted. Section 4 discusses the results obtained, and Section 5 presents the conclusions and future work.

2. Related Work

Sentiment analysis, often focused on classifying polarity (positive, negative, or neutral) [Cavicchio 2025], is widely used in classification tasks, especially in market contexts, such as app review analysis [Pereira 2021]. In parallel, the development of computational systems applied to health can be observed, as in [Schroeder et al. 2025], which proposes an ontology-based strategy for anxiety management grounded in contextual histories.

However, studies integrating sentiment analysis techniques with applications directly aimed at supporting healthcare are less common. Even rarer are those that address emotion detection, a task that seeks to identify specific emotional states, such as joy, anger, or fear [Cavicchio 2025]. Despite this, some studies approach this proposal by investigating disorders such as anxiety, post-traumatic stress disorder (PTSD), and depression [Batista et al. 2021].

For example, an approach based on large language models integrated with ontologies to support clinical decision-making in the identification of depression is proposed in [Foppa 2025]. For PTSD, neural networks are used to classify magnetic resonance imaging data, achieving an accuracy of 86.25% in [Fernandes et al. 2025]. In the field of sentiment analysis, Twitter texts are analyzed using different language models, achieving approximately an 80% F1-score [Henz et al. 2025].

Regarding emotion detection, studies focused on dataset construction and evaluating models stand out. SVM and BERT are applied for multi-label recognition in texts from platform X (formerly Twitter), with the TweetEmotions dataset being made available in the study proposed in [Mendes et al. 2024]. Similarly, the PlayStoreReviews dataset is presented, composed of reviews extracted from the Google Play Store and annotated for sentiments and emotions, indicating a strong imbalance with a predominance of negative emotions in [Siqueira et al. 2024]. BERTimbau, on the other hand, is evaluated for multi-label emotion classification, reporting an average macro F1-score of approximately 48% and using class-balanced loss reweighting to handle class imbalances [Hammes 2021].

In general, although there are advances in machine learning applied to healthcare and in emotion detection in Portuguese, a gap remains in the integration of these areas in an applied context. Thus, this work differs by investigating emotion detection models in texts, using the datasets presented in [Mendes et al. 2024] and [Siqueira et al. 2024], in addition to a translated version of GoEmotions [Demszky et al. 2020], focusing on the integration of the service with the chat of A&S.

3. Methodology

Figure 2 presents an overview of the methodological process adopted. The process consists of five main steps: (1) selection of datasets, (2) text preprocessing, (3) data splitting, (4) experimental phase, and (5) final evaluation and comparison.

3.1. Datasets

Three Portuguese-language datasets were used, chosen for their availability in the literature and for presenting text characteristics similar to those expected in chat, in addition to distinct structures regarding the distribution of emotions. The TweetEmotions dataset contains 12,419 documents annotated with 16 emotions. Unlike the approach proposed

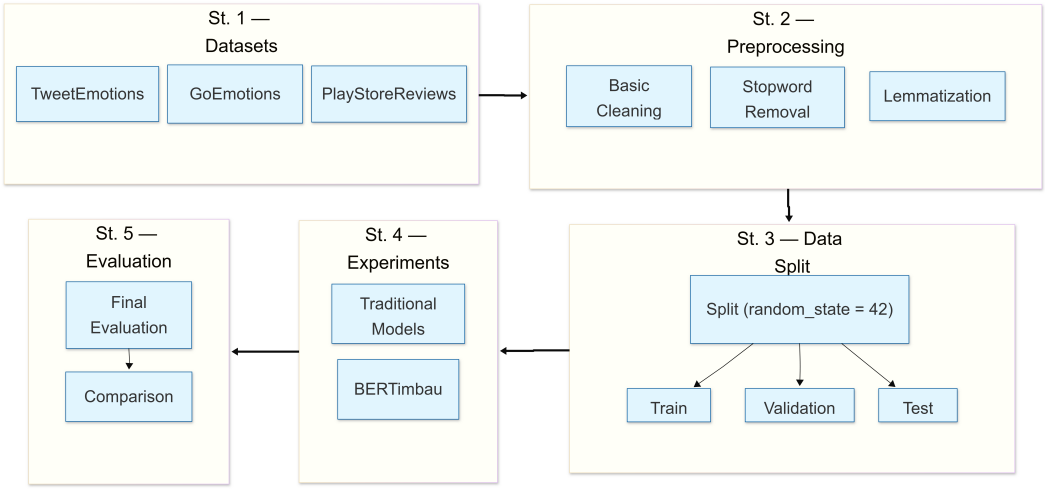

Figure 2. Methodological process.

in [Mendes et al. 2024], the test was performed directly on the dataset itself, and primary and secondary emotions were treated as single labels, characterizing a multi-class task. The PlayStoreReviews dataset contains 3,006 documents with 6 emotions. The dataset shows a strong imbalance, with a predominance of negative emotions. The "surprise" class, with a frequency of less than 0.2% of the data, was removed because its low representation would make splitting into 5 folds in cross-validation unfeasible. Finally, GoEmotions[3] has 54,234 annotated documents in multi-label format, with 28 emotions. The same filter described in [Demszky et al. 2020] and [Hammes 2021] was applied.

3.2. Preprocessing

After removing empty documents from all datasets, a cleaning step was applied, consisting of the removal of HTML tags, URLs, and email addresses; normalization of whitespace and accented characters; removal of tokens containing digits; removal of punctuation characters (excluding hyphens and apostrophes); and conversion to lowercase. In sequence, the texts were tokenized. Tokens corresponding to punctuation, spaces, numeric expressions, or with a length of two characters or less were removed. Thus, five textual representations were generated:

- Original Text: used for extracting additional linguistic attributes;
- Base Text: normalized and token-filtered text without stopwords removal or lemmatization;
- Text Without Stopwords: Base Text without stopwords;
- Lemmatized Text: Base Text with lemmatization;
- Lemmatized Text Without Stopwords: a combination of stopwords removal and lemmatization.

Initially, experiments were conducted with the Base Text. Subsequently, the results were compared with other versions to verify the impact of preprocessing on performance.

[3]https://www.kaggle.com/datasets/antoniomenezes/go-emotions-ptbr

3.3. Data division

Each dataset was divided into training (70%), validation (15%), and test (15%), using a fixed random state of 42 for reproducibility. The training and validation sets were used for experimental decision-making (preprocessing selection, balancing, representations, and inclusion of linguistic attributes) and for adjusting the BERTimbau model. The test set was reserved exclusively for final evaluation and comparison between models.

3.4. Experimental configuration

The experimental phase was organized into two flows, as illustrated in Figure 3. For the first, the following classifiers were evaluated: Logistic Regression, Linear SVC, Random Forest, and XGBoost. These models were chosen because they have demonstrated good performance in text classification tasks in the literature, with low computational cost when compared to deep neural networks and language models. Regarding text representation, the Bag-of-Words (BoW) and TF-IDF vectorization techniques were compared.

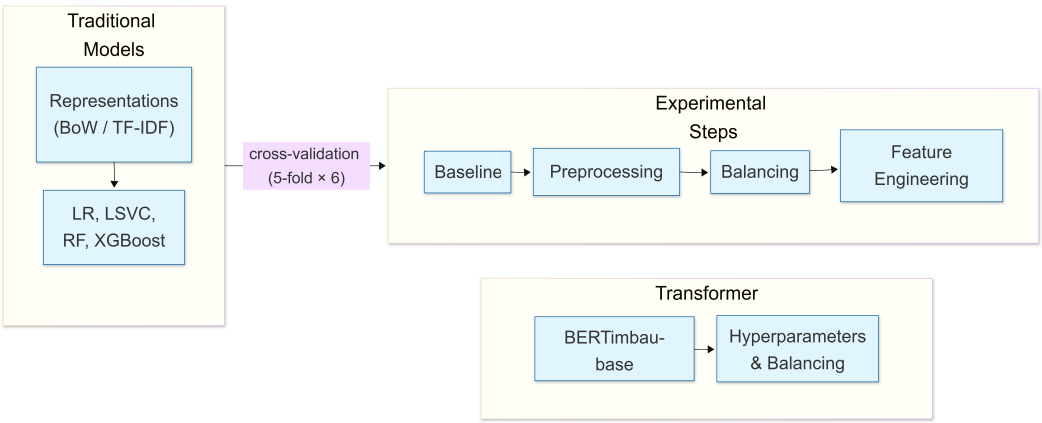

Figure 3. Experimental flow.

All experiments with traditional models were conducted using 5-fold cross-validation repeated 6 times, totaling 30 runs. The results were reported as mean ± standard deviation. Cross-validation was applied only to the training set and was used to compare preprocessing versions, evaluate balancing impact, test inclusion of linguistic attributes, and compare BoW and TF-IDF.

Regarding balancing, for the multi-class datasets (TweetEmotions and PlayStoreReviews), the SMOTE technique was applied to the training data. In the case of GoEmotions, since it is a multi-label problem, balancing was performed by treating each emotion as an independent binary problem, a common strategy in multi-label classification based on binary decomposition (Binary Relevance) [Zhang et al. 2018]. For each emotion, a binary classifier was trained with SMOTE applied individually. The predictions were then aggregated to reconstruct the multi-label matrix and calculate the global metrics.

In addition to the textual representations, 14 linguistic attributes were extracted from the Original Text, including number of words, sentences, and syllables; average words per sentence; estimated reading time; average number of verbs and auxiliary verbs per sentence; amount of punctuation; proportion of capital letters; character repetition;

and Flesch Index. In particular, the Flesch Index for Portuguese was calculated as[4]:

$$Flesch = 226 - 1,04 \cdot \frac{Number\,of\,words}{Number\,of\,sentences} - 72 \cdot \frac{Number\,of\,syllables}{Number\,of\,words}$$

These attributes were concatenated to the text vectors to evaluate their impact on performance. If they did not show a consistent gain in cross-validation, they were discarded in the final configuration, as were the preprocessing and balancing techniques. At the end of the cross-validation, the best combination involving model, representation, preprocessing version, use or not of SMOTE, and use or not of additional attributes was selected for each dataset. This combination was then retrained using the entire training set and subsequently evaluated on the test set.

During the experiments, accuracy, precision, recall, and F1-score metrics were computed in their micro, macro, and weighted variants. Although all metrics were analyzed, model comparisons primarily relied on the macro F1-score, as it better reflects performance across all classes equally, mitigating the influence of label imbalance. This is particularly relevant for the GoEmotions dataset, which exhibits significant class imbalance within its multi-label setting. Furthermore, to assess the consistency of the observed improvements across different configurations, the Wilcoxon[5] statistical test for paired samples was applied, considering the results obtained in the repetitions of the cross-validation. This procedure allowed us to verify whether the observed differences in the metrics were statistically significant ($p < 0.05$), even when numerically small.

In addition to traditional models, the pre-trained BERTimbau-base model, specific to Portuguese, was evaluated. The model was fitted using the training set, with monitoring on the validation set. Unlike traditional models, BERTimbau did not receive linguistic attributes and did not use SMOTE. Instead, class imbalance was addressed using Class Balanced Loss as per [Hammes 2021], which assigns weights to each class based on the effective number of samples. The weighting scheme was computed using $\beta = 0.999$. These weights were incorporated into a binary cross-entropy loss function. After fitting, the final model was evaluated exclusively on the test set, without cross-validation.

3.5. Final evaluation and comparison

The comparison between models was performed exclusively on the test set, ensuring impartiality in the evaluation. The following metrics were reported: Accuracy, Precision (macro), Recall (macro), and macro F1-score. In the case of GoEmotions, the metrics were calculated considering the aggregation of binary predictions for reconstruction of the multi-label scenario. Thus, the experimental process establishes two distinct methodological flows: traditional models evaluated with cross-validation for experimental decision analysis and BERTimbau with adjustment via a validation set. Both flows were compared in the final evaluation using the same test set.

4. Results and Discussion

About implementation, preprocessing was conducted using the unidecode and spaCy libraries with the large Portuguese model. Traditional models were employed with their

[4]https://legibilidade.com/

[5]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html

default configurations from scikit-learn and XGBoost. The SMOTE technique was also applied using default parameters. For the fine-tuning of BERTimbau, the neuralmind/bert-base-portuguese-cased model was adopted with a learning rate of 2e-5, batch size of 16, four epochs, and a maximum sequence length of 128. Training employed a linear scheduler with 20% warmup, no weight decay, and FP16 precision. Model selection was based on the macro F1-score. The source codes and the results are publicly available[6]. Also, the results of the cross-validation for each dataset are presented in Table 1, which summarizes the best combination of traditional model, textual representation, and use of additional attributes.

Table 1. Best combination and its results for traditional models.

Dataset	TweetEmotions	PlayStoreReviews	GoEmotions
Model	XGBoost	Logistic Regression	Logistic Regression
Representation	BoW	TF-IDF	TF-IDF
Stopwords or Lemma.?	No	Lemmatization	No
Additional Attributes?	No	Yes	Yes
Macro F1-score	0.9631 ± 0.0044	0.479 ± 0.0291	0.3996 ± 0.0064

4.1. TweetEmotions

The TweetEmotions dataset showed the best overall performance. The best configuration achieved an macro F1 of 0.9631 ± 0.0044, indicating high classification performance and low variability. It was observed that, in this dataset, the BoW representation slightly outperformed TF-IDF when balancing was added. While with TF-IDF the XGBoost achieved an macro F1 of 0.9594, with BoW the value was 0.9631. The removal of stopwords and lemmatization caused small reductions in absolute terms (around 1% in the macro F1), but these differences were statistically significant according to the Wilcoxon test. That is, the numerical impact was small but consistent across the repetitions of the cross-validation.

Balancing produced variations of less than 0.02% in the macro F1, confirming a small impact, given that the dataset was already balanced. The inclusion of linguistic attributes also slightly reduced the macro F1 (reductions close to 0.12%), again with a statistically significant difference. Even so, attributes such as average number of syllables per word and the Flesch Index were among the most relevant in the training, indicating that, although they did not improve aggregate performance, they capture linguistic signals associated with emotions. It should be noted that the high values obtained in prediction may be associated with the construction of the dataset itself, which used explicit synonyms of emotions as a collection criterion [Mendes et al. 2024]. Such a procedure may encourage the learning of lexical patterns and increase lexical overlap between training and testing splits, potentially leading to overly optimistic evaluation results.

4.2. PlayStoreReviews

In the PlayStoreReviews dataset, the best macro F1-score was 0.479, a value significantly lower than that observed in TweetEmotions. The greater variability (standard deviation

[6]https://github.com/LuanBarbs/tcc-emotion-detection-nlp/tree/sbcas-article

of 0.0291) also indicates lower predictive stability. Unlike what was observed for TweetEmotions, lemmatization showed consistent improvement, with average gains of 1.76% (BoW) and 0.94% (TF-IDF) in the macro F1. The removal of stopwords, however, did not produce relevant variations (differences less than 0.5%). Initially, BoW (0.4435 ± 0.0231) showed superior performance to TF-IDF (0.4267 ± 0.0309), however, after balancing, TF-IDF began to show better performance, reaching 0.4671 ± 0.026, in contrast to 0.4485 ± 0.027 with BoW.

The inclusion of linguistic attributes contributed positively, with an approximate 1.2% increase in the macro F1 for the TF-IDF. Unlike the previous dataset, here the complementary attributes seem to partially compensate for the limited data volume. However, despite the improvements with balancing and linguistic attributes, the highest accuracy observed in the cross-validation was 0.5983 ± 0.0136, indicating limitations for direct application in A&S, especially considering that emotion-based decisions can influence intervention strategies. As the dataset was previously preprocessed in [Siqueira et al. 2024], attributes related to punctuation and capitalization were not considered. Even so, reading time, number of words per sentence, number of letters, and Flesch Index remained relevant among the most important variables.

4.3. GoEmotions

In the GoEmotions dataset, the multi-label nature of emotions and the use of translated texts make the task more complex. A drop in the macro F1-score from approximately 0.36 (English) to 0.33 (Portuguese) was observed with BoW, highlighting the negative impact of translation, an aspect already discussed in [Hammes 2021]. TF-IDF obtained lower initial results, achieving 0.34 (English) and 0.30 (Portuguese) for the macro F1. In the final configuration, the best result was 0.3996 ± 0.0064 with TF-IDF and balancing. Lemmatization and stopwords removal reduced the macro F1-score by more than 4.5%.

Balancing was especially relevant for TF-IDF, increasing the average macro F1-score by 4.26%. The inclusion of linguistic attributes did not substantially raise the macro F1 for BoW, but it did raise the TF-IDF results from 0.3823 ± 0.0058 to 0.3996 ± 0.0064 when using Logistic Regression. Comparatively, [Hammes 2021] reported an macro F1 of 0.48 with BERTimbau-based, while [Demszky et al. 2020] found 0.40 ± 0.18 in cross-validation. Although the average value obtained in this work is lower, the standard deviation is significantly smaller, suggesting greater consistency between partitions.

4.4. Evaluation

Table 2 presents the final comparison between the traditional models and BERTimbau. With TweetEmotions, both models showed high performance, with BERTimbau achieving a slight advantage across all metrics (a difference of 0.0054 in the macro F1-score). For PlayStoreReviews, BERTimbau with CB Loss (0.6053) showed a modest gain in accuracy compared to Logistic Regression (0.5965), along with a slight improvement in macro F1-score (0.4658 vs. 0.4532), indicating a more balanced performance across classes.

With GoEmotions, BERTimbau outperformed the traditional model in both accuracy and macro F1-score. Considering that this is a complex multi-label task, values close to 0.40 indicate competitive performance for application in the chat context. The ability to

identify multiple emotions simultaneously may be particularly relevant for mental health support systems, where coexisting emotions can provide additional information about the patient's emotional state.

In general, among the traditional models, XGBoost and Logistic Regression showed the best performance in their respective datasets. BERTimbau was superior in all evaluations in the test set, although with a marginal difference in TweetEmotions and a more significant difference in GoEmotions. In PlayStoreReviews, a modest gain was observed in both accuracy and macro F1-score. It should be noted that the traditional models did not undergo hyperparameter optimization and vectorization, which may have restricted their comparative performance.

Furthermore, in the GoEmotions dataset, complementary behavior was observed between the models: Logistic Regression showed greater sensitivity (0.5527), while BERTimbau showed greater precision (0.5430), indicating possible future strategies for combining models. Finally, although the PlayStoreReviews dataset presents limitations evidenced by the macro F1-score and the accuracy in a multi-class scenario, its results do not rule out its use as an auxiliary evaluation component in the emotional detection system.

Table 2. Comparison of traditional models and BERTimbau in the test set.

Dataset	Accuracy	Precision	Recall	F1	Model
TweetEmotions	0.9581	0.9590	0.9583	0.9586	XGBoost
	0.9635	**0.9644**	**0.9639**	**0.9640**	BERTimbau (CB)
PlayStoreReviews	0.5965	0.4510	0.4570	0.4532	Logistic Regression
	0.6053	**0.4594**	**0.4936**	**0.4658**	BERTimbau (CB)
GoEmotions	0.1874	0.3096	**0.5527**	0.3881	Logistic Regression
	0.3929	**0.5430**	0.3548	**0.4013**	BERTimbau (CB)

Moreover, Table 3 shows the F1-score obtained per emotion across all datasets for the best model among the traditional ones and BERTimbau. For TweetEmotions, the per-class results remain consistently high across all emotions, which is aligned with the strong overall performance observed in the aggregate metrics. For PlayStoreReviews, most classes follow a pattern similar to that observed for the macro F1-score, with relatively modest performance across emotions. However, exceptions are observed for fear and neutral, which present particularly low F1-scores. This behavior is likely associated with class imbalance and the difficulty in distinguishing these categories, especially considering that neutral texts may still contain implicit emotional cues.

Regarding GoEmotions, more pronounced differences between emotions can be observed. Although an overall macro F1-score of 0.40 is limited for practical healthcare applications, the models show more consistent performance for emotions such as gratitude, amusement, love, admiration, remorse, and neutral, possibly due to their higher frequency or more well-defined linguistic patterns. On the other hand, there are evident difficulties in predicting less frequent emotions, such as realization, disappointment, and relief, which may be associated with dataset imbalance and greater semantic ambiguity.

Table 3. F1-score per emotion, where LR refers to Logistic Regression, XGB to XGBoost, and BT to BERTimbau. Dashes (–) indicate emotions not present in the corresponding datasets.

Emotion	TweetEmot.		PlayStoreR.		GoEmotions	
	XGB	BT	LR	BT	LR	BT
aggressiveness	0.93	0.92	–	–	–	–
anger	0.98	0.98	0.61	0.62	0.35	0.41
anticipation	0.99	**1.00**	–	–	–	–
confidence	0.94	0.95	–	–	–	–
contempt	0.99	0.99	–	–	–	–
deception	0.94	0.97	–	–	–	–
disgust	0.99	0.97	0.58	0.59	0.22	0.32
fear	0.99	0.99	**0.00**	**0.00**	0.53	0.54
intimidation	0.99	**1.00**	–	–	–	–
joy	0.94	0.94	0.64	**0.71**	0.37	0.45
love	0.95	0.95	–	–	0.82	0.78
optimism	0.98	0.99	–	–	0.45	0.55
remorse	0.99	0.98	–	–	0.70	0.61
sadness	**0.87**	0.89	0.65	0.63	0.39	0.47
submission	0.97	0.97	–	–	–	–
surprise	0.94	0.94	–	–	0.50	0.55
admiration	–	–	–	–	0.60	0.66
amusement	–	–	–	–	0.81	0.78
annoyance	–	–	–	–	0.04	0.08
approval	–	–	–	–	0.16	0.25
caring	–	–	–	–	0.31	0.40
confusion	–	–	–	–	0.23	0.33
curiosity	–	–	–	–	0.23	0.30
desire	–	–	–	–	0.44	0.48
disappointment	–	–	–	–	0.02	0.04
disapproval	–	–	–	–	0.15	0.23
embarrassment	–	–	–	–	0.23	0.34
excitement	–	–	–	–	0.11	0.19
gratitude	–	–	–	–	0.87	**0.89**
grief	–	–	–	–	0.10	0.10
neutral	–	–	0.24	0.24	0.49	0.57
nervousness	–	–	–	–	0.39	0.35
pride	–	–	–	–	0.38	0.48
relief	–	–	–	–	0.04	0.06
realization	–	–	–	–	0.00	**0.00**

5. Conclusions and Future Work

Harmful alcohol consumption is associated with significant health impacts, including the worsening of chronic diseases and premature death. In this context, Álcool & Saúde proposes digital tools to support consumption reduction, including psychologist-led chat interventions. Considering that difficulties in emotional regulation can interfere with adherence to and continuity of these interventions, this work investigated the feasibility of

integrating an emotion detection service to support consultants' decision-making. To this end, a comparative evaluation of emotion classification models in Portuguese was carried out, aiming to support their future incorporation into the A&S chat.

Among the analyzed datasets, GoEmotions proved to be the most suitable, mainly due to its multi-label classification capability, which is more aligned with the real context of interactions. The results indicated good performance of the Logistic Regression, XGBoost, and BERTimbau models across the different datasets. In general, preprocessing techniques such as lemmatization and stopwords removal did not prove advantageous, while SMOTE balancing consistently improved results. The inclusion of linguistic attributes had a variable impact. Although their relative contribution was smaller when combined with vectorized text, they frequently appeared among the most relevant variables in Random Forest importance analysis, indicating complementary potential.

It is important to highlight that this study represents an initial step toward the development of emotion-aware support tools, focusing on the pre-training and comparative evaluation using publicly available datasets. The results—such as the macro F1-score of 0.40 for GoEmotions—should be interpreted with caution, especially in the healthcare context. The datasets do not directly represent real therapeutic chats, which may introduce domain bias and limit external validity.

Future work focuses on advancing this stage toward real-world applicability, including the annotation of emotional data from actual chat interactions to enable domain adaptation via transfer learning. Strategies such as emotion grouping may also be explored to reduce annotation complexity and improve robustness. Further investigations consider alternative model architectures, integration of similarity-based features with emotionally referenced texts (golden texts), and strategies for integrating these attributes into BERTimbau-based models. The exploration of large language models is also a promising direction. Finally, implementing the service and evaluating its usability and perceived usefulness are essential to validate its impact in real clinical workflows.

6. Acknowledgments

The authors thank the support provided by CAPES, CNPq (grant 313452/2025-3), FAPEMIG (grants APQ-03313-22 and APQ-01832-22), PPGCC, PPGMC, and UFJF.

References

Batista, H. M. C. d., Paim, A. B., Siqueira, B. S., Ebecken, N. F. F., and Dias, A. C. (2021). Factors that can trigger depression. *P2P INOV.*, 7(2):164–185.

Cavicchio, F. (2025). *Emotion Detection in Natural Language Processing*. Springer Nature Switzerland.

Cortiz, D., Silva, J., Calegari, N., Freitas, A., Soares, A., Botelho, C., Rêgo, G., Sampaio, W., and Boggio, P. (2021). A weakly supervised dataset of fine-grained emotions in portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 73–81, Porto Alegre, RS, Brasil. SBC.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In Jurafsky, D., Chai, J., Schluter,

N., and Tetreault, J., editors, *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics.

Duarte, L. C. F. (2019). Reconhecimento automático de emoções em texto com recurso a emojis. Doutorado, Universidade de Coimbra, Portugal.

Fernandes, R., Carvalho, R., Junior, O. F., Portugal, L., and Ramos, T. (2025). Um classificador explicável para transtorno de estresse pós-traumático utilizando redes neurais convolucionais tridimensionais. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 236–247, Porto Alegre, RS, Brasil. SBC.

Foppa, Alexandre e Barbosa, J. (2025). Um modelo computacional para análise de depressão em dados de redes sociais. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 92–103, Porto Alegre, RS, Brasil. SBC.

Hammes, Luiz e Freitas, L. (2021). Utilizando bertimbau para a classificação de emoções em português. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 56–63, Porto Alegre, RS, Brasil. SBC.

Henz, M., Heckler, W., and Barbosa, J. (2025). Uma avaliação da capacidade de modelos de linguagem para análise de sentimentos em um contexto de saúde mental. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 293–304, Porto Alegre, RS, Brasil. SBC.

Matei-Mitacu, L.-M., Huțul, T.-D., Karner-Huțuleac, A., Huțul, A., and Dobria, C.-A. (2024). The role of alcohol consumption motives in the relationships between psychological distress, emotional dysregulation, and problematic alcohol consumption. a mediation model. *Current Psychology*, 43(48):36831–36845.

Mendes, R., Tavares, S., Campos, L., and Araújo, F. (2024). Mineração de emoções multirrótulo em textos curtos. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 445–450, Porto Alegre, RS, Brasil. SBC.

Ministério da Saúde (2023). Vigitel brasil 2006-2023: tabagismo e consumo abusivo de álcool.

OECD (2021). *Preventing Harmful Alcohol Use*. OECD Publishing, Paris.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artif. Intell. Rev.*, 54(2):1087–1115.

Schroeder, G., Paula, L., Francisco, R., and Barbosa, J. (2025). A-track: An ontological approach to assisting anxiety management. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 44–55, Porto Alegre, RS, Brasil. SBC.

Siqueira, V., Costa, R. H., Soares, T., Lunardi, G., and Silva, W. (2024). Dataset anotado de sentimentos a partir de comentários de aplicativos móveis. In *Anais do VI Dataset Showcase Workshop*, pages 65–76, Porto Alegre, RS, Brasil. SBC.

World Health Organization (2024). Global status report on alcohol and health and treatment of substance use disorders.

Zhang, M.-L., Li, Y.-K., Liu, X.-Y., and Geng, X. (2018). Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202.