AKCIT at SemEval-2025 Task 11: Investigating Data Quality in Portuguese Emotion Recognition

Iago A. Brito, Fernanda B. Färber, Julia S. Dollis, Daniel M. Pedrozo, Artur M. A. Novais, Diogo F. C. Silva, Arlindo R. Galvão Filho
Advanced Knowledge Center for Immersive Technologies (AKCIT)
Federal University of Goiás (UFG)
Correspondence: iagoalves@discente.ufg.br

Abstract

This paper investigates the impact of data quality and processing strategies on emotion recognition in Brazilian Portuguese (PTBR) texts. We focus on data distribution, linguistic context, and augmentation techniques such as translation and synthetic data generation. To evaluate these aspects, we conduct experiments on the PTBR portion of the BRIGHTER dataset, a manually curated multilingual dataset containing nearly 100,000 samples, of which 4,552 are in PTBR. Our study encompasses both multi-label emotion detection (presence/absence classification) and emotion intensity prediction (0 to 3 scale), following the SemEval 2025 Track 11 setup. Results demonstrate that emotion intensity labels enhance model performance after discretization, and that smaller multilingual models can outperform larger ones in low-resource settings. Our official submission ranked 6th, but further refinements improved our ranking to 3rd, trailing the top submission by only 0.047, reinforcing the significance of a data-centric approach in emotion recognition.

1 Introduction

Data quality plays a critical role in enabling machine learning models to generalize effectively and generate meaningful predictions (Budach et al., 2022). On the other hand, poor data quality (e.g., a high level of noise, inconsistencies, imbalanced distributions, or annotation errors) can lead to biased models and unreliable outcomes (Sambasivan et al., 2021). Several studies have shown that well-constructed datasets significantly enhance model performance in NLP tasks (Mishra et al., 2020; Longpre et al., 2024). In the context of emotion recognition, high-quality data is essential for capturing subtle emotional nuances and reflecting variations across different contexts and linguistic structures.

In this paper, we investigate the impact of data quality on emotion recognition in Brazilian Portuguese (PTBR) texts, focusing on data distribution, linguistic context (e.g., word-sentiment lexicons), and augmentation strategies such as translation and synthetic data generation. To evaluate these aspects, we conduct experiments on the PTBR portion of BRIGHTER dataset (Muhammad et al., 2025a), a multilingual large-scale dataset manually curated comprissing nearly 100,000 samples, which 4,552 are in PTBR. The dataset originates from SemEval 2025 Track 11 (Muhammad et al., 2025b), and we participate in both sub-track A (Multi-label Emotion Detection; classifying either if the emotion is or is not present in the sentence) and sub-track B (Emotion Intensity; classifying the intensity from 0 to 3 of the emotion in the sentence) sub-tracks. Our findings highlight the importance of a data-centric approach, measuring the impact of different data processes alternatives, as the official submission achieved 6th in Portuguese emotion recognition. However, further analysis and post-competition refinements established a new 3rd place ranking, trailing the top-ranked submission by only 0.047.

Our main contributions can be summarized as follows:

- **Data-Centric Analysis:** We demonstrate the impact of data-centric techniques on model performance, highlighting the role of data distribution, augmentation, and linguistic context.
- **Style-Domain Translation:** We introduce a translation approach that preserves the emotional content of the original text while adapting it to the target style and domain using few-shot learning, achieving 3% improvement in Macro F1 compared to traditional literal translation, underscoring its effectiveness in enhancing emotion recognition performance.
- **Label Granularity Effect:** We show that models trained with Emotion Intensity labels

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2305–2310

outperform those trained with binary labels in multi-label emotion classification. This result suggests that modeling emotion intensity improves generalization.

The rest of this paper is organized as follows: Section 2, reviews the related work. Section 3 describes the methods and Section 4 the Experimental Setup, while experiments results are discussed in Section 5. Section 6 brings the conclusions.

2 Related Work

Sentiment analysis has been widely studied across various domains, including movie (Bodapati et al., 2019) and product reviews (Reddy et al., 2024), with social media platforms also receiving significant attention (Singh et al., 2021) due to the vast amount of user-generated content. Baziotis et al. (2017) propose a bidirectional LSTMs with attention mechanisms to predict sentiment polarity, categorizing tweets as positive, negative, or neutral. Their model achieved first place in SemEval-2017 Task 4. However, since it follows a single-label classification approach, it cannot effectively capture the overlapping and co-occurring emotions often present in social media text.

da Silva et al. (2018) introduced a corpus of tweets from Brazilian investors, annotated with emotional labels. The dataset was constructed by collecting tweets that referenced stocks from the Brazilian stock exchange (IBOVESPA) and manually labeling them according to Plutchik's model, which categorizes emotions into eight distinct types. Their work underscores the importance of high-quality annotated datasets for training machine learning models in Portuguese, a task that remains challenging due to the limited availability of labeled data.

To address the challenges of multilabel emotion classification, (Kim et al., 2018) proposed an attention-based convolutional neural network (CNN) model capable of handling multiple emotion labels per instance. Their system integrates self-attention mechanisms to enhance sentence representation, allowing emotions to be classified independently within a single sentence. Evaluated on SemEval-2018 Task 1, their approach ranked first in Spanish and fifth in English, demonstrating its effectiveness across languages.

3 Methods

3.1 Data

The Brazilian Portuguese subset of the BRIGHTER dataset (Muhammad et al., 2025a) comprises 4,652 social media posts annotated with six fundamental emotions: anger, disgust, fear, joy, sadness, and surprise. The dataset follows a multi-label classification setup, where each instance can express zero, one, or multiple emotions. The data is split into 2,226 training instances, 200 validation instances, and 2,226 test instances. Each post was annotated by five independent raters, generating probabilistic emotion distributions rather than discrete labels. Table 1 provides an overview of the label distribution within the training set.

Emotion	Number of samples	Percentage
Anger	718	32.26%
Disgust	75	3.37%
Fear	109	4.90%
Joy	581	26.10%
Sadness	322	14.47%
Surprise	153	6.87%
No emotion	632	28.39%

Table 1: Number of samples per emotion in the PTBR portion of the dataset.

3.2 Style-Domain Text Translation

To enhance model robustness and generalization, we expanded the dataset through cross-lingual translations. However, these translations introduce domain and stylistic discrepancies. For instance, Brazilian Portuguese data primarily originates from social media, while Algerian Arabic and Mozambican Portuguese include content from literature and news, leading to factual information in Mozambican Portuguese that is absent in Brazilian Portuguese. Even within social media, stylistic variations exist: Brazilian Portuguese posts rarely use emoticons, whereas Latin American Spanish contains emoticons in nearly 22% of samples.

To mitigate these inconsistencies, we introduce Style-Domain Text Translation, a methodology designed to simulate Brazilian Portuguese social media discourse. We guide the LLM-based translation process to preserve both the literal meaning and original emotional content, while incorporating colloquial expressions, abbreviations, and informal linguistic patterns typical of Brazilian social media. The translation model is additionally regularized to avoid excessive formality. Although this approach

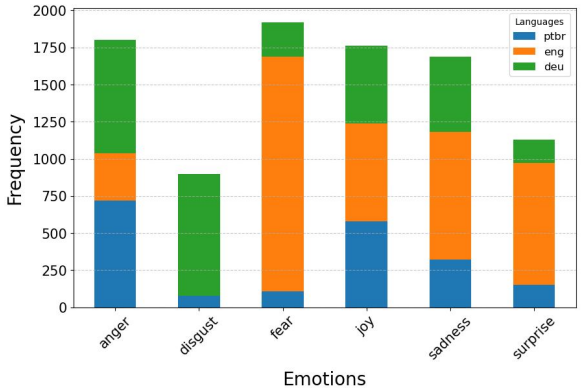

Figure 1: Proportion of each emotion in the dataset by adding English and Deutsch translation

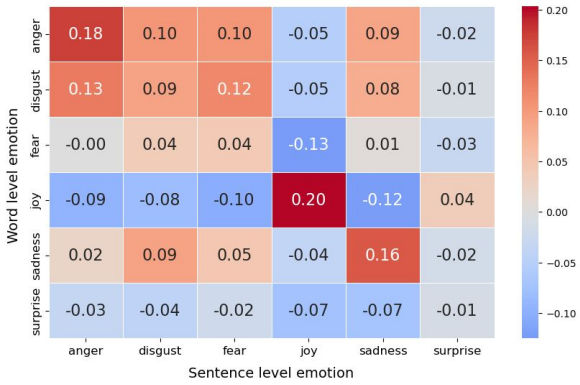

Figure 2: Pearson correlation between the emotion scores of individual words and the overall sentence emotion.

was applied across all languages in the dataset, our experiments primarily focus on English and German, allowing us to perform extensive evaluations with low-cost computational resources.

3.3 Synthetic data generation

LLM-based synthetic data augmentation has been widely explored across multiple domains, including social media text generation (Hosseini et al., 2024). However, existing studies predominantly focus on English, which dominates the pre-training and fine-tuning corpora of large-scale models. Our approach addresses this linguistic gap by generating synthetic informal social media posts in Brazilian Portuguese, incorporating realistic variability in emotion, text length, and sentiment intensity.

To ensure diversity in the synthetic samples, we adopt a few-shot prompting strategy that conditions the generation process on multiple aspects of the data. Emotions are dynamically sampled based on the original dataset, allowing for upsampling of underrepresented emotions while preserving the dataset's multi-label structure. Additionally, the generated texts mirror the natural length distribution of real social media posts, ensuring structural authenticity. By varying sentiment intensity levels within prompts, we enable the model to produce content with fine-grained emotional variation, improving alignment with real-world language use.

For text generation, we employ the GPT-4o mini model (Achiam et al., 2023), chosen for its cost efficiency and strong performance in instruction-following tasks (Kim et al., 2024). This approach allows us to conduct extensive experiments while maintaining a computationally efficient training pipeline.

3.4 Data Balance

Our analysis of the dataset revealed a class imbalance in the Portuguese training set, where disgust and fear were underrepresented. To address this issue, we improved emotion category distribution by augmenting the dataset through cross-lingual translation from English and German into Portuguese. This approach equalized class proportions across the six emotion categories, as shown in Figure 1, while preserving linguistic diversity in the samples.

While alternative strategies such as data duplication and LLM-based synthetic data generation were considered, translation provided the most effective solution for expanding low-frequency classes while maintaining natural language variability. As a result, our approach led to a more uniform label distribution and improved model performance across all emotion categories.

3.5 Word-based Emotion Lexicon

We also utilized the Portuguese version of the NRC Emotion Lexicon (Mohammad and Turney, 2010, 2013), which comprises a comprehensive set of words annotated with their associated emotions. Although the sentiment of an individual word does not necessarily imply that the entire sentence conveys that emotion, we examined the correlation between the sentiment of each word and the overall sentence emotion, which is shown in Figure 2. Furthermore, we incorporated this lexicon information into the model input and measured the impact of this information when training an LLM to identify emotions.

4 Experimental Setup

Implementation. All experiments were conducted using a single RTX 4090 GPU with a batch size of 8, a linear learning rate schedule of 2e-5, and a weight decay of 0.01. After identifying the best-performing strategy, we scaled up the approach and applied it to a larger model running on a single NVIDIA A100 GPU. All experiments were performed with 4-bit quantization and LoRA (Hu et al., 2022).

Model. We initially evaluated the Qwen 2.5 7B (Yang et al., 2024), LLaMA 3.1 8B (Dubey et al., 2024), and Phi-4 14B (Abdin et al., 2024) instruct models under few-shot and fine-tuning scenarios. Based on its superior performance in Brazilian Portuguese, we selected Qwen 2.5 7B as the base model for further experiments. The final tests were conducted using the Qwen 2.5 70B Instruct model (Yang et al., 2024), which shares pre-training and fine-tuning procedures similar to those of the selected base model.

5 Results

5.1 Model Selection

To identify the most suitable model for Brazilian Portuguese emotion recognition, we evaluated multiple architectures under few-shot and fine-tuning settings. Table 2 presents the performance results, highlighting the potential of the Qwen 2.5 architecture. Although its performance is slightly lower than that of the Phi model, Qwen 2.5 is half the size, offers scalable larger versions, and achieves better results than LLaMA 3.1 8B, making it a compelling choice for our study.

	Model	Macro F1	Micro F1
Few-shot	LLaMA 3.1 8B	0.35	0.44
	Phi 4 14B	**0.51**	**0.57**
	Qwen 2.5 7B	0.44	0.52
Fine-tuning	LLaMA 3.1 8B	0.46	0.66
	Phi 4 14B	**0.54**	**0.72**
	Qwen 2.5 7B	0.53	0.71

Table 2: Comparing base models in few-shot (4 shots) and fine-tuning on Brazilian Portuguese dataset portion.

5.2 Data processing

We examined both style-domain text translation and synthetic data generation. Table 3 shows that style-domain translation improved the macro-F1 score from 0.53 to 0.56, whereas traditional translation had no measurable impact. Although synthetic data generation increased the number of samples across all six emotions, it did not improve classification performance. This result suggests that while LLMs effectively translate existing samples into the target style and domain, they struggle to generate sufficiently diverse new samples from few-shot prompts, ultimately leading to a decline in overall performance.

Additionally, we investigated the correlation between NRC word-level emotions and sentence-level emotions. Our analysis revealed a low overall correlation, indicating that word-level sentiment features are not strong predictors of sentence-level emotion. For instance, the correlation between word-level and sentence-level "disgust" is lower than the correlation between "disgust" and "anger" (ranging from 0.09 to 0.13), as shown in Figure 2. The highest observed correlation was 0.20 for "joy", reinforcing that word-level sentiment signals provide limited value for sentence-level emotion prediction.

Data	Macro F1	Micro F1
Original baseline	0.53	0.71
NRC Lexicon	0.53	**0.73**
Traditional translation	0.53	0.64
Style-Domain Translation	**0.56**	0.67
Synthetic	0.49	0.68

Table 3: Performance impact of various data processing approaches on the original dataset. All methods were applied in combination with the original data, and results are reported in terms of Macro and Micro F1 scores.

5.3 Final Results

Following an extensive data analysis, we fine-tuned the Qwen 2.5 70B model (Yang et al., 2024) using the optimal hyperparameter configuration and data augmentation strategy on sub-tracks A and B. The final training dataset combined the original training set with English and German style-domain translations.

In our official submission, we ranked 6th place in sub-track A. However, further analysis revealed that using the model trained on sub-track B, which incorporated emotion intensity labels, and binariz-

ing the outputs (considering an emotion present if any intensity was detected), improved results by 3 percentage points, increasing the macro-F1 score from 0.61 to 0.64. When evaluated in the emotion intensity prediction task (sub-track B), the model achieved an average Pearson correlation (r) of 0.65.

6 Conclusion

Our study investigates how data quality and processing techniques influence emotion recognition in Brazilian Portuguese texts. We explored Style-Domain Text Translation, synthetic data generation, and the integration of a word-based emotion lexicon. Our findings show that cross-lingual translation improves the balance of low-frequency emotion classes while preserving linguistic diversity, leading to better model generalization.

Additionally, our experiments demonstrate that training with emotion intensity labels, rather than binary labels, enhances performance when the outputs are binarized. Model selection results suggest that smaller models can sometimes outperform larger ones in this task, emphasizing the importance of architecture choice in low-resource settings. Finally, our post-competition refinements led to significant performance gains, reinforcing the role of fine-grained data processing strategies in improving emotion recognition models.

Limitations

Despite the promising results, our work has several limitations. First, our experiments are constrained by the size and diversity of the available Brazilian Portuguese data, which may not capture all linguistic nuances. Second, while the style-domain translation methodology shows potential, it relies on few-shot learning and may require further validation across different domains and larger datasets. Third, the observed performance improvements when using intensity labels indicate potential sensitivity to label binarization methods, suggesting that additional samples with higher emotions intensities could beneficiate the model. Finally, our analysis primarily focused only in PTBR portion of BRIGHTER dataset, and results might vary when applied to other domains or languages. Future work should address these limitations by incorporating more extensive and diverse datasets and refining our data processing techniques.

Acknowledgments

This work has been fully funded by the project Research and Development of Algorithms for Construction of Digital Human Technological Components supported by the Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI grant number 057/2023, signed with EMBRAPII.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *International Workshop on Semantic Evaluation*.

Jyostna Bodapati, N. Veeranjaneyulu, and Nagur Shareef Shaik. 2019. Sentiment analysis from movie reviews using lstms. *Ingénierie des systèmes d information*, 24:125–129.

Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. 2022. The effects of data quality on machine learning performance. *arXiv preprint arXiv:2207.14529*.

Fernando José Vieira da Silva, Norton Trevisan Roman, and Ariadne Carvalho. 2018. Building an emotionally annotated corpus of investor tweets.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mohammad Javad Hosseini, Andrey Petrov, Alex Fabrikant, and Annie Louis. 2024. A synthetic data approach for domain generalization of NLI models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2212–2226, Bangkok, Thailand. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2024. Evaluating language models as synthetic data generators. *arXiv preprint arXiv:2412.03679*.

Yanghoon Kim, Hwanhee Lee, and Kyomin Jung. 2018. Attnconvnet at semeval-2018 task 1: Attention-based convolutional neural networks for multi-label emotion classification. In *International Workshop on Semantic Evaluation*.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. 2020. Dqi: Measuring data quality in nlp. *arXiv preprint arXiv:2005.00816*.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, et al. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Pebbeti Charitha Reddy, Pallala Indrani, Polidasu Janaki, Padala Gayathri, Padubandla Chandrahasini, G Apparao, and Rajeshwari. 2024. Product review sentiment analysis. *International Journal For Multidisciplinary Research*.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Mrityunjay Singh, Amit Kumar Jakhar, and Shivam Pandey. 2021. Sentiment analysis on the impact of coronavirus in social life using the bert model. *Social Network Analysis and Mining*, 11.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.