VerbaNexAI at SemEval-2025 Task 11: A RoBERTa-Based Approach for the Classification of Emotions in Text

Danileth Almanza Gonzalez, Edwin Puertas, Juan Carlos Martinez-Santos
Universidad Tecnológica de Bolívar
Cartagena, Colombia
{daalmanza,epuerta,cmartinezs}@utb.edu.co

Abstract

Emotion detection in text has become a highly relevant research area due to the growing interest in understanding emotional states from human interaction in the digital world. This study presents an approach for emotion detection in text using a RoBERTa-based model, optimized for multi-label classification of the emotions joy, sadness, fear, anger, and surprise in the context of the SemEval 2025 - Task 11: Bridging the Gap in Text-Based Emotion Detection competition. Advanced preprocessing strategies were incorporated, including the augmentation of the training dataset through automatic translation to improve the representativeness of less frequent emotions. Additionally, we implemented a loss function adjustment mechanism to mitigate class imbalance, enabling the model to enhance its detection capability for underrepresented categories. The experimental results reflect competitive performance, with a macro F1 of 0.6577 on the development set and 0.6266 on the test set. The model ranked 70th in the competition, demonstrating solid performance against the challenge posed.

1 Introduction

Emotion recognition in text has gained significant relevance with the rise of social networks, facing the challenge of identifying explicit and implicit emotions. The basic emotions most studied include joy, sadness, fear, anger, disgust, and surprise. Advances in NLP have driven the development of more accurate models, such as deep neural networks and pre-trained models like RoBERTa, improving emotion classification in various fields such as business, psychology, and security (Sboev et al., 2021; Faisal et al., 2024). This study addresses emotion detection in SemEval 2025 Subtask A - Task 11: Bridging the Gap in Text-Based Emotion Detection. The central problem motivating this research lies in the complexity of accurately detecting and classifying emotions in texts through multi-label prediction. The main objective of this task is to identify perceived emotions in five main categories (joy, sadness, fear, anger, and surprise) using a multi-label classification approach (Muhammad et al., 2025b). Although the competition addresses multiple languages, this study focuses exclusively on English. For this purpose, we employed the pre-trained RoBERTa model, leveraging its ability to capture advanced linguistic representations and improve emotion detection in complex texts. Furthermore, we implemented strategies to address the class imbalance, optimizing the model's overall performance and maximizing its classification accuracy. The results obtained were competitive, as our system reached the 70th position in the competition, achieving an acceptable performance. However, we identified challenges related to the representation of less frequent emotions, affecting the accuracy of minority classes. These findings highlight the importance of continuing to explore class balancing techniques and model tuning to improve emotion detection in texts. The repository is available via the following link: (available after reviewing.) [1]

2 Background

The dataset used in this task comes from BRIGHTER, a collection of multilingual datasets for text-based emotion recognition that spans 28 languages. For its construction, authors collected texts from various sources, including social media (Reddit, Twitter, YouTube, Weibo), speeches, personal narratives, literature, and news. Depending on the language, the data were obtained from specific platforms or manually generated by native speakers. Subsequently, the texts were curated and annotated by experts and collaborators through crowdsourcing platforms such as Amazon Mechanical Turk and Toloka and specialized tools like La-

[1] https://github.com/VerbaNexAI/SemEval2025

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1192–1197

belStudio and Potato. The annotation process included assigning multiple emotion labels and the intensity of each emotion on a scale from 0 to 3.

The author selected only English for this study, with data primarily extracted from Reddit. The dataset distribution is as follows: 2,768 instances for training, 116 samples for development, and 2,767 records for testing. the structured each instance into three primary columns: ID, text, and emotional labels corresponding to joy, sadness, fear, anger, and surprise. Table 1 shows the Distribution of emotions across the dataset subsets, evidencing an imbalance. It is worth noting that the emotion "disgust" was excluded from the English dataset due to its scarce representation in the collected texts. Unlike the other emotions, there were not enough examples labeled with "disgust", which prevented its inclusion and analysis within this corpus. In Figure 1, a representative sample of the training set is presented, illustrating examples of texts along with their emotional labels (Muhammad et al., 2025a).

Id	Text	Anger	Fear	Joy	Sadness	Surprise
01676	I have plenty more.	0	0	1	0	0
01903	That's messed up.	1	0	0	0	1

Figure 1: Sample of the training dataset.

Emotion	Training	Test	Development
Anger	333	322	16
Fear	1611	1544	63
Joy	674	670	31
Sadness	878	881	35
Surprise	839	799	31

Table 1: Distribution of each emotion in the training, test, and development sets.

Emotion analysis in text has been widely studied in Natural Language Processing (NLP), using approaches ranging from rule-based methods to advanced deep learning models. In particular, studies derived from Task 1 of SemEval-2018 have explored various strategies, highlighting Bi-LSTM networks with deep self-attention and transfer learning (Baziotis et al., 2018). Other works integrated the NRC VAD Lexicon, Transformer, and GRU to recognize emotions in code-mixed conversations, using techniques such as Emotion Flip Reasoning (EFR) and Emotion Recognition in Conversation (ERC) on the MELD and MaSaC datasets (Pacheco et al., 2024). Likewise, logistic regression with syntactic dependency graphs has been implemented to analyze emotional causality, although with limited results that motivate the use of more advanced models such as Transformers (Garcia et al., 2024). Finally, feature extraction techniques (TF-IDF, FastText, BERT) and predictive models (Naïve Bayes, SVM, Random Forest, Gradient Boosting, and neural networks) have been applied on the ISEAR dataset (Esfahani and Adda, 2024).

3 System Overview

The based the proposed system for emotion detection in texts on an architecture that combines the pre-trained RoBERTa model with additional classification and data balancing mechanisms (Hartmann, 2022). We designed the architecture to handle the inherent complexity of emotion classification, allowing for precise differentiation of emotions such as anger, fear, joy, sadness, and surprise. Figure 2 illustrates the system's structure, showing the processing flow from data input to prediction generation. To delve into the internal functioning of the system, we described some details of its components below.

3.1 Tokenization and DataLoader

These phases convert preprocessed texts into a numerical representation that the model can process. We used the RoBERTa tokenizer to split inputs into tokens, assign indices, and apply padding and truncation. Then, the EmotionDataset associates the encodings with emotional labels. Finally, the DataLoader creates mini-batches, optimizes efficiency, and prevents biases in training.

3.2 Model

We based the model on a pre-trained transformer (RoBERTa), which generates contextual representations for each token in the input sequence. Before selecting RoBERTa, we tested other pre-trained models during the experimentation phase, such as BERT. However, RoBERTa performed better in the task, partly due to its training providing specific advantages for emotion classification in English text. An aggregation mechanism is applied once the tokenizer segments the text and RoBERTa produces a sequence of embeddings. In this case, the

mechanism consists of computing the mean of the representations generated by RoBERTa along the sequence dimension. It serves as a way to consolidate the scattered information of each token into a single vector representing the general content of the text. We chose this technique because it can generate stable and balanced representations, avoid dependence on individual words, as occurs with max pooling, and reduce computational cost compared to more sophisticated methods such as attention pooling.

3.3 Attention Layer

We implemented an attention layer to assign specific weights to each token. This layer takes as input the vector representations for each token in the text. Subsequently, it uses an aggregation mechanism based on averaging to consolidate these representations, thus generating a single vector that summarizes the most relevant information from the entire text. Attention was indirectly implemented through this aggregation mechanism, allowing the model to automatically prioritize the most influential words in the final prediction. This approach was chosen due to its numerical stability, computational simplicity, and proven effectiveness in emotion classification tasks.

3.4 Classification Layer

We used a linear layer to transform the high-dimensional vector into a five-dimensional space. Each of these dimensions corresponds to one of the target emotions: anger, fear, joy, sadness, and surprise. The classification layer, which operates on the regularized vector, produces the logits for each emotion. These logits are subsequently interpreted through the sigmoid function in the loss computation phase (BCE With Logits Loss), transforming the results into probabilities that indicate the presence or absence of each emotion.

3.5 Class Balancing

Considering that the dataset presents an imbalanced distribution among the target emotions (anger, fear, joy, sadness, and surprise), a specific balancing strategy was incorporated into the loss function used during training. Specifically, the BCEWithLogitsLoss function from PyTorch was used with the pos_weight parameter adjusted according to the relative frequency of each emotional class in the training set. This weight was calculated by dividing the most frequent class by each of the other classes,

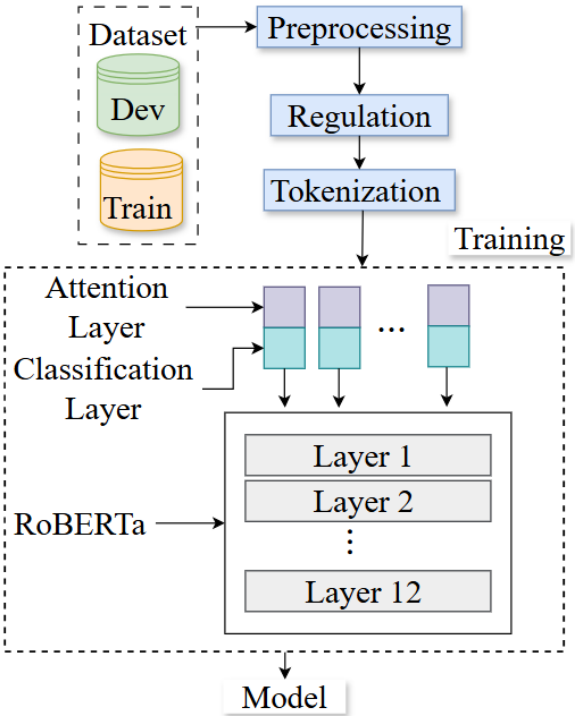

Figure 2: Architecture system.

thus ensuring that errors made on less represented emotions were penalized more heavily during training. This approach improved the model's sensitivity to minority categories and led to more balanced and accurate predictions overall. The technical implementation consisted of transforming these calculated weights into a tensor (class_weights_tensor), which was directly integrated into the loss function as a parameter during the model optimization process.

4 Experimental Setup

We designed the experimental configuration to ensure a robust and generalizable emotion detection model. The dataset was initially divided into training and testing sets, assigning 80% for training and the remaining 20% for testing. The original dataset consisted of 2,768 instances in English, which we considered insufficient to capture the complexity of emotional phenomena fully. To address this limitation, we incorporated 7,597 instances from German and Brazilian Portuguese datasets. We translated the instances into English using the DeepL Pro API (version v2), which based the learning on the Linguee service, an extensive database of phrases and text fragments translated by humans (DeepL, 2025). Thanks to this data source, the API achieves high accuracy in translations. This

strategy enriched the training set and improved the model's generalization ability across various structures and linguistic contexts. Moreover, the dataset expansion through high-quality translation not only increased data diversity but also contributed to a significant improvement in the model's performance, particularly in detecting underrepresented emotions. With the dataset duly expanded, a rigorous preprocessing procedure was applied to ensure the uniformity and quality of the textual information.

4.1 Preprocessing

This stage begins with the normalization and cleaning the text, which is fundamental to ensuring the uniformity of the corpus. First, Unicode normalization is applied, converting all characters to canonical forms and eliminating discrepancies due to different encodings and formats. This step is crucial for correctly handling accented characters and special symbols. Then, regular expressions are employed to remove undesired patterns, such as digits, redundant punctuation, and special characters that do not contribute semantic meaning. Additionally, we transformed all text to lowercase, ensuring we treated identical words in different formats uniformly, which reduces variability and noise in the data.

Once we completed the initial cleaning, we implemented a more advanced transformation pipeline, which included replacing specific text elements. We replaced URLs, mentions, and hashtags with generic tokens ([URL], [MENTION], and [HASTAG], respectively), which helps to preserve the semantic structure without overloading the model with unnecessary details. Moreover, emojis are identified and tagged by inserting a marker ([EMOJI]), allowing us to control the emotional information implicit in these symbols. Finally, this preprocessing chain is integrated in an automated fashion, ensuring that each corpus instance goes through the same cleaning and transformation steps before tokenization with the RobertaTokenizer. This standardization is essential for generating consistent and robust representations that facilitate learning and subsequent classification in the emotion detection model. This preprocessing strategy significantly contributed to improving the model's performance. These results clearly demonstrate that the implemented techniques have a direct and positive impact on the overall accuracy of the model and enhance its ability to correctly identify emotions in complex texts.

4.2 Pipeline Configuration and Hyperparameter Tuning

In this stage, we integrated preprocessing processes into a unified pipeline that prepares each input for the model. Initially, we performed tokenization using RobertaTokenizer from Hugging Face. We split the text into minimal units (tokens), which we converted into indices according to the model's predefined vocabulary. We set a maximum limit of 400 tokens per instance, ensuring the capture of relevant semantic information without introducing redundancies or excessive noise. Manual hyperparameter tuning was conducted during training to optimize the model's performance. Different learning rates ($2e-5$, $3e-5$, $4e-5$) were experimented with, as well as various batch sizes [8,16,32] and numbers of epochs (initially 10, then 50 and 100). After evaluating the performance of each configuration, we selected the values that provided the best performance: a learning rate of $2e-5$, a batch size of 8, and a total of 100 epochs. Combined with a weight decay of 0.01 and rigorous preprocessing, we made these adjustments to optimize the quality of the information provided to the model, facilitating its convergence during the training phase.

4.3 Evaluation and Performance Metrics

We evaluated the system using standardized metrics for comprehensive performance quantification. We employed accuracy, precision, recall, and F1-score measures, calculated as both micro and macro averages, providing a detailed view of the model's ability to identify each target emotion correctly.

These configurations have enabled the development of a robust and replicable emotion detection system, in which each stage, from the expansion and preprocessing of the corpus to the fine-tuning of the model, contributes significantly to improving the generalization and precision of the classification.

5 Results

In the results obtained in the different evaluation phases, as shown in Tables 3, 4, and 5 in the development and test sets, during training, the model showed a progressive improvement, reaching a micro precision of 0.9958 and a micro F1-score of 0.9917 in the last epoch. However, the average over all epochs (precision of 0.8694 and loss of

Text	Gold label					Prediction label				
	Anger	Fear	Joy	Sadness	Surprise	Anger	Fear	Joy	Sadness	Surprise
I slammed my fist against the door and yelled, Open up!	1	1	0	0	1	1	1	0	0	0
My heart dropped and I just replied “No.	0	1	0	1	0	0	1	0	0	0
Man, I can’t believe it	0	1	1	0	1	0	0	0	0	1

Table 2: Model error analysis

Metric	Development	Test
Macro F1	0.6577	0.6266
Micro F1	0.6571	0.6787

Table 3: Metrics obtained for the development and test sets.

Metric	Anger	Fear	Joy	Sadness	Surprise
F1	0.720	0.676	0.626	0.738	0.526

Table 4: Model performance on the development set

Metric	Anger	Fear	Joy	Sadness	Surprise
F1	0.495	0.787	0.6324	0.572	0.645

Table 5: Model performance on the test set

0.0962) indicates variability, suggesting the need for balancing. In the development set, the performance was lower (macro F1: 0.6577, micro F1: 0.6571), with differences among emotions, highlighting good performance in “Sadness” (0.7385) and lower in “Surprise” (0.5263). In the test set, the metrics were similar (macro F1: 0.6266, micro F1: 0.6787), with “Fear” obtaining the highest score (0.7874) and “Anger” the lowest (0.4950), suggesting difficulties in certain emotions due to data distribution.

The error analysis reveals that the model struggles to identify surprise, sadness, and fear accurately. False positives in surprise suggest that the model confuses emphatic expressions, even when they convey fear or disbelief. Similarly, false negatives in sadness indicate that the model does not adequately capture the emotional subtext when we use metaphors or figurative expressions without explicit terms like "sad" or "devastated." Additionally, the model has issues with ambiguous phrases where the emotion depends on the context, leading to omissions in detecting fear and joy. Table 2 shows that the model tends to confuse similar emotions in context, such as fear and surprise in disbelief or sadness and fear in distressing scenarios. Although the model performs better in classifying joy and anger, we also observed that it still makes errors, suggesting that it struggles to correctly interpret emotional language when it depends on contextual nuances or idiomatic expressions.

6 Conclusion

Automatic Emotion Classification in Text Analysis has applications in social networks, digital platforms, the business sector, and education. Given its relevance, it is essential to improve the accuracy and robustness of the models, ensuring their adaptability to multiple languages and domains. The proposed system has demonstrated effectiveness in detecting complex emotional nuances, achieving solid results in various categories. However, challenges persist in the identification of minority emotions such as anger and surprise. To address them, it was proposed to implement advanced class balancing strategies and automatic hyperparameter tuning techniques. Future work will consider search methods to optimize hyperparameters in several pretrained models and will explore hybrid approaches that integrate complementary architectures.

Acknowledgments

The authors express their gratitude to the Call 933 "Training in National Doctorates with a Territorial, Ethnic and Gender Focus in the Framework of the Mission Policy — 2023" of the Ministry of Science, Technology and Innovation (Minciencia). In addition, we thank the team of the Artificial Intelligence Laboratory VerbaNex [2], affiliated with the UTB, for their contributions to this project.

References

Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth S. Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *CoRR*, abs/1804.06658.

DeepL. 2025. La API de DeepL traduce y mejora contenido a gran escala. https://www.deepl.com/es/products/api.

Seyed Hamed Noktehdan Esfahani and Mehdi Adda. 2024. Classical machine learning and large models for text-based emotion recognition. *Procedia Computer Science*, 241:77–84.

Moshiur Rahman Faisal, Ashrin Mobashira Shifa, Md Hasibur Rahman, Mohammed Arif Uddin, and Rashedur M. Rahman. 2024. Bengali banglish: A monolingual dataset for emotion detection in linguistically diverse contexts. *Data in Brief*, 55:110760.

Santiago Garcia, Elizabeth Martinez, Juan Cuadrado, Juan Martinez-santos, and Edwin Puertas. 2024. VerbaNexAI lab at SemEval-2024 task 10: Emotion recognition and reasoning in mixed-coded conversations based on an NRC VAD approach. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1332–1338, Mexico City, Mexico. Association for Computational Linguistics.

Jochen Hartmann. 2022. Emotion english distilroberta-base. Available at: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, and 29 others. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *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, and 2 others. 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.

Victor Pacheco, Elizabeth Martinez, Juan Cuadrado, Juan Carlos Martinez Santos, and Edwin Puertas. 2024. VerbaNexAI lab at SemEval-2024 task 3: Deciphering emotional causality in conversations using multimodal analysis approach. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1339–1343, Mexico City, Mexico. Association for Computational Linguistics.

Alexander Sboev, Aleksandr Naumov, and Roman Rybka. 2021. Data-driven model for emotion detection in russian texts. *Procedia Computer Science*, 190:637–642. 2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: Eleventh Annual Meeting of the BICA Society.

[2] https://github.com/VerbaNexAI