NLP_goats at SemEval-2025 Task 11: Multi-Label Emotion Classification Using Fine-Tuned Roberta-Large Tranformer

Vijay Karthick Vaidyanathan
Sri Sivasubramaniya Nadar College of Engineering
vijaykarthick2210930@ssn.edu.in

Srihari V K
Sri Sivasubramaniya Nadar College of Engineering
srihari2210434@ssn.edu.in

Mugilkrishna D U
Sri Sivasubramaniya Nadar College of Engineering
mugilkrishna2210314@ssn.edu.in

Saritha M
Sri Sivasubramaniya Nadar College of Engineering
sarithamadhesh@ssn.edu.in

Abstract

Bridging the gap in text-based emotion detection has received significant attention due to the diverse ways in which emotions are explicitly conveyed in written text. Digital communication platforms often present complex emotional expressions which are a challenge to conventional analysis methods. This paper presents a two-track approach to address these challenges in English: Track 1 (Multi-label Emotion Detection) and Track 2 (Emotion Intensity) are described. The method primarily revolves around sophisticated textual mining techniques and fine-tuning transformer-based language models to generate powerful semantic features. A multi-label classification approach is applied on Track 1 for capturing common emotional states, and regression models are used on Track 2 to estimate emotion strengths. The developed system achieved competitive rankings of 31 and 17 in both tracks, highlighting the promise of the approaches used to improve the precision and robustness of text-based emotion detection.

1 Introduction

Text-based emotion recognition has become one of the core elements of contemporary natural language processing, which has changed the way we perceive and use digital communication. Today's social media and instant messaging culture often allows people to communicate their nuanced, often unacknowledged, emotional states through the written word. This rich tapestry of affective expression not only influences personal interactions but also drives applications in customer service, mental health monitoring, and social analytics. Nevertheless, conventional sentiment analysis methods usually lack the depth of human emotion spectrum and nuances.

Track 1 is concerned with Multi-label Emotion Detection and Track 2 is concerned with Emotion Intensity estimation.

Track 1 promotes the creation of models able to detect shared emotional states in a text, accommodating the complex dimension of human affect. Simultaneously, Track 2 asks the researchers to measure the fine nuances in the intensity of emotion, beyond dichotomizing between emotion intensities, to provide a more nuanced description of affective expression. Taken together, these tasks strive for the current state-of-the-art by encouraging novel solutions that can cope with the complexity and richness of real-world textual emotions.

This paper is structured as follows: Section 2 surveys the relevant related works in textual emotion detection and sentiment analysis, discussing both historical trends and recent progress. Section 3 gives a detailed explanation/ task description such as the dataset and the evaluation metrics. Section 4 consists of the adopted methodology, highlighting the preprocessing techniques, model architectures, and training procedures. Section 5 presents the experimental results along with a comparative analysis of the approaches used. Section 6 discusses the error analysis, to identify potential areas for improvement. Finally, Section 7, the conclusion section, contains a summary of key findings and insights into future research directions.

By addressing these challenges, the gap between conventional sentiment analysis methods and the complex, multidimensional nature of human emotion is bridged. Through advanced textual analysis and innovative modeling techniques, the aim is to enhance the reliability and depth of emotion detection systems, ultimately contributing to more empathetic and effective digital communication platforms. The code for the tasks is available on GitHub at this repository.

2 Related Work

Emotion analysis on textual data has been extensively researched in Natural Language Processing

(NLP) with applications from multi-label emotion classification to intensity regression of emotions. Conventional methods were based on lexicon-based approaches, in which words were assigned to pre-defined emotional categories in terms of resources like the NRC Emotion Lexicon (Mohammad and Turney, 2013). While these approaches provided useful insights, they lacked contextual understanding and struggled with complex linguistic patterns.

Multi-label emotion classification is to assign multiple emotion labels to a text. Early ML-based approaches have been applied to TF-IDF and n-gram features with Support Vector Machines (SVMs) and Random Forests (Strapparava and Mihalcea, 2008), respectively. Nevertheless, the traditional methods have been surpassed by deep learning methods like Long Short-Time Memory (LSTM) networks and Convolutional Neural Networks (CNNs) (Felbo et al., 2017). The introduction of the transformers, especially BERT (Devlin et al., 2019) and Roberta (Liu et al., 2019), has improved multi-label classification considerably by exploiting the contextual embeddings. Recent works have optimized thresholding techniques, such as adaptive thresholding (Pérez-Rosas et al., 2020) and focal loss, to handle label imbalances in multi-label classification.

Emotion intensity prediction (i.e., how much of an emotional expression is there in a given text) has been addressed from both lexicon-based and deep-learning points of view. Early works relied on affective lexicons such as the NRC Affect Intensity Lexicon (Mohammad, 2018) to assign predefined intensity scores. Nevertheless, those approaches could not learn the dynamic emotion expression in real-world text. Recurrent neural networks (RNNs) and BiLSTMs have been applied to model sequential dependencies in text, improving intensity prediction (Baziotis et al., 2018). Transformer-based architectures (e.g., BERT, Roberta) have also significantly helped this domain by training models to perform regression by tuning the model parameters for regression tasks using loss functions such as Mean Squared Error (MSE) or Huber Loss (Goel et al., 2021).

In this paper, we focus both on the multi-label emotion classification and the emotion intensity regression, both using transformer-based models. The classification problem is approached as a multi-label problem, using Binary Cross-Entropy with Logits (BCEWithLogitsLoss) and threshold-tuning methods to enhance emotion detection. In the regression task, the model is trained to predict the intensity of emotion by MSE loss, with the goal of optimal fine-grained emotion strength detection. By integrating recent advancements in transformers and loss function optimization, this work aims to enhance both the classification and regression aspects of emotion analysis in textual data.

3 Task Description

We focus on two related yet distinct tasks aimed at analyzing the emotional content in text: Multi-label Emotion Detection and Emotion Intensity Prediction. Both tasks contribute to a deeper understanding of affective computing, particularly in the context of social media, dialogues, and opinionated text. The dataset is due to the efforts of (Muhammad et al., 2025) and (Belay et al., 2025).

3.1 Track A: Multi-label Emotion Detection

The goal of this task is to classify a given text snippet into multiple perceived emotions. Specifically, for a given text, we determine whether each of the following six emotions applies: joy, sadness, fear, anger, surprise, or disgust. Since emotions are not mutually exclusive, a text may exhibit multiple emotions simultaneously. The model outputs a binary decision for each emotion: 1 if the emotion is present, 0 otherwise.

3.2 Track B: Emotion Intensity Prediction

This task extends the analysis by predicting the intensity of a given emotion in a target text. Given a text and a specified target emotion (one of joy, sadness, fear, anger, surprise, or disgust), the model predicts the emotion's intensity on a four-point ordinal scale: 0 (no emotion), 1 (low intensity), 2 (moderate intensity), and 3 (high intensity).

4 Methodology

This paper aims at two fundamental tasks of emotion analysis: emotion classification with multiple labels and regression of emotion intensity. We adopt transformer-based models, i.e., fine-tuned BERT and RoBERTa models, to better solve both tasks. The methodology is composed of data preprocessing, model structure, training approach, and evaluation metrics.

4.1 Data Preprocessing

The dataset is preprocessed by removing special characters, URLs and redundant whitespaces. By keeping stopwords to maintain the context integrity, the Byte-Pair Encoding (BPE) tokenizer is applied for tokenization. For multi-label classification, labels are one-hot encoded to enable discrete probability assignment to each emotion. On the other hand, for regression, the intensities of the emotion are normalized in the range of [0,1] so that they scale consistently and train stably (Mohammad, 2018). In addition, to address data sparsity, lowercasing and lemmatization are used as text normalization techniques and padding is performed to a fixed size so that the batches are uniform.

4.2 Model Architecture

In the field of multi-label classification, we adopt a pre-trained transformer model (RoBERTa) which utilizes multiple self-attention layers and can capture the inherent contextual dependencies within the whole textual information. The transformer encoder is applied to the input text that has been tokenized by Byte-Pair Encoding (BPE). The output from the last hidden layer is fed into a fully connected dense layer and each neuron represents an emotion label. As an instance may have several emotions at the same time, sigmoid activation is applied to each of the output neurons separately to produce probability scores for each of the labels. A threshold (e.g., 0.5) is applied to classify whether an emotion exists. Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) is employed for training as it aims to optimize predictions for each emotion category rather than assume mutual exclusivity.

For emotion intensity regression, the same architecture based on RoBERTa is employed as a feature extractor but instead of having a dense layer with multiple outputs and sigmoid activation, we make the last layer consist of a single neuron for each emotion class with linear activation. This setup allows the model to predict continuous intensity values rather than categorical labels. To minimize errors in continuous predictions, the Mean Squared Error (MSE) loss function is used, as it penalizes large deviations and ensures smoother optimization (Goel et al., 2021). Additionally, we introduce a dropout layer before the final output to reduce overfitting by randomly deactivating neurons during training, improving the model's generalization ability. We use a single model with five neurons in the final layer, where each neuron corresponds to one emotion, enabling the classification of all emotion classes.

In both tasks, the last transformer layer's hidden states are first passed through a pooling mechanism (CLS token embedding or mean pooling) before being input to the final output layer. This is done to ensure that the most important features are extracted and used effectively. Layer normalization and weight decay regularization are also used to stabilize training and avoid overfitting.

4.3 Training Strategy

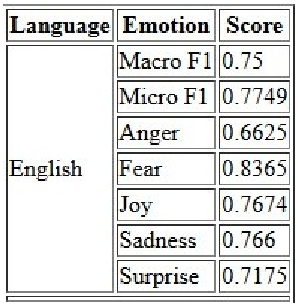
Fine-tuning is performed with the AdamW optimizer and a learning rate of 2e-5, and a linear scheduler with warm-up stages to avoid extreme weight changes in the early training stages. In the case of classification, threshold tuning after training is applied to refine decision boundaries in order to solve the multi-label assignment (Pérez-Rosas et al., 2020). Models are trained using batch sizes of 16 and 32 for classification and regression, respectively, in an attempt to maximize GPU memory use.

For multi-label classification, the sigmoid-activated logits are thresholded at an evolving value, learned from validation set analysis, to achieve the best performance trade-off in terms of precision-recall. Training is carried out for 10-15 epochs with early stopping using a validation loss based criterion to prevent overfitting. To regularize updates and improve convergence, in particular for very large batch training, the gradient accumulation method is employed.

In emotion intensity regression, The RoBERTa model is fine-grainedly trained using the MSE loss function. The dropout probability is fixed to 0.1 to further regularize learning and prevent overfitting. The dynamic learning rate scheduler is adaptive to provide the convergence. Model checkpoints are saved at the highest validation performance, guaranteeing that the final evaluation be performed on the most optimized state.

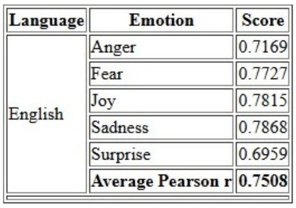
4.4 Evaluation Metrics

For classification (Track A), evaluation is conducted using F1-Score to assess label co-occurrence and retrieval effectiveness. We achieved a macro F1-score of 0.75. In the regression task (Track B), Pearson correlation is employed to measure the strength of linear associations (Strapparava and Mihalcea, 2008). In the second track we achieved a Pearson correlation of 0.75. The detailed results for Track A, and Track B are shown in Figure 1 and Figure 2 respectively.



Language	Emotion	Score
English	Macro F1	0.75
	Micro F1	0.7749
	Anger	0.6625
	Fear	0.8365
	Joy	0.7674
	Sadness	0.766
	Surprise	0.7175

Figure 1: F1-Score for track-A



Language	Emotion	Score
English	Anger	0.7169
	Fear	0.7727
	Joy	0.7815
	Sadness	0.7868
	Surprise	0.6959
	Average Pearson r	**0.7508**

Figure 2: Pearson correlation for track-B

5 Error Analysis and Results

In Track A, emotion classification with a multi-class classification method, the model obtained a Macro F1 of 0.75. While this reflects a good overall performance, scores across emotions significantly differ. The best-performing class was Fear (F1 = 0.8365), while the worst was Anger (F1 = 0.6625). This discrepancy indicates that the model has trouble separating Anger from other emotions, perhaps due to similar linguistic patterns with emotions like sadness or frustration. The class imbalance could also have affected the performance of some emotions. The Micro F1 score of 0.7749 reflects that the model performed better in classifying instances that occur frequently but struggled with less frequent or ambiguous emotional expressions.

For Track B, a multi-label regression task for emotion intensity prediction, the model achieved an average Pearson correlation of 0.7508, indicating a strong relationship between predicted and actual emotion intensities. However, the correlation varied across emotions, with Sadness (0.7868) and Joy (0.7815) being predicted more accurately than Surprise (0.6959). The lower correlation for Surprise suggests that the model found it challenging to predict its intensity, likely because of the subtlety of this emotion and context dependence. Another possible source of inaccuracy is the occurrence of co-occurring emotions within the text, which would give rise to underestimation or overestimation of certain intensities. One could make further enhancements by using more emotion-specific contextual embeddings or treating ambiguous instances better using contrastive learning mechanisms.

6 Conclusion

Track A and Track B results demonstrate strong performance in classification and regression tasks. In Track A, the model achieved a Macro F1 score of 0.75, which indicates a well-balanced performance across all emotion categories. The Micro F1 score of 0.7749 suggests that the model handles overall classification instances effectively. The model in Track B achieved a mean Pearson correlation value of 0.7508, indicating very high correspondence between predicted and ground truth emotion intensity. The figures indicate the model's power in both the classification of discrete emotion and prediction of continuous intensity and its strength in dealing with sensitive emotional expressions from the text.

References

Christos Baziotis, Nikolaos Athanasiou, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 Task 1: Predicting Affect Dimensions and Valence-Arousal in Tweets Using Attentive RNNs. In *Proceedings of SemEval-2018*, pages 245–251.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP 2017*, pages 1615–1625.

Pranav Goel, Shrey Agarwal, and Amir Hussain. 2021. Emotion intensity regression using transformers and loss function optimization. In *ACL 2021 Findings*, pages 1432–1440.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of ACL 2018*, pages 34–45.

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, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Verónica Pérez-Rosas, Rada Mihalcea, and Ken Resnicow. 2020. Predicting emotion intensity in text using contextualized embeddings. In *LREC 2020*, pages 3992–3998.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. *ACM Transactions on Speech and Language Processing*, 5(1):1–23.