NITK-VITAL at SemEval-2025 Task 11: Focal-RoBERTa: Addressing Class Imbalance in Multi-Label Emotion Classification

Ashinee K[1] **G Venkata Ravi Ram**[2] **B Chaithanya Swaroop**[3] **G Ram Mohana Reddy**[4]
Artificial Intelligence, Department of Information Technology
National Institute of Technology Karnataka, Surathkal, India
[1]ashineekesanam@gmail.com, [2]toraviram2003@gmail.com,
[3]cs.chaithanyaswaroop@gmail.com, [4]profgrmreddy@nitk.edu.in

Abstract

This paper presents our approach to SemEval Task 11, which focuses on multi-label emotion detection in English textual data. We experimented with multiple methodologies, including traditional machine learning models, deep learning architectures, and transformer-based models. Our best-performing approach employed RoBERTa with focal loss, which effectively mitigated class imbalances and achieved a macro F1-score of 0.7563, outperforming other techniques. Comparative analyses between different embedding strategies, such as TF-IDF, BERT, and MiniLM, revealed that transformer-based models consistently provided superior performance. The results demonstrate the effectiveness of focal loss in handling highly skewed emotion distributions. Our system contributes to advancing multi-label emotion detection by leveraging robust pre-trained models and loss function optimization.

1 Introduction

Understanding emotions in text is a crucial aspect of natural language processing (NLP) with applications in sentiment analysis, mental health monitoring, and human-computer interaction. Emotions are inherently complex, nuanced, and subjective, making their automatic detection a challenging task. The SemEval 2024 Task 11 focuses on multi-label emotion detection, where the goal is to determine the perceived emotions conveyed by a speaker in a given text snippet. This task is particularly challenging due to variations in individual emotional expression, cultural differences, and the ambiguity of language. (Muhammad et al., 2025b)

In this paper, we present our approach for Track A (English language), where we predict whether a given text snippet expresses one or more of the following emotions: joy, sadness, fear, anger, or surprise. Our method leverages deep learning-based models, incorporating pre-trained transformer architectures such as BERT, RoBERTa, and T5, along with fine-tuning strategies tailored for multi-label classification.

Through our participation in the task, we observed several key insights. Firstly, contextual embeddings significantly improve performance compared to traditional machine learning methods. Secondly, class imbalance poses a challenge, as certain emotions are underrepresented in the dataset, leading to biased predictions. Despite these challenges, our system achieved competitive performance, ranking among the top-performing models in the competition.

The rest of this paper is structured as follows: Section 2 discusses related work in emotion detection, Section 3 details our system architecture, Section 4 presents the experimental setup, Section 5 analyzes the results, and Section 6 concludes with future directions.

2 Related Works

A key challenge in multi-label emotion detection is class imbalance, where certain emotions are underrepresented. To address this, Lin et al. (Lin et al., 2017) introduced Focal Loss, which dynamically adjusts the cross-entropy loss to focus on harder examples while down-weighting easier ones. Though originally developed for object detection, its principle of emphasizing difficult samples has inspired applications in other domains. Jiang et al. (Jiang et al., 2021), for instance, extended this idea to frequency components in image reconstruction. This relevance extends to emotion detection, where rare emotions are similarly difficult to capture and benefit from targeted loss optimization.

In textual emotion detection, researchers have explored methods that incorporate emotion interdependencies. Chochlakis et al. (Chochlakis et al., 2022) leveraged label correlations through pairwise

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1077–1081

constraints as regularization terms, helping models better capture co-occurring emotions. Alhuzali and Ananiadou (Alhuzali and Ananiadou, 2021) proposed SpanEmo, reframing emotion classification as a span-prediction task to capture overlapping emotional cues within text. Additionally, Choudhary and Chakraborty (Choudhary and Chakraborty, 2020) combined deep learning features with handcrafted features to improve the granularity of emotion recognition, showing that hybrid approaches can enhance model interpretability and accuracy.

In multilingual and code-mixed contexts, Gupta et al. (Gupta et al., 2021) introduced SENTIMOJI, a dataset designed for multi-label emotion and sentiment classification in diverse linguistic settings. Their work highlights the complexities introduced by language mixing and the need for adaptable models. Although transformer-based architectures like RoBERTa have shown promise in emotion detection tasks, few studies have explored the integration of focal loss with such models. Our approach distinguishes itself by combining focal loss with a fine-tuned RoBERTa and a multi-head attention layer, leading to improved detection of infrequent emotions and better overall performance.

3 Data

The dataset used for this task consists of **2768** text samples, each labeled with one or more emotions. The dataset contains **7 columns**, including the text snippet and five binary labels representing the presence or absence of the following emotions: **joy, sadness, fear, anger, and surprise**. The text data serves as the primary input for classification. The evaluation metric for this task is **macro-averaged F1-score (F1-macro)**, which ensures balanced performance across all emotion categories, regardless of class imbalance. (Muhammad et al., 2025a)

4 System Overview

4.1 Machine Learning-Based Approaches

To tackle the multi-label emotion classification task, we Initially, implemented models such as Logistic Regression, Random Forest, Support Vector Classifier (SVC), and Extreme Gradient Boosting (XGB). These classifiers were applied using two multi-label strategies: Classifier Chains and MultiOutputClassifier. For feature extraction, we employed TF-IDF and experimented with BERT. While these approaches provided a solid baseline, they struggled with capturing the intricate semantic and syntactic nuances necessary for accurate emotion detection.

4.2 Feedforward Neural Network (FNN)

We further explored a deep learning approach by implementing a Feedforward Neural Network (FNN) trained on TF-IDF features. The model consisted of two hidden layers with ReLU activations and dropout layers to prevent overfitting. The final output layer used a sigmoid activation function for multi-label classification. The FNN model demonstrated moderate improvements over traditional machine learning models, but it still lacked the ability to effectively capture contextual dependencies.

4.3 MiniLM-Based Embeddings with Part-of-Speech Features

To enhance context awareness, we utilized the lightweight transformer model MiniLM to generate dense sentence embeddings. These embeddings were enriched with Part-of-Speech (POS) features extracted using the SpaCy NLP library. We incorporated counts of key syntactic elements such as nouns, verbs, adjectives, and adverbs, which influence emotional expression in text. A neural network was then trained on the combined feature set, improving emotion recognition precision. While this approach showed promise, it was ultimately outperformed by transformer-based models with multi-head attention.

4.4 BERT and RoBERTa with Multi-Head Attention (Best Performing Approach)

Our best-performing model utilized transformer-based architectures: BERT and RoBERTa enhanced with multi-head attention mechanisms.

Preprocessing and Tokenization
We preprocessed the dataset by lowercasing text, removing special characters, and tokenizing sentences. All sequences were truncated to a fixed length of 128 tokens.

Model Architecture We utilized a pre-trained BERT or RoBERTa encoder to obtain contextualized word embeddings from the input text. To improve feature extraction, a multi-head attention layer was added, allowing the model to capture deeper dependencies between emotion-relevant words. The output was pooled using the `[CLS]` token to obtain a sentence-level representation. This was passed through a fully connected layer with sigmoid activation to predict the five emotion labels. Refer Fig. 1.

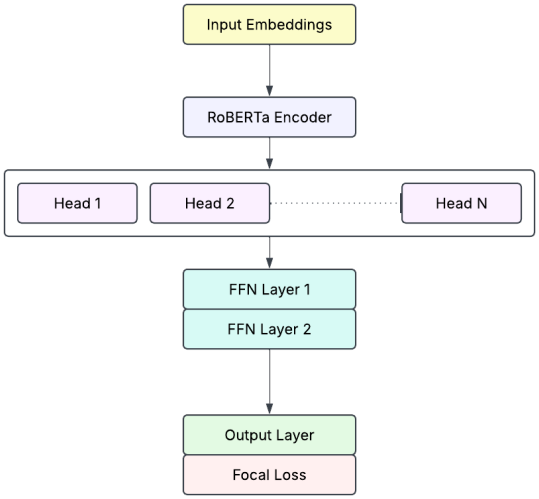

Figure 1: Focal-RoBERTa Architecture

Training Strategy and Loss Function To mitigate class imbalance, we replaced the standard Binary Cross-Entropy loss with Focal Loss, which down-weights easy examples and emphasizes harder ones. This helped the model better detect minority emotions like surprise and fear, which are often underrepresented compared to emotions like sadness and joy. The loss function was defined as:

$$FL(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t) \quad (1)$$

where α is the weighting factor (set to 1) and γ (focusing parameter) was set to 2.

5 Experimental Setup

5.1 Preprocessing

Preprocessing involved standard text cleaning techniques, including lowercasing, removal of special characters, and handling contractions. Tokenization was performed using the pre-trained BERT and RoBERTa tokenizers, ensuring compatibility with transformer-based models. Sequences were truncated or padded to a fixed length of 128 tokens to standardize input size.

5.2 Model Training and Hyperparameter Tuning

We trained multiple models, including traditional machine learning classifiers, Feedforward Neural Networks (FNN), and transformer-based architectures. Our best-performing model was a fine-tuned BERT and RoBERTa model with an additional multi-head attention layer.

For training, we employed the AdamW optimizer with a learning rate of 1×10^{-5}. The models were trained for three epochs with a batch size of 16, utilizing early stopping based on validation loss. To address class imbalance, we used Focal Loss instead of standard Binary Cross-Entropy (BCE) loss.

We used Focal Loss with = 1 and = 2. These values were selected based on the original work by Lin et al. (2017) and confirmed via preliminary experiments on the validation set. The value of 2 helps focus learning on harder examples, which is especially beneficial for underrepresented emotion classes like "surprise."

5.3 Implementation Details

Experiments were conducted on an NVIDIA A100 GPU using PyTorch and the Hugging Face Transformers library. We leveraged pre-trained weights from `bert-base-uncased` and `roberta-base` to initialize our models, fine-tuning them on the given dataset. Training and evaluation scripts were implemented in Python using the PyTorch framework.

6 Results and Analysis

We evaluated our models using the macro-F1 score.

6.1 Approach 1: Machine Learning Models

6.1.1 Results with TF-IDF Embeddings

The results obtained using traditional machine learning models with TF-IDF embeddings indicate that Classifier Chains with Logistic Regression achieved the highest F1-score of. The lowest Hamming Loss was observed with Binary Relevance using SVM, demonstrating its effectiveness in minimizing label misclassification. Ref Table. 4

6.1.2 Results with BERT Embeddings

When using BERT embeddings, the performance of machine learning models improved significantly. Classifier Chains with SVM achieved the highest F1-score of 0.6511 and the best accuracy of 0.3646. Binary Relevance with SVM demonstrated the highest precision at 0.7614, while Classifier Chains with SVM had the highest recall of 0.5728. Ref Table. 5

6.2 Approach 2: Custom Feedforward Neural Network

Our custom feedforward neural network, trained with TF-IDF embeddings, Ref Table.1

Metric	Score
Accuracy	0.2220
F1 Score (Macro)	0.4784
Hamming Loss	0.2574

Table 1: Performance of Neural Network with TF-IDF Embeddings

6.3 Approach 3: Lightweight MiniLM-based Model

The MiniLM-based model showed moderate performance with a micro-F1 score of 0.45. It excelled in detecting *Fear*, achieving an F1-score of 0.72, but struggled significantly with other emotions.

6.4 Approach 4: Multihead RoBERTa vs. Multihead BERT

Both models performed similarly, obtaining an F1-score of 0.72 for the *Fear* class but failing for other emotions. The overall macro-average F1-score was 0.14, indicating difficulties in handling class imbalances. Ref Table. 2

Label	BERT (F1)	RoBERTa (F1)
Anger	0.00	0.00
Fear	0.72	0.68
Joy	0.00	0.00
Sadness	0.00	0.10
Surprise	0.00	0.05
Macro F1	0.14	0.17

Table 2: Comparison of F1-scores between Multihead RoBERTa and Multihead BERT

6.5 Approach 5: BERT and RoBERTa with Focal Loss

Our final approach used focal loss, significantly improving results. RoBERTa achieved the highest macro-F1 score of 0.7563, outperforming BERT across all emotion categories. 3

Emotion	Focal-BERT (F1)	Focal-RoBERTa (F1)
Anger	0.6493	0.6808
Fear	0.8436	0.8481
Joy	0.7361	0.7655
Sadness	0.7483	0.7572
Surprise	0.6836	0.7299
Macro F1	0.7322	0.7563
Micro F1	0.7663	0.7825

Table 3: Comparison of F1-scores between BERT and RoBERTa with Focal Loss

This approach proved to be the most effective, leading to our final model submission. The application of focal loss allowed for better handling of class imbalances, making it superior.

6.6 Error Analysis

We analyzed misclassified examples and found that subtle distinctions between emotions like "joy" and "surprise" often caused confusion. For instance, sarcastic or ironic texts were misclassified due to implicit emotional cues. Moreover, "anger" was frequently mispredicted as "sadness," possibly due to shared lexical overlap. Adding interpretability methods like SHAP or attention visualization could help better understand model predictions.

7 Conclusion and Future Work

This work explored traditional machine learning, deep learning, and transformer-based models for multi-label emotion detection. Pre-trained transformers like BERT and RoBERTa notably boosted performance, with focal loss effectively addressing class imbalance—leading to the highest macro F1-score with RoBERTa.

Future work will focus on data augmentation methods, such as back-translation and synonym replacement, to improve generalization and handle class imbalance. Exploring ensemble approaches that combine multiple transformer models may further enhance performance. Finally, deploying the best-performing model in real-world applications like mental health monitoring could provide valuable insights.

References

Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1570–1581.

Georgios Chochlakis, Themis Exarchos, Rigas Kotsakis, and George Giannakopoulos. 2022. Multi-label emotion recognition by leveraging label correlations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6468–6475.

Shailza Choudhary and Tanmoy Chakraborty. 2020. A hybrid feature extraction approach for multi-label emotion classification. *IEEE Transactions on Affective Computing*.

Ishan Gupta, Aditya Joshi, Shubham Kumar, and Pushpak Bhattacharyya. 2021. Sentimoji: An emoji-powered learning approach for multilingual senti-

Model	F1 Score	Accuracy	Hamming Loss
Multilabel KNN	0.4279	-	0.2827
Binary Relevance - Logistic Regression	0.4839	0.2202	0.2556
Binary Relevance - Random Forest	0.4773	0.2040	0.2617
Binary Relevance - XGBoost	0.5058	0.2076	0.2625
Binary Relevance - SVM	0.5152	0.2419	0.2527
Classifier Chains - Logistic Regression	0.5551	0.2401	0.2621
Classifier Chains - Random Forest	0.4764	0.2202	0.2643
Classifier Chains - XGBoost	0.5551	0.2401	0.2621
Classifier Chains - SVM	0.5455	0.2419	0.2635

Table 4: Performance Comparison using TF-IDF Embeddings

Model	Precision	Recall	F1 Score	Accuracy
Binary Relevance - Logistic Regression	0.7020	0.6303	0.6642	0.3592
Classifier Chains - Logistic Regression	0.6830	0.6397	0.6606	0.3430
Binary Relevance - Random Forest	0.7117	0.4143	0.5237	0.2419
Classifier Chains - Random Forest	0.7209	0.4366	0.5439	0.2581
Binary Relevance - XGBoost	0.7274	0.5669	0.6372	0.3195
Classifier Chains - XGBoost	0.6992	0.5974	0.6443	0.3520
Binary Relevance - SVM	0.7614	0.5282	0.6237	0.3394
Classifier Chains - SVM	0.7543	0.5728	0.6511	0.3646

Table 5: Performance Comparison using BERT Embeddings

ment analysis of code-mixed text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5247–5260.

Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. 2021. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13919–13929.

Tsung-Yi Lin, Priya Goyal, Ross B Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

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. 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, 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.