Zero at SemEval-2025 Task 11: Multilingual Emotion Classification with BERT Variants: A Comparative Study

Revanth Gundam
IIIT Hyderabad
revanth.gundam@research.iiit.ac.in

Abhinav Marri
IIIT Hyderabad
abhinav.marri@research.iiit.ac.in

Radhika Mamidi
IIIT Hyderabad
radhika.mamidi@iiit.ac.in

Abstract

Emotion detection in text plays a very crucial role in NLP applications such as sentiment analysis and feedback analysis. This study tackles two tasks: multi-label emotion detection, where the goal is to classify text based on six emotions (*joy, sadness, fear, anger, surprise*, and *disgust*) in a multilingual setting, and emotion intensity prediction, which assigns an ordinal intensity score to each of the above-given emotions.

Using the BRIGHTER dataset, a multilingual corpus spanning 28 languages, the paper addresses issues like class imbalances by treating each emotion as an independent binary classification problem. The paper first explores strategies such as static embeddings such as GloVe with logistic regression classifiers on top of it. To capture contextual nuances more effectively, we fine-tune transformer based models, such as BERT and RoBERTa. Our approach demonstrates the benefits of fine-tuning for improved emotion prediction, while also highlighting the challenges of multilingual and multi-label classification.

1 Introduction

Emotion detection in text is a fundamental task in natural language processing (NLP), which has various important applications like sentiment analysis, feedback analysis, and chatbots. Understanding the emotions present in textual data is important in digital communication, where body language and facial expressions will not available. The increased textual interactions on social media, news, and online forums have emphasized the need for accurate emotion detection systems.

Recent breakthroughs and results in NLP have significantly enhanced the performance of emotion detection models (Belay et al., 2025). Current-age transformer architecture-based models (Vaswani et al., 2017) use large-scale language representation learning to model intricate semantic and syntactic relations in the text.

In this paper, we address the task of **multi-label emotion detection** (Muhammad et al., 2025b) as part of a shared workshop challenge. The task consists of two tracks:

- **Track A: Multi-label Emotion Detection** – Given a target text snippet, the goal is to predict the perceived emotion(s) expressed by the speaker. Each of the text samples is labelled with a binary classification for six emotions: *joy, sadness, fear, anger, surprise*, and *disgust*. The model needs to determine whether each emotion is either present (1) or absent (0) for each sample.

- **Track B: Emotion Intensity Prediction** – Given both, a target text snippet and a specific emotion in the six, the objective of this task is to predict the **intensity** of the perceived emotion on a scale from 0 (no emotion) to 3 (high intensity). This task is a more challenging one as it requires fine-grained understanding and ranking of emotional expressions.

Class imbalance is a common challenge when collecting data for emotion classification, mostly due to the natural distribution of real-world occurrences. The methodology discussed in this paper tries to adequately address the class imbalance and multi- label generated nature of the data. Rather than treating the task as a multi-label classification, we split it into separate independent binary classification tasks, where one classifier is trained for each emotion. This method ensures that the model learns to classify each emotion separately, thus trying to avoid class imbalance problems. We also finetune transformer models such as as BERT (Devlin et al., 2019), RoBERTa (Conneau et al., 2019), and other such models to leverage their contextual representations and enhance prediction performance. Through the exploration of both static embeddings and transformer models fine-tuned from

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1181–1186

scratch, we want to compare the efficiency of various feature representations in multi-label emotion classification and intensity estimation.

Past studies in this area have explored the topics of sentiment analysis and emotion detection using multiple different types of approaches. (V P et al., 2023) analyzed many Malayalam YouTube comments using ML models and deep learning techniques. The methodology provided in (Talaat, 2023) proposes a hybrid BERT model for the task, which shows improved contextual understanding. (Deho et al., 2018) uses word embeddings for sentiment analysis, clearly showcasing the role of pretrained representations. These works provide some important information about the evolution of emotion detection techniques.

Further sections provide a more detailed overview of the dataset, methodology, and experimental results.

2 Dataset

The BRIGHTER (Muhammad et al., 2025a) dataset is a collection of multilabeled emotion-annotated datasets in various low-resource languages from Africa and Asia and high-resource languages such as English. BRIGHTER covers text data in 28 different languages, annotated by expert annotators and fluent speakers based on the presence of six different emotions: anger, fear, sadness, joy, surprise and disgust. Data was mainly collected from social media, news, speeches and literature. Annotation is done with the help of crowd-sourcing platforms such as Amazon Mechanical Turk for largely spoken languages and directly recruited speakers for low-resource languages. Each instance of BRIGHTER consists of a sample ID, text snippet, and labels that indicate the presence of a particular emotion. Instances are multi-labeled and are labeled from 0 to 3 depending on the intensity of the emotion present in the sentence. For the initial task, these labels are simplified to depict the presence (0) or absence (1) of a specific emotion and intensity is ignored.

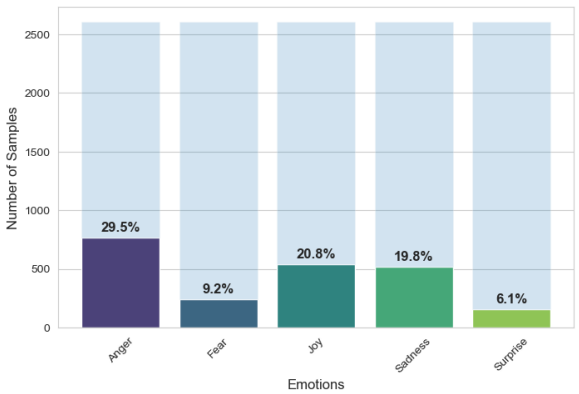

Figure 1: Percentage of positive samples in the English dataset.

We shall primarily use the English, Spanish, Russian, and Romanian datasets with the following emotions: anger, fear, joy, sadness, surprise and disgust. The English dataset does not have a label for disgust. Upon analyzing the class distribution, we observe an imbalance between positive and negative samples for nearly every emotion as depicted in Fig 1. The substantial variation in the class distribution comes from the method of choosing data from the named sources and also the amount of available data on platforms such as those. This is not only present across emotions in a language, but also across languages as expected.

3 Methodology

Due to the limited size of the dataset, it is rare to encounter samples that encompass all possible combinations of emotions across all the classes, as certain combinations may be inherently less likely to co-occur than others or may not be represented at all due to the natural distribution of real-world occurances. To address the observed class imbalances in the dataset, we split the classification task into independent binary prediction tasks to detect the presence of each emotion separately by utilizing distinct classifiers. The predictions are concatenated to obtain the final emotion representation vector.

3.1 Static and Frozen Embeddings

First, we explore static vector embeddings and frozen model embeddings to encode sentences as fixed-dimensional feature vectors. We then utilize these vectors as input data for logistic regression classifiers. A logistic regression classifier is a statistical model that predicts the probability of an input belonging to one of two classes. Here, it represents the presence or absence of a specific emotion. Two types of embeddings were tested, GloVe (Pennington et al., 2014) and BERT (Devlin et al., 2019):

- **Global Vectors for Word Representations (GloVe)**: Sentences are encoded as vectors

Model	Anger		Fear		Joy		Sadness		Surprise		Avg	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
GloVe	0.88	0.47	0.58	0.49	0.75	0.43	0.68	0.41	0.71	0.49	0.72	0.46
bert-base	0.89	0.69	0.77	0.76	0.83	0.76	0.80	0.76	0.80	0.75	0.82	0.74

Table 1: Performance of models using static and frozen embeddings.

Model	Anger		Fear		Joy		Sadness		Surprise		Avg	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
bert-base	0.89	0.76	0.76	0.76	0.86	0.80	0.81	0.76	0.81	0.78	0.83	0.79
RoBERTa	0.89	0.77	0.76	0.79	0.87	0.82	0.82	0.80	0.83	0.79	0.84	0.81
bert-large	0.90	0.77	0.80	0.80	0.83	0.79	0.84	0.81	0.84	0.80	0.84	0.82

Table 2: Performance of fine-tuned models.

by averaging the individual word vectors obtained from pre-trained GloVe representations.

- **Bidirectional Encoder Representations from Transformers (BERT - bert-base-uncased)**: Sentences are encoded using contextualized embeddings from the frozen pre-trained BERT model. Specifically, the `[CLS]` token representation from the final hidden layer is used as a fixed-dimensional sentence embedding, capturing contextual meaning and syntactic nuances more effectively than static word embeddings.

3.2 Fine-tuning Pre-trained Models

Static embeddings and pre-trained models are limited in the information they can capture. They are incapable of capturing task-specific variations. To overcome such limitations, we switch to fine-tuning models on our dataset, allowing them to adapt their representations to fit the patterns in the particular problem. We fine-tune the following models:

- **bert-base-uncased**: The base model built on the transformer architecture.
- **RoBERTa**: RoBERTa (Conneau et al., 2019) is a variant of BERT trained by eliminating the next sentence prediction task. The model has shown better performance on various NLP tasks.
- **bert-large**: bert-large-uncased is a larger model with 340 million parameters, compared to the base model with 110 million parameters. The increased number of layers allows us to capture more complex linguistic patterns.

3.3 Expanding to Multiclass Classification

To accommodate for multiclass classification, where the prediction can range from 0-3, which signifies the intensity of the emotion present, we modify the final layer of our models. Previously, we used a final layer of size 2 to represent classification between 0 and 1. Changing the size of this layer to 4 allows us to predict between classes 0-3.

3.4 Expanding to Multilingual Classification

To extend our model to multilingual classification so that it can process texts in several languages like Spanish, Russian, and Romanian along with English, we use Multilingual BERT (mBERT) also released by (Devlin et al., 2019). mBERT is a model of BERT trained on 104 languages on the basis of Wikipedia data and is thus fit for cross-lingual transfer learning. By utilizing mBERT, our model can handle text in various languages without the need for language-specific models. This is especially convenient for use cases where training data with the label is scarce in non-English languages.

3.5 Experimental Setup

The experiments were conducted using transformer-based models fine-tuned on the BRIGHTER dataset. The dataset was tokenized with a maximum sequence length of 128 using padding and truncation.

For optimization, we used the AdamW optimizer with a learning rate of 1×10^{-5}. Training was performed for 2 epochs with a batch size of 10.

4 Results

The evaluation of the models was conducted using different metrics based on the task at hand. For

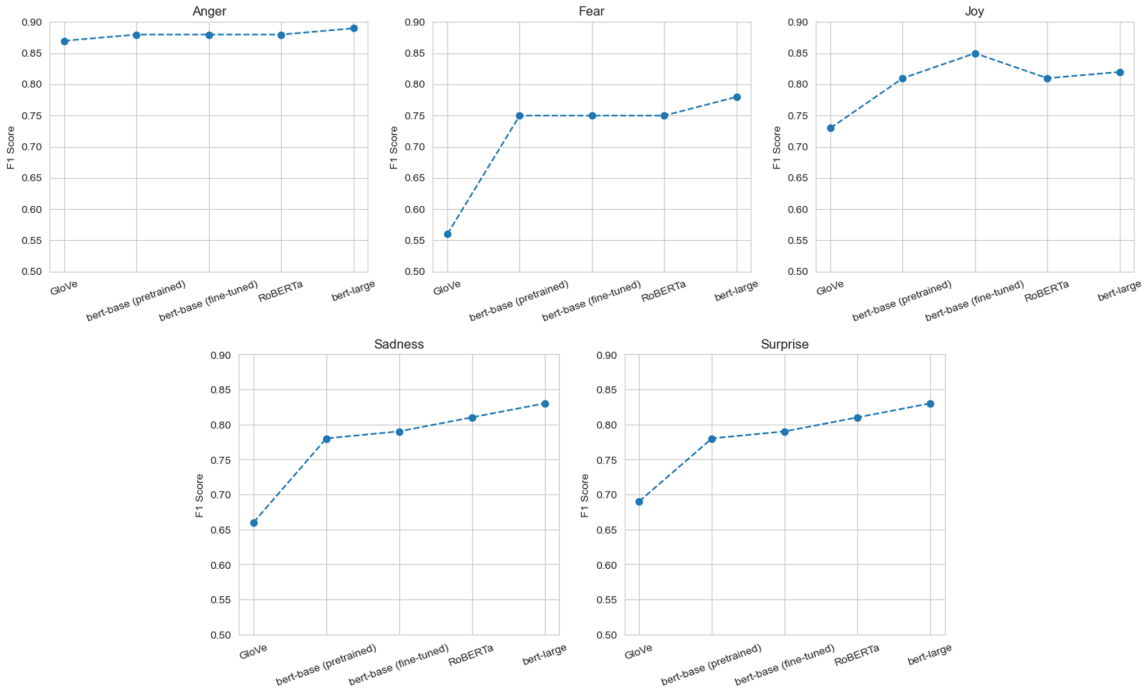

Figure 2: F1-scores across the models for each emotion.

binary emotion classification task, **accuracy** and **F1-score** were used. Accuracy provides an overall measure of the correct predictions, but since emotion classes are imbalanced in the present dataset, F1-score is a more informative metric as it takes into account both precision and recall.

For the task of emotion intensity prediction, the performance is measured using **Pearson correlation coefficient (r)**. This metric evaluates the linear relationship between the predicted values and the actual intensity values.

4.1 Binary Emotion Classification Performance

The results for binary classification are presented in Table 1 and Table 2. Results demonstrate that fine-tuning transformer-based models significantly improves performance over static or frozen embeddings. GloVe embeddings yield significantly lower F1 scores across all emotions, whereas fine-tuned BERT models achieve substantially higher F1 scores. The best-performing model, **bert-large**, attains the highest average F1 scores across all emotions, followed very closely by RoBERTa. This trend suggests that increasing model size and using contextualized embeddings contribute to better generalization in emotion classification. This can also be visualized in Fig 2.

4.2 Multiclass Emotion Intensity Prediction

For the task which involves predicting the intensity of a given emotion on a scale from 0 to 3, Table 3 presents the results for bert-large model on English. The **bert-large** model achieves the highest average Pearson correlation ($r = 0.6129$), outperforming frozen bert-base model embeddings ($r = 0.5315$). This indicates that fine-tuning enhances the model's ability to capture nuanced emotional intensity variations.

Language	Emotion	Score
English	Anger	0.2635
	Fear	0.7288
	Joy	0.6924
	Sadness	0.7457
	Surprise	0.6339
Average Pearson r		**0.6129**

Table 3: Performance of Bert Large in Track B on English Language

4.3 Multilingual Emotion Classification

The multilingual classification results, as shown in Table 4, show variations in performance across languages. Romanian and Russian data sets show

Language	Anger	Disgust	Fear	Joy	Sadness	Surprise	Micro F1	Macro F1
Russian	0.4961	0.6115	0.4031	0.6404	0.0	0.0	0.4798	0.3585
Spanish	0.1551	0.1364	0.8558	0.6824	0.8016	0.5505	0.5674	0.5303
Romanian	0.0	0.0	0.7398	0.8326	0.6223	0.0	0.5376	0.3658

Table 4: Emotion Classification Metrics for Different Languages

lower F1 scores for emotions such as surprise and anger, while the Spanish show higher scores, especially for fear and sadness. This discrepancy can be attributed to how mBERT itself is trained, particularly the quality and quantity of training data originally used for the model. Since mBERT is pre-trained on large-scale multilingual corpora, its effectiveness varies across languages depending on their representation in the training data. The macro F1 scores indicate that Spanish achieves the highest overall performance, suggesting that better model pre-training for certain languages leads to improved emotion classification results.

4.4 Overall Analysis

Across all experiments, fine-tuned transformer models evidently outperform static embeddings, reinforcing the importance of task-specific adaptation. Larger models like **bert-large** and **RoBERTa** demonstrate superior performance, benefiting from deeper contextual representations. The imbalance in dataset labels remains a challenge, particularly for low-resource languages, impacting overall classification efficacy.

5 Conclusion

The evaluation of transformer-based architecture models for the tasks of emotion detection and intensity prediction, highlights the advantages of fine-tuning over using static embeddings. BERT-based models, especially bert-large, consistently outperformed other models in both the tasks, achieving the highest F1-scores and Pearson correlations on average. RoBERTa also demonstrated competitive performance, particularly in the binary classification task, due to its optimized pre-training approach. In multilingual classification, mBERT facilitated cross-lingual generalization, though performance did vary depending on language representation in the pre-training corpus.

Across all tasks, larger models with deeper contextual representations provided the better results, reinforcing the impact of the size of models and training methods. These findings bring out the effectiveness of transformer models in emotion classification and suggest that more advancements in model architecture and quality of pre-training data could yield even better results.

Limitations

Despite the improvements achieved through fine-tuning transformer-based models, several limitations persist in our approach. One major challenge is **class imbalance** within the dataset. Certain emotions, particularly those less frequently expressed, have significantly fewer training samples. This imbalance leads to biased learning, where the model performs better on more common emotions while struggling with underrepresented ones. In future work, a more balanced dataset with uniform representation across emotions could help mitigate this issue. Additionally, techniques such as oversampling, under-sampling, and synthetic data generation could be explored to enhance model robustness.

Another limitation is that the study mainly relies on **BERT-based models**. While models like **BERT, RoBERTa**, and **bert-large** show good results, using more advanced architectures could further improve performance. Models such as **DeBERTa**, which introduces disentangled attention, and **T5 or GPT-based models**, which utilize generative learning strategies, might be better suited for capturing the complex emotional nuances.

Furthermore, the paper's current approach depends on **supervised learning** which requires labeled data. In low-resource settings, obtaining high-quality annotated datasets is quite challenging. Future research could explore semi-supervised and self-supervised learning techniques to leverage unlabeled data effectively. Pre-training on larger, diverse emotion-rich corpora before fine-tuning on task-specific data might enhance model adaptability across languages and emotional contexts.

References

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

B Oscar Deho, A William Agangiba, L Felix Aryeh, and A Jeffery Ansah. 2018. Sentiment analysis with word embedding. In *2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST)*, pages 1–4. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Amira Samy Talaat. 2023. Sentiment analysis classification system using hybrid BERT models. *J. Big Data*, 10(1):110.

Abeera V P, Dr. Sachin Kumar, and Dr. Soman K P. 2023. Social media data analysis for Malayalam YouTube comments: Sentiment analysis and emotion detection using ML and DL models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.