Trans-Sent at SemEval-2025 Task 11: Text-based Multi-label Emotion Detection using Pre-Trained BERT Transformer Models

Zafar Sarif[1], Md Sharib Akhtar[1], Abhishek Das[1], Dipankar Das[2]
[1] Aliah University, Kolkata
[2] Jadavpur University, Kolkata
zsarifau@gmail.com

Abstract

This paper presents Trans-Sent, our system developed for Track A: Multi-label Emotion Detection of SemEval-2025 Task 11: Bridging the Gap in Text-based Emotion Detection. The aim of this task is to predict the emotions that a speaker may convey i.e. perceived emotions. A target text snippet is given and our goal is to label the text by 1 and 0 as 'yes' and 'no' for different emotional classes like joy, sadness, fear, anger, surprise and/or disgust. Trans-Sent is a Transformer-based sentiment extraction model. It uses various pre-trained Bidirectional Encoder Representations from Transformers or BERT models for individual language to generate emotions and classify the emotion-labels. We have participated in the task for 7 different languages and achieve competitive results. Our system gives best result for Russian language, as it is ranked ninth among all ranked teams.

1 Introduction

Research on natural language processing is getting a lot of attention recently. Nevertheless, the majority of research is still limited to a few languages with plenty of available training data. Emotions are both recognizable and mysterious. We all express and control our emotions on a daily basis. However, these are intricate, subtle, and occasionally challenging to describe (Wiebe et al., 2005, Mohammad et al., 2018). Some basic emotions are —joy, sadness, anger, fear, surprise, and disgust etc. Emotion recognition is a broad term encompassing tasks like detecting a speaker's emotions, identifying emotions in text, and recognizing emotions evoked in a reader. Since people express and perceive emotions in complex, variable ways, it is impossible to determine one's feelings with absolute certainty (Mohammad, 2022). Here the goal of this task is to determine perceived emotion i.e. what emotion most people will think the speaker may be feeling given a sentence or short text snippet uttered by the speaker.

Two popular methods for processing categorical data in machine learning are multi-class and multi-label classification. In multi-class classification, each instance is assigned to only one category from a predefined set of mutually exclusive classes. One instance of a multi-class problem is sentiment analysis, which involves classifying a text as either good, negative, or neutral. Labels are not mutually exclusive in multi-label classification, however, since an instance can be a part of more than one category at the same time (Yen et al., 2016). This is especially helpful for jobs like emotion recognition, where a single text can simultaneously convey several emotions, like surprise and joy. Multi-label classification frequently uses sigmoid activation, whereas multi-class classification usually uses softmax activation.

In this paper, we have introduced Trans-Sent, a Transformer-based Sentiment extraction model to extract multi-label emotion from the text data. It uses various pre-trained Bidirectional Encoder Representations from Transformers or BERT models for individual language to generate emotions (Acheampong et al., 2021) and classify the emotion-labels. Some preprocessing measures are used on training data, then it is oversampled, as mostly the dataset is imbalanced for various emotion-labels. Then, feature extractions and feature engineering are performed to extract sentence-level contexts. Finally, different pre-trained BERT models are used to extract emotions.

2 Dataset

BRIGHTER (Muhammad et al., 2025) is a large-scale collection of multi-label emotion recognition datasets covering 28 languages, with a strong focus on low-resource languages from Africa, Asia, Eastern Europe, and Latin America. Unlike many

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 893–898

existing datasets that primarily focus on high-resource languages, BRIGHTER was specifically developed to bridge this gap. The dataset includes text samples sourced from various domains, including social media platforms like Twitter, Reddit, and Weibo, as well as news articles, literary texts, personal narratives, and speeches. Each text instance is annotated by fluent speakers and labeled with one or more of six basic emotions—joy, sadness, anger, fear, surprise, and disgust—along with a neutral category. The dataset has been carefully curated through preprocessing, quality control measures, and reliability assessments to ensure accuracy. By making this dataset publicly available, the authors hope to encourage research on emotion recognition across diverse languages and cultural contexts, improving the inclusivity and effectiveness of NLP applications (Muhammad et al., 2025).

EthioEmo (Belay et al., 2025) is another multi-label emotion classification dataset specifically designed for four Ethiopian languages: Amharic, Afan Oromo, Somali, and Tigrinya. It includes six core emotions—anger, disgust, fear, joy, sadness, and surprise—along with a neutral class. The dataset was constructed using text collected from multiple sources, including Twitter (X), Facebook comments, YouTube comments, and news headlines, ensuring a diverse range of emotionally expressive content. To maintain high annotation quality, native speakers were employed to label the data.

In this context, the task organizers have proposed a baseline model in the task description paper (Muhammad et al., 2025) for all the languages. But we have participated in the task for 7 languages – Amharic, German, English, Hindi, Marathi, Russian, Romanian. In the dataset, there are train, dev and test set of text-data. Initially, training data was labelled with the emotions, but dev and test data had only un-labelled texts. Later, the organizers published the labelled version of dev set too. We have used both the labelled train and dev data to train our model and tested the performance of the model by test set of data.

3 System Overview

Trans-Sent, our system is built on BERT (Bidirectional Encoder Representations from Transformers), fine-tuned for multi-label emotion classification. Initially, we experimented with traditional machine learning techniques (Siam et al., 2022), using TF-IDF embeddings with models such as Logistic Regression, Decision Trees, and Random Forest etc. To handle the multi-label nature of the task, we applied Label Powerset and Classifier Chaining (Dembczynski et al., 2012) techniques, which, despite providing structured predictions, were ineffective due to their inability to capture the semantic meaning and context of words. The architecture of our proposed Trans-Sent model is shown on Figure 1.

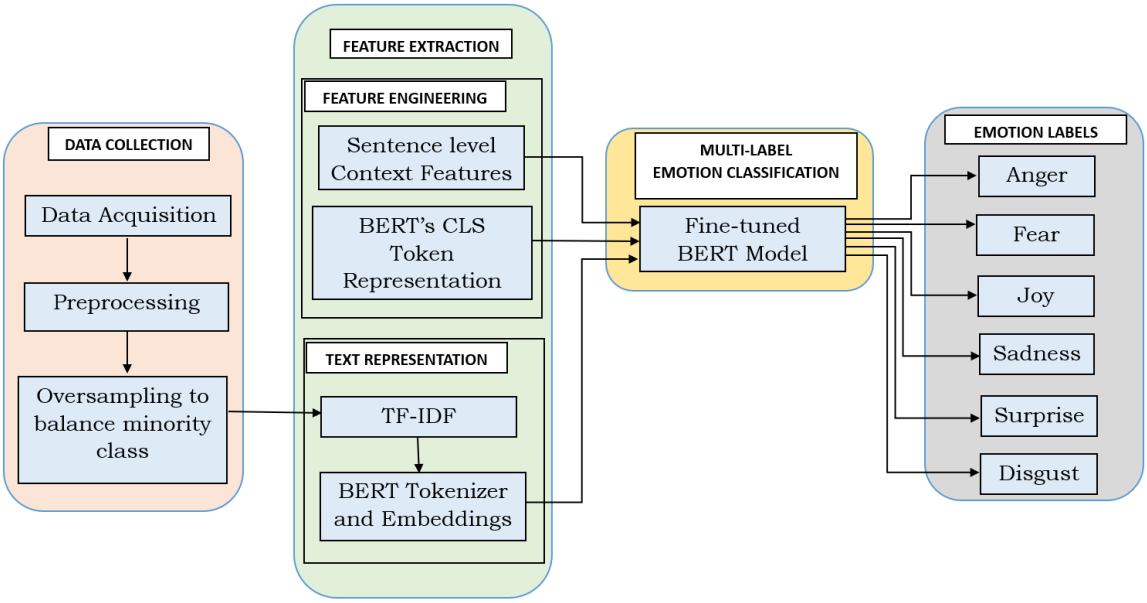

Figure 1: The architecture of the proposed approach, Trans-Sent

Input Sentence: "I am feeling very sad and angry today."

Tokenized input:

['[CLS]'	'i'	'am'	'feeling'	'very'	'sad'	'and'	'angry'	'today'	'[SEP]']

Label Encoding: [0, 1, 0, 0, 1, 0] # Corresponding to [Anger, Joy, Disgust, Fear, Sadness, Surprise]

Figure 2: Tokenization, attention-mask and label-encoding with an example.

The primary issue we encountered was that traditional models relied on word frequency-based representations, which struggled with incomplete or ambiguous text samples. Given that emotions are highly context-dependent and often co-occur, we needed a model capable of understanding semantic nuances and bidirectional dependencies. This led us to transition to a BERT-based approach, leveraging the pre-trained models like bert-base-uncased, ai-forever/ruBert-base etc and fine-tuning it for our specific task. The methodology behind working of Trans-Sent is explained below.

3.1 Data Preprocessing

Few data preprocessing techniques are applied before feeding the text into the model, such as data cleaning, tokenization, input formatting and level encoding. Removal of special characters, lowercasing, punctuation normalization are used for text cleaning and tokenizers, such as WordPiece tokenizer for bert-base-uncased model, are used for tokenization. Each text sample or token is converted into 'input_id's and 'attention-masks'. Tokenized sentences are converted into numerical form and attention-mask is introduced to indicate valid tokens (1) and padding (0). Finally, one-hot encoding is done to get the multi-label emotions i.e. a single sentence can have multiple 1s and 0s for different emotions. Various steps in data preprocessing are explained in Figure 2 with an example.

3.2 Oversampling

Training data set for various languages are imbalanced i.e. some labels are occurring far more frequently than others. Oversampling technique is used to improve performance and erase biasness of models trained by these data. As an example, train set of English (eng) data is highly imbalanced and it is shown in Figure 3. One of the most common method, Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) is used by us to perform oversampling. It focuses on generating synthetic samples near the decision boundary to improve class separability. This eliminates the possibility of our model to become biased for any label.

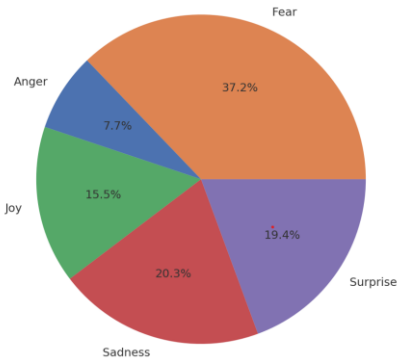

Figure 3: Imbalanced training data for English

3.3 BERT-Model Architecture

As a BERT is a transformer-based model (Tang et al., 2020) that utilizes bidirectional self-attention, enabling it to process words in context rather than in isolation. The architecture of BERT (Devlin et al., 2019) consists of multiple layers of self-attention mechanisms, making it highly effective for tasks requiring deep linguistic understanding. Our implementation uses the BertForSequence Classification model, which consists of a BERT Encoder, a Dropout Layer and a Classifier Head, a fully connected linear layer that maps the 768-dimensional BERT embeddings to five/six output neurons, corresponding to the five/six possible emotion labels. Different pre-trained BERT models[1] used for seven different languages are mentioned in Table 1.

[1] https://huggingface.co/models

Text in Language	BERT Model Used
Amharic (amh)	rasyosef/bert-medium-amharic
German (deu)	dbmdz/bert-base-german-cased
English (eng)	bert-base-uncased
Hindi (hin)	l3cube-pune/hindi-bert-scratch
Marathi (mar)	l3cube-pune/marathi-bert
Russian (rus)	ai-forever/ruBert-base
Romanian (ron)	dumitrescustefan/bert-base-romanian-cased-v1

Table 1: List of Pre-trained BERT models used for different languages.

4 Experimental Setup

To fine-tune the BERT-based model for multi-label emotion classification, we employed a structured training strategy designed to optimize performance while preventing overfitting. The optimization process was carried out using the AdamW optimizer, an improved variant of Adam that includes weight decay correction to mitigate over-regularization of important parameters. The learning rate was set to 2e-5, a commonly used value for fine-tuning transformer-based models, ensuring a balanced trade-off between convergence speed and stability. A batch size of 8 was chosen, considering the computational constraints of training large-scale transformer models while maintaining a sufficient number of samples for each gradient update.

To ensure effective learning, the model was trained for 3 epochs, as preliminary experiments indicated that performance plateaued beyond this point, and additional epochs led to minimal improvement while increasing the risk of overfitting. Weight decay was set to 0.01 to regulate the model's parameters and reduce the effect of less significant features, thereby enhancing generalization.

5 Results and Analysis

5.1 Evaluation Metric

For evaluation, we implemented an epoch-wise evaluation strategy, where the model's performance was assessed at the end of each epoch using standard classification metrics such as macro F1-score and micro F1-score. This approach allowed us to monitor the model's learning progression, detect potential overfitting or underfitting, and adjust hyperparameters if necessary. The final model selection was based on the best-performing checkpoint, ensuring optimal generalization to unseen data.

5.2 Results

Initially, training is done on train data and results are checked for dev set of data, the results are submitted through portal and then a score is generated by the organizer as per macro F1 and shown on the leaderboard. Once this phase is over, the final labelled version of dev data set for all languages are also released. For the final phase, we have used train and labelled dev set of data for training our model and testing its performance for test data. The performance of Trans-Sent is shown in a tabular form in Table 2.

Language	Anger	Disgust	Fear	Joy	Sadness	Surprise	Micro F1	Macro F1
Amharic (amh)	0.6485	0.7452	0.0010	0.7042	0.7011	0.4828	0.6821	**0.547**
German (deu)	0.7389	0.6743	0.0981	0.6563	0.5533	0.1576	0.62	**0.4797**
English (eng)	0.5629	-	0.8184	0.7209	0.7234	0.6936	0.7439	**0.7038**
Hindi (hin)	0.8185	0.8273	0.8522	0.8289	0.7665	0.8531	0.823	**0.8244**
Marathi (mar)	0.7648	0.9091	0.8148	0.747	0.7773	0.8227	0.7932	**0.8029**
Russian (rus)	0.9061	0.8571	0.9327	0.9155	0.8192	0.871	0.8855	**0.8829**
Romanian (ron)	0.5536	0.6627	0.8512	0.9562	0.7544	0.4735	0.7309	**0.7086**

Table 2: Performance of Trans-Sent for different languages on test data

In the result, it shows the performance of our model for different labels and micro F1 as well as macro F1. The organizers have chosen Macro F1 score as final score of any model. And on that basis, final ranking is published by them. Trans-Sent has performed reasonably well as ranked best for Russian language and stood 9th rank amongst all rank holders.

5.3 Analysis

From the result Table 2, we observe that different languages exhibit varying performance in multi-label emotion classification. The Macro F1 scores indicate that languages like Russian (0.8829), Marathi (0.8029), and Hindi (0.8244) perform better, whereas languages like German (0.4797) and Amharic (0.547) lag behind. Several factors contribute to these differences:

A) Morphological Complexity

Some languages have complex grammar and morphology, making it harder for models to generalize. For example, German has compound words and flexible word order, which could make sentence structures more ambiguous for emotion classification models. On the other hand, Hindi has more predictable syntactic patterns, which contributed to better results.

B) Pretrained Model Support

Languages with stronger NLP resources and pre-trained embeddings (like English, Russian, and Hindi) benefit from better contextual understanding. In the contrary, for Marathi language the pre-trained model misses various contextual understandings, which leads to the performance drop for this language.

C) Emotion Representation Across Cultures

Emotional expression varies by culture. Some languages may have clearer distinctions between emotions, while others might use the same words for multiple emotional states. For example, in Russian and Hindi, emotion-laden words might be more explicitly defined, aiding classification.

D) Tokenization Challenges

Languages with rich inflections and complex scripts (like Amharic) face difficulties in tokenization, leading to suboptimal embeddings. In contrast, languages with simpler tokenization (like Russian and Hindi) allow the model to capture meaning more effectively.

6 Conclusion

We introduced Trans-Sent, a Transformer-based model for multi-label emotion detection in SemEval-2025 Task 11, leveraging pre-trained BERT models to classify joy, sadness, anger, fear, surprise, and disgust across seven languages. Our approach combined data preprocessing, SMOTE-based oversampling, and fine-tuning, achieving competitive results, with Russian ranking ninth overall. Performance varied across languages due to morphological complexity, availability of pre-trained models, cultural nuances, and tokenization challenges. While Russian, Marathi, and Hindi performed well, German and Amharic faced difficulties due to grammatical complexity and ambiguous sentence structures. We have also highlighted the impact of data imbalance, as some emotions appeared more frequently than others, influencing classification accuracy. Traditional machine learning models struggled with the nuances of multi-label classification, reaffirming the effectiveness of Transformer-based architectures.

Despite its strengths, Trans-Sent has room for improvement. Future work could explore better data augmentation, cross-lingual learning, and ensemble models to refine classification. Additionally, incorporating context-aware modeling and multimodal data could enhance accuracy. Expanding to low-resource languages through domain-specific fine-tuning is another promising direction.

References

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, *39*, 165-210.

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018, June). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 1-17).

Mohammad, S. M. (2022). Best practices in the creation and use of emotion lexicons. *arXiv preprint arXiv:2210.07206*.

Yen, I. E. H., Huang, X., Ravikumar, P., Zhong, K., & Dhillon, I. (2016, June). Pd-sparse: A primal and

dual sparse approach to extreme multiclass and multilabel classification. In *International conference on machine learning* (pp. 3069-3077). PMLR.

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, *54*(8), 5789-5829.

Muhammad, S. H., Ousidhoum, N., Abdulmumin, I., Wahle, J. P., Ruas, T., Beloucif, M., de Kock, C., Surange, N., Teodorescu, D., Ahmad, I. S., Adelani, D. I., Aji, A. F., Ali, F. D. M. A., Alimova, I., Araujo, V., Babakov, N., Baes, N., Bucur, A.-M., Bukula, A., … Mohammad, S. M. (2025). BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages. arXiv.

Belay, T. D., Azime, I. A., Ayele, A. A., Sidorov, G., Klakow, D., Slusallek, P., Kolesnikova, O., & Yimam, S. M. (2025). Evaluating the capabilities of large language models for multi-label emotion understanding. Proceedings of the 31st International Conference on Computational Linguistics, 3523–3540. Association for Computational Linguistics.

Muhammad, S. H., Ousidhoum, N., Abdulmumin, I., Yimam, S. M., Wahle, J. P., Ruas, T., Beloucif, M., De Kock, C., Belay, T. D., Ahmad, I. S., Surange, N., Teodorescu, D., Adelani, D. I., Aji, A. F., Ali, F., Araujo, V., Ayele, A. A., Ignat, O., Panchenko, A., Zhou, Y., & Mohammad, S. M. (2025). SemEval Task 11: Bridging the gap in text-based emotion detection. Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025). Association for Computational Linguistics.

Siam, A. I., Soliman, N. F., Algarni, A. D., Abd El-Samie, F. E., & Sedik, A. (2022). Deploying machine learning techniques for human emotion detection. *Computational intelligence and neuroscience*, *2022*(1), 8032673.

Dembczyński, K., Waegeman, W., & Hüllermeier, E. (2012). An analysis of chaining in multi-label classification. In *ECAI 2012* (pp. 294-299). IOS Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Tang, T., Tang, X., & Yuan, T. (2020). Fine-tuning BERT for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access*, *8*, 193248-193256.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

A Appendix

A.1 Result on dev Set

As mentioned earlier, the organizers of the task has provided dataset, which consists of train, dev and test data set. Initially, we trained our model on train set and then tested the model on dev set. It eventually becomes labeled and generates a .csv file. To check the performance i.e. accuracy, micro F1 and macro F1 etc of the model, this file was submitted to the portal. Finally, a score was generated. Based on those scores our model was updated technically. During that we have achieved the result mentioned in Table 3. We have participated in 4 languages only.

It is observed that the results are quite similar for both dev set and our final result, which we got for test set. Only for Marathi text a significant drop of performance is seen as macro F1-score gets decreased. For test set we achieved macro F1-score 0.8029, but on dev set it was 0.9066. Though it's very tough to analyze this performance drop, we pointed out a few probable reasons for this in the analysis section. Lack of contextual awareness of the pre-tuned model about Marathi language, cultural differences on emotions for various Marathi spoken people are some reasons we can list out here.

Language	Amharic (amh)	English (eng)	Hindi (hin)	Marathi (mar)
Anger	0.637	0.6207	0.8276	0.88
Disgust	0.703	-	0.6667	0.9524
Fear	0.001	0.7883	0.9333	0.9286
Joy	0.6556	0.6667	0.72	0.9189
Sadness	0.658	0.6875	0.6667	0.8
Surprise	0.3429	0.7143	0.8421	0.96
Micro F1	0.6453	0.724	0.7862	0.9006
Macro F1	0.4994	0.6955	0.7761	0.9066

Table 3: Performance of Trans-Sent for different languages on dev set