Tarbiat Modares at SemEval-2025 Task 11: Tackling Multi-Label Emotion Detection with Transfer Learning

Sara Bourbour[1], Maryam Gheysari[2], Amin Saeidi Kelishami[2]
Tahereh Talaei[2], Fatemeh Rahimzadeh[3], Erfan Moeini[2]
[1]School of Industrial and Systems Engineering, Tarbiat Modares University,
[2]Computer Engineering Department, Sharif University of Technology,
[3]School of Electrical and Computer Engineering, University of Tehran,
s.bourbour@modares.ac.ir, Maryamgheysari75@gmail.com, amin.saeidi.1997@gmail.com
taheretalaei@gmail.com, fatemehra10@gmail.com, emoeini@ce.sharif.edu

Abstract

The SemEval-2025 Task 11 addresses multi-label emotion detection, classifying perceived emotions in text. Our system targets Amharic, a morphologically complex, low-resource language. We fine-tune LaBSE with class-weighted loss for multi-label prediction. Our architecture consists of: (i) text tokenization via LaBSE, (ii) a fully connected layer with sigmoid activation for classification, and (iii) optimization using BCEWithLogitsLoss and AdamW. Ablation studies on class balancing and data augmentation showed that simple upsampling did not improve performance, highlighting the need for more sophisticated techniques. Our system ranked 14th out of 43 teams, achieving 0.4938 accuracy, 0.6931 micro-F1, and 0.6450 macro-F1, surpassing the task baseline (0.6383 macro-F1). Error analysis revealed that anger and disgust were well detected, while fear and surprise were frequently misclassified due to overlapping linguistic cues. Our findings underscore the challenges of multi-label emotion detection in low-resource languages.

1 Introduction

Emotion detection is a key task in NLP with applications ranging from social media monitoring to mental health and human-computer interaction. Unlike sentiment analysis, it identifies nuanced emotions such as anger, joy, sadness, fear, disgust, and surprise. SemEval-2025 Task 11 (Muhammad et al., 2025a,b) focuses on perceived emotions shaped by cultural and contextual cues.

We tackle this challenge in Amharic, a low-resource and morphologically complex language, by exploring the effectiveness of transfer learning in this setting.

We fine-tune LaBSE (Feng et al., 2020), a multilingual sentence embedding model, using a dropout-enhanced dense layer and BCEWithLogitsLoss. Our method avoids hand-crafted features, instead relying on pre-trained embeddings.

Our decision to focus on Amharic was motivated by both strategic and linguistic considerations. As native speakers of Persian—a low-resource and morphologically rich language—we were naturally drawn to Amharic, which shares similar linguistic challenges and characteristics. Although Persian was not included in the competition's dataset, Amharic was available, and we saw this as an opportunity to engage with a language that, like Persian, is often underrepresented in NLP research. Furthermore, we anticipated that the low-resource nature of Amharic might lead many teams to overlook it, which strengthened our motivation to select it and address the gap.

Despite solid performance on high-confidence emotions, our model struggled with overlapping expressions and class imbalance. It ultimately ranked 14th out of 43 teams.

Code is available at: https://github.com/Amin-Saeidi/SemEval2025-Task11.

2 Background

Emotion detection has been widely explored in NLP, with recent advancements driven by deep learning and transformer-based models. Unlike early lexicon-based approaches (Mohammad and Turney, 2013), modern methods leverage contextual embeddings from models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLM-R (Conneau et al., 2020). The SemEval-2025 Task 11 (Muhammad et al., 2025b) aims to advance perceived emotion detection by providing multilingual datasets and a structured evaluation framework. The task consists of three tracks. Track A, Multi-label Emotion Detection, involves assigning zero, one, or two emotion labels to a given text snippet. Track B, Emotion Intensity Estimation, focuses on predicting the intensity of a given emotion.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1168–1173
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Track C, Cross-lingual Emotion Detection, aims to transfer knowledge between languages, facilitating emotion recognition across linguistic boundaries.

Our work focuses on Track A, where each text snippet is labeled with multiple emotions as binary values (1 for presence, 0 for absence). This multi-label classification setup poses unique challenges due to emotion co-occurrence and ambiguous expressions. The dataset includes multiple languages, but we specifically study Amharic, leveraging the dataset from Belay et al. (2025), which provides a benchmark for emotion detection in low-resource languages.

Several studies have contributed to advancing multi-label emotion detection. Zhang et al. (2020) introduced a multi-modal framework capturing label dependencies, while Firdaus et al. (2020) proposed a dataset for emotion recognition in dialogues. A comprehensive review by Nandwani and Verma (2021) emphasized the role of cultural context in emotion perception.

Recent research has explored multi-label classification techniques to improve accuracy. Ni and Ni (2024) demonstrated models that effectively capture emotion correlations, while Wang et al. (2023) showed that modeling label dependencies enhances emotion recognition. For low-resource languages, Belay et al. (2025) introduced EthioEmo, highlighting the challenges of applying pre-trained models to Amharic.

For Amharic-specific emotion classification, Bayu et al. (2024) focused on deep learning for analyzing social media comments, while Birara (2024) evaluated LSTM, BiLSTM, CNN, and GRU models for multi-label classification. These studies demonstrate the potential of deep learning but also expose the need for better adaptation of transformer-based models for Amharic.

Our research builds on these foundational works, focusing on low-resource Amharic and applying transfer learning through LaBSE embeddings.

For model training, we utilize publicly available datasets and pre-trained models. Our experiments are based on the dataset *Evaluating the Capabilities of Large Language Models for Multi-label Emotion Understanding* (Belay et al., 2025). For feature extraction and classification, we employ LaBSE, a language-agnostic sentence embedding model, available via Hugging Face's `sentence-transformers/LaBSE` repository.

3 System overview

Our system is based on the Language-agnostic BERT Sentence Embedding (LaBSE) model, a multilingual transformer-based model designed for cross-lingual sentence representations (Feng et al., 2020). Given the nature of the SemEval-2025 Task 11 as a multi-label classification problem, we introduced specific modifications to adapt LaBSE for effective emotion detection. This section outlines our model architecture, preprocessing pipeline, and key challenges, along with the solutions implemented to address them.

3.1 Model Architecture

Our model architecture follows a structured pipeline, as shown in Figure 1. We fine-tuned LaBSE for multi-label classification by adding a fully connected layer atop the transformer encoder. To mitigate overfitting, we incorporated a dropout layer ($p = 0.4$) and applied early stopping based on validation loss. While we experimented with data augmentation and upsampling for class balancing (Section 5), their impact on final performance was minimal. However, dropout and early stopping effectively stabilized training and prevented excessive memorization of majority-class patterns.

The fully connected layer contains 768 × 6 neurons, where 768 is LaBSE's hidden size and 6 corresponds to the number of emotion categories. A sigmoid activation function outputs probability scores for each label. The architecture generates a contextualized embedding h from an input sentence x using LaBSE, passing it through additional layers to compute the final probability distribution:

$$h = LaBSE(x) \tag{1}$$

$$y = \sigma(Wh + b) \tag{2}$$

where $W \in R^{6\times768}$ is the fully connected layer's weight matrix, $b \in R^6$ is the bias term, and σ represents the sigmoid activation function. This setup enables independent probability estimation for each emotion label, making it well-suited for multi-label classification.

3.2 Loss Function and Optimization

We employed Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) for this multi-label classification problem. This loss function computes the error between the predicted probabilities and the true labels. The loss is computed as follows:

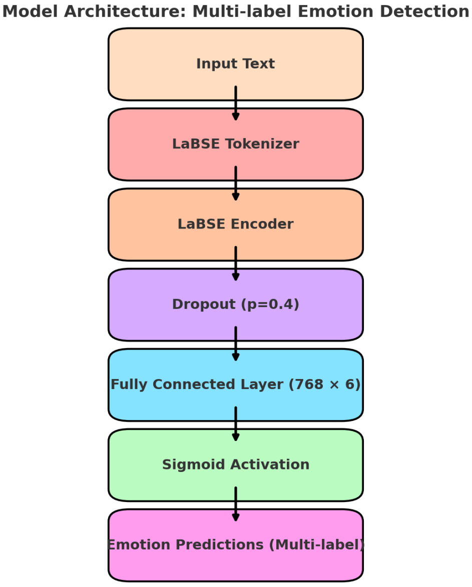
Model Architecture: Multi-label Emotion Detection
Input Text
LaBSE Tokenizer
LaBSE Encoder
Dropout (p=0.4)
Fully Connected Layer (768 × 6)
Sigmoid Activation
Emotion Predictions (Multi-label)

Figure 1: Schematic of our multi-label emotion detection model. The LaBSE encoder generates sentence embeddings, followed by dropout, a fully connected layer, and sigmoid activation for classification.

$$\mathcal{L} = -\sum_{i=1}^{6} [\hat{y}_i \log \sigma(y_i) + (1 - \hat{y}_i) \log(1 - \sigma(y_i))] \tag{3}$$

For optimization, we used the AdamW optimizer (Loshchilov and Hutter, 2019), which decouples weight decay from the optimization process. Unlike the standard Adam optimizer, which applies weight decay directly within the update rule, AdamW applies weight decay independently, preventing unintended updates to the learning rate. This decoupling results in better generalization and faster convergence, especially for transformer-based models, making AdamW a more suitable choice compared to the original Adam optimizer for this task.

3.3 Evaluation Metrics

We used several evaluation metrics to assess the performance of our model, which are well-suited for multi-label classification tasks and provide a comprehensive understanding of model performance. These metrics include Micro F1-score, which aggregates contributions of all classes and calculates the F1-score globally; Macro F1-score, which computes the F1-score for each class independently and averages them; Accuracy, which evaluates the percentage of correctly predicted labels; and the Confusion Matrix, which provides insight into class-wise predictions and misclassifications. Together, these metrics help us understand the overall performance of the system and identify areas where the model may be struggling.

3.4 Addressing Class Imbalance

Class imbalance was a significant challenge, with certain emotion labels underrepresented in the training data. To address this, we employed multiple strategies. First, we used class-weighted BCEWithLogitsLoss, assigning higher weights to underrepresented classes to enhance classification performance.

Additionally, we expanded the dataset using an Amharic sentiment dataset[1] containing approximately 9.4k Amharic tweets. Since it was originally annotated for sentiment analysis rather than multi-label emotion classification, we re-annotated the tweets using large language models (LLMs), specifically DeepSeek (DeepSeek-AI et al., 2025) and ChatGPT.

Due to the lack of publicly available multi-label emotion datasets in Amharic, we employed LLMs as practical tools for semi-automatic annotation. Their multilingual and culturally aware capabilities made them suitable for generating preliminary labels, which were manually reviewed by a trained Amharic educator. These models were not used as evaluators, but solely as bootstrapping tools to expand the training set under limited annotation resources.

Finally, we applied oversampling to improve minority-class representation and balance the emotion categories. The label distributions before and after augmentation and upsampling are presented in Tables 1, 2, and 3.

Table 1: Distribution of Emotion Labels in Main Data - TrainSet

Emotion	Not Exist (0)	Exist (1)
Anger	2360	1188
Disgust	2280	1268
Joy	3000	548
Fear	3439	109
Sadness	2777	771
Surprise	3397	151

[1] https://github.com/liyaSileshi/amharic-sentiment-analysis/tree/main

Table 2: Distribution of Emotion Labels After Data Augmentation - TrainSet

Emotion	Not Exist (0)	Exist (1)
Anger	3062	1574
Disgust	3237	1399
Joy	3997	639
Fear	3729	907
Sadness	3522	1114
Surprise	4298	338

Table 3: Distribution of Emotion Labels After Upsampling - TrainSet

Emotion	Not Exist (0)	Exist (1)
Anger	6891	3474
Disgust	6651	3714
Joy	8721	1644
Fear	10038	327
Sadness	8052	2313
Surprise	9912	453

3.5 Experiments with Alternative Architectures

We compared LaBSE with multilingual alternatives, including XLM-RoBERTa (Conneau et al., 2019) and a SentenceTransformer fine-tuned for Amharic retrieval (Belay et al., 2025). LaBSE outperformed both in F1-score, likely due to its robust cross-lingual representations, and was thus chosen as our final model.

4 Experimental Setup

4.1 Dataset and Splits

For this task, we used the dataset provided by SemEval-2025 Task 11, specifically focusing on the Amharic subset from *SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection* (Muhammad et al., 2025a). The dataset was divided into three splits: the training set contained 3,549 samples, the test set included 1,774 samples, and the validation set consisted of 592 samples. Table 4 presents the distribution of emotion labels across the dataset splits.

During hyperparameter tuning, we trained on the training set and used the validation set for evaluation and parameter selection. For the final model, we excluded the validation set and trained solely on the original training data, evaluating performance on the test set.

Table 4: Distribution of emotion labels across dataset splits

Emotion	Train	Dev	Test
Anger	1188	207	582
Sadness	771	127	355
Joy	549	93	276
Fear	109	22	54
Surprise	151	27	82
Disgust	1268	209	628

4.2 Preprocessing

Preprocessing was minimal, as the LaBSE tokenizer inherently handles multilingual text. We tokenized the input text using the LaBSE tokenizer without additional processing such as lowercasing, punctuation removal, or stopword filtering. The tokenized sequences were padded or truncated to a maximum length of 180 tokens to maintain computational efficiency while preserving meaningful context.

4.3 Model Training and Hyperparameter Tuning

We tuned hyperparameters iteratively based on validation F1-score and accuracy. The final model used a batch size of 64, a sequence length of 180, a learning rate of 1×10^{-5}, and AdamW optimizer with 1×10^{-6} weight decay. Training ran for 11 epochs using BCEWithLogitsLoss and was performed on CUDA.

Hyperparameter tuning was conducted through manual iterative adjustments based on validation performance. While this approach yielded competitive results, future work could explore more systematic tuning methods, such as grid search or Bayesian optimization, to refine parameter selection further.

4.4 External Tools and Libraries

The implementation relied on several external libraries for model training, preprocessing, and evaluation. The `transformers` library from Hugging Face (v4.36.1) was used for model loading and fine-tuning. PyTorch (v2.1.0) was utilized for deep learning computations, while Scikit-learn (v1.3.0) was employed for evaluation metrics such as F1-score, accuracy, and confusion matrices. Additionally, Pandas (v2.1.1) and NumPy (v1.25.0) were used for data handling and numerical computations.

Table 5: Performance metrics of various models on the dev set

Model	Using main data	Add data	Upsampling	Accuracy	F1-score (micro)	F1-score (macro)
LaBSE	Yes	No	No	0.512	0.701	0.652
LaBSE	Yes	Yes	No	0.472	0.673	0.632
LaBSE	Yes	No	Yes	0.478	0.646	0.611
xlm-r-retrieval-am	Yes	No	No	0.433	0.632	0.559
xlm-r-retrieval-am	Yes	Yes	No	0.433	0.632	0.559
xlm-r-retrieval-am	Yes	No	Yes	0.425	0.627	0.586
xlm-roberta-base	Yes	No	No	0.449	0.637	0.586

5 Results

5.1 Error Analysis and Observations

Table 6 summarizes the confusion matrix findings. The model performed well in classifying anger and disgust, achieving high true positives and low false negatives. However, it struggled with fear and surprise, which had low true positives and high false positives, indicating difficulty in distinguishing these emotions. Joy and sadness showed moderate performance, with room for improvement in reducing misclassifications.

Table 6: Error patterns across emotion categories

Observation	Finding
Strong perf.	Anger, Disgust (High TP, Low FN)
Weak perf.	Fear, Surprise (Low TP, High FP)
Moderate perf.	Joy, Sadness (Decent TP, some errors)

Class imbalance remained a challenge, as some emotions appeared far less frequently than others, limiting generalization. While automatic evaluation provided insights, a qualitative evaluation was conducted to assess the quality of LLM-generated labels.

A trained Amharic educator reviewed 100 augmented samples, comparing annotations from ChatGPT and DeepSeek based on contextual relevance, linguistic coherence, and cultural fit. DeepSeek was preferred in 72% of cases for its better handling of idiomatic expressions and emotional nuance, supporting its use for dataset expansion.

5.2 Quantitative Findings

Our best-performing model was evaluated using the official SemEval-2025 Task 11 metrics. As shown in Table 7. One possible factor affecting our performance is the complexity of multi-label emotion detection in Amharic, a morphologically rich and low-resource language.

We performed an ablation study to assess the impact of class balancing and data augmentation, as summarized in Table 5. The results indicate that upsampling the minority classes did not improve performance. In fact, models trained without class balancing achieved slightly better accuracy and F1 scores. This suggests that simple oversampling may have introduced redundant or noisy examples, which did not enhance generalization.

Table 7: Final model performance on the test set

Metric	Score
Accuracy	0.4938
F1-score (Micro)	0.6931
F1-score (Macro)	0.6450

6 Conclusion

This work explored multi-label emotion detection in Amharic using a transfer learning approach. We fine-tuned LaBSE with class-weighted loss and evaluated its performance on the SemEval-2025 Task 11 dataset. Our model achieved competitive results, ranking 14th out of 43 teams, with an accuracy of 0.4938, a micro-F1 score of 0.6931, and a macro-F1 score of 0.6450.

Through ablation studies, we found that simple upsampling for class balancing did not improve performance, suggesting the need for more effective data augmentation techniques. Error analysis revealed strong classification performance for anger and disgust but difficulties distinguishing fear and surprise, highlighting the challenge of context-dependent emotional expressions.

Future work can explore paraphrasing-based augmentation, adversarial training, and more adaptive loss functions to better handle class imbalance. Additionally, incorporating context-aware embeddings and expanding the dataset with high-quality labeled examples could further enhance multi-label emotion detection in Amharic.

References

Yeshimebet Bayu et al. 2024. Multi-label emotion classification on social media comments using deep learning.

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.

Tadesse Birara. 2024. *Deep Learning-Based Emotion Classification for Amharic Text*.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Preprint*, arXiv:1308.6297.

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. *arXiv 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.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81.

Yingying Ni and Wei Ni. 2024. A multi-label text sentiment analysis model based on sentiment correlation modeling. *Frontiers in Psychology*, 15:1490796.

Peiying Wang, Sunlu Zeng, Junqing Chen, Lu Fan, Meng Chen, Youzheng Wu, and Xiaodong He. 2023. Leveraging label information for multimodal emotion recognition. *Preprint*, arXiv:2309.02106.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3584–3593, Online. Association for Computational Linguistics.