Amado at SemEval-2025 Task 11: Multi-label Emotion Detection in Amharic and English Data

Girma Yohannis Bade [1,a], Olga Kolesnikova[2,a], José Luis Oropeza[3,a], Grigori Sidorov[4,a], Mesay Gemeda Yigezu[5,a]
[a]Centro de Investigaciones en Computación(CIC),
Instituto Politécnico Nacional(IPN), Miguel Othon de Mendizabal,
Ciudad de México, 07320, México.

Abstract

Recently, social media has become a platform for different human emotions. Although most existing works treat the user's opinions into a single emotion, the reality is that one user can have more than one emotion at a time, representing multiple emotions at the same time. Multi-label emotion detection is a more advanced and realistic approach, as it acknowledges the complexity of human emotions and their overlapping nature. This paper presents multi-label emotion detection in Amharic and English data. The work is part of SemEval2025 shared task 11, where tasks and datasets are offered by task organizers. To accomplish the aim of the given task, we fine-tune transformers base BERT model, passing through all different workflow pipelines. On unseen test data, the model evaluation achieved 0.6300 and 0.7025 an average macro F1-score for Amharic and English, respectively.

Keywords: Social media, Emotion, Multi-label, BERT

1 Introduction

In the current digital era, people can openly share their thoughts, sentiments, disagreements, opinions, and attitudes on websites, microblogs, and social media platforms. For a number of reasons, including decision-making, product analysis, customer feedback analysis, political promotions, marketing research, and social media monitoring, there is more interest in obtaining these user's attitudes and emotions all around (Belay et al., 2025). However, all those feedbacks come from different feelings and thoughts, thus becoming complicated and needs those emotions to be organized in certain sort of correlations (Bade et al., 2024b).

Emotion detection is a critical area of research in natural language processing (NLP) that focuses on identifying and understanding the emotional states expressed in text. Emotions play a vital role in human communication, influencing how messages are interpreted and how individuals respond to one another. Automatically detecting emotions from text has significant applications in various fields, such as mental health analysis, customer feedback evaluation, social media monitoring, and human-computer interaction. By understanding emotions, systems can provide more personalized and empathetic responses, enhancing user experience and decision-making processes.

In emotion detection, a single emotion refers to the identification of one dominant emotion in a given text. For example, a sentence like "I am so happy today!" would be classified as expressing the emotion "joy." This approach assumes that each text conveys only one primary emotion, which simplifies the task but may not fully capture the complexity of human expression. In reality, people often express multiple emotions simultaneously, as emotions are rarely isolated and can coexist in nuanced ways.

This is where multi-label emotion detection comes into play. Unlike single-emotion classification, multi-label emotion detection recognizes that a single text can express multiple emotions at once. For instance, a sentence like "I feel excited but also a bit nervous about the upcoming event" could be labeled with both "joy" and "fear." Multi-label emotion detection is a more advanced and realistic approach, as it acknowledges the complexity of human emotions and their overlapping nature. It allows for a richer and more accurate representation of the emotional content in text, making it particularly valuable for applications requiring a deeper understanding of human sentiment.

This paper presents our contribution to SemEval 2025 Shared Task 11 (Muhammad et al., 2025b), where the tasks and gold standard dataset were provided by the organizers. The shared task consists of three tracks; however, our team participated exclusively in Track_A, which focuses on detect-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1406–1410

ing multiple emotions from textual inputs across various languages (Belay et al., 2025; Muhammad et al., 2025a). Through this participation, we aim to advance the development of more inclusive and accurate emotion detection systems for Amharic and English data in detailed contexts.

2 Related Works

Recent researches on emotion detection have leveraged various approaches. Serrano-Guerrero et al. (2022) proposed a deep learning architecture based on a bidirectional gated recurrent unit with a multichannel convolutional neural network layer to tackle the issue of identifying numerous emotions from patient reports, collecting a sizable patient viewpoints from a website, and identified several emotions from these evaluations, achieving average accuracy of 95.82%. Deng and Ren (2020) addresses the multiple emotions detection in online social networks from a user-level view, finding emotion labels correlations, social correlations, and temporal correlations from an annotated Twitter data set. They Use a factor graph-based emotion recognition model to detect these multiple emotions, and finally it outperformed the baselines. For multi-label emotion classification, Ameer et al. (2023) examined the application of LSTMs as well as the refinement of Transformer Networks using Transfer Learning in conjunction with a single-attention network and a multiple-attention network. According to the experimental results, their innovative transfer learning models that used pre-trained transformers with and without multiple attention mechanisms were able to achieve an accuracy of 62.4%, surpassing the state-of-the-art on RoBERTa-MA (RoBERTa-Multi-attention).

Rathnayaka et al. (2019) offer a unique Pyramid Attention Network (PAN) based approach for microblog emotion identification, emphasizing the approach's benefit in capturing many emotions present in a single text by evaluating words from several angles, with an accuracy of 58.9%.

To help with context-based multi-label multi-task emotion detection, Bendjoudi et al. (2021) suggests a new deep learning architecture that emphasizes three key modules: body features extraction, scene features extraction, and fusion-decision. Furthermore, a new loss function called multi-label focal loss (MFL) was chosen to handle imbalanced data after comparing three continuous and three categorical loss functions to highlight the significance of synergy between loss functions in multi-task learning. It outperformed the state of the art on the less frequent labels and produced better results than any other combination. Likewise, Zhang et al. (2020) show the methods that have three parts: the general representation module, the emotion representation module, and the adversarial classifier. After incorporating the link between various emotions using emotion descriptors, the model employs adversarial training to avoid over-injecting emotion-relevant data into the shared layer, achieving a macro-average F1 scores of 50.21%, 41.33%, and 40.24% on the Chinese, English, and Indonesian datasets, respectively.

Belay et al. (2025) presented EthioEmo, a multi-label emotion classification dataset for four Ethiopian languages—Amharic (amh), Afan Oromo (orm), Somali (som), and Tigrinya (tir)—alongside the English dataset acquired from SemEval 2018 Task 1 to assess encoder-only, encoder-decoder, and decoder-only language models using zero and few-shot approaches of LLMs to fine-tune smaller language models, concluding that accurate multi-label emotion classification is still insufficient, even for high-resource languages like English, and that there is a significant performance gap between high-resource and low-resource languages.

Muhammad et al. (2025a) introduce BRIGHTER– a set of datasets with multilabeled emotion annotations in 28 languages. With instances from a variety of domains annotated by fluent speakers, BRIGHTER primarily covers low-resource languages from Africa, Asia, Eastern Europe, and Latin America. It also describes the processes involved in data collection and annotation, as well as the difficulties in creating these datasets. It reports various experimental results for intensity-level emotion recognition and monolingual and crosslingual multi-label emotion identification, and it concludes that the results of the investigation, both with and without the use of LLMs, showed significant variability in performance across languages and text domains.

3 Task and Dataset Descriptions

We used datasets that were especially intended for this job in order to detect multi-label emotions in both Amharic and English (Muhammad et al., 2025a; Belay et al., 2025). The class label conains multilple labels including anger, fear, joy, sadness,

and surprise except 'disgust' label belongs to only Amharic data. The task aims to categorize the given input text into either of these class labels, the input text may fall in more than one labels.

Text	Anger	Fear	Joy	Sadness	Surprise
You know what happens when I get one of these stupid ideas in my head	1	1	0	0	0
They don't fear death, and it seems they believe in reincarnation	0	1	0	0	1
My stomach even started giving me fits	0	1	0	1	0
Victim-less, non-violent crimes shouldn't be handled so harshly	1	0	0	0	0

Table 1: Four sample instances of input texts with their correspondence multi-labeled classes, while 1 denotes existance of emotion in that particular label and 0 not.

While Table 1 shows the overview of tasks, Table 2 shows the data sets statistics include training, validation, and test data with comprehensive distributions.

4 Methodology

In this section a comprehensive set of methods and methodologies is described. Our approach optimizes and fine-tunes transformer-based models of BERT-base to improve emotion detection performance. Figure 1 shows the methodology workflow.

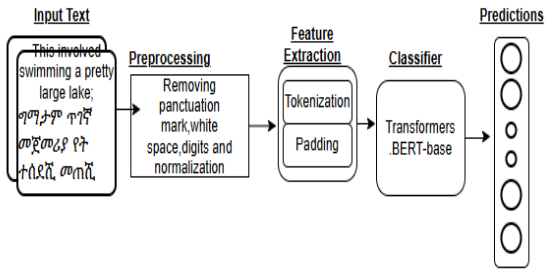

Figure 1: Abstract process of multi-label emotion detection in Amharic and English data.

4.1 Preprocessing and Feature Extraction

In order to standardize the input data and establish necessary circumstances for successful model training, we employed a number of preprocessing approaches. To maintain uniformity (Bade and Seid, 2018), the processing stage entailed converting all textual data to lowercase and eliminating tags, punctuation, and numbers (Bade et al., 2024d). In addition, the tokenization of the texts and the padding of the tokenized pieces in equal length (max-length = 128) are considered in the preprocessing.

Datasets		Languages
Train classes	**English**	**Amharic**
#Anger	333	1,188
#Disgust	–	1,268
#Fear	1,611	109
#Joy	674	549
#Sadness	878	771
#Surprise	838	151
#ideal total*	4,334	4,036
#Actual total^{+}	2,768	3,549
Dev classes		
#Anger	16	207
#Disgust	–	209
#Fear	63	22
#Joy	31	93
#Sadness	35	127
#Surprise	31	27
#Ideal total*	176	685
#Actual total^{+}	116	592
Test classes		
#Anger	322	582
#Disgust	–	628
#Fear	1,544	54
#Joy	670	276
#Sadness	881	355
#Surprise	799	82
#Ideal total	4,216	1,977
#Actual total^{+}	2,767	1,774
#Total^{+}	**5,651**	**5,915**

Table 2: Class-wise dataset statistics of training, development, and testing where '*' denotes the emotions that are overlapped one another and '+' denotes the actual record (row-wise) dataset. '–' indicates that the English data doesn't contain a 'disgust' label.

4.2 Tramsformers Base-Model

The use of a transformer-based model, known for its effectiveness and resilience in managing a variety of NLP tasks, to identify multilabel emotions in two languages was the main point of our methodology (Bade et al., 2024c, 2025). We used the transformers library to tokenize the text data, integrated early stopping, and made dynamic learning rate adjustments to prevent overfitting and speed up convergence to an ideal model state to fine-tune the transformer-based BERT-base model in the data set. Using Hugging Face's Trainer API, training was carefully carried out, utilizing techniques including validation-based tuning and batch size optimization to guarantee the model's efficacy (Mersha et al.,

2024). With regular assessments to modify training parameters based on real-time performance data, the models were trained for a maximum of five epochs. To guarantee improved generalization and dependability on unknown data, the model was verified using a separate validation data set (Bade et al., 2024a).

4.3 Result and Discussion

We observed the ability of the model that was fine-tuned based on the BERT to detect the multilabel emotion in both Amharic and English. The measurement of the performance included the average macro of accuracy (Acc), recall (Rec), precision (Pre), and F1-score.

Accuracy: Measures the proportion of correctly predicted labels over all labels.

Recall:Measures the proportion of correctly predicted positive labels out of all actual positive labels.

Precision:Measures the proportion of correctly predicted positive labels out of all predicted positive labels.

F1-score:The harmonic mean of precision and recall, providing a balance between the two.

However, as a convention and for ranking purpose, the average macro F1-score was used to measure the model performance. The macro-average calculates metrics for each label and then takes the unweighted mean. Thus, our developed BERT-base model achieved the average-macro F1-score of 0.6300 for amharic and 0.7025 for English, respectively. The Table3 gives the details of the results.

Language	Acc	Mac Rec	Mac Pre	Mac F1
Amharic	0.4696	0.6781	0.5910	**0.6300**
English	0.4405	0.7272	0.6822	**0.7025**

Table 3: Performance of the classifier on the test dataset.

From Table 3, it is evident that the selected model exhibits a stronger performance on English data compared to Amharic. While the model achieves an accuracy of 0.4405 and a macro F1 score of 0.7025 for English, it also demonstrates reasonably good results for Amharic, with an accuracy of 0.4696 and a macro F1 score of 0.6300. This suggests that the model is more biased toward English, likely due to factors such as the availability of more training data, better language representation, or inherent linguistic complexities in Amharic. Nevertheless, the model's performance on Amharic is still commendable, indicating its potential for multilingual applications with further optimization and fine-tuning.

5 Error Analysis and Limitation

The investigation demonstrates that the task of classifying emotions is difficult and even complex for human competence. This demonstrates the importance of this task in assessing the current models and noting that, even for high-resource languages like English, multi-label emotion categorization requires further research. The inability to pinpoint the writers' precise emotions—which requires context—and the overlap between some emotion classes, like anger and disgust (for instance, anger and disgust may appear together, and joy and surprise may also exhibit similarities) are some of the challenges. Despite many constraints, our model has demonstrated a comparable performance in detecting multilabel emotions. Table 4 compares the performance of our model against the SemEval base model which used RemBERT, using macro F1 score.

Models	Languages	
	Amharic	English
BERT-base (ours)	0.6300	0.7025
RemBERT (Muhammad et al., 2025a)	0.6383	0.7083
Differences	0.0083	0.0058

Table 4: Comarsion of the performance of our model against SemEval baseline model, reflecting the existance of non significant differences, indicating that almost BERT performed equal to baseline model.

6 Conclusion and Future Work

This study investigated the use of a transformer-based model for multilabel emotion recognition in both Amharic and English. According to the study, the transformer-based BERT model produced the top F1-scores for both the English and Amharic datasets, indicating an effective outcome. The study's findings support the notion that transformer models are highly capable of classifying text in multilabel emotions. To improve the accuracy rate, the next stage should focus on utilizing large datasets, improved fine-tuning techniques, and contextual factors.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-

S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024a. Social media hate and offensive speech detection using machine learning method. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 240–244.

Girma Yohannis Bade, O Koleniskova, José Luis Oropeza, Grigori Sidorov, and Kidist Feleke Bergene. 2024b. Hope speech in social media texts using transformer. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS. org*.

Girma Yohannis Bade, Olga Kolesnikova, and Jose Luis Oropeza. 2024c. Evaluating the quality of data: Case of sarcasm dataset.

Girma Yohannis Bade, Olga Kolesnikova, José Luis Oropeza, and Grigori Sidorov. 2024d. Lexicon-based language relatedness analysis. *Procedia Computer Science*, 244:268–277.

Girma Yohannis Bade and Hussien Seid. 2018. Development of longest-match based stemmer for texts of wolaita language. *International Journal on Data Science and Technology*, 4(3):79–83.

Girma Yohannis Bade, Muhammad Tayyab Zamir, Olga Kolesnikova, José Luis Oropeza, Grigori Sidorov, and Alexander Gelbukh. 2025. Girma@ dravidianlangtech 2025: Detecting ai generated product reviews. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 133–138.

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.

Ilyes Bendjoudi, Frederic Vanderhaegen, Denis Hamad, and Fadi Dornaika. 2021. Multi-label, multi-task cnn approach for context-based emotion recognition. *Information Fusion*, 76:422–428.

Jiawen Deng and Fuji Ren. 2020. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. *IEEE Transactions on Affective Computing*, 14(1):475–486.

Melkamu Abay Mersha, Girma Yohannis Bade, Jugal Kalita, Olga Kolesnikova, Alexander Gelbukh, and 1 others. 2024. Ethio-fake: Cutting-edge approaches to combat fake news in under-resourced languages using explainable ai. *Procedia Computer Science*, 244:133–142.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, and 1 others. 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, and 2 others. 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.

Prabod Rathnayaka, Supun Abeysinghe, Chamod Samarajeewa, Isura Manchanayake, Malaka J Walpola, Rashmika Nawaratne, Tharindu Bandaragoda, and Damminda Alahakoon. 2019. Gated recurrent neural network approach for multilabel emotion detection in microblogs. *arXiv preprint arXiv:1907.07653*.

Jesus Serrano-Guerrero, Mohammad Bani-Doumi, Francisco P Romero, and Jose A Olivas. 2022. Understanding what patients think about hospitals: A deep learning approach for detecting emotions in patient opinions. *Artificial Intelligence in Medicine*, 128:102298.

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.