SyntaxMind at SemEval-2025 Task 11: BERT Base Multi-label Emotion Detection Using Gated Recurrent Unit

Md. Shihab Uddin Riad and Mohammad Aman Ullah
Dept. Of Computer Science & Engineering
International Islamic University Chittagong
shihab.riadn@gmail.com and aman_cse@iiuc.ac.bd

Abstract

Emotions influence human behavior, speech, and expression, making their detection crucial in Natural Language Processing (NLP). While most prior research has focused on single-label emotion classification, real-world emotions are often multi-faceted. This paper describes our participation in SemEval-2025 Task 11, Track A (Multi-label Emotion Detection) and Track B (Emotion Intensity). We employed BERT as a feature extractor with stacked GRUs, which resulted in better stability and convergence. Our system was evaluated across 19 languages for Track A and 9 languages for Track B.

1 Introduction

Emotions play a significant role in shaping human behavior, speech patterns, and body language. Natural Language Processing (NLP) plays a crucial role in analyzing and extracting valuable information based on emotions. Earlier research on sentiment and emotion analysis has primarily focused on single-label classification, where a piece of text is assigned just one emotion or sentiment category, like “happy" or “sad". However, human emotions are rarely that simple, people often experience and express multiple emotions at once. For example, a movie review might convey both excitement and disappointment, or a social media post might reflect anger and fear simultaneously. Multi-label emotion classification addresses this complexity by allowing a system to identify and tag multiple emotions within the same text, providing a more accurate and nuanced understanding of human emotional expression. To address this challenge, we present our submission for SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection(Muhammad et al., 2025b) which is based on “BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages”(Muhammad et al., 2025a) and “Evaluating the Capabilities of Large Language Models for Multi-label Emotion Understanding”(Belay et al., 2025). The task is divided into three tracks and we participated in Track A: Multi-label Emotion Detection, and Track B: Emotion Intensity. While participating in the task, we observed that training the model presented several challenges, particularly with overfitting and data imbalance. Specifically, when using pre-trained embeddings like GloVe(Pennington et al., 2014), the model over-fitted quickly, likely due to its tendency to memorize the training data rather than generalize to unseen examples. To address the unbalanced dataset, we experimented with SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002), but this approach did not yield favorable results, possibly because it introduced synthetic samples that failed to capture the true emotional context of the data. Additionally, we attempted to augment our dataset with data from the SemEval-2018 Affect in Tweets(Mohammad et al., 2018) task to enrich the training set. However, this also did not improve performance. Finally we used BERT as feature extractor with two stacked layer of Gated Recurrent Units(GRUs) to overcome unstable training and achieve better convergence on overall (per language) dataset. We participated in a total of 19 languages for Track A and 9 languages for Track B.

2 Related Works

In the past most of emotion or sentiment analysis related work heavily relied on machine learning based approaches (Mullen and Collier, 2004);(Jain et al., 2017). Such work critically depended on hand-crafted features.

(Kar et al., 2017), provides two different methodologies to work on sentiment analysis on financial data. They used both machine learning based approach and deep learning technique in their study.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1450–1455

In latter one, they incorporated Convolution Neural Network (CNN) (LeCun et al., 1989) and GRU to predict sentiment of financial data.

(Baziotis et al., 2018), at SemEval-2018 Task 1, proposed a Bidirectional Long short-term memory(Hochreiter and Schmidhuber, 1997) (Bi-LSTM) architecture equipped with a multi-layer self attention mechanism. They used a set of word2vec word embeddings that were enhanced by a set of word emotional attributes and trained on an extensive collection of 550 million Twitter messages.

(Ameer et al., 2023), proposed models, based on Bi-LSTM with multiple attention layers, implemented n independent attention mechanisms for n emotion labels, where each attention mechanism learns information specific to its corresponding emotion label. The study also implemented transformer models with multiple attention (MA) layers, including XLNet-MA, DistilBERT-MA, and RoBERTa-MA. Multiple attention mechanisms were incorporated into the output of these Transformer models, and the models were fine-tuned on the datasets.

3 System Overview

3.1 Logarithmic Weights Calculation

To address class imbalance in the dataset, logarithmic weighting is employed to adjust the contribution of each class. The logarithmic weights are determined using the formula:

$$w_i = \log\left(1 + \frac{\text{total samples}}{\text{class totals}_i}\right) \quad (1)$$

Here, each class weight is derived by taking the natural logarithm of the inverse class frequency, scaled by the total sample count. This approach ensures that underrepresented classes receive higher weights, mitigating the effects of class imbalance during model training.

To maintain a relative scale, the computed weights are normalized by dividing by the minimum weight value:

$$w_i = \frac{w_i}{\min(w)} \quad (2)$$

This ensures that the smallest weight is set to 1 while preserving relative differences among classes.

3.2 BERT Embedding

BERT (Bidirectional Encoder Representations from Transformers), introduced by (Devlin et al., 2019), is a transformer-based model designed for natural language processing tasks. Unlike traditional models that process text uni-directionally, BERT leverages bidirectional context, pre-training on large corpora to capture deep semantic and syntactic relationships. For English, we used bert-base-uncased as the feature extractor, while for Chinese, we employed bert-base-chinese, and for German, we utilized bert-base-german-cased. For all other languages, we relied on multilingual BERT (mBERT) from Hugging Face. We are using these models for the task of multi-label emotion detection and emotion intensity prediction.

3.3 Gated Recurrent Unit (GRU)

Gated Recurrent Units (GRUs), proposed by (Cho et al., 2014), are a type of recurrent neural network (RNN) designed to model sequential data efficiently. GRUs simplify traditional RNNs by using update and reset gates to control information flow, mitigating issues like vanishing gradients while maintaining performance comparable to Long Short-Term Memory (LSTM) units. In a bidirectional GRU (Bi-GRU), the model processes sequences in both forward and backward directions, capturing past and future context simultaneously. This bidirectional approach enhances the model's ability to understand dependencies in text, making it particularly effective for tasks requiring comprehensive sequence comprehension, such as emotion detection. We used 128 hidden states in the Bi-GRU for this task to balance model capacity and computational efficiency.

3.4 Output Layer

The output layer of the model is designed to transform the processed features into predictions for the target emotion labels. It consists of a dense layer that takes an input dimensionality equal to twice the hidden dimension, reflecting the combined forward and backward representations from the bidirectional GRU. This layer maps these features to a set of output scores, where each score corresponds to one of the emotion categories in the multi-label task.

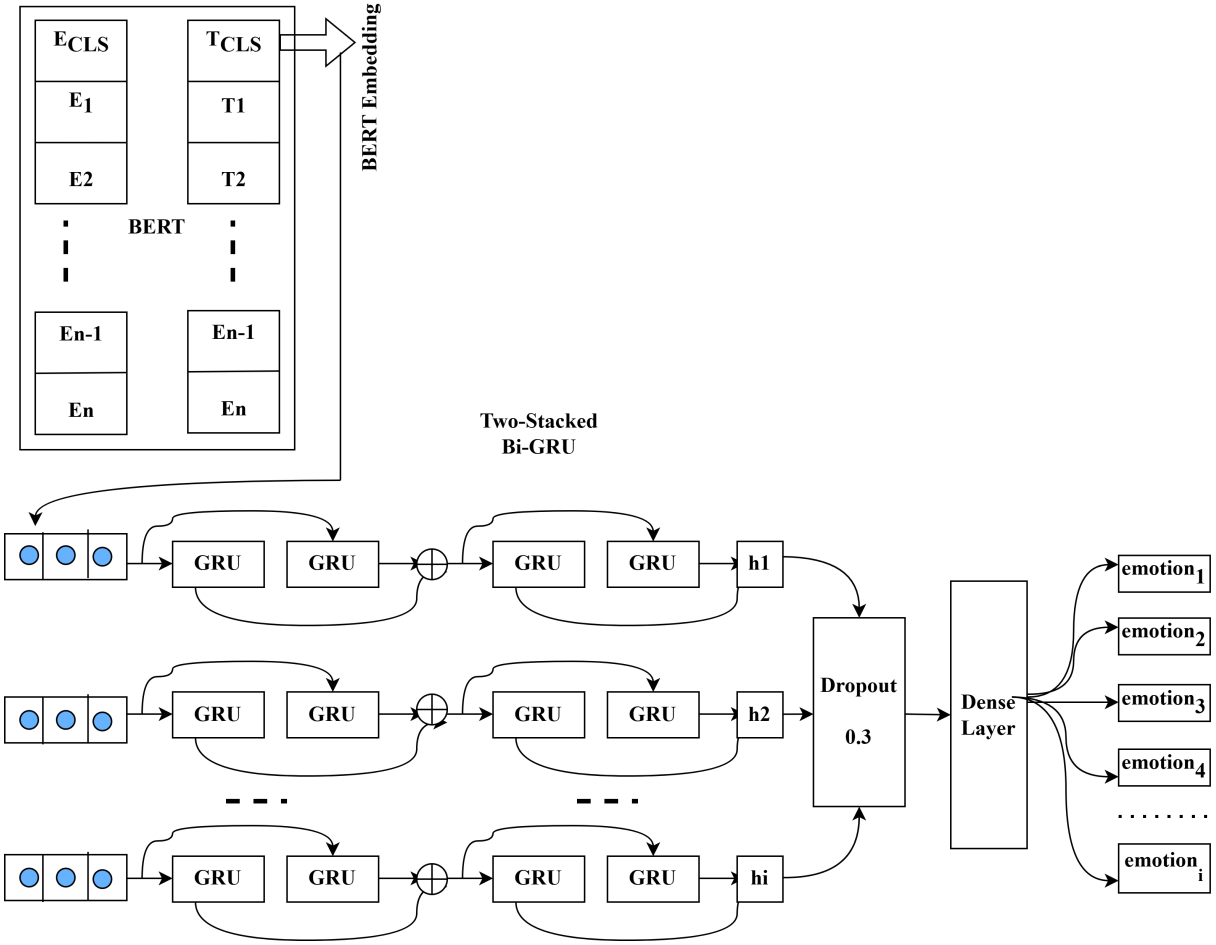

Figure 1: Proposed Model (BERT + Bi-GRU)

4 Experimental Setup

Preprocessing was minimal across the languages studied. For languages other than English, no preprocessing steps were applied. For English, only basic operations were performed, including lowercasing text and expanding contractions, to standardize the input data. To train the model, we utilized BCEWithLogitsLoss as the loss function for multi-label classification, COnsistent RAnk Logits (CORAL) (Cao et al., 2020) as the loss function for emotion intensity, employed the AdamW (Loshchilov and Hutter, 2019) optimizer for efficient parameter updates, and implemented early stopping to prevent overfitting.

Parameters	Track A	Track B
Batch size	2	16
Learning rate	1×10^{-5}	1×10^{-5}
Loss function	BCEWithLogitsLoss	CORAL
Optimizer	AdamW	AdamW
Dropout	0.3	0.3
Hidden Units	128	256

Table 1: Hyperparameter values

5 Results

5.1 Track A

Table 2 presents the macro F1 scores for different models evaluated on the test dataset for SemEval-2025 Task 11 Track A. Our model, denoted as **Ours (SyntaxMind)**, is compared against PAI, PA-oneteam-1, and the SemEval Baseline across 19 languages.

Among the 19 languages, our model achieved competitive results in several cases but lagged behind the top-performing models. For high-resource languages such as English (eng), Spanish (esp), and Hindi (hin), our model achieved macro F1 scores of 0.6646, 0.5739, and 0.6508, respectively. While these results are reasonable, they remain lower than the best-performing model (PAI), which achieved 0.823, 0.8488, and 0.9197, respectively. Similarly, our model performed moderately well on

Language	PAI	PA-oneteam-1	Ours (SyntaxMind)	SemEval Baseline
afr	0.6986	0.6092	0.3649	0.3714
arq	0.6687	0.6623	0.4567	0.4141
ary	0.6292	0.621	0.3733	0.4716
chn	0.7094	0.6877	0.5578	0.5308
deu	0.7399	0.7355	0.4868	0.6423
eng	0.823	0.821	0.6646	0.7083
esp	0.8488	0.8454	0.5739	0.7744
hin	0.9197	0.9194	0.6508	0.8551
mar	0.8843	0.9058	0.7245	0.822
ptbr	0.6833	0.6735	0.3142	0.4257
ptmz	0.5477	0.5033	0.3706	0.4591
ron	0.7943	0.7794	0.6171	0.7623
rus	0.8823	0.9087	0.6596	0.8377
sun	0.5414	0.5072	0.3556	0.3731
swa	0.3848	0.3504	0.2408	0.2265
swe	0.6262	0.6162	0.4331	0.5198
tat	0.8459	0.837	0.4912	0.5394
ukr	0.7256	0.7199	0.315	0.5345
yor	0.4613	0.457	0.2614	0.0922

Table 2: Comparison of macro F1 scores across 19 languages on the test dataset for SemEval-2025 Task 11 Track A

Language	PAI	PA-oneteam-1	Ours (SyntaxMind)	SemEval Baseline
arq	0.6497	0.6338	0.1576	0.0164
chn	0.7224	0.6946	0.4791	0.4053
deu	0.7657	0.7654	0.3886	0.5621
eng	0.8404	0.8339	0.5537	0.6415
esp	0.808	0.7797	0.3916	0.7259
ptbr	0.71	0.6932	0.2363	0.2974
ron	0.726	0.7196	0.3682	0.5566
rus	0.9254	0.9175	0.5259	0.8766
ukr	0.7075	0.6773	0.1912	0.3994

Table 3: Comparison of Pearson Correlation scores across 9 languages on the test dataset for SemEval-2025 Task 11 Track B

German (deu) and Russian (rus), obtaining 0.4868 and 0.6596, respectively.

In low-resource languages such as Yoruba (yor), Swahili (swa), and Sundanese (sun), the performance of all models declined significantly. Our model achieved macro F1 scores of 0.2614, 0.2408, and 0.3556, respectively.

For Arabic dialects, including Algerian Arabic (arq) and Moroccan Arabic (ary), our model obtained scores of 0.4567 and 0.3733, whereas PAI achieved 0.6687 and 0.6292, respectively. A similar trend was observed for Portuguese variants, where our model's performance on Brazilian Portuguese (ptbr) and Mozambican Portuguese (ptmz) was 0.3142 and 0.3706, lower than the leading model's 0.6833 and 0.5477, respectively.

Our model demonstrated moderate performance in languages such as Romanian (ron), Ukrainian (ukr), and Tatar (tat), with macro F1 scores of 0.6171, 0.315, and 0.4912, respectively. Despite this, the highest-performing models achieved significantly better scores.

Though we have beaten the SemEval Baseline model results in the arq, chn, swa, and yor languages.

5.2 Track B

Table 3 shows the results of our system indicate that while it performs moderately well in some languages, there is a significant gap compared to the

top-performing systems. The highest Pearson Correlation score out model achieved in 0.4791 for Chinese (chn), while the lowest is 0.1576 for Arabic Algerian (arq). Across all nine languages, our system consistently lags behind PAI and PA-oneteam-1, suggesting limitations in capturing the nuances of emotion intensity. Notably, performance is particularly weak for Arabic Algerian (arq), Ukrainian (ukr), and Brazilian Portuguese (ptbr), indicating potential challenges in handling certain linguistic structures or data limitations. Compared to the SemEval Baseline, our system performs better in most cases but still requires significant improvements.

6 Conclusion

In this paper, we demonstrate our proposed model (BERT + GRU) for tackling the multi-label emotion challenge. Although our performance was not optimal, we intend to improve our model in the coming days. We also aspire to participate in Track 3 (Cross-lingual Emotion Detection) in the future.

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.

Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 245–255, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Preprint*, arXiv:1406.1078.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Swapnil Jain, Shrikant Malviya, Rohit Mishra, and Uma Shanker Tiwary. 2017. Sentiment analysis: An empirical comparative study of various machine learning approaches. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 112–121, Kolkata, India. NLP Association of India.

Sudipta Kar, Suraj Maharjan, and Thamar Solorio. 2017. RiTUAL-UH at SemEval-2017 task 5: Sentiment analysis on financial data using neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 877–882, Vancouver, Canada. Association for Computational Linguistics.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

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

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

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.

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 412–418, Barcelona, Spain. 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.