AGHNA at SemEval-2025 Task 11: Predicting Emotion and Its Intensity within a Text with EmoBERTa

Moh. Aghna Maysan Abyan
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
13521076@std.stei.itb.ac.id

Abstract

This paper presents our system that have been developed for SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection. The system is able to do two sub-tasks: Track A, related to detecting emotion(s) in a given text; Track B, related to calculate intensity of emotion(s) in a given text. The system will have EmoBERTa as the model baseline, despite some minor differences used in the system approach between these tracks. With the system designed above, Track A achieved a Macro-F1 Score of 0.7372, while Track B achieved Average Pearson r Score of 0.7618.

1 Introduction

SemEval-2025 is the 19th edition of SemEval. SemEval-2025 presents 11 different tasks, one of which is the task titled as "SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection". This task focuses on text-based emotion recognition. In the repository provided, there are 3 sub-tasks that can be done, which will be referred to as "Tracks", namely:

- **Track A:** Multi-Label Emotion Detection
- **Track B:** Emotion Intensity
- **Track C:** Cross-lingual Emotion Detection

Due to time constraints, we were only able to build the system for two different tracks, that being Track A and Track B. For the English language, there are five available emotions: anger, fear, joy, sadness, and surprise. Track A focused on predicting emotions within a text by assigning label to each of the five emotions, with either 0 (no emotion detected) or 1 (emotion detected). Track B focused on predicting emotions within a text by assigning label to each of the five emotions, with either 0 (no emotion), 1 (low degree of emotion), 2 (moderate degree of emotion), or 3 (high degree of emotion).

There are several applications on Text-Based Emotion Detection (TBED) in the modern world. First, TBED can help to detect or diagnose a user's mental health through their posts on social media (Saffar et al., 2022). Second, integrating TBED into an AI system allows for better understanding and interaction between the AI or computers and humans (Machová et al., 2023). And the last example being its integration to business and finances allows data analysts to understand customer reviews more efficiently (Kusal et al., 2022).

In this paper, we propose a system named AGHNA (Automated Generalized Human-emotion detection with a Neural Approach). AGHNA utilizes EmoBERTa as the base model for both tasks mentioned before (Track A & Track B). EmoBERTa is a model developed or fine-tuned from RoBERTa to achieve better results specifically in Emotion Recognition in Conversation (ERC) tasks. EmoBERTa is able to generate better performance compared to other ERC models, such as DialogXL and CESTa (Kim & Vossen, 2021). To further enhance EmoBERTa's performance for both tasks, AGHNA incorporates several additional approaches to the system. These approaches are: Experimenting different feature extraction methods, applying optimization with AdamW, using Binary Cross Entropy (for Track A) and Mean Squared Error (for Track B), and other approaches that will be elaborated even further throughout the paper.

2 Related Works

Related work in the field of TBED has produced various new methods and approaches, but several aspects are still not perfect. The biggest challenge is the high computing resources required due to the complexity of the models being built and the large amount of data that must be trained before testing.

One of the related studies is a research on creating an annotated corpus for Bangla multi-label

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1057–1063
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

emotion detection (Banshal et al., 2023) by collecting data in the form of comments totaling 136,583 data taken from 11 different news stories on Facebook.

The implemented approach uses feature extraction methods such as tokenization and TF-IDF. After that, various methods were applied, using Machine Learning (ML) algorithm (Logistic Regression, Random Forest, Multinomial Naive Bayes, Support Vector Machine, and K-Nearest Neighbors), Deep Learning (DL) algorithm (LSTM, BiLSTM, and hybrid CNN-BiLSTM and CNN-LSTM), and transformer-based algorithms (BanglaBERT, mBERT, Bangla-Bert-Base, and Bangla-Electra). The results are as follows:

- The MNB algorithm achieved the best performance among ML algorithms with an accuracy of 82.64
- BiLSTM provided the best performance among DL algorithms with an accuracy value of 79.14
- Bangla-Bert-Base provided the best performance among transformer-based algorithms with an accuracy of 83.23

There are several advantages offered from the results of this research. These advantages include MONOVAB's contribution to providing a TBED in Bangla which was previously minimal, the use of various approaches (ML, DL, and Transformers) to obtain the most optimal results, as well as an annotation process using a context-based approach. However, there are also limitations in this research. First, the high complexity due to the implementation of various approaches requires high computing resources. Second, there are data that can't be adapted to the corpus because the data cannot be processed using a context-based approach, suggesting for a more suitable lexical-based approach.

Another related research discuss about emotion prediction in text and multi-turn conversations by Combining Advanced NLP, Transformers-based Networks, and Linguistic Methodologies (Singh et al., 2024) which was carried out based on tasks from "WASSA 2022 Shared Task: Predicting Empathy, Emotion and Personality in Reaction to News Stories" and "WASSA 2023 Shared Task: Empathy, Emotion and Personality Detection in Conversation and Reactions to News Articles", both of which are related to emotion prediction.

Split	WASSA 2022	WASSA 2023
Training	1,860	792
Test	525	136
Validation	270	208

Table 1: Data frequency for WASSA 2022 and WASSA 2023 datasets.

The dataset used comes from WASSA 2022 and WASSA 2023, with the statistics provided in the Table 1.

The approach taken in this research involves using a Feedforward Neural Network (FFNN) with ReLU activation and PyTorch, experimenting with various embedding models as input to the neural network, hyperparameter tuning, overcoming data imbalances, utilizing lexicon features, and an ensemble method using two SVR models to model the relationship between text features and emotions. The final results of the study showed an average score increase of 33.59% over the baseline for WASSA 2022 and 64.02% over the baseline for WASSA 2023.

There are several advantages obtained from the results of this research. These advantages include the use of transformers that are integrated with various linguistic features and ensemble methods that can mitigate bias by combining multiple predictions. However, there are also drawbacks in this research. The most noticeable drawback in this research is the high complexity of the model due to the combination of various approaches that requires high computational resources.

3 Dataset

This research will use the dataset provided by the organizers of this SemEval task. There will be an equal amount of data and texts given for both Track A and Track B in the English language, with the statistics provided in Table 2.

Split	# of rows
Train	1,860
Dev	525
Test	270

Table 2: Data frequency for SemEval-2025 Task 11 English dataset.

There are two stages of system development during the process: Development stage, where participants use the Dev data to make predictions; Testing

Emotion	# Emotion Frequency
Anger	333
Fear	1,611
Joy	674
Sadness	878
Surprise	839
Total	4,335
Avg. emotion frequency/text	1.566

Table 3: Emotion frequency for SemEval-2025 Task 11 English Train dataset.

stage, where participants use the Test data to make predictions.

Since the task involves a multi-label dataset, there are multiple instances where a sentence may have more than one detected emotion, either in the Train dataset or as the result of system's predictions. After a quick analysis, each emotion's frequency in the Train dataset are provided in Table 3.

There is a noticeable discrepancy in terms of emotion frequency within the given dataset. For example, the emotion fear appears in 1,611 different texts, whereas the emotion anger appears in only 333 different texts, approximately five times less than fear. This, in return, causes data imbalance and may lead into biases in the system's predictions.

4 Benchmark

The benchmark used for evaluating the performance of the proposed system in this research is based on the official baseline scores provided by the task organizers. These baseline scores are derived from the organizers' own research effors related to the task and serves as the reference for the evaluation of our system's final performance. The baseline system utilizes the RemBERT model, a model that can be used for multiple tasks including text classification, the main topic of this research. In detail, the baseline has a Macro-F1 Score of 0.7083 for Track A, and an Average Pearson Correlation Coefficient (r) Score of 0.6415 for Track B (Muhammad et al., 2025).

5 System Overview

Although several adjustments are required to handle Track A and Track B separately, it is important to note that, due to the similar nature of processing for both tracks (making predictions on given texts),

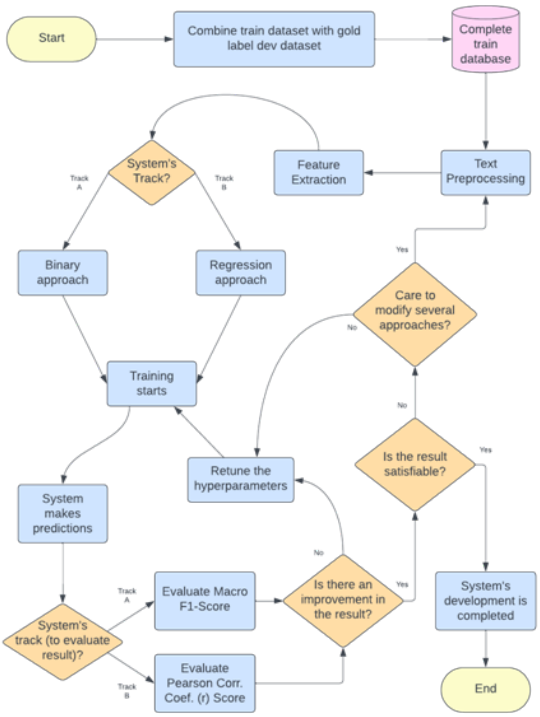

Figure 1: AGHNA's Architecture Design

the overall system design remains similar to that shown in Figure 1.

The system designed for this task will be based on EmoBERTa. Unlike most existing works on ERC that combine different kinds of neural network architectures, and therefore deemed too complex, EmoBERTa simply utilizes the existing RoBERTa model while encoding the speaker's information along with multiple utterances.

EmoBERTa demonstrated very good results when tested with MELD (Multimodal EmotionLines Dataset) and IEMOCAP (Interactive Emotional Dyadic Motion Capture) datasets outperforming several ERC models. On MELD, EmoBERTa achieved a weighted F1-Score of 65.61%, slightly better than the next model, COSMIC (Ghosal et al., 2020) with 65.21%. And, on IEMOCAP, EmoBERTa achieved a weighted F1-Score of 68.57%, slightly better than the next model, CESTa (Wang et al., 2020) with 67.1% (Kim & Vossen, 2021). After understanding the beneficial performance of EmoBERTa for this task, we aim to improve its performance by implementing additional methods into our final system. The additional methods will focused on hyperparameter tuning in several areas, such as the number of epoch, batch size, learning rate, etc.

5.1 Track A

Track A focuses on predicting whether if a specific emotion is present in a given text. Since this is a binary classification task, Binary Cross Entropy (BCE) loss function will be used to analyze the model's performance. For the preprocessing stage, we will utilize Term Frequency–Inverse Document Frequency (TF-IDF) as a method to calculate the importance of words in a text. TF-IDF analyzes several key terms in a document relative to the corpus.

Inside the main training process, focal loss will be utilized to handle class imbalance. To optimize model training, AdamW Optimizer will be used to maintain model stability and performance by decoupling weight decay, alongside learning rate schedulers to adjust the learning rate throughout the training process.

5.2 Track B

Track B focuses on predicting the intensity of an emotion in a given text. Unlike Track A that focuses on binary classification issues, Track B deals on regression classification, as the label may spanning in a real value between 0 and 3. Therefore, they system can't use BCE for this track. Instead, We will use Mean Squared Error (MSE) as the alternative loss function. MSE works similarly to BCE in calculating training errors but it is designed for regression tasks instead of binary tasks.

While the main approach for Track B is mostly similar to Track A, we also put an experimentation on new methods during our research for Track B. For example, in the preprocessing stage, we explored using TextBlob as it is provides more capabilities at feature extraction. We've also integrated attention mechanism to help the system focus more on specific/relevant parts of the text, improving accuracy. Lastly, to handle class imbalance, we also implemented data augmentation to the system, giving a more diverse training examples. Unfortunately, due to time constraints, we were unable to add these new approaches to Track A.

6 Result

For the training dataset, we combined 2,768 data from the provided training dataset with an additional 116 data from the Dev dataset. The Dev dataset were included because they had been assigned gold labels by the organizers, meaning they had been manually reviewed and correctly labeled. Giving a total number of 2,884 data prepared to be trained before the system starts to make predictions into the Test dataset.

During the training process, we conducted several experiments to fine-tune the system's performance. After analyzing the results, we identified and collected the optimal combination of hyperparameters that may yielded the best results for the system. The hyperparameters' values are provided in Table 4.

Hyperparameter	**Value**
seed	42
model	tae898/emoberta-large
max. sequence length	128
batch size	16
epoch	3
learning rate	4e-5
warmup ratio	0.1
gradient clipping	max_norm = 1
(tf-idf)	
max_features	2000
ngram_range	(1, 3)
min_df	2
max_df	0.95
alpha	0.5
gamma	2
weight	based on class imbalance
feature extraction	sentiment_polarity
	sentiment_subjectivity
	text_length
	word_count
	uppercase_ratio
	exclamation_count
	question_count

Table 4: Model Hyperparameters

To minimize the system's execution time, we opted to ran the system for both tracks using NVIDIA A100 GPU on Google Colab, as it is the fastest accelerator compared to other accelerators in Google Colab. After running the system for both tracks independently, we also run the system for two other different models with the intention of model comparison: BERT (bert-large-uncased) and RoBERTa (roberta-large). The system's final results/performance submitted to CodaBench for the SemEval competition, alongside comparisons to other models, are presented in Tables 5 and 6.

Emotion	F1-Score (%)
Anger	68.12
Fear	80.05
Joy	73.70
Sadness	74.10
Surprise	72.63
Macro-F1	**73.72**
Micro-F1	**75.19**

Table 5: System's final result for Track A (EmoBERTa only).

Emotion	emoberta-large	roberta-large	bert-large-uncased
Anger	72.15	69.18	67.21
Fear	77.14	76.22	74.54
Joy	80.41	77.83	77.00
Sadness	79.50	75.96	75.92
Surprise	71.72	68.90	66.46
Average	**76.18**	**73.62**	**72.23**

Table 6: System's final result for Track B between three different models (%).

For the purpose of ranking and evaluation from the organizers, Macro-F1's Score will be used as the main metric for Track A, while Average Pearson Correlation Coefficient (r)'s Score will be used for Track B.

In Track A, our system AGHNA achieved a Macro-F1 Score of 73.72%, placing 38th out of 97 participants, placing it within top 40% of all submissions. This result represents a small improvement of 4.08% over the baseline score of 70.83% provided by the task organizers. Meanwhile, for Track B, AGHNA achieved an Average Pearson Correlation Coefficient (r) Score of 76.18%, placing 11th out of 43 participants, placing it within top 26% of all submissions. This result represents a major improvement of 18.75% over the baseline score of 64.15% provided by the task organizers.

From Table 6, it can be seen that EmoBERTa (emoberta-large) outperforms BERT (bert-large-uncased) and RoBERTa (roberta-large) in Track B by an average margin of 2-4%. Unfortunately, we lost the experiment logs for Track A, and since the system was updated after the SemEval test phase ended, we are unable to re-run the models in their pre-update versions before the paper submission deadline, hence why we only showed the EmoBERTa-only result for Track A in Table 5. Nevertheless, we can confirm that EmoBERTa also outperforms both BERT and RoBERTa in the updated system, by combining approaches from Track B into Track A as shown in Table 7.

Data imbalance remains a significant challenge in Track A, provided by the notable difference in F1-Score between the emotions anger and fear. The gap between these two emotions is 11.93%, indicating several emotions are more accurately predicted than others. Such difference suggests the model struggles to generalize across every emotions in the dataset, most likely due to uneven distribution of emotions within the given dataset. Although with such problems faced in Track A, Track B doesn't seem to suffer as much, with the lowest and highest Pearson Correlation Coefficient (r) Score, performed by surprise and joy respectively, differs by only 8.69%. This statistics further highlight Track B's overall success while also giving a massive understanding the need of improvement for Track A.

Due to the time constraints given by the task organizers, our research was unable to implement several Track B's methods to Track A. Although, by how successful the result for Track B is compared to Track A, we have integrated several methods from Track B (such as the implementation of TextBlob) into Track A after the SemEval test phase ended, and we plan to add more features for both tracks in the future.

7 Conclusion

As seen in the results and ranking statistics from the previous chapter, AGHNA demonstrates strong capabilities in predicting emotions, both in binary

Emotion	emoberta-large	roberta-large	bert-large-uncased
Anger	62.55	63.75	56.21
Fear	84.70	84.06	82.91
Joy	78.48	75.82	76.07
Sadness	75.43	76.55	73.93
Surprise	72.81	70.71	68.84
Average	**74.79**	**74.18**	**71.59**

Table 7: Updated system's final result for Track A between three different models (%).

classification detection (Track A) and regression classification for intensity (Track B). This is evident from AGHNA's performance outperforming two baselines (one for each tracks) set by the organizers and achieved a top-half ranking in both tracks, including an almost top-quarter ranking in Track B.

Despite these results, there are still several room for improvements in both tracks, especially Track A. Therefore, we hope that in future research, we, or others interested in further improving the system, can develop way better solutions to bring another improvement for EmoBERTa's performance in emotion detection. Although improvements are desired for both tracks, we believe Track A warrants more in-depth analysis, as it shows bigger potential for improvement.

Acknowledgments

First, we would like to express our deepest gratitude to God, The One and Only, for granting us the strength, faith, and determination to complete this research. Secondly, we extend our appreciation to the organizer of this SemEval task for providing us the opportunity to develop an NLP-related system, applying our knowledge in this field of computation. Thirdly, we would like to express our thankfulness our supervisor, Dr. Fariska Zakhralativa Ruskanda, S.T., M.T., for her guidance and support throughout this research. And lastly, we want to thank our friends and families for supporting us during the period of research.

References

Sumit Kumar Banshal, Sajal Das, Shumaiya Akter Shammi, and Narayan Ranjan Chakraborty. 2023. Monovab: an annotated corpus for bangla multi-label emotion detection. *arXiv preprint arXiv:2309.15670*.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2022. A review on text-based emotion detection–techniques, applications, datasets, and future directions. *arXiv preprint arXiv:2205.03235*.

Kristína Machová, Martina Szabóova, Ján Paralič, and Ján Mičko. 2023. Detection of emotion by text analysis using machine learning. *Frontiers in Psychology*, 14:1190326.

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.

Alieh Hajizadeh Saffar, Tiffany Katharine Mann, and Bahadorreza Ofoghi. 2023. Textual emotion detection in health: Advances and applications. *Journal of Biomedical Informatics*, 137:104258.

Manisha Singh, Divy Sharma, Alonso Ma, and Nora Goldfine. 2024. Towards more accurate prediction of human empathy and emotion in text and multi-turn conversations by combining advanced nlp, transformers-based networks, and linguistic methodologies. *arXiv preprint arXiv:2407.18496*.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*, pages 186–195.