Project#3: Text Based Emotion Detection

Tafazzul Nadeem, Riyansha Singh, Suyamoon Pathak
232110401, 232110601, 241110091
CSE, CSE, CSE
{tafazzul23, riyansha, suyamoonp24}@cse.iitk.ac.in

Abstract

The report outlines our approach to SemEval 2025 Task 11: Bridging the Gap in Text-Based Emotion Detection, focusing on multi-label emotion classification using transformer-based models. Our work addresses challenges such as class imbalance, intensity classification, and multilingual adaptability. We leveraged various transformer models, including Llama3-8B, which achieved the highest F1 scores, and Cardiffnlp/twitter-roberta-large-emotion-latest, which demonstrated superior accuracy. Initial experiments highlight the effectiveness of fine-tuning and data augmentation in enhancing model performance. This report presents insights into the dataset, preliminary results, and a roadmap for further improvements, including ensemble methods and multilingual adaptation strategies.

1 Introduction

1. Learning the emotion of the text is very important for business problems like customer review analysis, social media sentiment, translation, and more (Acheampong et al., 2021). The text can imply multiple emotions, so we are framing this a multi-label problem. The whole problem can be divided into these four parts- finding a good emotion representation, text pre-processing, contextual and semantic understanding, and finally multi-label classification.
2. Three tracks have been provided for Task 11- Track A: Multi-label Emotion Detection, Track B: Emotion Intensity, Track C: Cross-lingual Emotion Detection. The given emotions for English language are joy, sadness, fear, anger, and surprise (Abdulmumin, 2024).
3. Text Based Emotion Detection (TBED) is a classic multi-label classification problem. So, researches have tried algorithms like Naive Bayes (Azmin and Dhar, 2019), Support Vector Machines (SVM) (Hasan et al., 2014), and others which required manual feature preprocessing. Text processing is a sequential problem. So, researchers have also tried using neural network approaches like Recurrent Neural Networks (RNNs), Long Short Term Memory (LSTMs), and Gated Recurrent Units (GRUs)(Kusal et al., 2022).
4. Transformers have been proven to be the most successful in understanding the contextual and semantic information of the text (Acheampong et al., 2021). Models like BERT have performed exceedingly well in capturing the contextual and semantic representations of the text snippets (Gillioz et al., 2020).
5. Our approach is also to use a transformer-based model. We are still researching on the best way to understand the context and semantics. Bidirectional approaches are the way to go, but limitations of fixed length need to be solved. Other than that, attention with respect to each word will be studied. And, we will be trying to build a model that appropriately sets the attention to the most relevant word, and also the latent meaning (if there is any).

2 Project Github Link

https://github.com/suyamoonpathak/text-based-emotion-detection-semeval-2025-task-11

3 Problem Definition

A text snippet is capable of conveying emotions as shown in Table 1 and this problem of emotion detection using text has been tackled previously using several approaches.

We approach emotion recognition as a multi-label classification problem (Track A). Let $t \in V^*$ represent an input text and consider a set of M possible emotions. Our objective is to learn a function $\psi : V^* \rightarrow Q^M$ that maps t to independent

Text	Emotion
I can't believe this is happening.	Disbelief, Surprise
You really did it! I knew you could!	Pride, Joy
Should I say something? Or stay silent?	Hesitation

Table 1: Text snippets with corresponding emotion.

probabilities for each emotion $z_1, z_2, \ldots, z_M$. For Track B, we need to predict the intensity (0, 1, 2, 3) associated with each of the classes for a given input text. Here 0 indicates no emotion and 3 corresponds to a high degree of emotion. For the last track C - Cross lingual emotion detection in text, the problem is formulated as a classification problem. Let $\mathcal{L}$ represent the set of all languages, and $\mathcal{E} = \{e_1, e_2, \ldots, e_k\}$ be the set of k predefined emotions. The goal of cross-lingual emotion detection is to map a given text t in any language $l \in \mathcal{L}$ to one or more emotions in $\mathcal{E}$.

4 Related Work

The landscape of human emotions has been described through various taxonomies and frameworks. Ekman (Ekman and Friesen, 1971) identified six core emotions expressed through facial expressions that are universally recognized across cultures: joy, sadness, anger, surprise, disgust, and fear. Later, finer-grained taxonomies have been developed, capturing a broader range of up to 600 emotions and employing machine learning to cluster emotion concepts (Cowen and Keltner, 2019). These frameworks highlight the complex, culturally influenced nature of emotions expressed through vocalization (Cowen and Keltner, 2018), music, and facial expressions.

Previous approaches to text-based emotion detection have primarily utilized machine learning (ML) techniques. For instance, Wikarsa et al. and Ameer et al. (Ameer et al., 2021) focused on multi-label emotion classification for code-mixed SMS messages in Roman Urdu and English, employing classical ML methods (SVM, J48, Naive Bayes, etc.) and deep learning models (LSTM, CNN, etc.) on a new dataset. Their results indicated that classical ML methods outperformed both ML and deep learning models. Similarly, Polignano et al. (Polignano et al., 2020) developed a model combining Bi-LSTM, Self-Attention, and CNN for emotion detection, finding that word embeddings significantly improved performance. Their experiments across the ISEAR, SemEval-2018 (Mohammad et al., 2018), and SemEval-2019 datasets demonstrated that the ISEAR dataset yielded the best precision and recall.

In recent years, research on text-based emotion detection has increasingly utilized transformer-based pre-trained language models. For instance, Acheampong et al. (Acheampong et al., 2020) conducted comparative analyses of models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and XL-Net (Yang et al., 2019) for emotion recognition using the ISEAR dataset. Asalah et al. (Thiab et al., 2024) proposed an ensemble deep learning approach for emotion detection in textual conversations, CNN-based model, and transformer-based models, including BERT, RoBERTa, and XLNet. They utilize hard and weighted majority voting methods to enhance prediction accuracy. Their method demonstrates superior performance, achieving a micro-averaged F1-score of 77.07% on the SemEval-2019 Task 3: EmoContext dataset (Chatterjee et al., 2019), outperforming previous baseline results. More recently, a participation (Nedilko, 2023) in the shared task for multi-label and multi-class emotion classification organized as part of WASSA 2023 has used generative pretrained transformers (GPT) achieving the macro F1 score of 0.7038 and the accuracy of 0.7313 on the blind test set of code mixed Roman Urdu and English SMS text messages.

Although ensemble approaches tend to give good results but they increase the complexity of the training process since they require the training of multiple individual models, and then combining their predictions to obtain the final result. This in turn imposes an overhead on prediction time. Secondly, dataset imbalance presents a challenge in these methods. It can be addressed using different approaches such as undersampling and oversampling techniques.

5 Corpus/Data Description

There are 29 languages for which the dataset will be released for all the three tracks. The organizers will provide train, dev and test dataset for each language. So far datasets for seven languages has been released. The organizers have also stated to release a dataset paper very soon that describes the data collection, annotation process, and baseline experiments. We are planning to participate for English language for track A and B. Track C languages will be chosen at a later stage of the project by analysing the results of the models we train for

track A. Key points about the English language dataset are as follows:

1. **Train set**: Track A dataset contains 2768 samples with five binary emotion labels (joy, sadness, fear, anger and surprise). Track B dataset contains emotion intensities (0,1,2 and 3) for all five labels with same number of samples as track A. There is no specific dataset for Track C, one has to use datasets of track A for each language they participate in.

2. **Dev set:** It contains 116 samples without any labels for all tracks. For validation one has to submit the predicted labels of their model on the competition page (Codabench) for getting the F1 scores. The submissions for dev set will be acceptable till 10 Jan 2025.

3. **Test set:** Test set contains 2767 samples. The submissions for test set will start from 10 Jan 2025 and ends on 15 Jan 2025 on the competition page.

4. Figure 4 shows the histogram of the emotions present in the dataset. It was plotted to know the frequency of each emotions to assess any imbalance between emotion labels. Any imbalance needs to be taken care of by

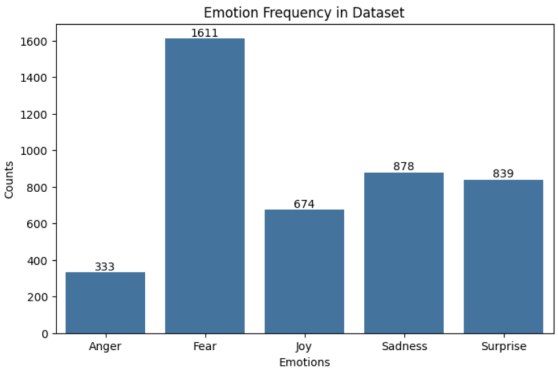

Figure 1: Emotion Frequencies in the provided dataset

techniques like Resampling, Class Weighting, Data Augmentation etc since there is no restriction in using other datasets apart from the one provided by the organizers.

5. The **co-occurence matrix** of the labels for English dataset was plotted to gain insights about the correlation between the labels as shown in Figure 2.

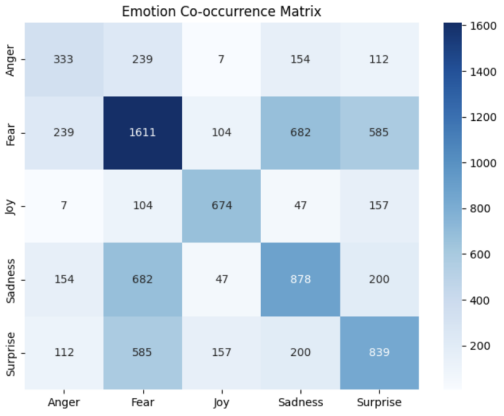

Figure 2: Co-occurence matrix of emotion labels

6. We have also visualized the box plot for length of the text snippets showing the interquartile range (Q1, Q2 etc.) in Figure 3 to gain insights about the context length or prompt length we require while selecting a transformer model. The maximum length was found out to be 94 words. Hence any model with 512 token size can be used since no. of tokens = 4 x no. of words (widely used estimate).

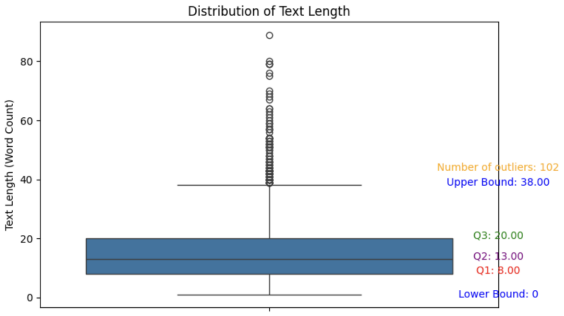

Figure 3: Boxplot of text snippet length

For effective emotion classification, the original dataset consisted of only 2,768 training samples, which poses challenges for robust model training. To enhance the dataset, we explored data augmentation through the integration of external datasets. The XED (Öhman et al., 2020) dataset was evaluated for this purpose. Our approach involved similarity analysis between datasets and an empirical performance comparison using trained models. The dataset consists of emotion annotated movie subtitles from OPUS. Plutchik's 8 core emotions are used to annotate. The data is multilabel. The original annotations have been sourced for mainly English and Finnish, with the rest created using

Model Name	Track	Accuracy	Micro F1	Macro F1	Weighted F1
SamLowe/roberta-base-go_emotions	A	0.21	0.45	0.44	0.42
cardiffnlp/twitter-roberta-large-emotion-latest	A	**0.29**	0.54	0.53	0.50
Emanuel/twitter-emotion-deberta-v3-base	A	0.16	0.45	0.40	0.45
meta-llama/Meta-Llama-3-8B-Instruct	A	0.24	**0.58**	**0.59**	**0.58**
SamLowe/roberta-base-go_emotions	B	0.11	-	-	-
cardiffnlp/twitter-roberta-large-emotion-latest	B	0.10	-	-	-
meta-llama/Meta-Llama-3-8B-Instruct	B	**0.18**	-	-	-

Table 2: Results without fine-tuning on the shortlisted transformer models

annotation projection to aligned subtitles in 41 additional languages, with 31 languages included in the final dataset (more than 950 lines of annotated subtitle lines).

Number of annotations	24164 + 9384 neutral
Number of unique data points	17530 + 6420 neutral
Number of emotions	8 (+pos, neg, neu)
Number of annotators	108 (63 active)

Table 3: Statistics for XED

6 Proposed Approach

1. Since transformer based models have shown promising results in classification tasks lately, we decided to first find some suitable models for the multilabel classification task. From recent works ((Lowe, 2022), , (Barbieri et al., 2022), (Antypas et al., 2023), (Huber, 2021), (Dubey and et al., 2024)) we shortlisted the following huggingface models to check their performance on the provided dataset:
   (a) **Small models**:
      i. SamLowe/roberta-base-go_emotions
      ii. cardiffnlp/twitter-roberta-large-emotion-latest
      iii. Emanuel/twitter-emotion-deberta-v3-base
      iv. cardiffnlp/twitter-xlm-roberta-base-sentiment
   (b) **Large models**:
      i. meta-llama/Meta-Llama-3-8B-Instruct

   The models were evaluated on the training data for english language with full fine-tuning. The results are shown in Table 4.

2. We are also searching for more pre-trained models which can give better results on the provided datasets.

3. We used multilingual models Twitter-xlm-roberta-base-sentiment for Track C

4. We further aim to explore the ensemble approach of the models we worked on.

7 Experiments and Results

For all the tracks of the competition we had to train on the provided training data with labels and make predictions on the dev set. The predictions had to be then submitted to the competition webpage to get the evaluation score. Evaluation metric for different tracks are as follows:
Track A & C: Multi-label accuracy (Jaccard score), Micro F1 score and Macro F1 score
Track B:Average Pearson rating

7.1 Track A

1. **Full Fine-tuning of Off-the-Shelf Models**: We selected some of the best-performing models from preliminary experimentation: cardiffnlp/twitter-roberta-large-emotion-latest and SamLowe/roberta-base-go_emotions. Full fine-tuning was performed on these models, and the results are as follows:
   - **cardiffnlp/twitter-roberta-large-emotion-latest:**
     – Multi-label accuracy (Jaccard score): 0.43
     – Micro F1 score: 0.60
     – Macro F1 score: 0.49
   - **SamLowe/roberta-base-go_emotions:**
     – Multi-label accuracy (Jaccard score): 0.44
     – Micro F1 score: 0.61
     – Macro F1 score: 0.49

2. **Training of an Added Classifier Layer Only**:

Since the off-the-shelf models had more emotion categories than our dataset, we added an extra fully connected (FC) layer with five neurons (equal to our dataset's emotion labels). The base model was frozen, and only this FC layer was trained. The results on cardiffnlp/twitter-roberta-large-emotion-latest were:

- Multi-label accuracy (Jaccard score): 0.57
- Micro F1 score: 0.73
- Macro F1 score: 0.69

3. **Training of an Added FC Layer Along with the Last Classifier Layer**:
   Next, we trained both the added FC layer and the last classifier layer of the base model. Using cardiffnlp/twitter-roberta-large-emotion-latest, the following scores were achieved on the development set:

   - Multi-label accuracy (Jaccard score): 0.62
   - Micro F1 score: 0.77
   - Macro F1 score: 0.75

4. **Data Augmentation**
   To address the limitation of a small dataset (2,768 training samples), we implemented a data augmentation strategy. This was necessary to increase the dataset's size and variability, improving model generalization.

   Data Augmentation Strategy:
   We used ChatGPT-4o to generate synthetic data. The steps were:

   (a) Randomly selected 100 samples from the original dataset.
   (b) Created 50 new synthetic entries based on these samples.

   The synthetic entries were designed to maintain a similar style, tone, and emotional distribution as the original data, ensuring consistency while adding variability.

   Prompt for Synthetic Data Generation:
   The following prompt was used to instruct the AI model for generating high-quality synthetic data. To ensure it fits within the column, it is presented in a compact, inline format:

   ```
   You are an AI model specializing in text-based emotion detection. Generate synthetic data in the format: "text, Anger, Fear, Joy, Sadness, Surprise." Labels are binary (1 or 0). Using 100 provided samples, create 50 entries with realistic language and emotion combinations while maintaining variability. Example: "I can't believe this happened, everything feels hopeless.,0,1,0,1,0".
   ```

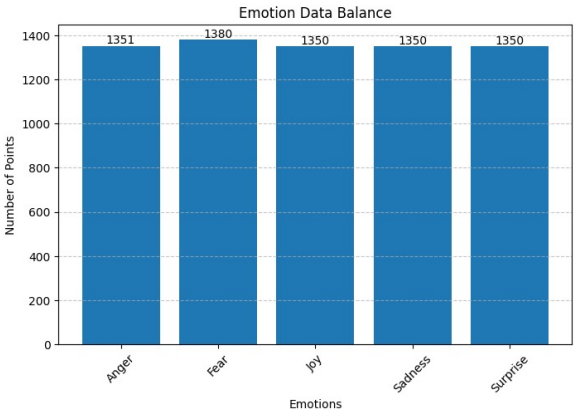

Figure 4: Class Distribution after Data Augmentation

Results of Data Augmentation:
After data augmentation, we evaluated the performance of three transformer-based models fine-tuned on the augmented dataset. The results are as follows:

- **SamLowe/roberta-base-go_emotions:**
  - Accuracy: 0.40
  - Micro F1 score: 0.59
  - Macro F1 score: 0.48
  - Weighted F1 score: 0.55
- **cardiffnlp/twitter-roberta-large-emotion-latest:**
  - Accuracy: **0.61**
  - Micro F1 score: **0.76**
  - Macro F1 score: **0.74**
  - Weighted F1 score: **0.72**
- **Emanuel/twitter-emotion-deberta-v3-base:**
  - Accuracy: 0.24
  - Micro F1 score: 0.49
  - Macro F1 score: 0.44
  - Weighted F1 score: 0.45

5. **Training the current best model on Entailment Approach**: The dataset is converted to Premise-Hypothesis pair dataset with a label 0 or 1 for Hypothesis being in contradiction/neutrality of the Premise and 1 for Hypothesis being entailment of the Premise respectively for every emotion: Example:

   Original Sample:
   Text: But not very happy.
   Labels: [0.0, 0.0, 1.0, 1.0, 0.0] ('Anger','Fear','Joy','Sadness','Surprise')

   Converted to:
   premise: But not very happy.
   Hypothesis: The speaker is feeling Anger
   Labels: [0.0] (Neutral or Contradiction)

   Hence for every input sample 5 samples are created for every emotion. So the new dataset is 5x of the original size. The current best model so far (cardiffnlp/twitter-roberta-large-emotion-latest) is trained on this dataset with added binary classifier for Contradiction/Entailment prediction. The scores obtained are as follows:
   Multi-label accuracy (Jaccard score): 0.60
   Micro F1 score: 0.73
   Macro F1 score: 0.73

6. **Oversampling minority class with replacement- Method-1**: The original dataset provided by the organizers was highly imbalanced. The Fig-4 denotes the imbalance between the labels in which fear being represented highest. So we oversampled the minority classes with replacement to balance the datasets. We selected the class with lowest representation and selected those samples which have 1 in the label and all others 0. We randomly sampled 100 samples from it and then again checked for lowest class representations and repeated the procedure till we got balanced samples. We trained the model on this dataset on cardiffnlp/twitter-roberta-large-emotion-latest with classifier and added extra FC layer trainable and all other parameters fixed. Scores obtained are as follows:
   Multi-label accuracy (Jaccard score): 0.59
   Micro F1 score: 0.73
   Macro F1 score: 0.73

7. **Oversampling minority class with replacement- Method-2**: In the previous oversampling method we only selected those samples which have sole labels as 1 of the selected class and all others 0. In this method we sampled those rows which do not have 1 in the majority label class (like Anger which had 1611 samples). We trained on the same model with added FC layer with same configuration. The scores on dev set are as follows:
   Multi-label accuracy (Jaccard score): 0.64
   Micro F1 score: 0.78
   Macro F1 score: 0.77

7.2 Track B

1. **classification head of base model and added linear layer**Since the last configuration (training classification head of base model and added linear layer was giving the best results on the Track-A dataset, we tried to train the same configuration on the TrackB dataset. The model used to train was cardiffnlp/twitter-roberta-large-emotion-latest. Since the base model was trained on dataset with 0/1 labels and ours was an intensity prediction problem, we first converted the intensities into range 0 to 1 with the following rule:
   0: 0.0, 1: 0.6, 2: 0.75 and 3: 1.0.

   After prediction the score was reconverted to desired labels with another set of rules as follows: Score< 0.5: label-0
   0.5 <= Score < 0.7: label-1
   0.7 <= Score < 0.8: label-2
   0.8 <= Score <= 1.0: label-3

   The score obtained for this setting was:
   Anger: 0.8012
   Fear: 0.424
   Joy: 0.7242
   Sadness: 0.5947
   Surprise: 0.5202
   Average Pearson r: 0.6129

7.3 Track C

1. **Finetuning** A multilingual XLM-roBERTa-base model, cardiffnlp/twitter-xlm-roberta-base-sentiment trained on 198M tweets and finetuned for sentiment analysis where sentiment fine-tuning was done on 8 languages

(Ar, En, Fr, De, Hi, It, Sp, Pt) is used to fine tune on our dataset. Since no seperate dataset is provided for Track C, Russian language dataset from track A is used to fine tune the model. English dataset from the dev set of track C is later used for evaluation. The scores obtained are:
Multi-label accuracy (Jaccard score): 0.10
Micro F1 score: 0.17
Macro F1 score: 0.13.

Further, the compatibility of the XED dataset with the original dataset was assessed using cosine similarity. Similarity score between the two datasets was computed to be 0.26, indicating a moderate level of similarity. While not highly similar, the overlap suggests potential for useful augmentation.

The XED dataset was utilized for fine tuning the best-performing model ((cardiffnlp/twitter-roberta-large-emotion-latest)from the original dataset. The model trained achieved comparable accuracy to the results obtained using the original dataset. This outcome indicates that the emotion categories and distribution patterns in XED align with those of the original dataset to a significant extent, despite differences in the specific vocabulary or contextual framing of emotions.

8 Error Analysis

Each of the models had its own advantages and drawbacks likely due to the differences in the pre-training data used by each of the models. The performance of each of the models was observed separately on English language over the development set except for track C where Russian language is used for training. It can be seen that certain models performed better depending on the task and data (language) given.

9 Individual Contributions

Table 6 presents a breakdown of the contributions made by individual members to the project so far.

10 Future Directions

1. **Improved Data Augmentation:** Future work can explore more advanced data augmentation techniques, including leveraging external datasets with similar emotion annotations. Techniques like back-translation or generative methods (e.g., using large language models) could help create richer datasets to address class imbalance and improve model robustness.
2. **Ensemble Models:** Combining multiple transformer models through ensemble approaches could improve overall prediction accuracy. Strategies like majority voting, weighted averaging, or stacking classifiers can be investigated to leverage the strengths of different models.
3. **Custom Model Architectures:** Building custom transformer-based architectures tailored for multi-label emotion detection could be explored. This includes designing architectures that focus on better handling of contextual and semantic nuances, such as hierarchical attention mechanisms or task-specific embeddings.
4. **Cross-Lingual Emotion Detection:** Expanding to multilingual tracks by fine-tuning pre-trained multilingual models like XLM-Roberta on cross-lingual datasets can address the challenges of emotion detection in diverse languages. Incorporating techniques like zero-shot or few-shot learning might also improve performance on low-resource languages.
5. **Real-Time Applications:** Developing models optimized for deployment in real-time applications, such as customer support chatbots or social media monitoring tools, can be a valuable extension. Efforts can focus on reducing model size and latency while maintaining accuracy.
6. **Emotion Intensity Prediction:** Further research can delve into refining the intensity prediction tasks by utilizing regression-based approaches instead of classification for more granular understanding of emotional intensity.
7. **Explainability and Interpretability:** Future directions could focus on making the emotion detection models more interpretable. Techniques like attention visualization or explainable AI methods can provide insights into how and why specific emotions are predicted.
8. **Domain-Specific Models:** Fine-tuning models for specific domains, such as healthcare, education, or entertainment, could improve performance by focusing on domain-specific emotional expressions and contexts.
9. **Integration with Multimodal Systems:** Combining text-based emotion detection with

Model Name	Track	Accuracy	Micro F1	Macro F1	Weighted F1
SamLowe/roberta-base-go_emotions	A	0.21	0.45	0.44	0.42
cardiffnlp/twitter-roberta-large-emotion-latest	A	0.29	0.54	0.53	0.50
Emanuel/twitter-emotion-deberta-v3-base	A	0.16	0.45	0.40	0.45
meta-llama/Meta-Llama-3-8B-Instruct	A	0.24	0.58	0.59	0.58
cardiffnlp/twitter-roberta-large-emotion-latest (Full Fine-tuning)	A	0.43	0.60	0.49	-
SamLowe/roberta-base-go_emotions (Full Fine-tuning)	A	0.44	0.61	0.49	-
cardiffnlp/twitter-roberta-large-emotion-latest (Added Classifier Layer Only)	A	0.57	0.73	0.69	-
cardiffnlp/twitter-roberta-large-emotion-latest (Added FC Layer Along with the Last Classifier Layer)	A	0.62	0.77	0.75	-
cardiffnlp/twitter-roberta-large-emotion-latest (Entailment Approach)	A	0.60	0.73	0.73	-
cardiffnlp/twitter-roberta-large-emotion-latest (Oversampling minority class with replacement-Method-1)	A	0.59	0.73	0.73	-
cardiffnlp/twitter-roberta-large-emotion-latest (Oversampling minority class with replacement-Method-2)	A	**0.64**	**0.78**	**0.77**	-
After Data Augmentation					
SamLowe/roberta-base-go_emotions	A	0.40	0.59	0.48	0.55
cardiffnlp/twitter-roberta-large-emotion-latest	A	0.61	0.76	0.74	0.72
Emanuel/twitter-emotion-deberta-v3-base	A	0.24	0.49	0.44	0.45
cardiffnlp/twitter-xlm-roberta-base-sentiment	C	**0.10**	**0.17**	**0.13**	**0.15**

Table 4: Results for Track A , B and C

Model Name	Anger (Pear R.)	Fear (Pear R.)	Joy (Pear R.)	Sadness (Pear R.)	Surprise (Pear R.)	Average Pearson r
SamLowe/roberta-base-go_emotions	A	0.21	0.45	0.44	0.42	

Table 5: Results for Track B

other modalities like audio and visual cues can create more holistic emotion recognition systems. Multimodal approaches could capture nuances missed in text alone.

11 Conclusion

In this report, we worked on SemEval 2025 Task 11 (Bridging the Gap in Text-Based Emotion Detection) , which focuses on detecting emotions from text. We looked at three main challenges: identifying multiple emotions in one text, measuring how strong the emotions are, and working with different languages. Our experiments used advanced models like transformers, and we found that some models worked better than others. We also used techniques like data augmentation to handle the small dataset size and improve results. We explored the possibility of augmenting our data with external data sources. Our findings show that fine-tuning smaller models can sometimes perform as well as, or even better than, larger models. This means it is possible to improve accuracy without needing very big models. We also explored ways to deal with unbalanced data, such as oversampling, and discussed future improvements like creating custom models or using ensembles. Overall, this project gave us a better understanding of how text-based emotion detection works and what can be done to make it even better.

References

Idris Abdulmumin. 2024. Semeval2025-task11.

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.

Frank A Acheampong, Henry Nunoo-Mensah, and Wei Chen. 2020. Transformer models for text-based emotion detection: A comparative analysis. *arXiv preprint arXiv:2004.13704*.

Asim Ameer, Sher Maqbool, and Ghulam Azam. 2021. Multi-label emotion classification on code-mixed roman urdu and english sms messages using machine

Task Done	Members Contribution
Literature survey for Text Based Emotion Detection	Equal contribution by all
Group discussions on literature survey	Equal contribution by all
Shortlisting the models based on literature survey	Equal contribution by all
Initial experimentation with the pretrained models	Equal contribution by all
Creation of presentation slides, Project Document and Mid-Term Project Report	Equal contribution by all

Table 6: Member Contribution

learning and deep learning approaches. In *2021 International Conference on Computer and Communication Technologies (ICCCT)*, pages 80–86. IEEE.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research.

S. Azmin and K. Dhar. 2019. Emotion detection from bangla text corpus using naïve bayes classifier. In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–5, Khulna, Bangladesh.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Ankush Chatterjee, Khyathi Raghavi Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48.

Alan S Cowen and Dacher Keltner. 2018. Vocal expression of emotion reveals cross-cultural recognition, stereotypes, and differentiation. *Nature Human Behaviour*, 2(6):360–372.

Alan S Cowen and Dacher Keltner. 2019. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909.

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

Abhimanyu Dubey and et al. 2024. The llama 3 herd of models.

Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled. 2020. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183, Sofia, Bulgaria.

Maryam Hasan, Elke A. Rundensteiner, and Emmanuel O. Agu. 2014. Emotex: Detecting emotions in twitter messages.

Emanuel Huber. 2021. Emanuel/twitter-emotion-deberta-v3-base.

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.

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. *arXiv preprint arXiv:1907.11692*.

Sam Lowe. 2022. Samlowe/roberta-base-go$_e motions$.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17.

Andrew Nedilko. 2023. Generative pretrained transformers for emotion detection in a code-switching setting. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 616–620, Toronto, Canada. Association for Computational Linguistics.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *The 28th International Conference on Computational Linguistics (COLING 2020)*.

Michele Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2020. Emotion recognition through a multi-model approach: A study of different word embedding techniques for emotion detection in textual data. In *Proceedings of the 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 45–51. IEEE.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Asalah Thiab, Luay Alawneh, and Mohammad AL-Smadi. 2024. Contextual emotion detection using ensemble deep learning. *Computer Speech Language*, 86:101604.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.