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

This mid-semester report is presented for the SemEval2025-Task 11: Bridging the Gap in Text-Based Emotion Detection. In this work, we aim to build a multi-label emotion classifier using transformer-based models. This report presents the insights into the dataset, results of our preliminary experiments with multiple transformer-based models, and strategies for addressing challenges like class imbalance, intensity classification (0,1,2,3), and multi-lingual adaptation. Before fine tuning the models on our dataset, early results indicate the Llama3-8B model achieved the best F1 scores, whereas cardiffnlp/twitter-roberta-large-emotion-latest, a RoBERTa model trained on twitter dataset achieved the highest accuracy (cardiffnlp, 2021).

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 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.

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.

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 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.

3 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.

4 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 1 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

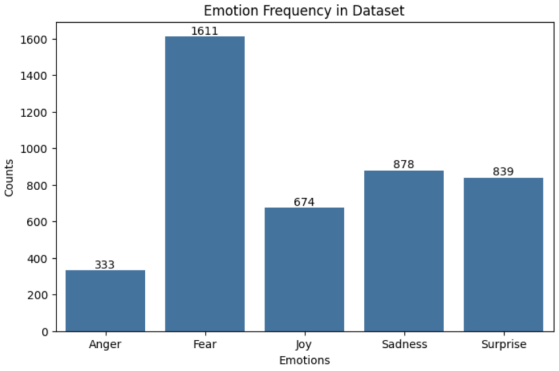

Figure 1: Emotion Frequencies in the provided dataset

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.

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).

5 Future Directions

1. Since transformer based models have shown promising results in classification tasks lately,

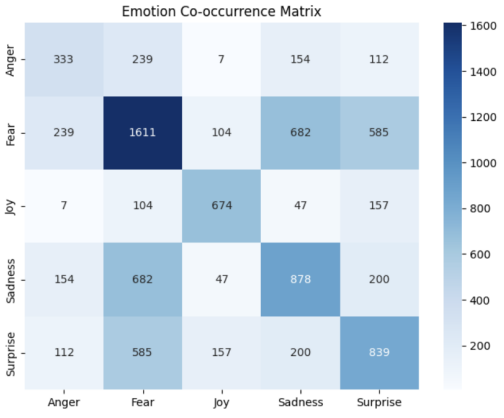

Figure 2: Co-occurence matrix of emotion labels

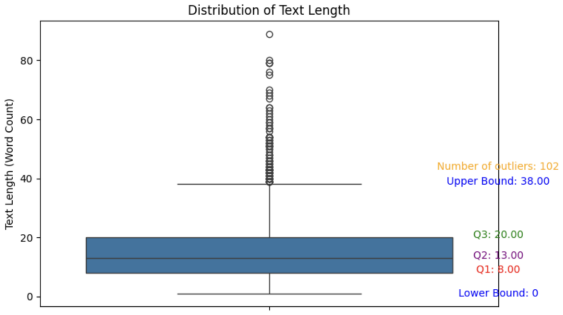

Figure 3: Boxplot of text snippet length

we decided to first find some suitable models for the multilabel classification task. From recent works ((Lowe, 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

(b) **Large models**:
   i. meta-llama/Meta-Llama-3-8B-Instruct

The models were evaluated on the training data for english language without any fine-tuning. The results are shown in Table 2.

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

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

provided datasets.

3. Once we have shortlisted sufficient number of models, we can fine-tune the shortlisted models with released data to check their performance. This will also help in exploring ensemble models using these models.

4. We are also planning to fine-tune some fine-grained multi-label classifier models with an extra classifier on top of some models. This can be done in two ways, either training all the parameters or only training the linear classifier on top and freezing the weights of the base model.

Future Task	Member Responsible and Timeline
Selection of some more models for track A and B	Every member-till 05/10/24
Selection of multi-lingual models for track C	Every member-till 05/10/24
Fine-tuning on all the shortlisted models	Every member-till 13/10/24
Exploring different combination of ensemble models	Every member-till 20/10/24
Building our own models for all the tracks	Every member-till 10/11/24
Fine-tuning our models	Every member-till 20/11/24
Improving and finalizing our models	Every member-till 05/01/25
Final submission on the competition webpage	Every member-till 15/01/25

Table 3: Timeline of planned future tasks

5. So far we have only used mono-lingual models but for Track C we require multi-lingual model (like XL-Net) trained on the languages we participate in the competition

6. After we are done with fine-tuning existing pre-trained models, we will also explore building our own model architecture for all the three tracks.

7. Table 3 gives a tentative timeline of our project.

6 Individual Contribution

Table 4 presents a breakdown of the contributions made by individual members to the project so far.

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 4: Member Contribution

7 Conclusion

This report provided a brief overview of the semEval Task 11 (Bridging the Gap in Text-Based Emotion Detection) where we discussed the problem statement of all tracks, datasets released by the organizer, other available datasets etc. In related works, we presented a literature review on Text based emotion detection from its early developments to the most recent advancements where we found that pretrained transformer models showed state of the art performance on the given tasks. The dataset was also analysed gaining insights about the structure of the dataset like emotion frequencies, length of text etc. Experiments on shortlisted transformer based models with the available datasets gave comparable performance for small models finetuned on emotion datasets (Roberta based) and large pretrained models (Llama3-8B) hinting architectural changes on small models can even lead to better performance than large models. Finally we discussed our road-map ahead and contribution so far made by each member.

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 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.

cardiffnlp. 2021. twitter-roberta-base-emotion-latest. `https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion-latest`.

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.

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*.