Lotus at SemEval-2025 Task 11: RoBERTa with LLaMA-3 Generated Explanations for Multi-Label Emotion Classification

Niloofar Ranjbar
Persian Gulf University / Bushehr, Iran
nranjbar@pgu.ac.ir

Hamed Baghbani
Persian Gulf University / Bushehr, Iran
baghbani.hamed@gmail.com

Abstract

This paper presents a novel approach for multi-label emotion detection, where LLaMA-3 is used to generate explanatory content that clarifies ambiguous emotional expressions, thereby enhancing RoBERTa's emotion classification performance. By incorporating explanatory context, our method improves F1-scores, particularly for emotions like fear, joy, and sadness, and outperforms text-only models. The addition of explanatory content helps resolve ambiguity, addresses challenges like overlapping emotional cues, and enhances multi-label classification, marking a significant advancement in emotion detection tasks.

1 Introduction

Emotion classification plays a crucial role in natural language processing (NLP) for applications like sentiment analysis and emotion-aware dialogue systems (Mohammad and Kiritchenko, 2018). The challenge lies in accurately identifying emotions from text, which are often subtle, multi-faceted, and context-dependent. Furthermore, emotions can be expressed simultaneously, making multi-label classification essential (Belay et al., 2025).

Despite advancements, emotion classification remains complex due to ambiguous emotional expressions and diverse contexts. Early keyword-based methods struggled with generalizing across languages and expressions (Wiebe et al., 2005), and even modern transformer models face challenges with short or under-explained sentences, particularly in multi-label tasks (Kusal et al., 2022; Mohammad and Kiritchenko, 2018).

To address these challenges, we propose a novel approach using Large Language Models (LLMs) to generate explanatory content, enhancing the understanding of ambiguous emotions. We fine-tuned a LLaMA-3 model to generate context-rich explanations for each sentence, improving emotion classification, especially for multi-label settings. The explanatory context significantly boosts performance, as shown in prior work on LLMs and commonsense reasoning (Yang et al., 2023; Xenos et al., 2024). The generated explanations were used with the original text to fine-tune RoBERTa (Liu et al., 2019) for multi-label emotion classification, enabling simultaneous emotion prediction.

We participated in SemEval 2025 Task 11, Subtask 1 (Muhammad et al., 2025b), which focuses on multi-label emotion detection across multiple languages, including English. The dataset consists of social media text annotated by 122 annotators, with multi-label annotations for five emotions: anger, fear, joy, sadness, and surprise. The training set has 2,768 samples, the development set has 116, and the test set includes 2,767 samples, all with binary labels indicating the presence or absence of each emotion. Our system, evaluated on English data, demonstrates that adding explanatory content significantly enhances model performance. Specifically, the Text + Explanation model achieved a Macro F1 score of 0.7396 with a standard deviation of 0.0016 over four runs, outperforming the Text-only model, which had a Macro F1 score of 0.7112 with a standard deviation of 0.0095 over four runs. This shows that explanatory context improves classification accuracy across different classes.

The BRIGHTER dataset (Muhammad et al., 2025a), which addresses the lack of high-quality emotion datasets, serves as the primary resource for this task. It provides labeled data in 28 languages and supports tackling challenges in emotion classification, such as ambiguous or complex emotional expressions.

The code and data used in this study are available for reproducibility[1].

[1] https://github.com/nranjbar/emotion_detection_LLM

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 431–439
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Emotion	**Training Data**	**Development Data**	**Test Data**
Anger	333	16	322
Fear	1611	63	1544
Joy	674	31	670
Sadness	878	35	881
Surprise	839	31	799
Total	**2768**	**116**	**2767**

Table 1: Class Distribution in Training, Development, and Test Data

2 Background

This section provides an overview of the emotion detection task, dataset, and related works, focusing on the use of large language models (LLMs) and contextual information for improving emotion classification.

2.1 Task and Dataset Details

We propose using Large Language Models (LLMs), specifically LLaMA-3, to generate explanations for ambiguous emotional expressions, which are then used to fine-tune RoBERTa for emotion classification. Our results show that the inclusion of explanatory context improves performance compared to using text alone. The emotion distribution across the datasets, shown in Table 1, illustrates the challenges of handling imbalanced classes in multi-label emotion detection.

2.2 Related Works

Recent advancements in emotion recognition have been driven by the use of Large Language Models (LLMs), particularly transformer-based architectures like RoBERTa. demonstrated that fine-tuning pre-trained models significantly improves emotion detection compared to traditional keyword-based methods, which often struggle to generalize across languages and diverse emotional expressions. Transformer models, including RoBERTa, have been successfully applied to fine-grained emotion classification tasks, as shown by Demszky et al. (2020) on the GoEmotions dataset, excelling in multi-label classification.

Efforts to further enhance LLMs for emotion detection have included integrating additional context or knowledge during fine-tuning. For example, Suresh and Ong (2021) proposed augmenting transformers with knowledge-embedded attention mechanisms using emotion lexicons, which improved the recognition of nuanced emotional expressions. Similarly, Xenos et al. (2024) showed that incorporating common-sense reasoning significantly enhances performance, particularly in multi-label contexts.

Specialized models like EmoLLMs, fine-tuned with multi-task affective analysis datasets, have also demonstrated promise in improving emotion detection across a range of domains (Liu et al., 2024). Additionally, DialogueLLM, fine-tuned with emotional dialogues, has improved emotion recognition in conversational contexts, where emotional expression varies depending on the interaction flow (Zhang et al., 2024).

Our work builds upon these approaches by leveraging LLaMA-3 to generate explanatory content that clarifies ambiguous emotional expressions, followed by fine-tuning RoBERTa for multi-label emotion classification. By incorporating explanatory context, we enhance the model's ability to capture complex emotional nuances, aligning with previous findings that emphasize the importance of context in emotion classification.

3 System Overview

The task of multi-label emotion detection in text is inherently complex, especially when emotions are expressed simultaneously in a single sentence. To address this, our system employs a two-phase pipeline: first, generating explanatory content to enhance the understanding of ambiguous emotional expressions, followed by fine-tuning a RoBERTa model for multi-label classification.

3.1 Phase 1: Explanation Generation with LLaMA-3

The first stage of our system employs LLaMA-3, a 7B-parameter language model fine-tuned to generate contextual explanations for text. We chose LLaMA-3 over alternatives like EmoLLMs and DialogueLLM due to its superior ability to produce coherent, general-purpose explanations without explicitly stating emotions—making it well-suited for disambiguating subtle or overlapping emotional cues in multi-label classification.

To prepare for fine-tuning, we randomly selected 150 sentences from the training data. GPT-4 generated explanations for these sentences using the following prompt:

"Read the given text and generate a short explanation of the emotional or situational context behind the sentence. The explanation should be concise and relevant to the sentence. Do not explicitly mention emotions but focus on the implications behind the sentence."

We then fine-tuned LLaMA-3 using these sentence–explanation pairs to ensure it could consistently produce high-quality, emotionally informative content. The resulting explanations were appended to the original inputs, enriching the dataset used to train RoBERTa for final classification.

3.2 Phase 2: RoBERTa Fine-Tuning for Multi-Label Emotion Classification

In the second stage, we utilized the RoBERTa model, a transformer-based architecture known for its high performance in text classification tasks. RoBERTa was fine-tuned on the training data enriched with the explanations generated by LLaMA-3. During this fine-tuning, both the original text and the generated explanations were concatenated with a space between them and then fed into RoBERTa. This approach allowed the model to learn the intricate relationships between emotions and their contextual expressions in the text.

RoBERTa was fine-tuned with binary labels (0 or 1) for each emotion in the dataset: anger, fear, joy, sadness, and surprise. These binary labels indicate the presence (1) or absence (0) of each emotion. The task is a multi-label classification, meaning multiple emotions can be predicted for a given text. This was crucial for handling complex emotional expressions where more than one emotion could be conveyed simultaneously.

3.3 Challenges and Solutions

Our system addressed three main challenges:

- **Ambiguous Emotional Expressions:** Emotion detection is challenging due to the subtle and complex nature of emotions in text. To resolve ambiguity, we used LLaMA-3 to generate additional explanatory context, providing the model with clearer, more explicit information that aids in correctly interpreting emotions, especially when they are not overtly expressed.

- **Multi-label Classification:** Emotions often overlap in natural language, and multiple emotions can be expressed simultaneously. Our system's multi-label classification approach enables it to predict multiple emotions for each input sentence, which is crucial for capturing real-world emotional expressions. This multi-label classification is essential for addressing the intricate and overlapping emotional cues that occur in natural language.

- **Imbalanced Dataset:** Emotion detection tasks often face class imbalance, where some emotions are more prevalent than others. While our system did not explicitly address this issue through over-sampling or under-sampling techniques, the explanatory context generated by LLaMA-3 helped mitigate this imbalance. By providing richer, more contextually informed inputs, LLaMA-3's explanations offered a way to enhance the recognition of less frequent emotions. This context made the model more sensitive to underrepresented emotions by providing additional clarifying information that could compensate for their lesser frequency in the dataset.

3.4 Code and Resources Used

The code for fine-tuning LLaMA-3 is available in the Unslothai GitHub repository. This repository contains the necessary scripts for fine-tuning LLaMA-3.

4 Experimental Setup

We evaluated our multi-label emotion detection approach by fine-tuning RoBERTa with explanatory content generated by LLaMA-3 on the BRIGHTER dataset.

Text preprocessing and tokenization were performed with the `RobertaTokenizer` from Hugging Face. In the first phase, LLaMA-3 generated explanations, which were concatenated with the original text. In the second phase, both the original text and the generated explanations were tokenized together, allowing the model to learn the emotional context.

For fine-tuning LLaMA-3, we used 4-bit quantization and LoRA, with a batch size of 2, gradient accumulation steps of 4, and a learning rate of 1×10^{-4} for 30 training steps. These explanations were then used in the second phase for emotion classification. RoBERTa was fine-tuned with binary emotion labels (0 or 1) for each emotion in the dataset, using a batch size of 8, a learn-

Method	**Macro**			**Micro**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Text + Exp (LLaMA-3) + RoBERTa	0.7421 ± 0.0047	**0.7433 ± 0.0011**	0.7396 ± 0.0016	0.7550 ± 0.0026	**0.7809 ± 0.0027**	0.7678 ± 0.0026
Text Only (RoBERTa)	0.7477 ± 0.0150	0.6831 ± 0.0216	0.7112 ± 0.0095	0.7650 ± 0.0201	0.7372 ± 0.0195	0.7412 ± 0.0209
Text Only (LLaMA-3)	0.7136	0.6563	0.6739	0.7145	0.7175	0.7160
Text + Exp (Mistral) + RoBERTa	**0.7719 ± 0.0051**	0.7206 ± 0.0135	**0.7436 ± 0.0068**	**0.7889 ± 0.0028**	0.7608 ± 0.0083	**0.7746 ± 0.0037**

Table 2: Overall performance comparison across different models.

ing rate of 5×10^{-5}, and 3 epochs. The model performance was evaluated using precision, recall, and F1-scores, including both Macro and Micro F1-scores to assess multi-label classification. Despite only 30 training steps, this light fine-tuning produced explanations that notably improved RoBERTa's downstream performance. All experiments were conducted on Kaggle's GPU resources, which provided the computational power for efficient fine-tuning.

5 Results

In this section, we present the performance of our system, Lotus, on the competition task. Using the Text + Explanation (RoBERTa) method, Lotus achieved a score of 0.7319, outperforming the SemEval Baseline (0.7083), but falling short of the top score of 0.823. Ranked 36th, Lotus performed competitively, although there is still room for improvement to reach the top positions.

5.1 Overall Performance Comparison Across Models

Table 2 provides a summary of the overall performance of Lotus across four methods:

Text + Explanation (LLaMA-3) + RoBERTa: In this approach, LLaMA-3 generates explanations, and these explanations are combined with the original text to fine-tune RoBERTa for emotion classification.

Text Only (RoBERTa): This model uses only the text (without any explanations) to fine-tune RoBERTa.

Text Only (LLaMA-3): This model fine-tunes LLaMA-3 directly with text for emotion classification.

Text + Explanation (Mistral) + RoBERTa: In this method, Mistral generates explanations, which are combined with the original text and used to fine-tune RoBERTa.

Among these methods, Text + Explanation (LLaMA-3) + RoBERTa achieved the best overall performance, with Macro F1 (0.7396) and Micro F1 (0.7678). This approach outperformed the other methods in both recall and F1-score, demonstrating the value of combining LLaMA-3's generative explanations with RoBERTa's emotion detection capabilities.

Text Only (RoBERTa) achieved the highest Macro Precision (0.7477) and Micro Precision (0.7650), indicating better selectivity in its predictions. However, it lagged behind in recall and F1-scores, particularly when compared to Text + Explanation (LLaMA-3) + RoBERTa.

Text Only (LLaMA-3) performed the weakest overall, especially in recall and F1-scores. This highlights the limitations of fine-tuning LLaMA-3 directly with text without the added benefit of explanations.

Text + Explanation (Mistral) + RoBERTa showed performance similar to Text + Explanation (LLaMA-3) + RoBERTa, with slight improvements in recall and F1-score. However, the difference between Mistral and LLaMA-3 was minimal and may not be significant, suggesting that both models can perform similarly when combined with RoBERTa for fine-tuning.

5.2 Performance Comparison for Individual Emotions

Table 3 compares performance across individual emotions, showing precision, recall, and F1-scores for each method.

Anger: Text + Explanation (LLaMA-3) + RoBERTa achieved precision (0.6695), recall (0.6304), and F1-score (0.6479). Text Only (RoBERTa) had the highest precision (0.6892), but lower recall (0.5116) and F1-score (0.5871). Text + Explanation (Mistral) + RoBERTa performed similarly to LLaMA-3, with precision (0.7196), recall (0.6056), and F1-score (0.6577).

Fear: Text + Explanation (LLaMA-3) + RoBERTa achieved the highest recall (0.8739) and F1-score (0.8343), outperforming Text Only (RoBERTa), which had lower recall (0.3200) and F1 (0.5149). Text + Explanation (Mistral) + RoBERTa showed slight improvements in recall (0.8601) and F1-score (0.8416), but the difference with LLaMA-3 was marginal.

Joy: Text + Explanation (LLaMA-3) +

Emotion	**Text + Exp (LLaMA-3) + RoBERTa**			**Text Only (RoBERTa)**			**Text Only (LLaMA-3)**			**Text + Exp (Mistral) + RoBERTa**		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Anger	0.6695	**0.6304**	0.6479	0.6892	0.5116	0.5871	**0.7337**	0.4193	0.5336	0.7196	0.6056	**0.6577**
Fear	0.7983	**0.8739**	0.8343	0.8009	0.8200	0.8149	0.7658	0.8387	0.8006	**0.8238**	0.8601	**0.8416**
Joy	0.7957	0.7291	0.7581	0.7587	0.6925	0.7232	**0.7971**	0.6567	0.7201	0.7687	**0.7687**	**0.7687**
Sadness	0.6831	**0.8127**	0.7423	0.7636	0.6935	0.7268	0.6624	0.7662	0.7105	**0.7743**	0.7321	**0.7526**
Surprise	**0.7625**	0.6702	**0.7132**	0.7248	**0.6877**	0.7039	0.6091	0.6008	0.6049	0.7378	0.6834	0.7096

Table 3: Performance comparison of individual emotions across models with highlighted maximum results.

RoBERTa led in recall (0.7291) and F1-score (0.7581), while Text Only (LLaMA-3) excelled in precision (0.7971). Despite LLaMA-3's higher precision, it had lower recall (0.6567) and F1 (0.7201), trailing behind Text + Explanation (LLaMA-3) + RoBERTa. Text + Explanation (Mistral) + RoBERTa showed similar performance with an F1-score of 0.7687.

Sadness: Text Only (RoBERTa) had the highest precision (0.7636), while Text + Explanation (LLaMA-3) + RoBERTa excelled in recall (0.8127) and F1-score (0.7423). Text + Explanation (Mistral) + RoBERTa showed slight improvements in recall (0.7321) and F1-score (0.7526), with minimal differences compared to LLaMA-3.

Surprise: Text + Explanation (LLaMA-3) + RoBERTa achieved the highest precision (0.7625), while Text Only (RoBERTa) had the highest recall (0.6877). Text + Explanation (Mistral) + RoBERTa showed improved precision (0.7378) and recall (0.6834). LLaMA-3 performed weakest with precision (0.6091), recall (0.6008), and F1-score (0.6049), likely due to its difficulty in capturing the nuances of Surprise compared to other emotions.

6 Discussion

We introduced Lotus, a multi-label emotion detection approach combining LLaMA-3's generative explanations with RoBERTa for emotion classification. This combination significantly improved performance, particularly for nuanced emotions like Fear (F1: 0.8343), Joy (F1: 0.7581), and Sadness (F1: 0.7423), surpassing text-only models.

Integrating LLaMA-3's explanations with RoBERTa effectively balanced precision and recall, outperforming Text Only (LLaMA-3), especially for complex emotions like Fear and Sadness, emphasizing the importance of explanatory context in capturing emotional nuances.

Although LLaMA-3 was initially chosen, smaller models like Mistral and Qwen faced no significant GPU constraints on Kaggle. After testing Mistral, the results were nearly identical to LLaMA-3, suggesting both models perform similarly when fine-tuned with RoBERTa. Further exploration of other models will provide more insights.

For further illustration, Table 4 in the Appendix presents input sentences, predicted emotions, and generated explanations, providing context to clarify emotional intent and improve classification accuracy.

7 Conclusion and Future Work

Lotus showed that combining generative explanations with emotion detection models significantly improves emotion classification, particularly for ambiguous emotions. Using LLaMA-3 for explanation generation and RoBERTa for emotion detection enhanced the system's ability to handle nuanced emotional expressions.

Future work will focus on improving detection of underrepresented emotions like Anger, refining the explanation generation process, and addressing imbalanced datasets. Expanding the model to support multiple languages and emotional contexts will enhance its generalizability. Additionally, we plan to compare Mistral with other models like Qwen and conduct ablation studies to assess their contributions. Further improvements will target challenging emotions like Anger and Surprise, with error analysis and model comparisons refining the system.

References

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.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

4040–4054, Online. Association for Computational Linguistics.

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*.

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. *Preprint*, arXiv:1907.11692.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5487–5496, New York, NY, USA. Association for Computing Machinery.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Varsha Suresh and Desmond C. Ong. 2021. Using knowledge-embedded attention to augment pretrained language models for fine-grained emotion recognition. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

Alexandros Xenos, Niki Maria Foteinopoulou, Ioanna Ntinou, Ioannis Patras, and Georgios Tzimiropoulos. 2024. Vllms provide better context for emotion understanding through common sense reasoning. *Preprint*, arXiv:2404.07078.

Daniel Yang, Aditya Kommineni, Mohammad Alshehri, Nilamadhab Mohanty, Vedant Modi, Jonathan Gratch, and Shrikanth Narayanan. 2023. Context unlocks emotions: Text-based emotion classification dataset auditing with large language models. *Preprint*, arXiv:2311.03551.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2024. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Preprint*, arXiv:2310.11374.

Response to Reviewers

We sincerely thank the reviewers for their valuable and constructive feedback. Below, we provide a point-by-point response outlining how we addressed each comment and question.

Reviewer 1

Comment: *The use of GPT-4 and LLaMA-3 in a master-student framework is interesting. I would have liked to see a comparison of master models. It is commendable that the approach outperforms the baseline and is competitive. The error analysis section was particularly insightful.*

Response: We appreciate the reviewer's positive feedback. In the revised version, we extended our experiments to include **Mistral** as another explanation-generating model alongside LLaMA-3. As shown in Tables 2 and 3, both models achieved comparable results when used with RoBERTa for final emotion classification, indicating that different LLMs can be viable choices for explanation generation. We also clarified this point in the *Discussion* section and plan to expand comparisons to **Qwen** in future work.

Question: *Was the model selection constrained by the GPU resources available on Kaggle? How would this approach fare with comparable models (e.g., Qwen)?*

Response: Yes, our initial selection of LLaMA-3 was partly influenced by the GPU constraints of Kaggle. However, as noted in the revised paper, we found that smaller models like **Mistral** can be used effectively within these constraints. We included Mistral in our updated experiments and observed results similar to LLaMA-3. These findings are now discussed in Section 6 (*Discussion*), and we intend to include **Qwen** in future work for a broader comparison.

Reviewer 2

Comment: *The methodology lacks detail, such as the number of parameters in LLaMA-3 and how the 150 samples were selected.*

Response: We added a clarification in Section 3.1. The **150 sentences** used for explanation fine-tuning were selected **randomly** from the training set to ensure diversity. We also briefly described the configuration of LLaMA-3 (7B parameter variant) and noted that **LoRA** and **4-bit quantization** were applied to make fine-tuning feasible on Kaggle.

Comment: *Results and impact of fine-tuning should be discussed. Did 30 steps significantly influence the model?*

Response: In Section 4 (*Experimental Setup*), we now elaborate on this. Although only **30 steps** of fine-tuning were applied, the generated explanations noticeably improved RoBERTa's performance when appended to the original text, as evidenced by the performance gains in Table 2. We interpret this as indicating that **even light fine-tuning**, when paired with a strong base model and a clear task-specific prompt, can be beneficial.

Comment: *Dataset information is incomplete. The emotion "disgust" appears in some languages but is not mentioned.*

Response: Thank you for pointing this out. We clarified in Section 1 that our experiments are strictly based on the **English portion** of the BRIGHTER dataset, which includes only **five emotions** (anger, fear, joy, sadness, surprise). We added a note explicitly stating that *"disgust"* was part of the multilingual dataset but was **not included** in the English subset we used.

We hope these revisions adequately address the concerns raised and improve the clarity, rigor, and completeness of our work. We again thank the reviewers for their thoughtful input.

A Examples of input texts

Table 4 shows input sentences from the dataset, along with the predicted emotions and the generated explanations for each sentence. These explanations provide additional context, helping to clarify the emotional intent behind the text and improving the model's ability to correctly classify emotions.

B Error Analysis

B.1 Misclassification of Anger

Anger is often misclassified due to subtle emotional cues or when it overlaps with related emotions like frustration or anxiety. For example:

- **Text:** "Man, I can't believe it." **Explanation:** "The speaker expresses surprise or frustration." **Predicted:** Anger = 0, Actual = 1.
- **Text:** "I could not summon up the courage to get up." **Explanation:** "The speaker conveys vulnerability or exhaustion." **Predicted:** Anger = 0, Actual = 1.

These examples indicate that anger is misclassified when the emotional reaction is subtle or related to emotions like frustration or exhaustion, which may not have the overt aggression typically associated with anger.

Additionally, anger is sometimes misclassified due to physical or emotional intensity, which the model may confuse with anxiety or frustration. For example:

- **Text:** "I felt fire in my stomach." **Explanation:** "The speaker describes a strong emotional or physical reaction." **Predicted:** Anger = 0, Actual = 1.
- **Text:** "There was no stopping the relentless torrent." **Explanation:** "The speaker describes an intense, unstoppable force." **Predicted:** Anger = 0, Actual = 1.

These misclassifications suggest that the system struggles to interpret emotional intensity related to anger, and may categorize it as anxiety or frustration instead.

Lastly, anger is sometimes misclassified as fear or sadness, especially when the emotional cue is indirect or combined with vulnerability:

- **Text:** "The weekend didn't live up to my storm standards." **Explanation:** "The speaker expresses disappointment and frustration." **Predicted:** Anger = 0, Actual = 1.
- **Text:** "She was growling, barking, snarling, foaming." **Explanation:** "The speaker describes an intense emotional state, possibly fear or anger." **Predicted:** Anger = 1, Actual = 0.

In summary, anger is misclassified due to the subtlety of its expression or its overlap with other emotions such as frustration or anxiety. Additionally, emotional intensity or indirect cues, especially when mixed with vulnerability, can confuse the model. Future improvements should focus on enhancing the model's ability to differentiate between anger and these overlapping emotional states, and better handle the more subtle or complex expressions of anger.

B.2 Misclassification of Surprise

Surprise is often misclassified due to subtle or ambiguous emotional cues in the text. For example:

- **Text:** "The lock was a dial-lock." **Explanation:** "The speaker describes a specific detail, focusing on the nature of the lock." **Predicted:** Surprise = 1, Actual = 0.
- **Text:** "I immediately started getting nervous and panic intensified." **Explanation:** "The speaker describes anxiety, which may be confused with surprise." **Predicted:** Surprise = 1, Actual = 0.

These examples show that Surprise is sometimes misclassified as confusion or anxiety, especially when the emotional reaction is subtle or combined with other emotions.

Additionally, Surprise is occasionally misclassified as fear or anger, particularly when unexpected events are associated with discomfort or frustration:

- **Text:** "She was growling, barking, snarling, foaming." **Explanation:** "The speaker describes an intense emotional state, possibly fear or anger." **Predicted:** Surprise = 1, Actual = 0.
- **Text:** "I almost got my hands on the door handle, when..." **Explanation:** "The speaker describes a moment of frustration or missed opportunity." **Predicted:** Surprise = 1, Actual = 0.

ID	Text	Emotions	Generated Explanation
1	But not very happy.	Joy and Sadness	The speaker conveys a sense of dissatisfaction or disappointment, but without strong emotion.
2	About 2 weeks ago I thought I pulled a muscle in my calf.	Fear and Sadness	The speaker recounts a minor injury, suggesting concern or discomfort.
3	Yes, the Oklahoma city bombing.	Fear, Anger, Sadness and Surprise	The speaker references a significant historical event, evoking a sense of tragedy or reflection.
4	Dad on the warpath.	Fear and Anger	The speaker conveys tension or anger, likely due to a confrontational situation.

Table 4: Examples of input text, emotions, and generated explanations

These misclassifications suggest that when surprise is combined with aggression, frustration, or physical tension, the system may confuse it with fear or anger.

Finally, Surprise is misclassified when there is a lack of clear emotional cues, particularly when surprise is related to unexpected information:

- **Text:** "My great-grandad was a full-blood Cherokee." **Explanation:** "The speaker introduces their ancestry with pride and a sense of revelation." **Predicted:** Surprise = 0, Actual = 1.

In summary, Surprise is misclassified due to subtle emotional cues, especially when it overlaps with other emotions like fear or anger, or when it is expressed in less overt ways. To improve the model, future work should focus on enhancing its sensitivity to these subtle cues and improving its ability to differentiate Surprise from overlapping emotions.