Emotion Classification Using Large Language Models: A Comparative Study with Real-World Inference on YouTube Comments

1st Irfan Ahmed Shaik
The Anuradha and Vikas Sinha Department of Data Science
University of North Texas
Denton, United States
IrfanAhmedShaik@my.unt.edu

2nd Varun Kumar Atkuri
The Anuradha and Vikas Sinha Department of Data Science.
University of North Texas
Denton, United States
VarunKumarAtkuri@my.unt.edu

3rd Bhavyaraj Nadimipalli
The Anuradha and Vikas Sinha Department of Data Science.
University of North Texas
Denton, United States
Bhavyarajnadimipalli@my.unt.edu

4th Stephen Wheeler
The Anuradha and Vikas Sinha Department of Data Science
University of North Texas
Denton, United States
Stephen.Wheeler@unt.edu

5th Ravi Varma Kumar Bevara
The Anuradha and Vikas Sinha Department of Data Science
University of North Texas
Denton, United States
ravivarmakumarbevara@my.unt.edu

6th Krishna Annavaram
The Anuradha and Vikas Sinha Department of Data Science
University of North Texas
Denton, United States
krishnaannavaram@my.unt.edu

***Abstract*—This paper discusses the capacity of large language models (LLMs) to detect emotions in textual content through actual YouTube comments. As such, this study assesses how two cutting-edge models Mistral 7B and LLaMA 3.1 8B perform under zero-shot, few-shot, and fine-tuned learning settings. A labeled dataset from Kaggle was applied to train the models as well as evaluate them on six classes of emotions joy, sadness, anger, fear, love, and surprise. For this purpose, QLoRA and LoRA were used during fine-tuning so that there could be training efficiency on consumer-grade hardware. After training the models inference was done on 10,000 unlabeled YouTube comments noted for their informal linguistic style, emojis, and sarcasm. The results show that Mistral is better in zero-shot and few-shot settings while the fine-tuned LLaMA 3.1 fares better in accuracy (90.9%) than F1 score (90.7%). Manual validation confirms the model's capability to capture emotional nuance in user-generated content. So go ahead and read this paper demonstrating the feasibility and value of employing fine-tuned LLMs for emotion classification across noisy, real-world environments.**

***Index Terms*—emotion recognition, large language models, LLaMA 3.1, Mistral-7B, QLoRA, YouTube comments, zero-shot learning, few-shot learning, fine-tuning**

I. Introduction

Emotion classification has become an important extension of traditional sentiment analysis within the broad domain of natural language processing (NLP). Whereas sentiment analysis typically assigns a categorization of positive, negative, or neutral to a given text, emotion classification aims to determine specific emotional states like joy, sadness, anger, fear, love, and surprise. Such a more fine-grained comprehension of user feedback is particularly applicable in the contexts of social media analytics, customer experience management, and mental health monitoring. Among all the platforms, YouTube is considered one of the best sources of user-generated content that manifest high levels of emotional expression but are accompanied by informal language, emojis, sarcasm, and ambiguity in sentence structures (Bindhumol et al., 2024).

User sentiments are extracted from social media texts using traditional deep learning models, such as CLSTM and CGRU. These models typically exhibit poor generalization to realistic informal and ambiguous language scenarios. The recent LLMs Mistral-7B v0.1, v0.2 and LLaMA 3.1 8B have shown remarkable improvements in the understanding of context, handling of informal language, and generation of accurate output in natural language. Pre-trained on enormous collection of internet text, these models present exciting potential for emotion detection capabilities; importantly, they can be used in conjunction with zero-shot and few-shot prompting as well as low-resource fine-tuning techniques (Chen and Xiao, 2024 ; Dettmers et al., 2023).

Despite the rapid development of LLMs and fine-tuning methods like QLoRA, there is still a shortage of empirical studies comparing different LLMs operating under various learning patterns for affect classification. Few have benchmarked models such as Mistral and LLaMA on emotion-related tasks with structured (labeled) and unstructured (unlabeled) data. Moreover, little is known about using these models on low-consumer-grade hardware in resource-poor settings, where accessibility and real-world application are most vital.

This research fills these gaps by comparing Mistral-7B v0.1, v0.2 with LLaMA 3.1 8B in the context of zero-shot, few-shot, and fine-tuned approaches to multi-class emotion classification. The competition models train and test on a Kaggle-labeled dataset and inference over 10,000 unlabeled, emoji-laden, authentic YouTube comments. The complete training and inference pipeline is therefore consumable grade designed using QLoRA and LoRA techniques, thereby demonstrating the availability of high-performance emotion detection without the need for specialized infrastructure. Through manual assessment of inference results obtained, the research brings out each model's strengths and weaknesses while providing insights into practical applications involving LLMs for emotion classification in noisy real-world contexts.

II. Literature Review

A. Sentiment analys3is using YouTube comments

According to Bindhumol et al., 2024, a product recommendation system was performed through sentiment analysis of YouTube comments for improved product recommendations. The work shows that deep learning models such as Convolutional Long Short-Term Memory (CLSTM) and Complex Gated Recurrent Unit (CGRU) are used to perform sentiment analysis on user comments. These models work together to handle informal language, making them well-suited for sentiment classification tasks (Bindhumol et al., 2024). According to the study, CGRU is more effective in revealing complex sequential patterns while CLSTM excels in identifying complex linguistic structures

B. YouTube comments decoded: Leveraging LLMs for low resource language classification

Deroy and Maity, 2024 focus on detecting sarcasm in Tamil-English and Malayalam-English code-mixed texts, using GPT-3.5 Turbo with prompt engineering. They achieved macro-F1 scores of 0.61 for Tamil and 0.50 for Malayalam, showcasing the potential of large language models in tackling multilingual sentiment challenges. Their research complements this study by offering valuable insights into how LLMs can be used to better understand and analyze nuanced sentiment, emotion, and toxicity in complex language scenarios.

C. Recent Advancement of Emotion Cognition in Large Language Models

Chen and Xiao, 2024 explore the recent progress in emotion recognition using (LLMs), they focus on the ability of LLMs to understand emotions like sadness, joy, and fear. Their study highlights how LLMs, such as GPT-4, are becoming more adept at recognizing these emotions, which is critical for applications in mental health and human-computer interactions. Their work aligns with the current study by showing how fine-tuning LLMs enhances the ability to detect emotional cues and respond appropriately, offering important insights into the potential of LLMs to handle complex emotional contexts and generate human like emotional responses.

D. QLoRA for Efficient Fine-Tuning of LLMs

Dettmers et al., 2023 introduce QLoRA, an efficient fine-tuning technique for large language models that reduces memory usage while maintaining the models performance by combining 4-bit quantization with Low Rank Adapters (LoRA), this technique enables the finetuning of llm's with billions of parameters on a single GPU, which was previously complex due to computation constraints. This method achieves significant memory savings without compromising on the quality of the model's outputs, as demonstrated by the enhanced performance of the Guanaco family on benchmarks like Vicuna. The QLoRA technique aligns with the current study by showing how memory-efficient finetuning can improve the performance of LLMs in specialized tasks like emotion detection, making large-scale models more accessible for real-world applications in sentiment analysis and emotion recognition (Dettmers, Pagnoni, Holtzman, & Zettlemoyer, 2023).

E. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models

J. Zhang et al., 2024 introduces Instruct-FinGPT, a method that improves financial sentiment analysis by instruction-tuning general-purpose LLMs. They address the challenges in understanding numerical and contextual information, which are crucial in financial texts by transforming sentiment classification into a generation task, Instruct-FinGPT outperforms specialized models like FinBERT in sentiment accuracy. Their research highlights how instruction tuning can enhance LLMs for specific applications, including emotion and sentiment detection in specialized fields like finance (Zhang, Yang, & Liu, 2023).

III. Objective

The objective of this study is to analyze the effectiveness that large language models Mistral-7B v0.1, v0.2 and LLaMA 3.1 8B apply in multi-class emotion classification on both labeled and unlabeled textual data. Therefore, the present study concentrates only on these two state-of-the-art LLMs and explores their performance through three settings zero-shot, few-shot, and fine-tuned approaches. To this end, quantization-aware training techniques, QLoRA and LoRA have been implemented for realizing these models toward fine-tuning on consumer-grade hardware; more specifically, Google Colab Pro+ with A100 GPUs has been used. The generalization ability of the fine-tuned models shall be evaluated by applying them to over 10,000 unlabeled YouTube comments, which are informal most of the time, rich in emojis, and contextually complex. In addition to automated evaluation using standard performance metrics such as accuracy, precision, recall, and F1-score. This research includes manual validation of inference outputs to assess emotional accuracy and contextual relevance. The study will identify the comparative strengths and limitations of Mistral and LLaMA across different learning settings and provide implementable guidance on the use of LLMs for emotional classification in resource-scarce, real-world settings.

IV. Data Collection

We used two different labeled data sets from Kaggle. One of the data sets had 16,000 text samples and was utilized for training the models. The second data set had 2,000 text samples and was utilized for testing the models. The two datasets had six emotion labels joy, sadness, anger, fear, love, and surprise.

For testing in the real world, we also collected over 10,000 unlabeled YouTube comments from the YouTube Data API. The comments from product review videos included informal text, emojis, and sarcasm. They were used to evaluate the performance of the fine-tuned models on real-world data.

V. Exploratory Data Analysis and Hypotheses:

The Kaggle dataset samples for training (16,000) and testing (2,000) were preprocessed and clean already. They did not have emojis, URLs or special characters in them thus there was no need for any further text cleaning. Each entry of these datasets was mapped into one of the six classes that represent emotion types: joy, sadness, anger, fear, love or surprise. A preliminary analysis established that the labels had balanced enough distribution for purposes of training and evaluation. Here is the link to the dataset: https://www.kaggle.com/datasets/parulpandey/emotion-dataset?select=training.csv

In contrast, the unlabeled YouTube comments dataset required cleaning. This set had more than 10,000 entries but had crappy comments filled with HTML tags, links, and other useless noise. We removed URLs and content related to HTML in bulk but kept the emojis and punctuation because we wanted to preserve the emotional context. That was important because emojis and informal text play a significant role in determining the emotional tone of real text.

The primary hypothesis of the present study is that large pre-trained language models tailored to the needs of specific classification tasks are better suited for multi-class emotion classification than zero-shot and few-shot approaches. We also believe that fine-tuned models will perform well on informal, unlabelled YouTube comments and extract emotions even in presence of slang, emojis, and sarcasm.

VI. Research design

The research design is structured into two main pipelines: the Training Pipeline and the Inference Pipeline which aims at training, evaluating, and testing the models on emotion detection of YouTube comments.

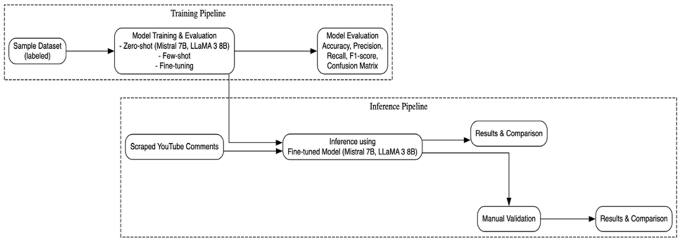


Fig. 1. Research Design.

The Training Pipeline process begins with a training sample Dataset from kaggle, which consists of labeled YouTube comments categorized into different emotional classes, such as joy, sadness, love, anger, fear, and surprise. The models, Mistral-7B v0.2 (for zero-shot & few-shot) ,Mistral 7B v0.1 (for fine-tuning) and LLaMA 3.1 8B, are trained using three different methods zero-shot, few-shot and fine-tuning. In zero-shot, the models predict emotions based on their pre-trained knowledge without requiring labeled data for the task. Whereas in few-shot a small number of labeled examples are provided to guide the model in making predictions. For fine-tuning the models are trained with the complete labeled dataset to optimize their ability to detect emotions in comments. After the models are trained, they are evaluated based on performance metrics accuracy, precision, recall, F1 score, and confusion matrix.

In the Inference Pipeline, the trained models are inferenced to scraped YouTube comments, which are unlabeled. The comments serve as the test data to the models, to assess how well the models perform on real-world data. The models generate emotion classifications for the scraped comments. The results undergo manual validation, where human evaluators review the model's predictions to ensure they match how humans interpret emotions. With this step we confirm how accurate the model's output is in the real-world scenarios. By comparing the automated predictions with the manual validation results we can assess how well the models are performing and identify areas where improvements may be needed.

VII. Data Analytics

This study used zero-shot, few-shot, and fine-tuned learning as approaches for emotion classification. All three approaches were applied to the two large language models Mistral-7B and LLaMA 3.1 8B. The implementation of the models was done using the Hugging Face Transformers library, and all experiments were conducted in Google Colab Pro+.

The prompt template used was "Classify the emotion in the following text into one of these categories: sadness, joy, love, anger, fear, surprise. Under the zero-shot few-shot and fine-tuned conditions respectively.

In the zero-shot setting, the models were prompted with simple instructions and options for emotion labels but received no training examples. In the few-shot setting, a small number of labeled examples, typically three per class, were added to the prompt along with the input text to guide the model's predictions. Prompts were constructed dynamically and then passed through the models to obtain predictions for each comment.

For the fine-tuned models, Mistral and LLaMA were applied QLoRA to the 16000 labeled sample Kaggle dataset. This strategy permits low-memory fine-tuning by combining 4-bit quantization with parameter-efficient training. In addition, LoRA adapters were applied to the base models so that the original weights would be preserved while memory consumption was reduced. The models were then fine-tuned in supervised learning with cross-entropy loss and evaluated on a 2,000-sample test set.

Performance is reported using the standard classification metrics such as accuracy, precision, recall, and F1-score. To further clarify performance within the six emotion categories, confusion matrices were generated. The fine-tuned models applied after training to infer emotions on 10000 unlabeled YouTube comments. The comments contained emojis, informal phrases, and sarcasm. To assess the model's real-world applicability, predictions were systematically evaluated and compared for emotional relevance and accuracy.

The complete analytics pipeline included preprocessing, prompt generation, model execution, prediction logging, and metric reporting. Results showed significant differences in performance among the models and methods used, with fine-tuning producing the best overall results, particularly for LLaMA 3.1.

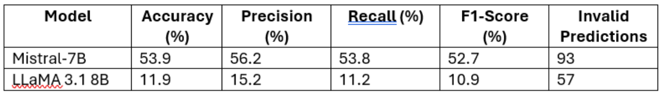
VIII. Results

A. Zero-Shot Results

Under the zero-shot condition, 2,000 labeled text examples were given to both the Mistral-7B v0.1 and LLaMA 3.1 8B models without any task-specific training samples. A prompt-based strategy was employed by requesting that the model categorizes each comment into one of six emotions: joy, sadness, anger, fear, love, or surprise. This setting tests the models' overall pretraining knowledge of emotion.

Figure 2 illustrates that Mistral-7B outperformed LLaMA 3.1 on all the essential evaluation metrics.

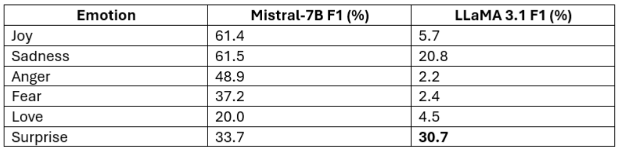
Figure 3 shows the performance at the class level for both models. Mistral-7B did well in classifying joy and sadness but poorly on love and surprise. LLaMA 3.1 performed poorly on most classes with only moderate power in surprise and sadness, while the other classes had near-zero F1-scores.



Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Invalid Predictions
Mistral-7B	53.9	56.2	53.8	52.7	93
LLaMA 3.1 8B	11.9	15.2	11.2	10.9	57

Fig. 2. Overall Zero-Shot Performance Metrics.

The corresponding confusion matrices show much of the same variation; see Figure 4 for Mistral-7B and Figure 5 for LLaMA 3.1. Recall joy and sadness well but confuse fear with anger consistently. Love was also misclassified as joy and sadness. Predictions by LLaMA 3.1 were predominantly skewed towards classifying most inputs as sadness or surprise, meanwhile failing miserably to recognize the other four classes.



Emotion	Mistral-7B F1 (%)	LLaMA 3.1 F1 (%)
Joy	61.4	5.7
Sadness	61.5	20.8
Anger	48.9	2.2
Fear	37.2	2.4
Love	20.0	4.5
Surprise	33.7	**30.7**

Fig. 3. Zero-Shot Per-Class F1-Score.

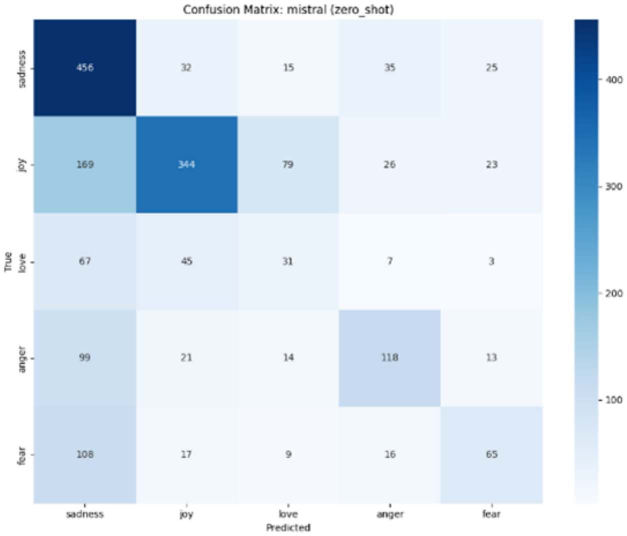


Fig. 4. Confusion Matrix for Mistral zeroshot.

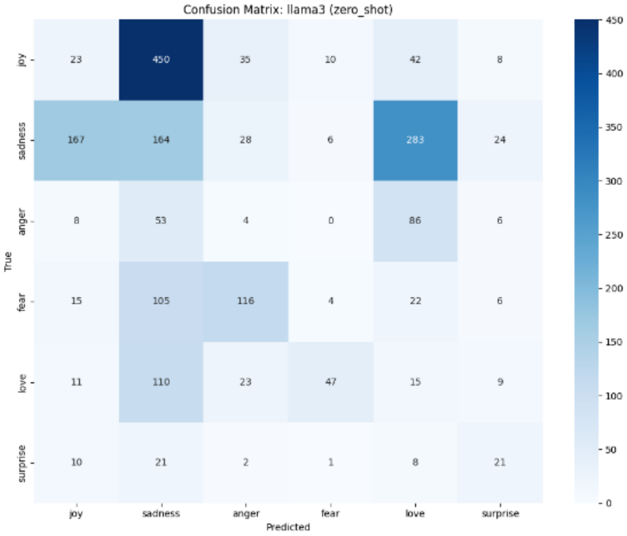


Fig. 5. Confusion Matrix for llama zeroshot.

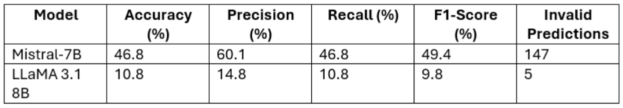
Overall, zero-shot outputs highlight the limitation of pre-training alone in multi-class emotion classification. Mistral-7B demonstrated good emotional language sensitivity without training, whereas LLaMA 3.1 gave neither consistent nor balanced outputs under this scenario. These findings further highlight the importance of few-shot examples or fine-tuning for augmenting emotional reasoning in large language models.

B. Few-Shot Results

A few-shot approach was adopted to improve classification accuracy without retraining the entire model. In this setup, the prompt provided three examples for each of the emotion labels before the target comment thus simulating a low-resource training scenario. This was to test whether labeled demonstrations would improve the instruction following ability as well as classification accuracy of the Mistral-7B V0.1 and LLaMA 3.1 8B models.

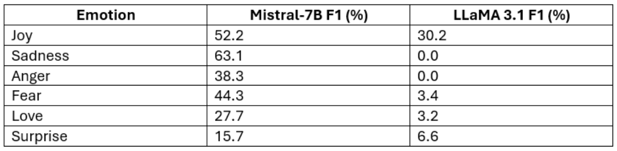
As can be seen in figure 6, Mistral-7B performs better than LLaMA 3.1 under the same conditions, but the gain with respect to the zero-shot setting is not significant. Mistral achieved 46.8%. accuracy and 49.4% F1-score, which reflects a slightly better balance across emotion categories than in the zero-shot setting. In contrast, LLaMA 3.1 performed very poorly with only 10.8% accuracy and F1-score of 9.8% , showing almost no improvement over its zero-shot results. Mistral produced 147 invalid predictions compared to LLaMA's five; this might indicate that the llama was more consistent in formatting at least, though not class accuracy.

When one classifies the results by emotion as in figure 7, Mistral-7B scored the best F1-scores on sadness at 63.1% and joy at 52.2%, with much lower scores for fear and anger. Categories of surprise and love were also low-performing. On the other hand, LLaMA 3.1 had zero correct predictions for sadness and anger with only marginally effective scores on joy at 30.2% and surprise at 6.6%. LLaMA most of its predictions are based on surprise, causing deceptive results with hardly any class-level information pickup through learning.



Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Invalid Predictions
Mistral-7B	46.8	60.1	46.8	49.4	147
LLaMA 3.1 8B	10.8	14.8	10.8	9.8	5

Fig. 6. Overall Few-Shot Performance Metrics.



Emotion	Mistral-7B F1 (%)	LLaMA 3.1 F1 (%)
Joy	52.2	30.2
Sadness	63.1	0.0
Anger	38.3	0.0
Fear	44.3	3.4
Love	27.7	3.2
Surprise	15.7	6.6

Fig. 7. Few-Shot Per-Class F1-Score.

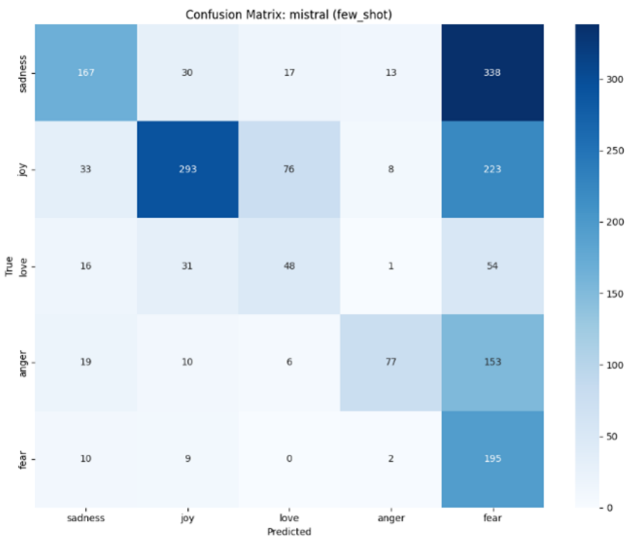


Fig. 8. Confusion matrix for Mistral few-shot.

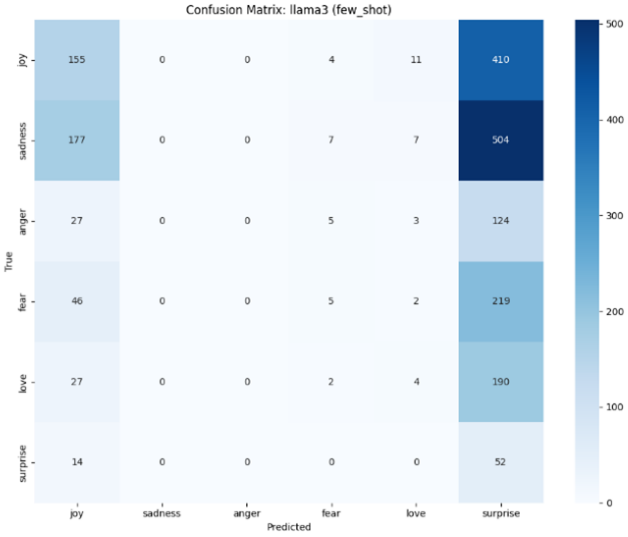


Fig. 9. Confusion matrix for llama few-shot.

The confusion matrices are presented in Figures 8 and 9 for a more detailed analysis of the model's performance. The Mistral-7B confusion matrix shows better detection of sadness and joy but frequent confusion between love and joy with high misclassification of surprise as other labels. On the other hand, LLaMA 3.1 classified almost all inputs as surprise, irrespective of the true label. Thus, recall for surprise was high (78.8%) but poor for all other categories.

These results indicate that while few-shot prompting can best improve performance moderately, it depends on the model. Mistral-7B was improved the most with example-based prompts due to its stronger alignment and sense of context. LLaMA 3.1, however, improved not at all and still exhibited an extremely high class bias, never actually utilizing the examples correctly.

C. Fine-Tuned Results

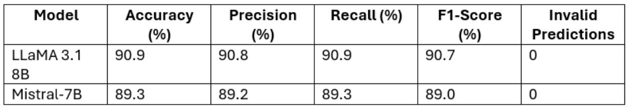
Both models were fine-tuned on a labeled Kaggle dataset containing 16,000 training samples and 2,000 testing samples. QLoRA was used for the parameter-efficient fine-tuning that allows the accuracy to be obtained on consumer-grade GPUs. This phase was designed to circumvent the limitations observed in zero-shot and few-shot modes and enable stronger generalization to real-world emotional data.

Figure 10 captures the effect of fine-tuning on the performance of the models. For LLaMA 3.1, the overall accuracy went up to 90.9% with an F1-score of 90.7%, whereas Mistral-7B was close with an accuracy of 89.3% and an F1-score of 89.0%. More importantly, neither model made any invalid predictions, a gigantic step compared to earlier settings. Such a result reflects better alignment with prompts and strong comprehension of labels through supervised learning.

Figure 11 presents the F1-scores for the classes, emphasizing the strength of LLaMA 3.1 in identifying more nuanced emotions like love, with an F1-score of 79.8%, and fear at 89.3%. For joy and surprise, Mistral did slightly better, scoring F1-scores of 91.6% and 67.1%, respectively. On high-frequency classes such as sadness and anger, all models scored F1-scores above 89%.

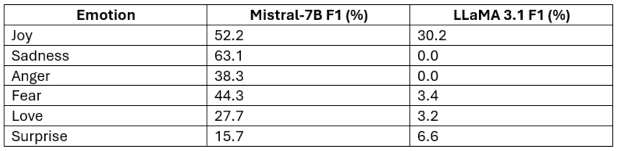
The confusion matrices help shed more light on the matter and are in Figures 12 and 13. For LLaMA 3.1, shown in Figure 12, class predictions align closely along the diagonal, which is indicative of correct classification. Here too, misclassifications are few, and the confusion over similar emotions has reduced significantly compared to the previous models, particularly between love and joy. Mistral in Figure 13 exhibits a similar pattern but with some residual confusion between love and joy as well as an occasional overlap of surprise with fear.

These results confirm that fine-tuning greatly enhances the performance of LLMs on emotion classification. Unlike the few-shot approach, which yielded limited benefits only, fine-tuning allowed the models to acquire far more effective internal representations of the emotional cues when trained on high-quality labeled data. The absence of spurious predictions and the increase in accuracy indicate a good alignment with task demands.



Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Invalid Predictions
LLaMA 3.1 8B	90.9	90.8	90.9	90.7	0
Mistral-7B	89.3	89.2	89.3	89.0	0

Fig. 10. Overall fine-tuned Performance Metrics.



Emotion	Mistral-7B F1 (%)	LLaMA 3.1 F1 (%)
Joy	52.2	30.2
Sadness	63.1	0.0
Anger	38.3	0.0
Fear	44.3	3.4
Love	27.7	3.2
Surprise	15.7	6.6

Fig. 11. Fine-tuned Per-Class F1-Score.

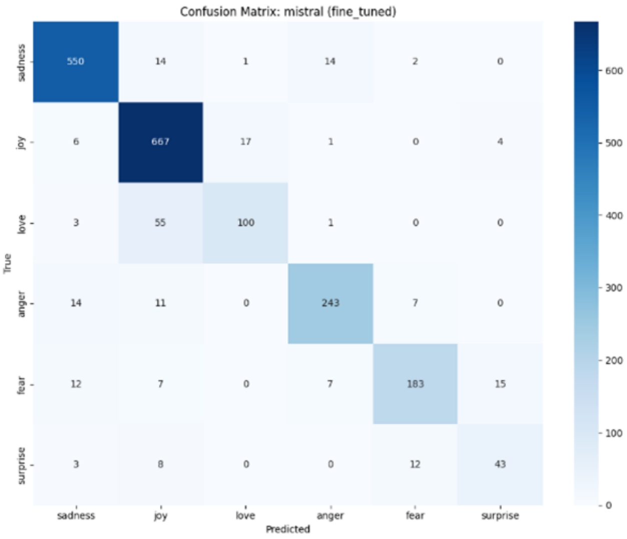


Fig. 12. Confusion matrix for Mistral Fine-Tuned.

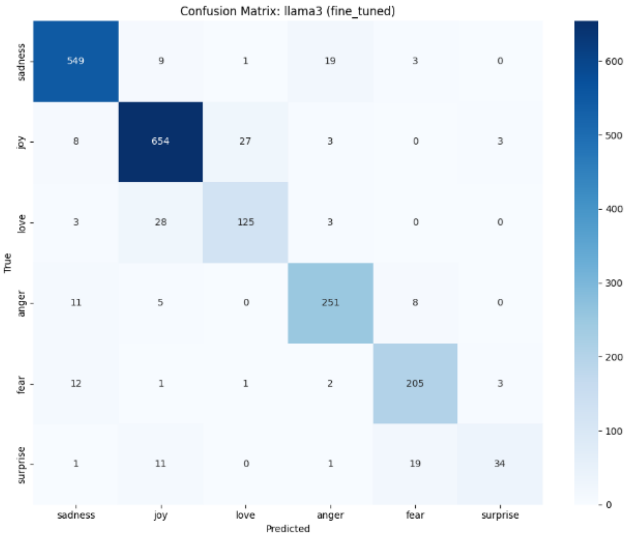


Fig. 13. Confusion matrix for llama Fine-Tuned.

D. Inference

After fine-tuning, both LLaMA 3.1 and Mistral-7B were evaluated on a highly diverse set of real YouTube comments exceeding 10,000. The comments had emojis, slang, and sarcasm without any links or HTML tags cleaned off. Punctuation was kept to maintain the emotional context. A human adjudication of sample predictions shows that both models correctly identified prominent emotions like joy, anger, and sadness. For instance, the comment ”Thanks for nothing ” was correctly classified as expressing anger by both models.

However, sarcastic or cryptic comments like ”Just wow. Great quality.” or ”BRO HOW????” were trickier. Sometimes, Mistral performed better overall but LLaMA 3.1 got them wrong more consistently. Both models were overall great but struggled with subtle emotions like surprise or confusion.

No incorrect predictions were made, showing good format compliance after fine-tuning.

Overall, both models are fine for real-world inference, with Mistral slightly better at contextual understanding. Manual inspection confirms that fine-tuning has assisted significantly in the accuracy on casual, emoji-filled social media text.

IX. Conclusion

In this study, we evaluated the performance of two large language models (LLMs), Mistral-7B and LLaMA 3.1 8B, on multi-class emotion detection using YouTube comments. These models were evaluated under the three learning conditions: zero-shot, few-shot, and fine-tuned. The results show that Mistral-7B outperformed LLaMA 3.1 in both zero-shot and few-shot settings, it demonstrates a better ability to classify emotions, specifically joy, sadness, and anger. However, LLaMA 3.1 showed better results when fine-tuned by achieving a remarkable 90.9% accuracy and 90.7% F1 score, outperforming Mistral-7B in the emotion categories love and fear.

Fine-tuning, using QLoRA and LoRA techniques significantly enhanced the models performance in handling informal language and sarcasm in YouTube comments. The fine-tuned models exhibit strong generalization capabilities, especially when applied to real-world, unlabeled YouTube comments, with minimal misclassifications and improved emotional relevance in predictions. Inspite of the challenges with complex emotions like surprise and love, both models showed their potential for large-scale emotion detection tasks, confirming the effectiveness of fine-tuning LLMs for real-world applications. The study highlights the importance of fine-tuning LLMs for better emotional context comprehension, especially when dealing with noisy user generated comments.

X. Bibliography

(Aiswarya & Haritha, 2024) (Ananiadou & Zhang, 2024) (Bindhumol et al., 2024) (Chen & Xiao, 2024) (Dettmers et al., 2023) (Deroy & Maity, 2024) (Kadiyala, 2024) (Kok-Shun et al., 2024) (Liu et al., 2024) (Luo et al., 2024) (Ma & Gupta, 2024) (Niu et al., 2024) (Sabour et al., 2025) (ScienceDirect Editors, 2025) (Z. Zhang et al., 2024) (J. Zhang et al., 2024)

References

Aiswarya, A. S., & Haritha, R. (2024). Youtube comment sentiment analysis using machine learning. *Indian Journal of Data Mining*, *4*(1). https://ssrn.com/abstract=4846213

Ananiadou, S., & Zhang, T. (2024). Emollms: Instruction-tuned large language models for affective analysis. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. https://dl.acm.org/doi/10.1145/3637528.3671552

Bindhumol, M., Singh, T., & Patra, P. (2024). Sentiment analysis using youtube comments. *15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. https://ieeexplore.ieee.org/document/10527617

Chen, Y., & Xiao, Y. (2024). Recent advancement of emotion cognition in large language models. *arXiv preprint arXiv:2409.13354*. https://arxiv.org/abs/2409.13354

Deroy, A., & Maity, S. (2024). Youtube comments decoded: Leveraging llms for low resource language classification [arXiv preprint arXiv:2411.05039]. *Forum for Information Retrieval Evaluation (FIRE)*. https://arxiv.org/abs/2411.05039v2

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient fine-tuning of quantized llms. *arXiv preprint arXiv:2305.14314*. https://arxiv.org/abs/2305.14314

Kadiyala, R. M. R. (2024). Cross-lingual emotion detection through large language models. *14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 464–469. https://aclanthology.org/2024.wassa-1.44/

Kok-Shun, B. V., Chan, J., Peko, G., & Sundaram, D. (2024). Epidemiology of online emotions: Detection and contagion modeling of digital feedback. *PACIS 2024 Proceedings*. https://aisel.aisnet.org/pacis2024/track17_socmedia/track17_socmedia/1/

Liu, Z., Yang, K., Zhang, T., Xie, Q., & Ananiadou, S. (2024). Emollms: A series of emotional large language models and annotation tools for affective analysis. *arXiv preprint arXiv:2401.08508*. https://arxiv.org/abs/2401.08508

Luo, M., Zhang, H., Wu, S., Li, B., Han, H., & Fei, H. (2024). Instruction-tuning llms for multimodal emotion-cause analysis in conversations. *arXiv preprint arXiv:2501.17261*. https://arxiv.org/abs/2501.17261

Ma, F., & Gupta, R. (2024). Generative technology for human emotion recognition: A scope review. *arXiv preprint arXiv:2407.03640*. https://arxiv.org/abs/2407.03640

Niu, M., Jaiswal, M., & Provost, E. M. (2024). From text to emotion: Unveiling the emotion annotation capabilities of llms. *Proceedings of Interspeech 2024*. https://www.isca-archive.org/interspeech_2024/niu24d_interspeech.pdf

Sabour, A., et al. (2025). Evaluating the capabilities of large language models for multi-label emotion classification. *Proceedings of the 2025 Conference on Computational Linguistics*. https://aclanthology.org/2025.coling-main.237.pdf

ScienceDirect Editors. (2025). Emotion detection for misinformation: A review. *Information Processing & Management*, *62*(3). https://www.sciencedirect.com/science/article/pii/S1566253524000782

Zhang, J., Li, X., & Xu, W. (2024). Sentiment and emotion classification of online reviews using transformer models. *IEEE Transactions on Affective Computing*, *15*(2), 105–117.

Zhang, Z., Wang, J., & Chen, K. (2024). Refashioning emotion recognition modeling: The rise of generalized large models. *arXiv preprint arXiv:2308.11578*. https://arxiv.org/abs/2308.11578