Emotion Detection Using Multilabel Classification with PEFT and Efficient Model Adaptation

GitHub - Jyothsna-Reddy-CJ/ML-Final-Project

Yasaswi IGS
Computer Science
UTD, Richardson,TX
gxi230000@utdallas.edu

Jyothsna Reddy C J
Computer Science
UTD, Richardson,TX
jxc220075@utdallas.edu

***Abstract* - Emotion detection of tweets is a vital component in NLP applications like sentiment analysis and social media monitoring. Traditional models such as FFNN, while effective, encounter challenges related to efficiency and scalability, especially in multilabel classification tasks. This research addresses these limitations by conducting a comprehensive comparative study across various architectures and optimization techniques. The study begins with a custom feed-forward neural network and progresses to pre-trained transformer models as benchmarks like DistilBERT, DistilRoberta and Flan T5. It then evaluates advanced transformer models like google/gemma-1.1-2b-it and Alibaba-NLP/gte-QWEN1.5 from Hugging Face for performance assessment.**

To enhance model efficiency, Parameter Efficient Fine-Tuning (PEFT) techniques, including LoRA (Low-Rank Adaptation) and IA3 are applied to pre-trained models. Additionally, quantization techniques like QLoRA (Quantized Low-Rank Adaptation) are utilized on models from the MTEB benchmark leaderboard to optimize performance in resource-constrained environments. Model performance will be evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrices to validate the models' ability to correctly classify emotions and to provide a comprehensive view of classification performance. This approach is expected to yield insights into optimizing transformer-based models for efficient and scalable emotion detection in real-world applications, demonstrating improvements as the architectures and techniques evolve.

Keywords: Emotion Detection, Multilabel Classification, Feed Forward Neural Networks (FFNN), Transformer Models, DistilBERT, RoBERTa, Flan T5, Parameter Efficient Fine-Tuning (PEFT), Low-Rank Adaptation (LoRA), IA3, Quantized Low-Rank Adaptation (QLoRA)

I. INTRODUCTION

Emotion detection in natural language processing (NLP) is becoming increasingly critical across various domains, including social media monitoring, customer feedback analysis, and mental health assessments. For instance, during significant events like the COVID-19 pandemic, real-time sentiment analysis on social media provided key insights that informed policy decisions and organizational responses. However, detecting multiple emotions from a single text input-known as multilabel classification-presents distinct challenges due to the high-dimensional nature of textual data. Traditional models, such as Long Short-Term Memory (LSTM) networks, have been effective for smaller datasets but face scalability and computational efficiency constraints when applied to larger, real-world datasets.

Transformer-based models, such as DistilBERT, distilRoBERTa, and FLAN T5, have set new benchmarks in NLP, particularly due to their ability to capture intricate contextual relationships within text. These models offer strong performance for tasks like emotion detection, especially in multilabel scenarios. However, their significant computational and memory requirements can limit their applicability in real-time or resource-constrained environments, such as mobile platforms or live customer feedback systems. This raises the need for solutions that can balance high accuracy with computational efficiency, enabling the deployment of these models in real-time applications.

To address these challenges, this project proposes leveraging Parameter Efficient Fine-Tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA), IA3 and Quantized LoRA (QLoRA), to enhance the scalability and efficiency of transformer models. These techniques aim to reduce the computational overhead typically associated with fine-tuning while maintaining high performance in emotion detection tasks. This approach has the potential to bridge the gap between the high accuracy of transformers and the practical efficiency required for real-time applications, facilitating the deployment of these models in environments with limited resources.

II. PROBLEM DEFINITION

Understanding emotions expressed in short text posts, such as tweets, is a critical component of social media analysis with significant real-world implications. Applications include mental health monitoring, where patterns of emotional distress or positivity can inform interventions, public sentiment analysis to gauge societal reactions to events or trends, and customer engagement, where analyzing emotional feedback enhances user experience. Multi-label emotion classification, where a single text can express multiple emotions simultaneously (e.g., joy and trust), poses unique challenges due to the overlapping nature of emotions and the ambiguity of short-text contexts.

Current approaches face several limitations. Traditional models such as Long Short-Term Memory (LSTM) networks, while effective at handling sequential data, struggle with computational efficiency and scalability when applied to large-scale datasets. These limitations make them impractical for real-world multi-label emotion classification tasks. Transformer-based models like DistilBERT, RoBERTa, and FLAN T5 offer significantly better performance due to their ability to capture nuanced contextual relationships. However, their high computational and memory requirements hinder their deployment in resource-constrained environments such as real-time sentiment monitoring or mobile applications.

To address these challenges, this work proposes a phased approach that combines transformer-based models with Parameter-Efficient Fine-Tuning (PEFT) techniques to balance computational efficiency and performance. Initially, a custom feed forward neural network model is built and evaluated to establish baseline metrics to multi-label tasks. Performed implementation of various basic and advanced transformer models and then fine-tuned using techniques like Low-Rank Adaptation (LoRA) and Instruction-Aware Adaptation (IA3), which reduce the number of parameters trained, significantly improving computational efficiency and scalability without requiring full re-training of the models. Additionally, Quantized LoRA (QLoRA) is employed to further reduce memory usage, enabling efficient performance in resource-limited environments. It is important to note that while PEFT techniques optimize resource usage, they do not inherently guarantee performance improvements and require careful implementation and tuning.

The solution addresses the specific challenges of multi-label emotion classification, including emotion overlaps, short - text ambiguity , and the need for specialized evaluation metrics like F1-macro, F1-micro, and Hamming Loss. A comparative study assesses performance improvements across models, including the baseline LSTM and the fine-tuned transformers. Key metrics such

as accuracy, precision, recall, and F1-score are used to validate the effectiveness of the proposed approach in real-world applications.

In summary, this work aims to optimize multi-label emotion classification by leveraging PEFT techniques to reduce computational overhead while maintaining high performance. By addressing the limitations of traditional and transformer-based models, this solution enables the practical deployment of robust and scalable models for applications like mental health monitoring, public sentiment analysis, and customer engagement.

A. Problem Formalization

The task of multi-label emotion classification in short texts, such as tweets, can be defined as follows: the system processes a dataset $T=\{t1,t2,\ldots,tn\}$, where each t_i is a short text instance, and a predefined set of emotion labels $L=\{l1,l2,\ldots,lm\}$. For each instance t_i, a subset of labels $L_i \subseteq L$ represents the emotions conveyed in t_i.

The objective is to develop a predictive $f: T \rightarrow P(L)$, where $P(L)$ represents the power set of L. The model predicts a subset L_i^* for each instance t_i that closely approximates the true subset L_i. The goal is to optimize key metrics like F1-macro and F1-micro while minimizing errors such as Hamming Loss, ensuring precise and well-balanced multi-label predictions.

This formalization provides a structured approach to tackling the unique challenges of multi-label classification in short-text contexts, focusing on accuracy, computational efficiency, and scalability.

B. Challenges in Multi-Label Emotion Classification

The challenges in this domain can be categorized into problem-level and solution-level challenges:

1. Problem-Level Challenges

- **Emotion Overlap**: Emotions in short texts often co-occur, with a single tweet potentially expressing multiple emotions (e.g., joy and trust). Capturing the interdependencies between labels is complex yet essential for accurate classification.
- **Short Text Ambiguity**: Tweets are inherently brief and frequently lack sufficient context, complicating the task of interpreting nuanced emotional expressions compared to longer text forms.
- **Label Imbalance**: Certain emotions are underrepresented in datasets, leading models to favor more frequent labels, which can reduce their ability to generalize across all classes.
- **Multi-Label Complexity**: The exponential growth of the prediction space with the number of possible labels increases the difficulty of both optimization and evaluation.

2. Solution-Level Challenges

- **Computational Constraints**: Training transformer-based models like RoBERTa and FLAN T5 demands extensive computational resources, which can make them impractical in resource-constrained scenarios.
- **Data Size Limitations**: Due to resource constraints, training is often performed on subsets of data, which can adversely affect model generalization and performance.
- **Model Complexity**: Advanced architectures like Google's Gemma and Alibaba Qwen require significant tuning and training time, adding to their computational expense.
- **Evaluation Instabilities**: Complex architectures occasionally face issues like NaN values in validation loss during training, indicating potential shortcomings in model or loss function design.
- **Scalability and Efficiency**: Deploying models efficiently on real-time systems, such as edge devices or mobile applications, requires scalable approaches that balance resource usage and performance.

C. Key Concepts in Proposed Solutions

- **Multi-Label Classification**: This approach enables the model to predict multiple labels for a single instance, addressing overlapping emotional states often found in short texts.
- **Parameter-Efficient Fine-Tuning (PEFT)**: Techniques such as Low-Rank Adaptation (LoRA), Instruction-Aware Adaptation (IA3), and Quantized LoRA (QLoRA) are applied to reduce the computational overhead of fine-tuning large transformer models. These methods allow for effective fine-tuning by training only a subset of parameters, significantly lowering memory and resource requirements.
- **Quantization Techniques**: QLoRA enhances the efficiency of large-scale models like Alibaba Qwen (7B parameters), enabling them to operate effectively in resource-limited environments without sacrificing significant performance.
- **Comparative Analysis**: A systematic comparison of baseline models (e.g., distilBERT, DistilRoberta, Flan-T5) and advanced transformers(e.g.,Google-gemma1.1, Alibaba-NLP/gte-QWEN1.5), fine-tuned PEFT techniques(e.g.,LORA, IA3, QLORA), highlights the trade-offs between computational efficiency and predictive performance.

III. RELATED WORK

A. Hochreiter and Schmidhuber [1] introduced the Long Short-Term Memory (LSTM) model, a significant advancement in Recurrent Neural Networks (RNNs). LSTMs are highly effective in handling long-term dependencies in sequential data, making them widely applicable to tasks such as sentiment analysis and emotion detection. Their ability to maintain information over extended sequences has made LSTMs a valuable tool in early emotion detection tasks. However, LSTMs face scalability challenges when applied to large-scale datasets, which our work addresses by exploring more scalable transformer-based models, enhanced with fine-tuning techniques like LoRA and QLoRA.

B. Zhou et al. [2] explored the use of RNNs and LSTMs for emotion detection, demonstrating their strength in processing sequential data and detecting emotional patterns. However, their study pointed out the limitations of LSTMs in handling larger datasets and complex multilabel classification tasks. Our work builds on this by leveraging transformer models, which are better suited for large-scale emotion detection, and optimizing them using LoRA and QLoRA to improve computational efficiency while maintaining high accuracy.

C. Hu et al. [3] proposed Low-Rank Adaptation (LoRA), a fine-tuning method designed to reduce the number of trainable parameters in large transformer models, significantly reducing computational costs without sacrificing performance. Their work laid the foundation for parameter-efficient fine-tuning, which we extend by applying LoRA and Quantized LoRA (QLoRA) to

transformer models like DistilBERT and RoBERTa in the context of emotion detection. These techniques allow us to maintain high performance while improving scalability and reducing memory requirements, making transformer models viable in real-time emotion detection applications.

D. Rezapour et al. [4] compared transformer-based models, such as BERT and RoBERTa, with traditional models like LSTMs for emotion detection, demonstrating the superior performance of transformers in capturing complex emotions from text. While their work highlighted the effectiveness of transformers in multilabel classification tasks, it also noted the computational intensity of these models. Our work builds on their findings by employing LoRA and QLoRA to make transformer models more efficient, thereby addressing the scalability concerns raised in their study and enabling real-time deployment.

E. Samghabadi et al. [5] focused on the use of hierarchical transformers for emotion recognition, emphasizing their ability to capture subtle emotional nuances more effectively than traditional models like LSTMs. While their research showcased the benefits of transformers in emotion detection, it also recognized the computational demands associated with these models. We address this challenge by applying LoRA and QLoRA to optimize transformer models for real-time, resource-constrained environments, ensuring that the models can be deployed effectively without sacrificing performance.

IV. BACKGROUND

In this section, we provide an overview of key concepts, terminology, and prior techniques essential for understanding emotion detection using Parameter Efficient Fine-Tuning (PEFT) techniques.

1. Multilabel Classification in Emotion Detection: Multilabel classification is a crucial aspect of emotion detection, where multiple emotions can be expressed in a single piece of text. Unlike single-label classification, this approach requires models to detect multiple emotions (e.g., joy, anger, surprise) from the same input. Traditional models, like Long Short-Term Memory (LSTM) networks, were effective at capturing sequential dependencies, making them useful in early emotion detection tasks. However, scaling these models for real-time, large-scale datasets posed challenges, particularly in terms of computational efficiency and memory usage. Transformer-based models, such as DistilBERT and RoBERTa, have since emerged as promising alternatives due to their ability to capture nuanced, context-rich relationships within the text, enhancing accuracy in multilabel tasks.

2. Terminology and Concepts

Understanding several key concepts is crucial for grasping the approach used in this project:

i) Emotion Detection: Identifying one or more emotions from textual data, such as joy, sadness, anger, or fear.

ii) Multilabel Classification: Assigning multiple labels (emotions) to a single text input.

iii) Low-Rank Adaptation (LoRA): A fine-tuning technique that reduces the number of parameters updated during training, making it more computationally efficient while retaining performance.

iv) Quantized LoRA (QLoRA): A further optimization of LoRA that applies quantization to reduce memory and computational demands even further, particularly useful in resource-limited environments.

v) Pre-trained Transformers: Models like DistilBERT, RoBERTa, and FLAN T5 are pre-trained on large datasets and then fine-tuned for specific tasks, such as emotion detection.

3. Previous Techniques in Emotion Detection

Earlier approaches to emotion detection relied on rule-based systems and later, LSTM networks**,** which could capture the temporal relationships in text. While these models provided a strong foundation for sequential tasks, they struggled to scale efficiently with increasing data sizes and computational demands, particularly in multilabel classification scenarios. The introduction of transformer-based models like BERT and its variants marked a significant improvement in accuracy and contextual understanding. Despite these advantages, full fine-tuning of transformers is resource-intensive, requiring substantial computational power, making it challenging to deploy in real-time systems or on resource-constrained devices.

4. Transformer Models and Domain-Specific Embeddings

Pre-trained models like DistilBERT, RoBERTa, and FLAN T5 have revolutionized emotion detection tasks by providing a more contextually accurate understanding of text. These models use embeddings-dense vector representations of text-to capture semantic similarities, which can enhance the model's ability to recognize subtle emotional expressions. LoRA and QLoRA aim to optimize these embeddings by reducing the number of parameters that need to be fine-tuned, making the models more efficient. By using domain-specific
embeddings-such as those fine-tuned for emotion detection the models can better understand the subtleties of emotional language, improving both accuracy and efficiency.

5. Fine-Tuning and Optimization Techniques

LoRA and QLoRA are fine-tuning methods designed to reduce the computational load typically associated with transformer models. LoRA focuses on reducing the number of trainable parameters by updating low-rank matrices, while QLoRA further reduces the memory footprint through quantization techniques. These approaches are particularly important for scaling transformer models to real-time emotion detection tasks in resource-constrained environments, where memory and processing power are limited. This makes LoRA and QLoRA effective in applications like social media monitoring or mobile-based emotion detection systems

V. A MOTIVATING EXAMPLE

To illustrate the application of our proposed multi-label emotion classification framework, consider a scenario where a mental health organization monitors public sentiment during a global crisis through tweets. For instance, a tweet reads:

"Feeling overwhelmed but grateful for my supportive family during these tough times."

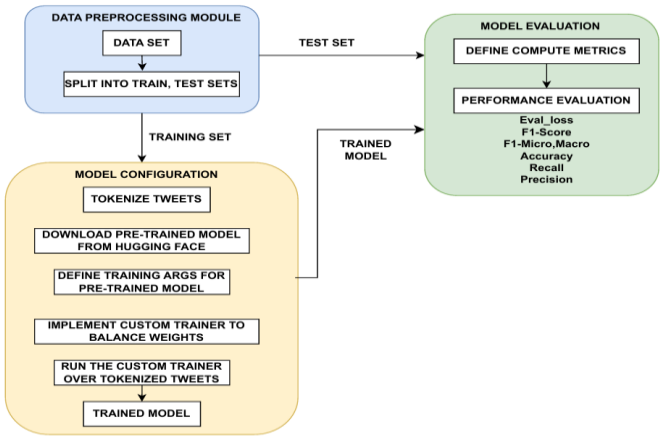
The system processes this tweet through the pipeline. First, the text is preprocessed, including tokenization and removal of noise, to prepare it for analysis. Using a fine-tuned transformer model (e.g., GEMMA or QWEN enhanced with LoRA, IA3 and QLoRA), the framework predicts multiple emotional labels. In this case, the output labels are "gratitude" and "anxiety", reflecting the complex emotional state conveyed in the text.

These results are then aggregated to inform actionable insights. For example, the organization might observe an increase in tweets expressing anxiety and gratitude, prompting targeted mental health interventions such as public awareness campaigns about coping mechanisms or resources for emotional support.

This example demonstrates the framework's ability to process short, ambiguous text, capture overlapping emotions, and provide valuable insights, highlighting its utility in real-world scenarios like mental health monitoring and public sentiment analysis.

VI. METHODOLOGY

i. CONTROL FLOW DIAGRAM



The Dataset is chosen from Kaggle, it contains 10985 rows where each row contains a tweet in text format followed by various 11 emotion labels (anger, happiness, joy, sadness, love and anxiety) as 0 or 1 per tweet.

A. **Data Preprocessing Module**
   a. **Load Dataset**: Use Pandas to load datasets [features vs label-11 emotions] (CSV/JSON) with tweet text and labels.
   b. **Text Cleaning**: Remove URLs, hashtags, mentions, special characters, and optionally apply lowercasing, stemming, or lemmatization.
   c. **Data Splitting**: Split the dataset into training, validation, and test sets using train_test_split [80:20] from scikit-learn.
   d.

B. **Model Configuration**
   a. **Tokenization**: Tokenize tweets with a tokenizer (auto-tokenizer from transformers module), pad, and truncate sequences for uniform input length.
   b. **Pre-Trained Model**: Import a pre-trained model (e.g., DistilBERTForSequenceClassification) and adjust the output layer for target classes.
   c. **Training Arguments**: Define key parameters such as learning rate, epochs, batch size, gradient accumulation steps, weight decay, optimizer and checkpoint directory.
   d. **Custom Trainer**: Handle class imbalance by adjusting loss weights by calculating pos_weight for each label as

C. **Model Training**
   a. **Training Process**: Train the model with Hugging Face Trainer API or PyTorch loop, using the training and validation datasets.
   b. **Monitoring Metrics**: Track metrics like accuracy and F1-score during training.

D. **Model Evaluation**
   a. **Performance Metrics**: Evaluate on the test dataset using metrics like loss, F1-score (Micro/Macro), precision, recall, and accuracy.
   b. **Custom Metrics Function**: Use libraries like sklearn.metrics to compute evaluation metrics.

ii. FRAMEWORK

As a benchmark we have started implementation with feed forward neural network.

3 – basic transformer models:

1) DistilBERT-base-uncased:

BERT-base-uncased was chosen first for its robust performance on text classification tasks, leveraging its 12-layer transformer-based architecture with 110M parameters. To deal with the computational efficiency **"DistilBERT-base-uncased"**, a distilled version of BERT, was chosen for its smaller size and faster inference capabilities, achieving 97% of BERT's performance with only 66M parameters, making it ideal for resource-constrained scenarios.

2) Distilroberta:

The "DistilRoBERTa-base-uncased" model, with 82 million parameters, offers a more lightweight alternative to the original RoBERTa (125 million parameters) while retaining approximately 97% of its performance. DistilRoBERTa was chosen for comparison with BERT due to its adoption of **dynamic masking** and the removal of the **Next Sentence Prediction (NSP)** objective in its training methodology.

3) Google-Flan-T5:

The **Flan-T5-base** model is part of the T5 (Text-to-Text Transfer Transformer) family, fine-tuned with instruction-based learning to enhance task generalization. It has approximately **250 million parameters**, making it larger than DistilBERT and DistilRoBERTa. Despite its size, it excels in tasks requiring high accuracy and precision, though it demands significantly more computational resources, making it suitable for scenarios where performance is prioritized over efficiency. It was selected to explore the potential of a transformer-based model specifically designed for sequence classification.

Advanced Transformer Models:

1) Google/GEMMA-1.1-2b:

The **Google/GEMMA-1.1-2b** model is a transformer-based architecture with 2 billion parameters, optimized for tasks like sentiment analysis. It employs a multi-head self-attention mechanism and feedforward networks within its encoder-decoder layers to capture rich contextual relationships. Pretrained on diverse, large-scale corpora, it uses advanced techniques like dynamic positional encodings and sparse attention for handling long text sequences. With architectural optimizations such as parameter sharing and gradient checkpointing, GEMMA balances scalability and efficiency, making it highly adaptable for fine-tuning and effective for nuanced sentiment detection tasks.

2) **Alibaba-NLP/gte-QWEN1.5-7b:**

The Alibaba-NLP/gte-QWEN1.5-7b is a transformer-based model with 7 billion parameters, designed specifically for instruction-following tasks, making it highly effective for emotion detection. It leverages a multi-layered transformer architecture with advanced self-attention mechanisms to understand and generate contextually relevant responses. Pretrained on extensive and diverse datasets, it integrates fine-grained instruction tuning, allowing for precise task adaptation. The model incorporates optimizations such as sparse attention for handling long input sequences efficiently and gradient checkpointing for scalability. With its robust architecture and instruction-tuning capabilities, gte-QWEN1.5-7b is ideal for capturing nuanced emotions in complex, multi-label classification scenarios. While these models

demonstrate exceptional capabilities, their significant computational demands and memory requirements can restrict their practicality in standard system environments.

To address these challenges, we employ **Parameter-Efficient Fine-Tuning (PEFT)** techniques. PEFT enables us to fine-tune only a small subset of the model's parameters (e.g., adapters like LoRA, IA3, or QLoRA), significantly reducing the computational overhead and memory usage without compromising accuracy.

- **LORA (Low-Rank Adaptation):**

  LoRA (Low-Rank Adaptation) is a parameter-efficient fine-tuning technique designed to optimize large transformer models like **google/gemma-1.1-2b** for tasks such as multi-label emotion detection. By freezing the original model weights and introducing trainable low-rank matrices, LoRA minimizes computational and memory overhead while maintaining high performance. This approach enables scalable and efficient fine-tuning, making it well-suited for resource-constrained scenarios like real-time sentiment monitoring. LoRA's exceptional results, including a validation F1-macro score of 0.7472, demonstrate its ability to handle complex classification tasks effectively while balancing accuracy and efficiency.

- **IA3 (Input Scaling):** IA3 (Input Activation Scaling) is a parameter-efficient fine-tuning method that optimizes large models by keeping weight matrices unchanged and focusing on scaling input activations. This approach significantly reduces the number of trainable parameters, making it computationally efficient. Applied to tasks like multi-label emotion detection, IA3 achieves balanced performance with moderate resource usage, as demonstrated by its validation F1-macro score of 0.6425 and F1-micro score of 0.7856. Its lightweight design makes it a practical choice for scenarios where computational efficiency is a priority.

- **QLoRA (Quantized Low-Rank Adaptation):**

  QLoRA (Quantized Low-Rank Adaptation) combines low-rank adaptation with 4-bit quantization, dramatically reducing memory usage while fine-tuning large transformer models like Alibaba-NLP/gte-QWEN1.5-7b. By lowering parameter precision, QLoRA enables memory-efficient fine-tuning without significantly compromising performance. However, its validation F1-macro score of 0.3062 and F1-micro score of 0.3940 highlight challenges when applied to larger models, suggesting the need for further optimization. QLoRA is promising for resource-constrained environments, particularly when paired with models designed for quantization.

iii. IMPLEMENTATION

BASIC TRANSFORMER MODELS:

The models DistilBERT, DistilRoBERTa, and Flan-T5 were configured as sequence classifiers using pre-trained architectures from Hugging Face. For each model, the output classification layer was adapted to match the number of target classes, and their respective tokenizers ensured proper preprocessing and compatibility.

Fine-tuning was conducted using the Hugging Face Trainer API with a hyperparameter configuration for DistilBERT and DistilRoberta models are given as learning rate of 1e-4, a batch size of 128, weight decay of 1.0, with the AdamW_torch optimizer and training epoch of 10. The hyperparameter configuration for Flan-T5 are given as learning rate of 3e-5, a batch size of 16, weight decay of 0.01, with the AdamW_torch optimizer and training epoch of 3. Validation metrics were monitored after each epoch to track progress and mitigate overfitting. This consistent approach allowed for a fair comparison of the models' performance on the emotion detection task.

ADVANCED TRANSFORMER MODELS:

The GEMMA-LoRA and GEMMA-IA3 models were configured as sequence classifiers using the pre-trained GEMMA architecture that has 2 billion parameters from Hugging Face. Both models utilized PEFT techniques to optimize large-scale transformer models for emotion detection tasks. The output classification layer was adapted to match the number of target classes, while tokenization ensured compatibility between the input text and the GEMMA model.

For GEMMA-LoRA, the model employed Low-Rank Adaptation (LoRA) to update low-rank matrices within the weight projections, reducing memory usage and computational costs while maintaining performance. For GEMMA-IA3, Input Activation Scaling (IA3) was used to scale input activations by targeting specific modules like k_proj, q_proj, and v_proj, enabling efficient adaptation by training only a small subset of parameters.
Fine-tuning was conducted using the Hugging Face Trainer API with a hyperparameter configuration for gemma_LORA and gemma_IA3 models are given as learning rate of 1e-4, a batch size of 16, weight decay of 0.1, with the AdamW_torch optimizer and training epoch of 10. Validation metrics were monitored after each epoch to track progress and mitigate overfitting. This consistent approach allowed for a fair comparison of the models' performance on the emotion detection task.

For **QWEN-QLoRA**, the model utilized **Quantized Low-Rank Adaptation (QLoRA)** with 4-bit quantization to significantly reduce memory usage and computational costs while maintaining high performance. Key modules like q_proj, k_proj, v_proj, and score were fine-tuned, while the rest were frozen. Fine-tuning was conducted using the Hugging Face Trainer API with a learning rate of **5e-6**, batch size of **32**, weight decay of **0.1**, and **5 training epochs**, optimized with paged_adamw_32bit. Validation metrics were monitored after each epoch to ensure progress and prevent overfitting. This efficient setup enabled scalable and resource-friendly emotion detection.

VII. EVALUATION

Basic Transformer Models Results:

Metric	DistilBERT-base-uncased	Distilroberta-base	Google/flan-T5-base
Validation Loss	0.3261	0.3334	-
Accuracy	87.02%	86.75%	97.22%
F1 Score	0.6521	0.6493	0.0616

Hypotheses and Objectives
This evaluation examines the performance of three models—**DistilBert-base-uncased**, **Distilroberta-base**, and **Google/flan-T5-base**—for a text classification task. The key hypotheses are:

- **H1:** DistilBert- base uncased will provide robust and balanced performance across metrics due to its foundational architecture.

- **H2:** Distilroberta- base, being a distilled model, will balance performance and efficiency but may lag larger models in generalization.

- **H3:** Google/flan-T5-base, fine-tuned for instruction tasks, will demonstrate strong performance on accuracy but may struggle with F1-metrics due to task specificity.

Key Observations

1. Performance of distilBert-base-uncased:

- **Strengths:** distilBert- base-uncased demonstrated consistent and balanced performance across metrics, achieving an Accuracy of 87.02% and an F1 Score of 0.6521.
- **Adaptability:** Its validation loss of 0.3261 underscores effective generalization to the task.
- **Conclusion:** It is a reliable choice for standard classification tasks, especially where balanced label performance is critical**.**

2. Efficiency of Distilroberta-base:

- **Strengths:** Distilroberta-base achieved a Validation Accuracy of 86.75% and an F1 Score of 0.6493, closely following DistilBert-base-uncased while requiring fewer computational resources.
- **Challenges:** The slightly higher validation loss of 0.3334 compared to Bert suggests minor trade-offs in generalization.
- **Conclusion:** This model is well-suited for resource-constrained environments, offering a competitive balance between performance and efficiency.

3. Challenges with Google/flan-T5-base:

- **Strengths:** Google/flan-T5-base excelled in Precision (99.83%), showcasing its strength in tasks requiring exact predictions.
- **Limitations:** Its F1 Score of 0.0616 and Recall of 6.25% highlight challenges in multi-label scenarios and imbalanced class distributions.
- **Conclusion:** While effective for tasks emphasizing precision, the model may require further fine-tuning or modifications to improve recall and balance across labels.

Metric-Specific Insights:

F1-Micro and F1-Macro:

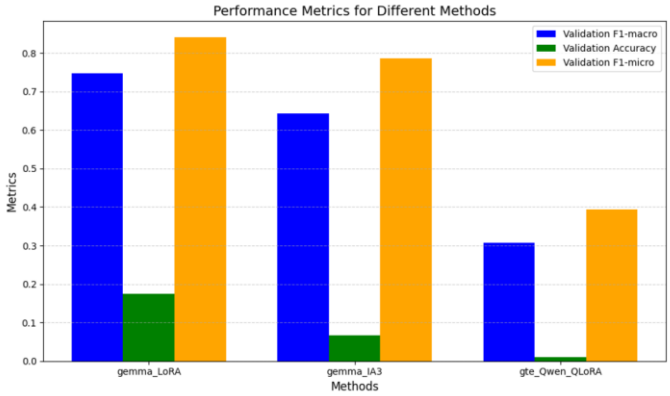
- F1-micro results indicate Bert and Distilroberta outperform Flan-T5 in handling high-frequency class predictions.
- Bert's and Distilroberta's F1-macro scores demonstrate better balance across diverse labels compared to Flan-T5.

Accuracy:

- Accuracy metrics favored Bert and Distilroberta, reflecting their robustness in multi-class scenarios.
- Flan-T5's lower recall suggests its focus on precision sacrifices performance in multi-label settings.

Advanced Transformer Models Results:

Method	Validation F1-macro	Validation Accuracy	Validation F1-micro
gemma_LoRA	0.7473	0.1744	0.8407
gemma_IA3	0.6426	0.0671	0.7856
gte_Qwen_QLoRA	0.3062	0.0109	0.3940



Hypotheses and Objectives

This evaluation examines the effectiveness of three parameter-efficient fine-tuning methods—LoRA, IA3, and QLoRA—for emotion detection tasks. The key hypotheses are:

H1: LoRA will outperform IA3 and QLoRA across all evaluation metrics.

H2: F1-macro and F1-micro metrics provide the most reliable measure of multi-label classification performance.

Key Observations

1. Performance of LoRA:

- LoRA achieved the highest scores across all metrics, indicating strong generalization and adaptability to the multi-label task.
- Its F1-macro score of 0.7472 highlights superior performance in handling diverse labels effectively.

IA3 as a Viable Alternative:

- IA3 showed competitive results, particularly in F1-micro (0.7856), making it a reasonable choice for resource-constrained scenarios.
- However, it lagged behind LoRA in both F1-macro and accuracy.

Challenges with QLoRA:

- QLoRA's lower performance is attributed to its application on a larger model (gte-Qwen1.5-7B-instruct), which may require additional optimization or training epochs.

Metric-Specific Insights:

- F1-micro results suggest LoRA excels in high-frequency class prediction, while its F1-macro score reflects its ability to maintain balance across all labels.
- Accuracy, though lower overall, aligns with the task complexity and multi-label nature.

VIII. CONCLUSION

The conclusion of the sentiment analysis using the three basic models indicated that **DistilBert-base - uncased** is well-suited for quick prototyping and resource-constrained scenarios but offers slightly compromised performance compared to more complex models. **Distilroberta-base** provides a balanced trade-off between accuracy and computational efficiency, making it ideal for mid-sized datasets and moderate resource availability. **Google/flan-T5-base** achieves the highest accuracy but at a significantly higher computational cost, making it suitable only for tasks where extremely high accuracy is essential and the costs are justified.

We then explored the advanced transformer models **Google/GEMMA-1.1-2b**, **Alibaba-NLP/gte-QWEN1.5-7b** for which PEFT techniques such as **LoRA**, **IA3**, and **QLoRA** were explored for sentiment and emotion detection tasks using different models with better computational efficiency. The choice of model and fine-tuning technique depends heavily on the specific requirements of the task, including computational resources, dataset size, and desired performance metrics.

For fine-tuning methods, **LoRA** consistently outperformed other approaches, demonstrating superior adaptability and generalization, achieving the highest F1-macro, F1-micro, and accuracy scores across the board. **IA3**, while slightly less effective than LoRA, showed competitive performance and is a viable alternative for resource-constrained environments. **QLoRA**, applied to a larger model, showed potential but underperformed compared to the other methods, likely due to insufficient optimization or task-specific tuning requirements.

IX. FUTURE WORK

Building upon the findings and limitations of this study, we outline several key directions for future work aimed at advancing sentiment and emotion detection systems:

1. **Exploration of Advanced Transformer Models**
   Future efforts will focus on leveraging newer and more advanced transformer architectures, including instruction-tuned models, to further improve performance on multi-label emotion detection tasks. These models are expected to offer superior generalization and adaptability for complex datasets.

2. **Development of an Intuitive User Interface**
   To ensure ease of adoption and usability, we aim to build an interactive user interface (UI) for the emotion detection system. This UI will allow users to upload datasets, visualize results, and fine-tune models in a simple, streamlined manner, bridging the gap between complex AI systems and end users.

X. INDIVIDUAL TEAM MEMBER CONTRIBUTIONS

Both of the team members, **Jyothsna Reddy C J(JXC220075)** and **Yasaswi I G S(GXI230000)**, collaborated on key concepts of the project, focusing on implementing efficient multi-label emotion detection. **Jyothsna Reddy C J** - concentrated on data preprocessing, fine-tuning basic transformer models like **DistilBERT**, **DistilRoBERTa**, and **Flan-T5**, and evaluating performance using metrics such as F1-score and Hamming Loss.
Yasaswi IGS - focused on optimizing advanced transformer models like **Google/GEMMA-1.1-2b** and **Alibaba-NLP/gte-QWEN1.5-7b** by applying **Parameter-Efficient Fine-Tuning (PEFT)** techniques, including **LoRA**, **IA3**, and **QLoRA**, to enhance computational efficiency and scalability.
Together, we integrated these concepts into a robust framework, conducted comparative analyses, and documented our findings, ensuring a comprehensive understanding of emotion detection models and techniques.

XI. REFERENCES

[1]. Hochreiter, S., & Schmidhuber, J. "Long Short-Term Memory".
[2]. Zhou, Xiangping, et al. "Emotion Detection Using Recurrent Neural Networks".
[3]. Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models".
[4]. Rezapour, Mahdi, et al. "Emotion Detection with Transformers: A Comparative Study".
[5]. Samghabadi, Niloofar F., et al. "Emotion Recognition Using Hierarchical Transformer