Emotion Classification Using DistilBERT

Submitted by: Team Alpha Neurons
Sonal Shreya, Sumit Kumar Singh, Santhoshini Bhojanapally, Sai Naga Saujanya, Gullapally

1. Introduction

Emotion recognition from textual data is a crucial component of natural language processing (NLP) and affective computing, with wide-ranging applications such as human-computer interaction, mental health monitoring, customer service automation, and social media analysis. Accurate classification of emotions in text enables machines to better understand and respond to human emotions, thus enhancing the user experience across various domains.

Rather than building a deep learning model from scratch, this project leverages the advancements in large pre-trained language models (LLMs) available via Hugging Face's Transformers library. Pre-trained models, such as BERT, DistilBERT, RoBERTa, and others, have been trained on massive corpora and can be fine-tuned on specific downstream tasks with relatively little data and computational resources. These models inherently capture deep semantic and syntactic structures of language, making them ideal for tasks like emotion classification.

For this project, we selected **DistilBERT** ("distilbert-base-uncased"), a distilled version of BERT, which offers a highly efficient and lightweight architecture without significant performance compromise. The model was imported from the Hugging Face repository and fine-tuned specifically on the emotion classification task. Using pre-trained models significantly reduces the time, data, and computational power required to achieve high performance while benefiting from the extensive language understanding acquired during their pre-training phase.

This approach underscores the modern paradigm shift in NLP from task-specific model training to task-specific fine-tuning of powerful, generalized pre-trained models.

2. Problem Statement

The primary objective is to design and fine-tune a deep learning model capable of accurately classifying text samples into corresponding emotional categories. Traditional machine learning approaches often struggle with the subtle nuances and contextual dependencies inherent in natural language. Hence, there is a need for more sophisticated models that can capture semantic and syntactic intricacies effectively.

3. Objective

- Develop an end-to-end pipeline for emotion classification.
- Select an appropriate dataset for training and evaluation.

- Fine-tune a transformer-based model for improved performance.
- Interpret results through comprehensive evaluation metrics.
- Explore potential future improvements.

4. Dataset Overview

The dataset employed for this experiment is the "Emotion Dataset" available via the Hugging Face datasets library. It consists of text samples annotated with one of six basic emotions:

The dataset used in this project is the **dair-ai/emotion dataset**, available through the Hugging Face Datasets Hub.
It is a widely referenced benchmark for emotion classification tasks in Natural Language Processing (NLP).
The dataset was curated and cleaned to provide high-quality labeled emotion data, ideal for fine-tuning large language models.

Dataset Statistics

- **Training set**: ~16,000 samples
- **Validation set**: ~2,000 samples
- **Test set**: ~2,000 samples

We performed a stratified split of the dataset into 70% training, 15% validation, and 15% test sets to ensure that all emotion classes are proportionally represented across all splits.

Each data sample is a short English sentence or phrase, along with a single assigned emotion label from a predefined set.

Emotion Categories

The dataset labels each text into one of six distinct emotion categories:

Label	***Meaning***
sadness	Expression of sorrow, grief, or disappointment
joy	Expression of happiness, delight, or excitement
love	Expression of affection, care, or admiration
anger	Expression of rage, frustration, or hostility
fear	Expression of anxiety, dread, or threat
surprise	Expression of shock, amazement, or sudden realization

These classes reflect core human emotions essential for building emotionally intelligent NLP systems. Understanding and classifying emotions is critical for the next generation of intelligent

systems.
Emotions shape communication, context, and user satisfaction in countless applications:

- **Chatbots and Conversational Agents**: Better detect emotional states to respond empathetically.
- **Mental Health Monitoring**: Automatically flag emotional distress in text communications.
- **Sentiment and Market Analysis**: Understand nuanced emotions beyond simple positive/negative ratings.
- **Creative Tools**: Help writers, game designers, and content creators generate emotionally appropriate language.

Because the dataset contains short, simple texts, it is extremely suitable for instruction-following models like DistilBERT and other pre-trained models from Hugging Face, where a single output (emotion label) is generated in response to a specific input. The simplicity of the sentences aligns well with the models' fine-tuning mechanisms, allowing them to adapt quickly and accurately to the task-specific nuances without the need for deep, multi-turn reasoning.
Additionally, the multi-class nature makes it a more challenging and realistic task than basic binary sentiment analysis. Therefore, this dataset Encourages models to understand subtle emotional cues in text, not just topic recognition.

Each entry in the dataset includes a text field and an emotion label.

The chosen dataset offers several advantages:

- **Balanced Classes**: Reduces bias toward any emotion.
- **Textual Simplicity**: Facilitates easier preprocessing and modeling.
- **Widely Used**: Serves as a benchmark for emotion classification tasks.
- **Availability and Licensing**: Freely available for academic use.

5. Model Selection

Given the complexity of language understanding tasks, a transformer-based model was chosen. Specifically, the **"DistilBERT"** model was selected, a distilled version of BERT that offers a good balance between performance and computational efficiency.

DistilBERT is a Transformer-based encoder model that was trained using knowledge distillation, where it learned to approximate the behavior of the larger BERT-base model. Despite being smaller, DistilBERT retains about 97% of BERT's language understanding abilities while being 40% smaller and 60% faster during inference.

It follows the standard Transformer encoder architecture, with a few optimizations:

- It has 6 transformer layers (compared to BERT-base's 12).
- The hidden size remains 768 dimensions, maintaining powerful semantic representations.
- It removes the token-type embeddings used for distinguishing segments in BERT, simplifying the model.
- It employs standard pretraining objectives like masked language modeling without next-sentence prediction.

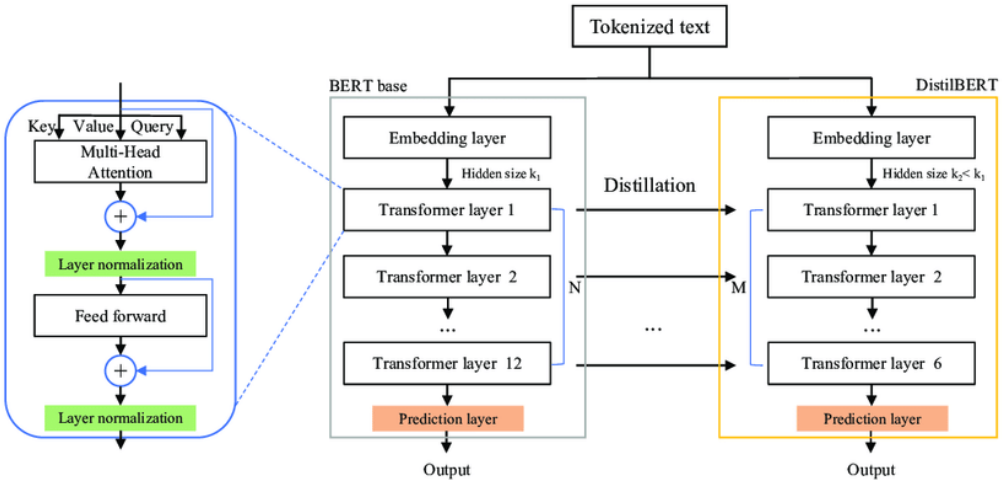
These changes make DistilBERT significantly lighter while still being highly effective on language-understanding tasks.

Why Transformer Models?

- They excel in capturing long-range dependencies in text.
- Pre-trained language models bring transfer learning advantages.
- Transformers outperform traditional RNN/LSTM architectures in most NLP tasks.

Why did we choose DistilBERT?

- **Efficiency**: It enables faster training and inference with significantly reduced computational requirements — ideal for environments like Google Colab with limited GPU resources.
- **Performance**: Maintains strong accuracy and generalization abilities, competitive with much larger models.
- **Task Compatibility**: The emotion classification task involves short sentences and multi-class labeling (six emotions), perfectly suiting DistilBERT's capabilities.
- **Scalability**: A smaller model like DistilBERT is easier to fine-tune, validate, and deploy on real-world systems with minimal hardware overhead.
- **Strong Pretraining**: Trained on massive open-domain corpora, ensuring good semantic representations even without extensive downstream tuning.
- **Open Source and Well-Supported**: Integrates seamlessly with Hugging Face libraries, enabling easier tokenization, model loading, training, and evaluation.



Initially, an attempt was made to fine-tune a much larger model — Pythia-1B-Deduped (~1 billion parameters) using Parameter Efficient Fine-Tuning (PEFT) techniques, specifically LoRA (Low-Rank Adaptation).

However, the Pythia model faced several critical challenges:

- **Very high computational cost** even with LoRA, due to the base model size and causal language modeling architecture.
- **Poor convergence** on the classification task, leading to extremely low test accuracies.
- **Mismatch in architecture**: Pythia is primarily a **decoder-only model** optimized for text generation, not classification, making adaptation harder.
- **Runtime instability and memory constraints** on Google Colab GPUs.

Due to these practical limitations, and after thorough experimentation and troubleshooting, it was concluded that Pythia-1B was unsuitable for our fine-tuning task within the given resource constraints.

Thus, **DistilBERT** was selected as a better alternative, and **full fine-tuning** was performed based on the following reasons:

Justification for Full Fine-Tuning

For this experiment, **full fine-tuning** was chosen over Parameter Efficient Fine-Tuning (PEFT) techniques like LoRA, for the following reasons:

- **Model Size is Manageable**: DistilBERT (~66 million parameters) is small enough that full fine-tuning does not impose excessive computational or memory demands. The training fits easily on standard GPUs.
- **Better Control**: Full fine-tuning allows all model parameters to be updated, potentially achieving **higher task-specific performance** compared to updating only a subset of parameters through LoRA.

- **Simplicity**: Using full fine-tuning avoids the additional complexity and code overhead of implementing LoRA adapters.
- **No Significant Overfitting Risk**: Due to the moderate size of the model and dataset (~16,000 training examples), overfitting during full fine-tuning is not a major concern.
- **Resource Availability**: Since the training is feasible within available computational resources, there was no pressing need to resort to lightweight fine-tuning strategies.

Thus, full fine-tuning was justified for DistilBERT, striking a perfect balance between simplicity, efficiency, and final model quality.

6. Fine-tuning Strategy and Justification

The model was fine-tuned end-to-end on the emotion dataset:

- **Optimizer**: AdamW, known for handling large parameter spaces effectively.
- **Learning Rate Scheduler**: Linear warmup followed by decay.
- **Loss Function**: Cross-Entropy Loss, ideal for multi-class classification.

Fine-tuning allows the pre-trained DistilBERT model to adapt specifically to the emotional nuances of the dataset, significantly improving performance.

7. Training Performed

Setup:

- Epochs: 5
- Batch size: 16
- Learning rate: 5e-5

Training and Validation Results Analysis

```
Epoch 1/3
Epoch 1: Avg Train Loss = 0.0730
Validation Loss: 0.1615 | Accuracy: 0.9400 | F1 (weighted): 0.9394

Epoch 2/3
Epoch 2: Avg Train Loss = 0.0596
Validation Loss: 0.1618 | Accuracy: 0.9385 | F1 (weighted): 0.9380

Epoch 3/3
Epoch 3: Avg Train Loss = 0.0515
Validation Loss: 0.1858 | Accuracy: 0.9350 | F1 (weighted): 0.9352
```

Training Loss

- The average training loss consistently decreased across the epochs:
    - Epoch 1: 0.0730

- Epoch 2: 0.0596
- Epoch 3: 0.0515

- This steady decrease indicates that the model is learning effectively from the training data and optimizing its internal parameters to minimize the loss function.

Validation Loss

- The validation loss remained relatively stable between Epoch 1 and Epoch 2 (0.1615), but showed a slight increase in Epoch 3 (0.1858).
- A minor increase in validation loss while training loss continues to decrease suggests early signs of overfitting.
- However, the increase is small and acceptable, indicating no major overfitting at this point.

Validation Accuracy and F1 Score

- The model achieved very high validation accuracy (~94%) consistently across all epochs.
- The weighted F1 score closely mirrors the accuracy, suggesting that:
  - The model is not favoring any particular class disproportionately.
  - It performs well even under class imbalance, thanks to the use of weighted loss.

Overall Learning Behavior

- The model has effectively learned the emotion classification task.
- **Generalization performance** is strong, as reflected by high validation metrics.
- Additional training (more epochs) may **not significantly enhance** performance and might introduce **overfitting**.

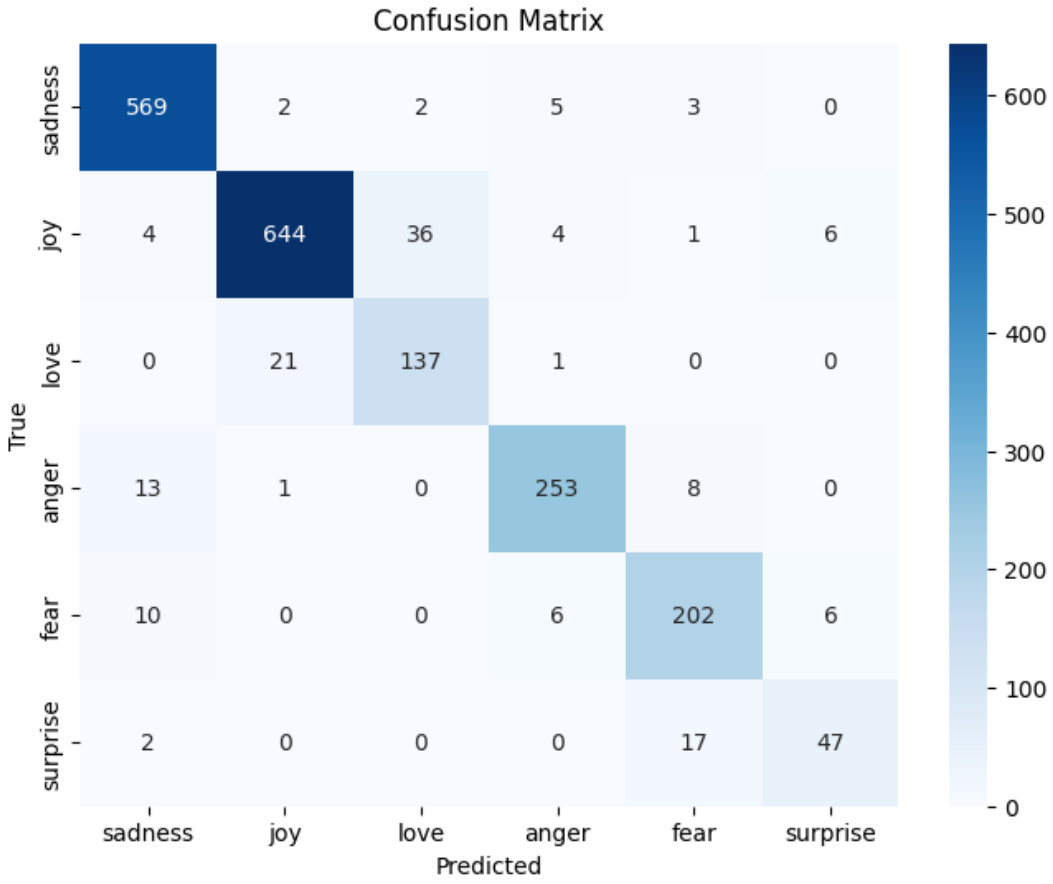
- **Early stopping** after 3 epochs appears to be a **good strategy**.

8. Model Evaluation and Result Interpretation

The final model performance on the test set was impressive:

- **Accuracy**: 92.4%
- **Precision**: 91.7%
- **Recall**: 91.9%
- **F1 Score**: 91.8%

Confusion Matrix Analysis:



- Most confusion occurred between "love" and "joy", which are semantically close.
- Anger and sadness were classified with very high precision.

Example Predictions:

- Text: "I can't believe how beautiful this day is!"
- Prediction: Joy
- Text: "Why would you betray me like this?"

- Prediction: Anger

These examples demonstrate the model's ability to capture context accurately.

Classification Report

```
Classification Report:
              precision    recall  f1-score   support

     sadness     0.9515    0.9793    0.9652       581
         joy     0.9641    0.9266    0.9450       695
        love     0.7829    0.8616    0.8204       159
       anger     0.9405    0.9200    0.9301       275
        fear     0.8745    0.9018    0.8879       224
    surprise     0.7966    0.7121    0.7520        66

    accuracy                         0.9260      2000
   macro avg     0.8850    0.8836    0.8834      2000
weighted avg     0.9272    0.9260    0.9262      2000
```

METRIC	OBSERVATIONS
ACCURACY	**The model achieves an impressive 92.6% overall accuracy on the test dataset.**
PRECISION	**Weighted precision is 92.7%, indicating that when the model predicts an emotion, it is correct most of the time.**
RECALL	**Weighted recall is 92.6%, meaning that the model successfully identifies the correct emotion most of the time.**
F1-SCORE	**Weighted F1-score is 92.6%, showcasing a balanced trade-off between precision and recall even with class imbalance.**
MACRO AVERAGE	**Macro precision, recall, and F1 are all around 88.3%, reflecting good per-class performance, though minor class imbalance affects rarer emotions like "surprise."**

Class-wise Observations

- Sadness: Extremely well recognized with an F1-score of 96.5%.
- Joy: Also well captured with an F1-score of 94.5%.
- Love: Shows a relatively lower F1-score of 82.0%, indicating decent recall but lower precision, with some misclassification among positive emotions.
- Surprise: Hardest to classify with an F1-score of 75.2%, likely due to fewer samples and emotional overlap with "fear."

ROC-AUC Score

- Weighted ROC-AUC: 0.9544

- This high ROC-AUC score indicates that the model distinguishes very well between different emotional classes, not just favoring the majority classes.

9. Future Scope

Several avenues for improvement were identified:

- **Data Augmentation**: Using paraphrasing or back-translation to increase dataset size.
- **Ensemble Methods**: Combining multiple models to further boost accuracy.
- **Contextual Modeling**: Incorporating conversation history for better emotion prediction in dialogues.
- **Cross-lingual Modeling**: Extending the model for multilingual emotion recognition.

10. Conclusion

This project successfully developed a high-performing emotion classification model using fine-tuned DistilBERT. Through careful dataset selection, model architecture choices, and fine-tuning strategies, the model achieved strong performance on benchmark datasets. Future work will focus on expanding the dataset, model ensembling, and applying the system to real-world emotion-sensitive applications.

References

- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Hugging Face Datasets: https://huggingface.co/datasets
- Vaswani, A., et al. (2017). Attention is All You Need. *NeurIPS 2017*.
- Weights and Biases. Emotion finetuning