DL■GENAi PROJECT REPORT

Multi■Label Emotion Classification

Name: Abhishek Saha

Roll: 23f1001572

1. ABSTRACT / EXECUTIVE SUMMARY

This project addresses the Multi■Label Emotion Classification problem from a Kaggle competition. The task involves predicting five possible emotions (anger, fear, joy, sadness, surprise) from short text inputs. Three different modeling approaches were implemented:

1) Fine■tuned Transformer Model (RoBERTa■Large)

2) Custom Deep Learning Model (Scratch BiLSTM + GRU + Multi■Head Attention)

3) TF■IDF + SGD Classifier (Linear Model)

The RoBERTa■Large model achieved the best macro■F1 score (0.8721), followed by the custom BiLSTM (0.7418) and TF■IDF + SGD (0.7791). Significant experiment tracking was performed using Weights & Biases (W&B;). The report presents preprocessing, tokenization, modeling choices, evaluations, and model comparisons.

2. INTRODUCTION

This competition required building models capable of multi■label emotion detection from text input. The dataset consisted of user■written sentences labeled with combinations of emotions. The project goal was to build three fundamentally different architectures, compare their performance, and understand trade■offs in deep learning models for NLP tasks.

This report is organized into sections covering data, preprocessing, tokenization, modeling, experiments, evaluation, and conclusions.

3. DATASET & PREPROCESSING

The dataset comprised 6827 training examples and 1707 test samples. Each sentence could contain zero, one, or multiple emotions. Preprocessing steps included:

- HTML unescaping
- Lowercasing
- Removing URLs, mentions, special characters
- Normalizing whitespace
- Reducing repeated characters

Additionally, custom preprocessing strategies were implemented for each model family.

Exploratory Data Analysis revealed:

- Strong class imbalance (fear most common, joy least)
- Varied text lengths (5–40 tokens on average)
- Emotion co■occurrences

4. TOKENIZATION STRATEGY

RoBERTa uses Byte■Pair Encoding (BPE). This tokenizer was selected because:

- It matches the pretrained checkpoint
- Handles subword decomposition effectively
- Reduces OOV issues

The custom BiLSTM model used a vocabulary built from scratch with a min frequency cutoff. TF■IDF used word and char analyzers.

5. MODELING & EXPERIMENTATION

5.1 FINE■TUNED TRANSFORMER (RoBERTa■Large)

Architecture: 24■layer transformer encoder, 1024■dim hidden states, multi■head attention.

Training:

– LR = 1e■5

– Batch = 8

– Epochs = 5

– Scheduler: Linear warmup

– Loss: Focal Loss γ=2.0

Achieved macro■F1 = 0.87213 (best).

5.2 CUSTOM DEEP LEARNING MODEL (Scratch BiLSTM)

Architecture:

– Embedding layer

– Spatial Dropout

– BiLSTM → BiGRU stack

– Multi■Head Attention (4 heads)

– Pooling (avg + max)

– Linear classifier

Trained for 12 epochs with AdamW + Cosine Warmup scheduler.

Achieved macro■F1 = 0.7418.

5.3 TF■IDF + SGD CLASSIFIER

Vectorization:

– Word■level TF■IDF (1–2 grams)

– Char■level TF■IDF (3–5 grams)

Features concatenated (60k total).

SGDClassifier (logistic loss) trained per label using partial_fit across 20 epochs.

Achieved macro■F1 = 0.7791.

6. PERFORMANCE & COMPARATIVE ANALYSIS

Model Comparison (Val Macro■F1):

– RoBERTa■Large: 0.8721

– TF■IDF + SGD: 0.7791

– Scratch BiLSTM: 0.7418

Observations:

– Transformer excels in nuanced contextual understanding.

– BiLSTM captures sequence patterns but is limited by vocabulary constraints.

– TF■IDF model is computationally cheap, interpretable, and still competitive.

Extensive W&B; visualizations were produced:

– Loss Curves

– F1 Evolution

– Threshold Optimization

– Model Agreement Heatmaps

– Positive Prediction Distribution

7. CONCLUSION & FUTURE WORK

Key learnings include:

– Transfer learning with large transformers dramatically improves emotion classification.

– Custom architectures require careful design and tuning.

– Lightweight models can remain competitive with proper preprocessing.

Future improvements:

– Ensemble of Transformer + TF■IDF for robustness

– Hyperparameter sweeps using W&B;

– Exploring DeBERTa■V3 or Longformer

– Data augmentation using back■translation or LLM■generated samples

8. REFERENCES

– Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.”

– Hochreiter & Schmidhuber. “Long Short■Term Memory.”

– TF■IDF, SGDClassifier — scikit■learn documentation.

– Hugging Face Transformers Library.

– Weights & Biases experiment tracking documentation.