Multilabel Emotion Classification

Karthikeya Hanu Prakash Kanithi (EE22BTECH11026)
Pandrangi Aditya Sriram (EE22BTECH11039)
Kondaparthy Anurag (AI22BTECH11011)

Indian Institute of Technology Hyderabad

Agenda

Dataset Overview

- Dataset from SemEval-2025 Task 11 competition.
- Multi-lingual text samples annotated for multiple emotions.
- Languages: English, Arabic, Chinese, Finnish, French, German, Italian, Portuguese, Spanish.
- Emotions for English: **anger, fear, joy, sadness, surprise**.
- Non-English languages have an additional label: **disgust**.
- English Dataset Size:
  - Train: 10,244 samples
  - Dev: 1,066 samples
  - Test: 2,345 samples

Model Architecture

- Base model: Pretrained transformer (e.g., **BERT-base**).
- Advanced model: XLM-Roberta based classfier

Feature	**bert-base-multilingual-cased**	**XLM-RoBERTa**
Architecture	BERT	RoBERTa (optimized BERT)
Pretraining Corpus	Wikipedia (104 languages)	CommonCrawl (2TB, 100 languages)
Tokenizer	WordPiece (cased)	SentencePiece (uncased)
Parameters	˜110M	˜270M
Training Objective	MLM + NSP	MLM only
Performance	Good on seen languages	Better multilingual performance

Vanilla model Training Details

- This is the first model that we implemented where we only considered English language dataset and have trained the model with train/eng.csv and validated with dev/eng.csv
- Loss Function: **Binary Cross-Entropy with Logits Loss**.

```
F1-macro: 0.6423326225861329
Precision-macro: 0.7417090384091819
Recall-macro: 0.6058960573476703
```

Detailed Classification Report:

	precision	recall	f1-score	support
anger	0.86	0.38	0.52	16
fear	0.68	0.79	0.73	63
joy	0.76	0.42	0.54	31
sadness	0.71	0.83	0.76	35
surprise	0.70	0.61	0.66	31
micro avg	0.70	0.66	0.68	176
macro avg	0.74	0.61	0.64	176
weighted avg	0.72	0.66	0.67	176
samples avg	0.61	0.58	0.58	176

Figure: Results of the Base Model

Further Improvements - 1

- **Hard Negative Mining via Focal Loss**
- Focus on difficult examples (hard negatives).
- **Focal Loss**: Down-weights easy examples and focuses on hard, misclassified ones.
- Formula:

$$\text{Focal Loss} = \alpha \cdot (1 - p_t)^{\gamma} \cdot \text{BCE}$$

  where:
  - p_t is the predicted probability for the true class.
  - γ is the focusing parameter ($\textit{typically} \geq 2$).
  - α is the weighting factor.
- Focal Loss finally supports class-wise alpha based on inverse class frequency.
- **Summary**: Easy samples are ignored; more effort is spent on hard negatives.

Further Improvements - 2

- **Dynamic Thresholding (DTT)**
- Each label (emotion) predicted independently using a sigmoid output.
- Common practice: fixed threshold (e.g., 0.5) to decide if label is active.
- DTT adapts thresholds for each class:
    - Compute Precision-Recall curve for each class.
    - Calculate F1 score at different thresholds.
    - Select the optimal threshold for highest F1 score.
- **Result:** Improved precision-recall balance and higher macro-F1 score.
- **Weighted Sampling** to oversample underrepresented emotions
- Weighted Sampling helps model learn the classes better whereas Dynamic Thresholding: Helps evaluate the classes better. **They are complementary, not canceling each other.**

Further Improvements - 3

- Early stopping based on macro F1
- Learning rate scheduling
- Per-class loss monitoring
- Per Epoch loss monitoring
- Optimizer: AdamW
- Scheduler: CosineAnnealingLR
- LR: 2e-5
- Epochs: 10

Optimized result on BERT-based for English Only

- As we can observe that the F1-macro increases from 0.642 to 0.6867

```
Epoch 5/10, Loss: 0.0327, Per-class Loss: [0.01512793 0.04494277 0.02731499 0.0428157  0.03344469]
F1-macro: 0.6867583929554211
Precision-macro: 0.647828394515701
Recall-macro: 0.7451472094214029
Detailed Classification Report:
              precision    recall  f1-score   support

       anger       0.64      0.56      0.60        16
        fear       0.62      0.94      0.75        63
         joy       0.61      0.74      0.67        31
     sadness       0.76      0.74      0.75        35
    surprise       0.61      0.74      0.67        31

   micro avg       0.64      0.80      0.71       176
   macro avg       0.65      0.75      0.69       176
weighted avg       0.65      0.80      0.71       176
 samples avg       0.65      0.74      0.66       176

Early stopping triggered.
```

Figure: Results of the Base Model on optimized training and validation environment

- We can increase the F1-macro even higher by choosing a more suited and complex model than BERT-Based i.e., XLM-RoBERTa

FINAL Optimized result on XLM-RoBERTa for English Only

- As we can observe that the F1-macro increases from 0.6867 to 0.7055

```
Epoch 5/10, Loss: 0.0001, Per-class Loss: [6.41689767e-05 6.76350974e-05 7.44218705e-05 1.20900535e-04
 1.00385812e-04]
Thresholds: [0.29381397 0.48461118 0.46552852 0.6718797  0.5450892 ]
F1-macro: 0.705502387444669
Precision-macro: 0.6638058098939464
Recall-macro: 0.7760931899641577
Classification Report:
              precision    recall  f1-score   support

       anger       0.48      0.75      0.59        16
        fear       0.65      0.97      0.78        63
         joy       0.80      0.65      0.71        31
     sadness       0.68      0.74      0.71        35
    surprise       0.71      0.77      0.74        31

   micro avg       0.66      0.81      0.73       176
   macro avg       0.66      0.78      0.71       176
weighted avg       0.68      0.81      0.73       176
 samples avg       0.67      0.73      0.68       176

Early stopping triggered.
```

Figure: Results of the RoBERTa Model on optimized training and validation environment

FINAL Optimized result on XLM-RoBERTa for English and Hindi

```
Epoch 9/10, Loss: 0.0095, Per-class Loss: [0.00647778 0.01600036 0.00779364 0.01480115 0.00963377 0.00235196]
F1-macro: 0.7327457124793372
Precision-macro: 0.7083521781005656
Recall-macro: 0.7633411727161726
Detailed Classification Report:
              precision    recall  f1-score   support

       anger       0.79      0.72      0.75        32
        fear       0.72      0.83      0.77        77
         joy       0.55      0.67      0.60        42
     sadness       0.65      0.79      0.71        52
    surprise       0.74      0.78      0.76        40
     disgust       0.80      0.80      0.80        10

   micro avg       0.69      0.77      0.73       253
   macro avg       0.71      0.76      0.73       253
weighted avg       0.69      0.77      0.73       253
 samples avg       0.65      0.64      0.63       253
```

- Early stopping triggered.

Figure: Results of the RoBERTa Model on optimized training and validation environment for English and Hindi

FINAL Optimized result on XLM-RoBERTa for Hindi and Marathi

- **Language Correlation:** We can observe that the F1-macro for hindi and marathi languages is very high when compared to other two languages, mostly because they both are highly similar / correlated.

```
Epoch 10/10, Loss: 0.0049, Per-class Loss: [0.00712453 0.00515499 0.00312707 0.00717585 0.00317921 0.00345204]
F1-macro: 0.8763740711828144
Precision-macro: 0.8874766053992973
Recall-macro: 0.8700553784732284
Detailed Classification Report:
              precision    recall  f1-score   support

       anger       0.90      0.93      0.92        30
        fear       1.00      0.86      0.93        29
         joy       0.77      0.77      0.77        30
     sadness       0.83      0.71      0.76        34
    surprise       0.95      0.95      0.95        21
     disgust       0.88      1.00      0.93        21

   micro avg       0.88      0.85      0.87       165
   macro avg       0.89      0.87      0.88       165
weighted avg       0.88      0.85      0.87       165
 samples avg       0.67      0.66      0.66       165
```

Figure: Results of the RoBERTa Model on optimized training and validation environment for Hindi and Marathi

Interpretability Analysis for single label

- **Technique used**: LIME.

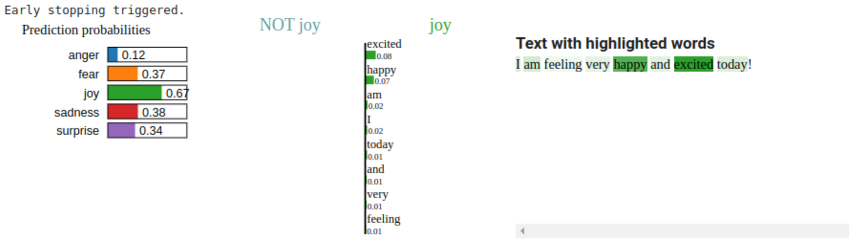

Figure: LIME analysis of a random sentence with single label after training

- From the above picture we can infer that
  - The model correctly predicts joy as the emotion.
  - It learned that words like happy and excited are strong signals for joy.
  - Less important words (like "am", "today", "very") didn't confuse it much.
  - The model works perfectly as its supposed to.

Interpretability Analysis for Multi label

- **Technique used**: LIME.

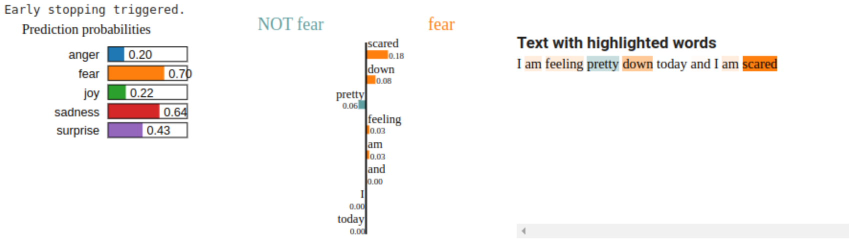

Figure: LIME analysis of a random sentence with single label after training

- From the above picture we can infer that
  - The model correctly predicts both sad and fear as the emotion.
  - It learned that words like down and scared are strong signals for sad and fear.
  - Less important words didn't confuse it much. although the word "pretty" pulls away from the fear, the down still is dominating and the model predicts ffear and sadness perfectly.

Conclusion

- Successfully built and fine-tuned a multilabel emotion classifier.
- testing many models and chose a basic model and an advanced model which give us significantly better performance.
- Achieved significant improvement over dataset baseline.
- Interpretability analysis provided insights into model behavior.
- Explored advanced loss functions like focal loss etc.