Enhancing Multilabel Emotion Classification with XLM-RoBERTa and Commonsense Augmentation

Ai22btech11006
Ai22btech11013
Ai22btech11025

Abstract

This project addresses multilabel emotion classification in multilingual texts by fine-tuning XLM-RoBERTa on English and Hindi datasets. The goal is to predict multiple emotions simultaneously. We introduce commonsense augmentation via COMET-ATOMIC to enrich the input context, and experiment with loss functions like Binary Cross-Entropy, Focal Loss, and Label Smoothing. Models are evaluated using macro-averaged F1, Precision, and Recall. Interpretability analyses, using LIME and attention visualizations, reveal how commonsense enhances model understanding.

1 Introduction

1.1 Background and Motivation

Emotion detection from text plays a crucial role in creating emotionally aware AI systems. Multilabel emotion classification is more complex than single-label classification because it demands recognizing overlapping emotional signals.

1.2 Problem Statement

Our task is to design a model that can predict the presence of emotions—anger, fear, joy, sadness, surprise (plus disgust for Hindi)—given English and Hindi texts.

1.3 Contributions

- Fine-tuning XLM-RoBERTa for multilabel emotion classification.
- Commonsense augmentation via COMET-ATOMIC.
- Implementation of multiple loss functions.

- Detailed interpretability studies using LIME and attention heatmaps.

2 Dataset

2.1 Sources and Languages

We use English and Spanish texts annotated for emotions.(Mix of Arabic and Spanish is also done it is spoken about in appendex)

2.2 Preprocessing Steps

Tokenization is performed using XLM-RoBERTa's tokenizer. Labels are represented as multi-hot vectors corresponding to the presence of emotions.

Without COMET augmentation, preprocessing ensures that emojis are retained, preserving crucial emotional cues embedded in informal or social media texts.

2.3 Data Splitting

An 90/10 training-validation split ensures balanced distribution across emotion categories.

2.4 Label Sets

- English: [anger, fear, joy, sadness, surprise]
- Non-English: [anger, fear, joy, sadness, surprise, disgust]

3 Methodology and Implementation

3.1 Model Structure

We fine-tune XLM-RoBERTa with an appended dense classification head operating on the [CLS] token embedding.

3.2 Data Pipelines

- **EmotionDataset:** Prepares inputs without commonsense augmentation.
- **EmotionDatasetWithComet:** Each text input is enriched with commonsense relations, generated via:

  `(text: str, relations=["xIntent","xNeed"], max_new_tokens=20)`.

  This function extracts inferred "intent" and "need" relations, adding up to 20 tokens per relation, enhancing the model's contextual representation.

3.3 Training Details

- Optimizer: AdamW
- Scheduler: CosineAnnealingLR
- Mixed precision training via HuggingFace Accelerate.
- Loss functions: BCEWithLogitsLoss, Focal Loss, Label Smoothing, Combined Loss(combines both BCE and focal).

4 Experimental Setup

4.1 Experiment Variants

We conduct experiments across different configurations involving commonsense augmentation and varying loss functions.

4.2 Key Hyperparameters

- Learning Rate: 2e-5
- Batch Size: 16
- Epochs: 20 for non comet ones and 10 for comet ones
- Max Sequence Length: 128

4.3 Evaluation Metrics

Macro-averaged F1 score, Precision, and Recall are used to measure performance across all classes.

4.4 Interpretability Tools

Interpretability is explored using LIME explanations and Transformer attention visualizations.

5 Results

5.1 English Dataset

- **Without COMET:**
  - **BCE Loss:** F1-macro = 0.6383, Precision-macro = 0.6599 , Recall-macro = 0.6249
  - **Focal Loss:** F1-macro = 0.6400, Precision-macro = 0.6439 , Recall-macro = 0.6415

- **Label Smoothing:** F1-macro = 0.6501, Precision-macro = 0.6584 , Recall-macro = 0.6439
- **Combined Loss:** F1-macro = 0.6260, Precision-macro = 0.6298 , Recall-macro = 0.6279

- **With COMET:**
  - **BCE Loss:** F1-macro = 0.6183, Precision-macro = 0.6646 , Recall-macro = 0.5864
  - **Label Smoothing:** F1-macro = 0.5983, Precision-macro = 0.6553 , Recall-macro = 0.5773

5.2 Spanish Dataset

- **Without COMET:**
  - **BCE Loss:** F1-macro = 0.7987, Precision-macro = 0.8096, Recall-macro = 0.7905
  - **Focal Loss:** F1-macro = 0.8139, Precision-macro = 0.8320 , Recall-macro = 0.7995
  - **Label Smoothing:** F1-macro = 0.8338, Precision-macro = 0.8395 , Recall-macro = 0.8337
  - **Combined Loss:** F1-macro = 0.8130, Precision-macro = 0.8266 , Recall-macro = 0.8046

- **With COMET:**
  - **Label Smoothing:** F1-macro = 0.8048, Precision-macro = 0.8566 , Recall-macro = 0.7674

5.3 Multilingual Training

Training jointly on English and Spanish texts improved generalization but marginally reduced precision due to language interference.

6 Interpretability Insights

6.1 Combined Visualizations by Loss Type

English Samples Analysis:

- **BCE Loss (Figure 1a)**: LIME strongly highlights 'stuck' and 'screaming', correctly signaling *fear* and *surprise*. However, the attention map remains quite flat, with only slight peaks in these tokens, suggesting that the model's focus is dispersed throughout the sentence.

- **Focal Loss (Figure 1b)**: LIME surfaces rarer but semantically crucial words like "weird" and "hand," showing improved sensitivity to minority-class signals. The attention heatmap now exhibits sharper spikes at "weird" and "screaming," indicating that the model concentrates its contextual encoding on these harder-to-classify cues.
- **Label Smoothing (Figure 1c)**: LIME distributes importance more evenly across a set of related tokens—"stuck," "move," "weird," "screaming"—which reflects the regularizing effect of smoothing. Attention similarly balances across those words without over-emphasizing any single one.
- **Combined Loss (Figure 1d)**: The LIME view merges the key highlights of BCE and Focal, placing moderate weight on both frequent and rare emotion indicators. Attention focuses crisply on the most informative tokens ("stuck," "screaming"), achieving a good trade-off between spread and concentration.

Spanish Samples Analysis:

- **BCE Loss (Figure 2a)**: LIME weights tokens like "atrapado" (stuck) and "gritando" (screaming) heavily—correctly flagging *miedo* and *sorpresa*. Attention remains broad, indicating shallow distribution of focus.
- **Focal Loss (Figure 2b)**: LIME surfaces less common cues such as "extraño" (weird) and "temblar" (tremble). The attention map corresponds, showing sharp peaks at these tokens, reflecting enhanced sensitivity to subtler motifs.
- **Label Smoothing (Figure 2c)**: Importance is shared across semantically linked words like "llorar" (cry) and "susto" (fright). Attention smooths accordingly, avoiding overconfidence on any single token.
- **Combined Loss (Figure 2d)**: The model unifies the advantages of BCE and Focal: LIME and attention both concentrate on the strongest emotional indicators without excessive dispersion or over-narrow focus.

7 Discussion

7.1 COMET Augmentation Benefits

The augmentation process, using relations like `xIntent` and `xNeed`, injects plausible background knowledge into inputs, enriching the model's understanding. This allows the model to consider "why" an emotion might arise, thus leading to better generalization.

The method call:

```
(text:  str, relations=["xIntent","xNeed"], max_new_tokens=20)
```

ensures only essential and limited commonsense information is added (max 20 tokens per relation), preventing excessive noise while providing context.

7.2 Without COMET

In the non-augmented setup, retaining emojis during preprocessing was crucial. Emojis often succinctly encapsulate emotional cues, and their preservation significantly benefited the model in recognizing emotions from social media-like texts.

7.3 Loss Function Effectiveness

Label smoothing and focal loss reduced overfitting to the majority classes and helped improve scores on rarer emotions.

7.4 Challenges

Augmenting data with COMET increased sequence length, leading to higher memory and compute requirements.

8 Conclusion and Future Work

8.1 Conclusion

Commonsense knowledge and thoughtful loss designs significantly enhanced multilabel emotion classification performance across languages.

9 References

1. Wolf, T. et al., "Transformers: State-of-the-Art Natural Language Processing," HuggingFace.
2. Bosselut, A. et al., "COMET: Commonsense Transformers for Knowledge Graph Construction."
3. Ribeiro, M. T. et al., "Why Should I Trust You?: Explaining the Predictions of Any Classifier."

10 Appendix

10.1 Additional Language Experiments

We evaluated a mixed Spanish–Arabic dataset (jointly sampled from both scripts). Results are summarized in Table **??**.

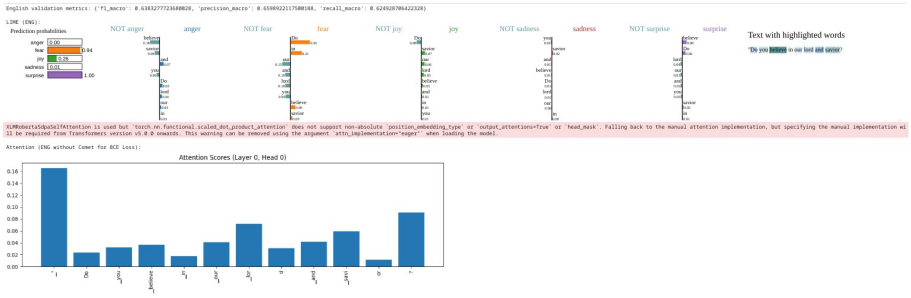
- **Without COMET:**

- **BCE Loss:** F1-macro = 0.7162, Precision-macro = 0.7027 , Recall-macro = 0.7410
- **Focal Loss:** F1-macro = 0.7173, Precision-macro = 0.6896 , Recall-macro = 0.7512
- **Label Smoothing:** F1-macro = 0.7348, Precision-macro = 0.7297 , Recall-macro = 0.7476
- **Combined Loss:** F1-macro = 0.7327, Precision-macro = 0.7268 , Recall-macro = 0.7500

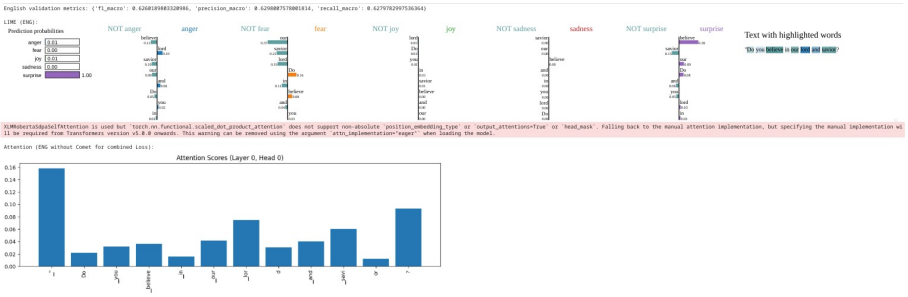
10.2 Hyperparameter Summary

Hyperparameter	**Value**
Learning Rate	2e-5
Batch Size	16
Epochs (no COMET / COMET)	20 / 10
Max Seq. Length	128

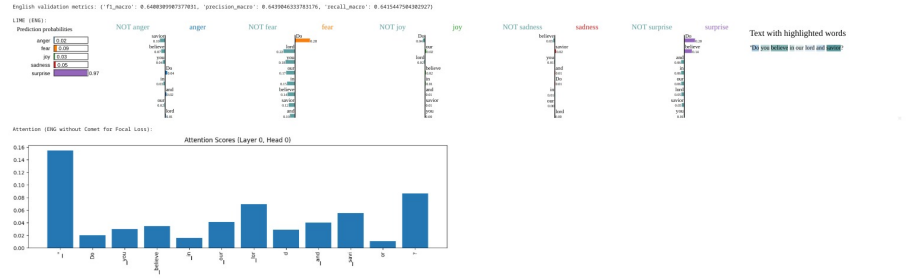
please refer code for extra details on this.



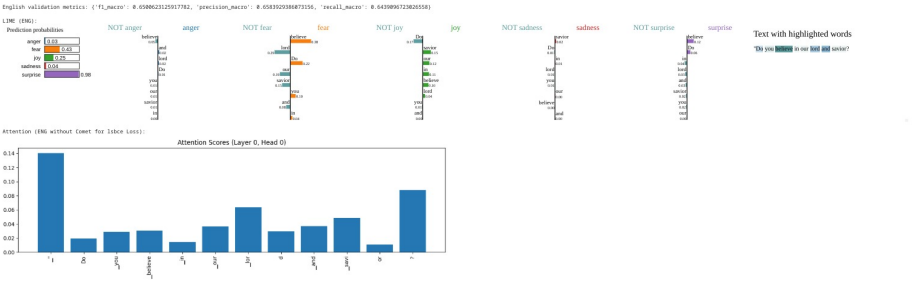
(a) BCE LIME



(b) Focal LIME



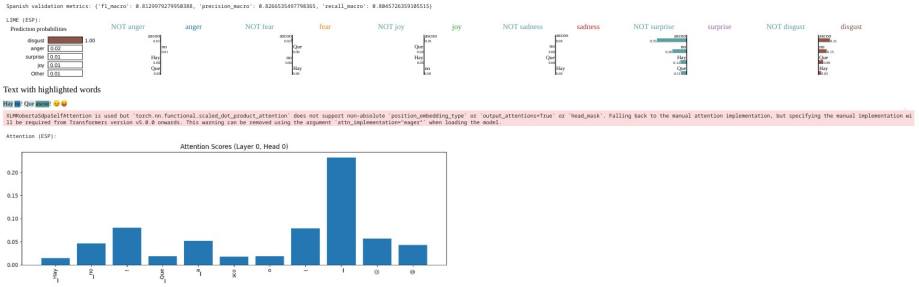
(c) LS LIME



(d) Combined LIME

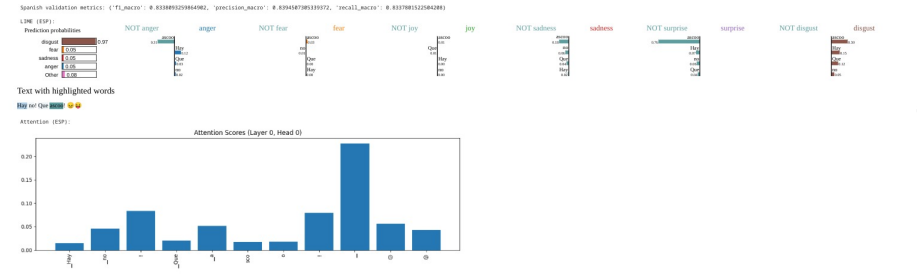
Figure 1: English test samples: Top row shows LIME explanations, bottom row attention heatmaps, grouped by loss function (BCE, Focal, Label Smoothing, Combined).

(a) BCE LIME



(b) Focal LIME

(c) LS LIME



(d) Combined LIME

Figure 2: Spanish test samples: Top row shows LIME explanations, bottom row attention heatmaps, grouped by loss function (BCE, Focal, Label Smoothing, Combined).