NLP Assignment-3

Team-32

Prasham Walvekar - CS21BTECH11047

Kallu Rithika - AI22BTECH11010

Armaan - CS22BTECH11051

Q.1 Multi-label Emotion Classification

Problem Statement

Classify text samples into appropriate emotion categories - such as anger, fear, joy, sadness, and surprise by developing a model that can identify the presence or absence of each emotion in multilingual datasets.

Languages Chosen

1. English (most commonly used, easy interpretability and understanding)
2. Hindi + Marathi
    a. Linguistic similarity - similar words, and similar script - devanagari
    b. Belong to the top 100 languages of the world which SOTA transformer models are trained on
    c. We understand both languages

ENGLISH

Architectural Approaches

1. Experimented with different pretrained transformers and classification techniques.
2. Transformers used: (from Hugging Face)
   a. bert-base-uncased
   b. roberta-base
   c. distilbert-base-uncased
   d. xlm-roberta-base
3. Classifiers used:
   a. MLP head with 1 hidden layer (num_layers = 1)
   b. MLP head with num_layers = 2
   c. BiLSTM (hidden_state = 64, num_lstm_layers = 2) + MLP head (num_layers = 2)
   d. Attention Pooling + MLP head (num_layers = 2) -> pools based on attention scores given to tokens (instead of using the class token) before passing to classification head

Training Approach

1. Dataset -
   a. train.csv (split: 90% train, 10% validation): 2768 samples
   b. test.csv: 2767 samples
2. Training approach -
   a. Phase-1: Transfer Learning (for first 3-5 epochs)
      i. Freeze all layers of transformer
   b. Phase-2: Fine-tuning (for rest of the epochs)
      i. Unfreeze a fixed number of last few layers of transformer
      ii. Dynamically unfreeze a few layers of transformer incrementally with each epoch
      iii. Unfreeze all layers of transformer (we found this to give best results)
3. Loss criterion -
   a. BCEWithLogitsLoss
      i. Helpful for multi-label classification tasks
      ii. Outputs logits (probability vector) with same length as number of labels / emotions
   b. Focal Loss
      i. Focal Loss helps NLP models focus on hard-to-classify texts (like rare classes) by down-weighting easy predictions, making it useful for imbalanced or multi-label text classification.
4. Other important features -
   a. Optimizer - AdamW (works best for fine-tuning of transformers)
   b. Learning Rate Scheduler (reduces/increases LR based on validation score for better convergence)

Results (BCE LogitsLoss)

1. Using bert-base-uncased

Model & Setup	F1 Macro	Precision Macro	Recall Macro
bert-base-uncased (unfreeze last 4 layers after epoch=3, MLP layers=1, epochs=7)	0.6867	0.7463	0.6522
bert-base-uncased (unfreeze all layers after epoch=3, MLP layers=1, epochs=7)	0.6944	0.7350	0.6702
bert-base-uncased (unfreeze all layers after epoch=3, MLP layers=2, LR scheduler, epochs=8)	0.7064	0.7219	0.6931
bert-base-uncased + BiLSTM (unfreeze all layers after epoch=3, MLP layers=2, LR scheduler, epochs=9, BiLSTM=128, 2 layers)	0.6681	0.7167	0.6321
bert-base-uncased (attention pooling, num_layers=2, epochs=7)	0.6999	0.7376	0.6749

Results (Contd.)

2. Using roberta-base: (Best Performance)

Model & Setup	F1 Macro	Precision Macro	Recall Macro
roberta-base (unfreeze all layers after epoch=3, MLP layers=2, LR scheduler, epochs=7)	0.7135	0.7474	0.6908
roberta-base (attention pooling, num_layers=2, epochs=7)	0.7068	0.7497	0.6847

3. Using xlm-roberta-base:

Model & Setup	F1 Macro	Precision Macro	Recall Macro
xlm-roberta-base (epochs = 10)	0.6374	0.6842	0.6082

Results (Contd.)

4. Using DistilBERT

Model & Setup	**F1 Macro**	**Precision Macro**	**Recall Macro**
distilbert-base (unfreeze all layers after epoch 3, MLP layers = 2, LR scheduler, epochs = 7)	0.6713	0.6975	0.6539
distilbert-base (unfreeze all layers after epoch 3, MLP layers = 2, LR scheduler, epochs = 9)	**0.6825**	0.6721	**0.7039**

Results (Focal Loss)

Experimental Results Using Focal Loss

Model & Setup	F1 Macro	Precision Macro	Recall Macro
bert-base-uncased (focal loss, epochs=7, num_layers=2)	0.6571	0.7880	0.5755
roberta-base (focal loss, epochs=7, num_layers=2)	0.6860	0.7908	0.6232

MULTI-LINGUAL

(HINDI + MARATHI)

Architectural Approaches

1. Experimented with different pretrained transformers and classification techniques.
2. Transformers used: (from Hugging Face)
   a. xlm-roberta-base
   b. xlm-roberta-large
   c. bert-base-multilingual-cased (mBERT)
   d. ai4bharat/indic-bert
3. Classifiers used:
   a. MLP head with 1 hidden layer (num_layers = 1)
   b. MLP head with num_layers = 2
   c. Attention Pooling + MLP head (num_layers = 2)

Training Approach

1. Dataset -
   a. Training: interleaved marathi and hindi datasets (split: 90% train, 10% validation)
      Total samples:
      i. Marathi: 2415
      ii. Hindi: 2556
   b. Testing: interleaved marathi and hindi datasets
      Total samples:
      i. Marathi: 1000
      ii. Hindi: 1010
2. Training Approach (Transfer learning + fine-tuning), Loss criterion, Optimizer, LR Scheduler - Same as we did for English

Results (BCEWithLogitsLoss)

1. Using xlm-roberta-base:

Model	Experiment Details	F1 Macro	Precision Macro	Recall Macro
XLM-RoBERTa-Base	num_layers=2, num_epochs=7	0.7589	0.8089	0.7299
XLM-RoBERTa-Base	attention pooling, num_layers=2, epochs=7	0.7695	0.8305	0.7304
XLM-RoBERTa-Base	attention pooling, num_layers=2, epochs=15	0.8246	0.8465	0.8044
XLM-RoBERTa-Base + BiLSTM	- (BiLSTM gave bad F1, ignored)	(bad)	(ignored)	(ignored)

Results (Contd.)

2. Using xlm-roberta-large: (best performance)

Model	Experiment Details	F1 Macro	Precision Macro	Recall Macro
XLM-RoBERTa-Large	num_layers=2, num_epochs=7	0.8730	0.8884	0.8627
XLM-RoBERTa-Large	attention pooling, num_layers=2, epochs=7	0.7584	0.9429	0.6943

3. Using bert-base-multilingual-cased (mBERT):

Model	Experiment Details	F1 Macro	Precision Macro	Recall Macro
mBERT	num_layers=2, num_epochs=7	0.7135	0.7662	0.6745

Results (Focal Loss)

Experimental Results Using Focal Loss

Model	Experiment Details	F1 Macro	Precision Macro	Recall Macro
XLM-RoBERTa-Large	focal loss, num_layers=2, epochs=7	0.8722	0.9355	0.8191
XLM-RoBERTa-Base	focal loss, num_layers=2, epochs=12	0.8105	0.8902	0.7479
IndicBERT	focal loss, num_layers=2, epochs=15	0.7278	0.8993	0.6166

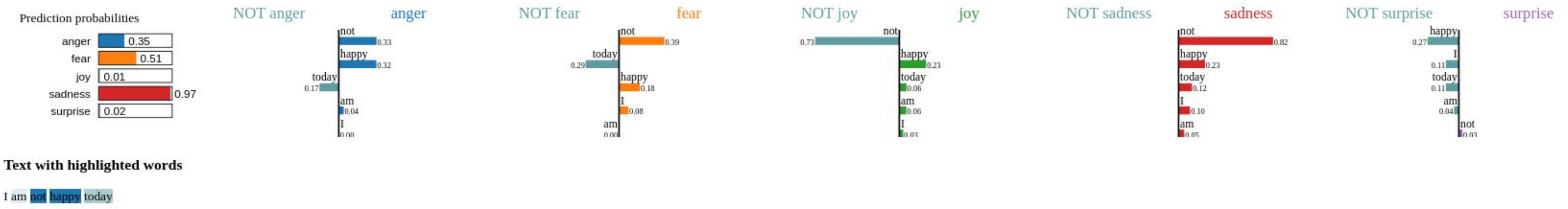
INTERPRETABILITY

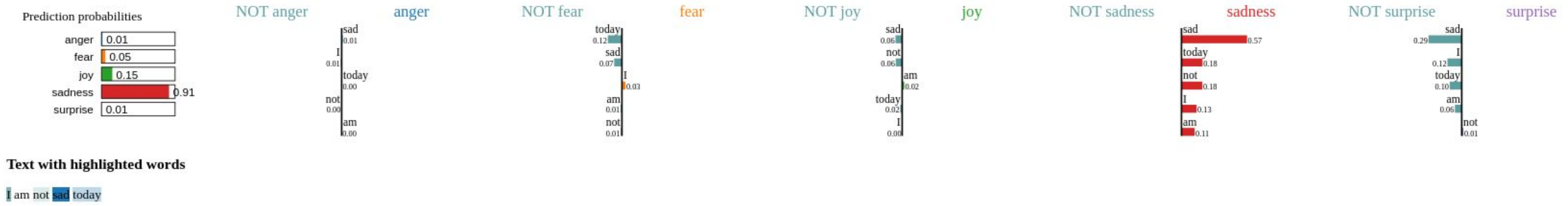
Interpretability and Faithfulness Check

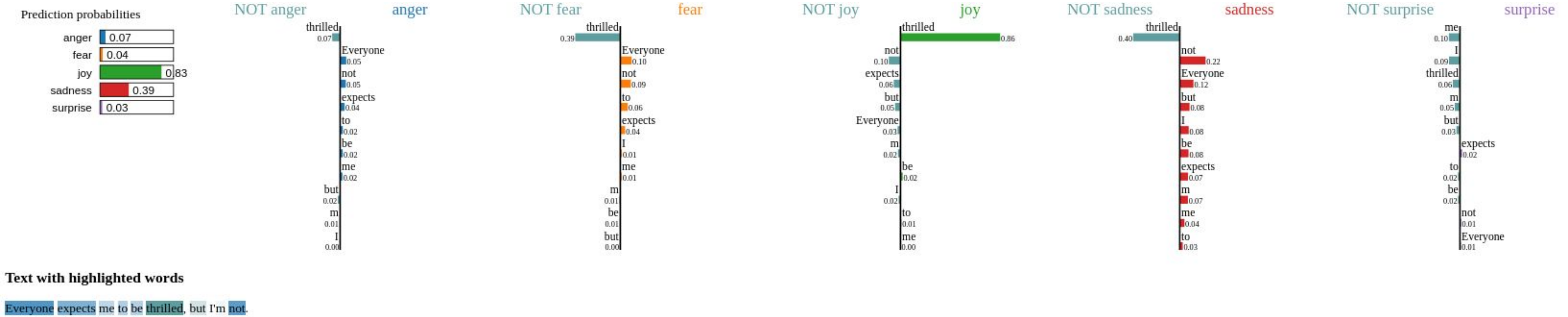
- **Tool Used :** LIME (Local Interpretable Model-agnostic Explanations)
- **Goal :** Check if the model relies on real emotional context or just keyword shortcuts.
- **Approach :**
  1. Used LIME to highlight important words per emotion.
  2. Tested sentence variants with negations, sarcasm, and misleading patterns, punctuations (ex: exclamations)
- **Findings :**
  1. Model relied on keywords like “sad” or “thrilled” for example, ignoring the context.
  2. Shows signs of undesirable shortcut learning.

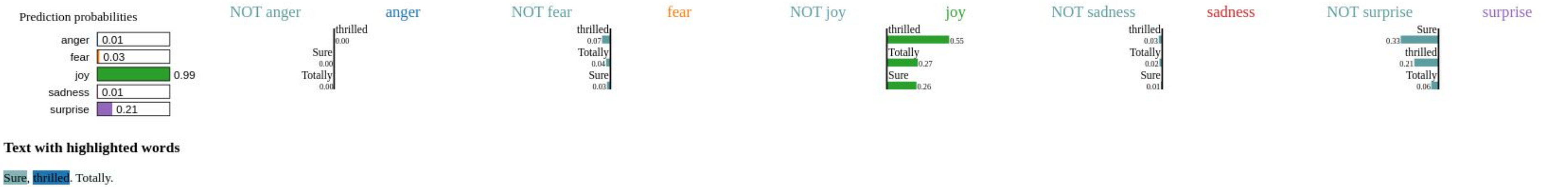
Results

Interpretability - Roberta-base Model on English Sentences

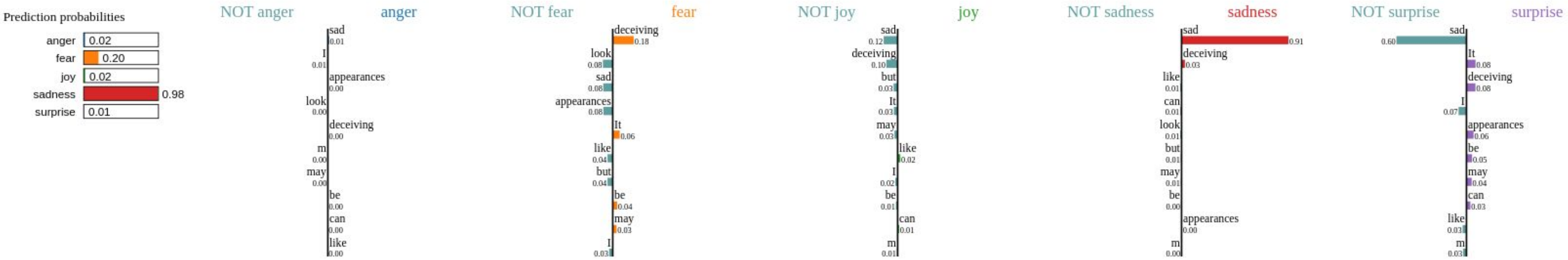





Prediction probabilities
anger 0.07
fear 0.04
joy 0.83
sadness 0.39
surprise 0.03
NOT anger anger
NOT fear fear
NOT joy joy
NOT sadness sadness
NOT surprise surprise
Text with highlighted words
Everyone expects me to be thrilled, but I'm not.


Prediction probabilities
anger 0.01
fear 0.03
joy 0.99
sadness 0.01
surprise 0.21
NOT anger anger
NOT fear fear
NOT joy joy
NOT sadness sadness
NOT surprise surprise
Text with highlighted words
Sure, thrilled. Totally.


Prediction probabilities
anger 0.05
fear 0.04
joy 0.16
sadness 0.91
surprise 0.02
NOT anger anger
NOT fear fear
NOT joy joy
NOT sadness sadness
NOT surprise surprise
Text with highlighted words
Why do you think I am sad? Infact I am quite the opposite.



Text with highlighted words

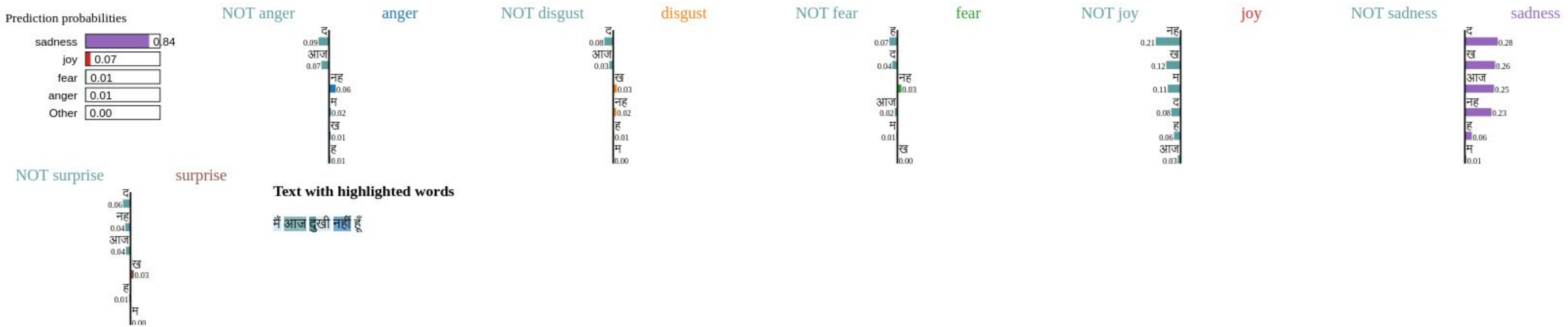
It may look like I'm sad, but appearances can be deceiving.

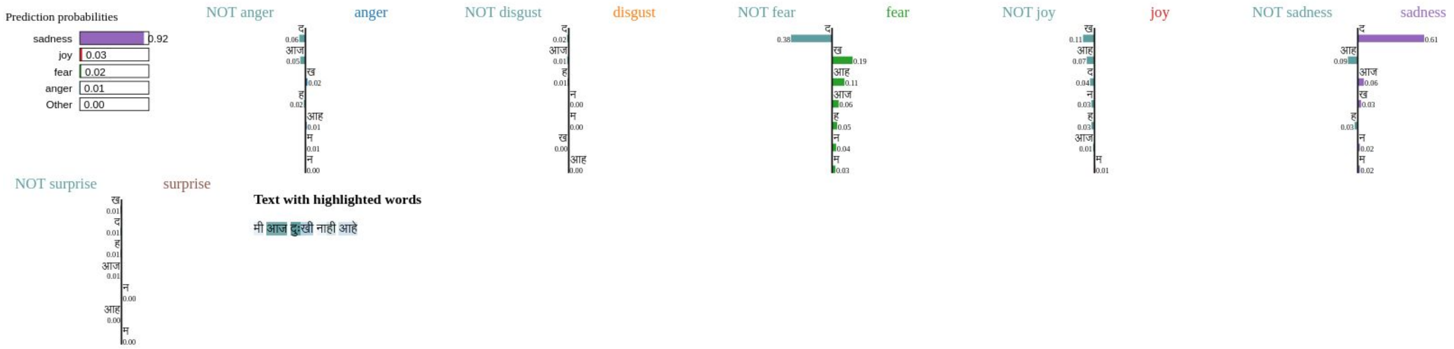
Text with highlighted words

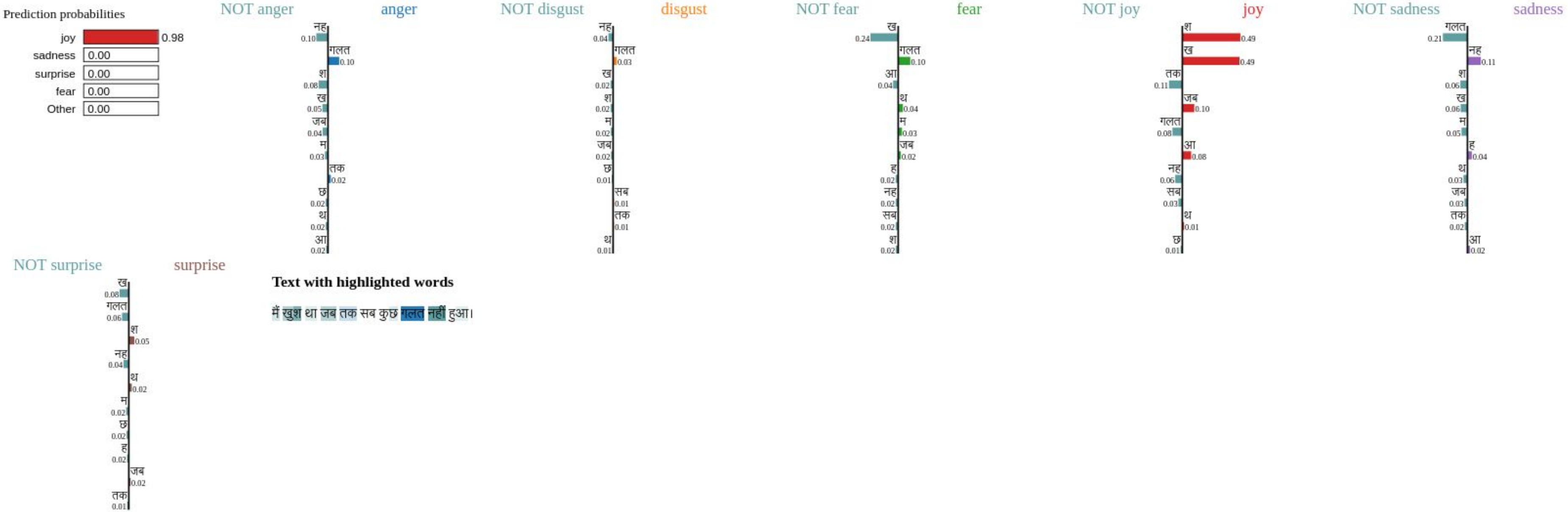
Sure, I'm sad.

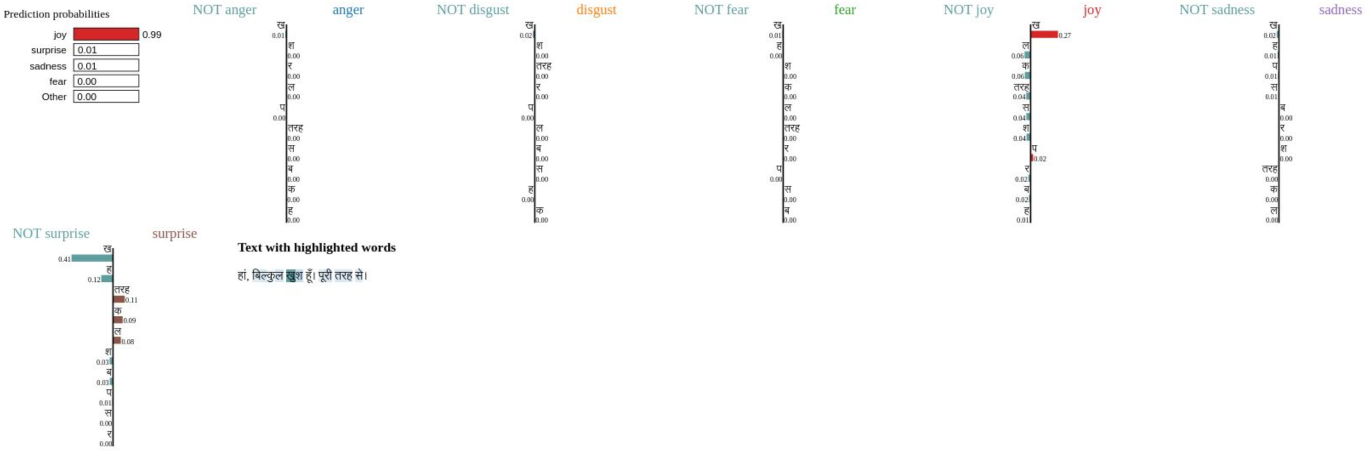
Results

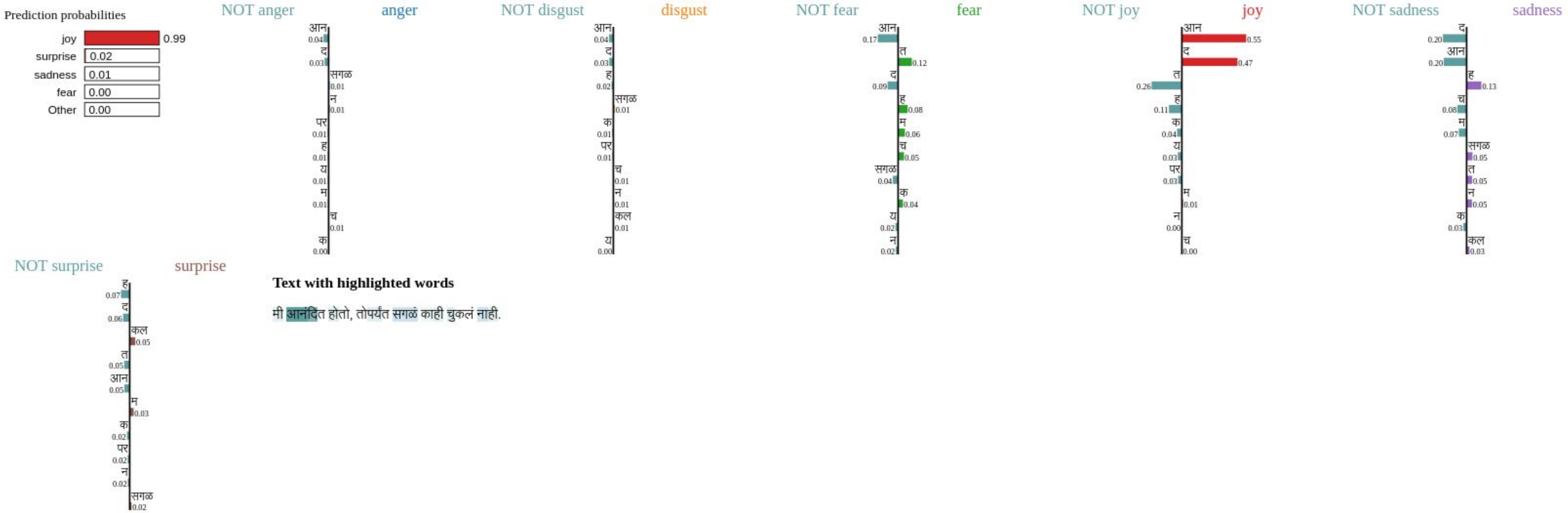
Interpretability - xlm roberta base Model on Hindi and Marathi Sentences

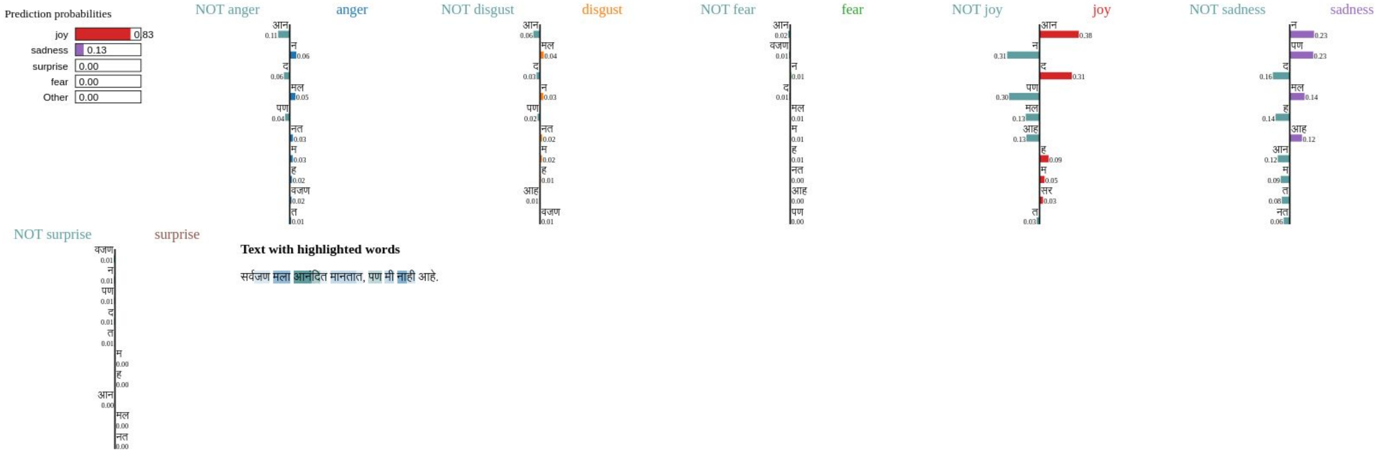





Prediction probabilities
joy 0.98
sadness 0.00
surprise 0.00
fear 0.00
Other 0.00
NOT anger anger
NOT disgust disgust
NOT fear fear
NOT joy joy
NOT sadness sadness
NOT surprise surprise
Text with highlighted words
मैं खुश था जब तक सब कुछ गलत नहीं हुआ।


Prediction probabilities
joy 0.99
surprise 0.01
sadness 0.01
fear 0.00
Other 0.00
NOT anger anger
NOT disgust disgust
NOT fear fear
NOT joy joy
NOT sadness sadness
NOT surprise surprise
Text with highlighted words
हां, बिल्कुल खुश हूँ। पूरी तरह से।


Prediction probabilities
joy 0.99
surprise 0.02
sadness 0.01
fear 0.00
Other 0.00
NOT anger
anger
NOT disgust
disgust
NOT fear
fear
NOT joy
joy
NOT sadness
sadness
NOT surprise
surprise
Text with highlighted words
मी आनंदित होतो, तोपर्यंत सगळं काही चुकलं नाही.

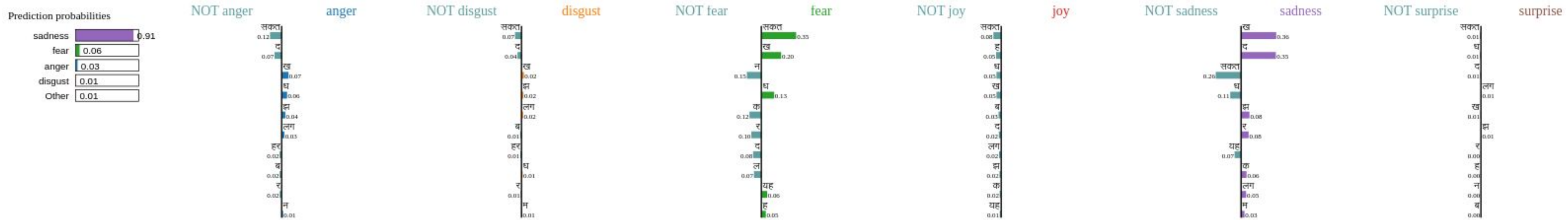

Prediction probabilities
joy 0.83
sadness 0.13
surprise 0.00
fear 0.00
Other 0.00
NOT anger
anger
NOT disgust
disgust
NOT fear
fear
NOT joy
joy
NOT sadness
sadness
NOT surprise
surprise
Text with highlighted words
सर्वजण मला आनंदित मानतात, पण मी नाही आहे.

Text with highlighted words

मैं आज दुखी नहीं हूँ

Text with highlighted words

तुम्हें क्यों लगता है कि मैं दुखी हूँ? दरअसल, मैं बिल्कुल उलट हूँ।

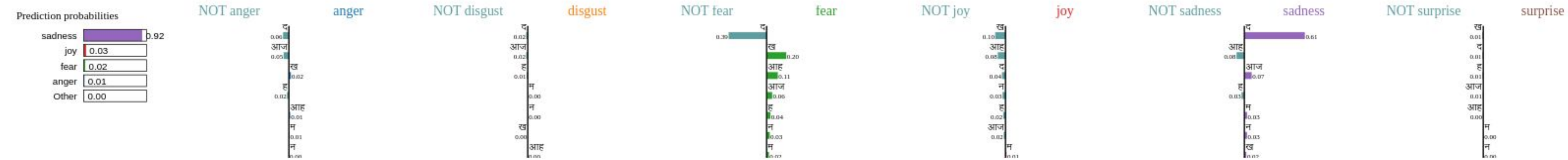


Text with highlighted words

यह हो सकता है कि मुझे दुखी लगे, लेकिन बाहरी रूप धोखा दे सकते हैं।

Text with highlighted words

हाँ, बिल्कुल, मैं पूरी तरह से दुखी हूँ!!

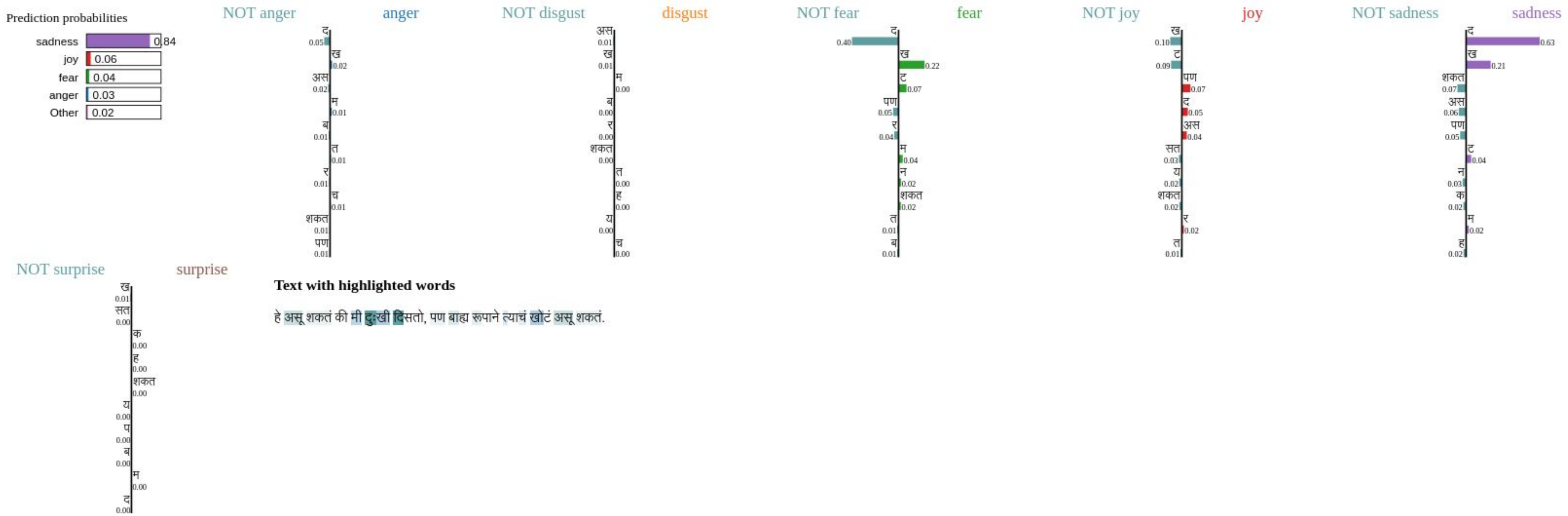


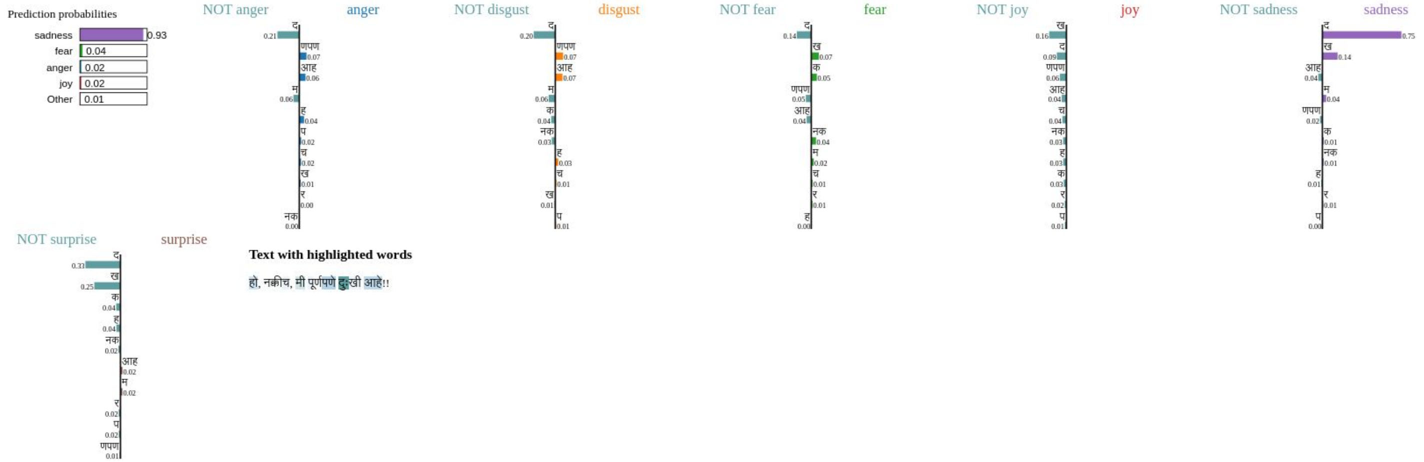
Text with highlighted words

मी आज दुःखी नाही आहे

Text with highlighted words

तुला का वाटतं की मी दुःखी आहे? खरंतर मी तर उलट आहे.


Prediction probabilities
sadness 0.84
joy 0.06
fear 0.04
anger 0.03
Other 0.02
NOT anger
anger
NOT disgust
disgust
NOT fear
fear
NOT joy
joy
NOT sadness
sadness
NOT surprise
surprise
Text with highlighted words
हे असू शकतं की मी दुःखी दिसतो, पण बाह्य रूपाने त्याचं खोटं असू शकतं.


Prediction probabilities
sadness 0.93
fear 0.04
anger 0.02
joy 0.02
Other 0.01
NOT anger
anger
NOT disgust
disgust
NOT fear
fear
NOT joy
joy
NOT sadness
sadness
NOT surprise
surprise
Text with highlighted words
हो, नक्कीच, मी पूर्णपणे दुःखी आहे!!

Summary of unfaithfulness (English)

Set1:
S4: "I am not sad today" -> predicts sadness due to emphasis on "sad"

Set2: (Undesirable shortcuts observed)
1. S2: "I was thrilled until everything went wrong." -> predicts joy based on "thrilled" (but the sentence indicates the opposite, indirectly)
2. S3: "Everyone expects me to be thrilled, but I'm not." -> predicts joy (the speaker is not feeling joy)
3. S4: "Sure, thrilled. Totally." -> predicts joy (can't recognize the sarcasm)

Set3: (Undesirable shortcuts observed)
1. S1: "I am not sad today" -> predicts sadness due to word "sad" (can't recognize negation here due to "not")
2. S2: "Why do you think I am sad? In fact I am quite the opposite." -> predicts sadness (but the speaker clearly indicates the opposite)
3. S3: "It may look like I'm sad, but appearances can be deceiving." -> predicts sadness (though the speaker suggests the opposite, indirectly)
4. S4: "Sure, I'm sad."-> predicts sadness (Can't recognize sarcasm)

Summary of unfaithfulness (Hindi + Marathi)

Set1:

Hindi:
Wrong prediction for "I am not sad" -> emphasis on word "sad" (undesirable shortcut to associate "dukhi" to sadness)

Marathi:
Wrong prediction for "I am not sad" -> emphasis on word "sad" (undesirable shortcut to associate "dukhi" to sadness)

Set2:

Hindi:

1. S2: "I was happy until everything went wrong" -> predicts joy based on word "khushi"

2. S4: "Sure! I am totally happy." -> predicts joy, can't understand sarcasm

Marathi:

1. S2 and S4: same prediction and reason as Hindi

2. S3: "I was happy, until everything went wrong" -> predicts joy based on the word "anandit"

Summary of unfaithfulness (Hindi + Marathi) (Contd…)

Set3:

Hindi:
1. S1: "I am not sad today." -> predicts sadness (can't capture negation)

2. S2: "Why do you think I am sad? In fact, I am quite the opposite." -> predicts sadness (even though it mentions in the end that it's the opposite)

3. S3: "It may look like I am sad, but external looks can be deceiving." -> predicts sadness (even though indirectly the person means the opposite)

4. S4: "Sure, I am sad." -> predicts sadness (can't capture sarcasm)

Marathi:

Same predictions and reasons mentioned above for hindi

Mitigation Strategy based on Interpretability Analysis

Contrastive Learning:
We can use contrastive loss to teach the model to differentiate between subtle language features like irony, sarcasm, or sarcasm-like patterns.
Or even patterns like negation.

Adversarial training:
Include samples which the model predicts wrongly (adversarial samples) based from interpretability analysis and re-train the model on the modified datasets

THANK YOU