Yermekov Danial, Nurbek Seilbek
Natural Language Processing, Daniyar Rakhimzhanov
Assignment 4

Part 1: Dataset Selection & Research Analysis

Dataset - Universal Joy(Combining low-resource with small)

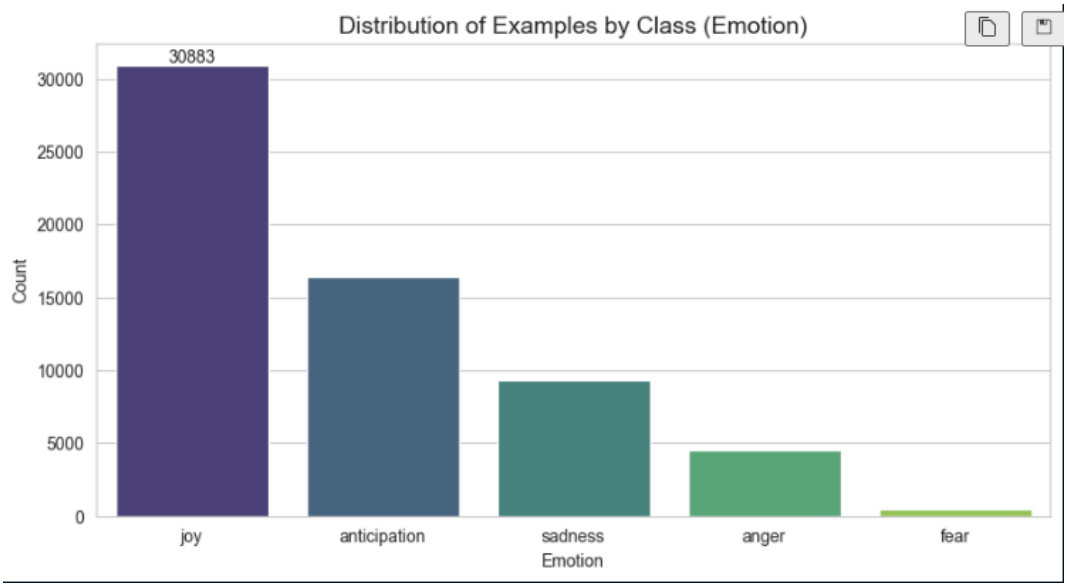
Low Resource (different size for every language)

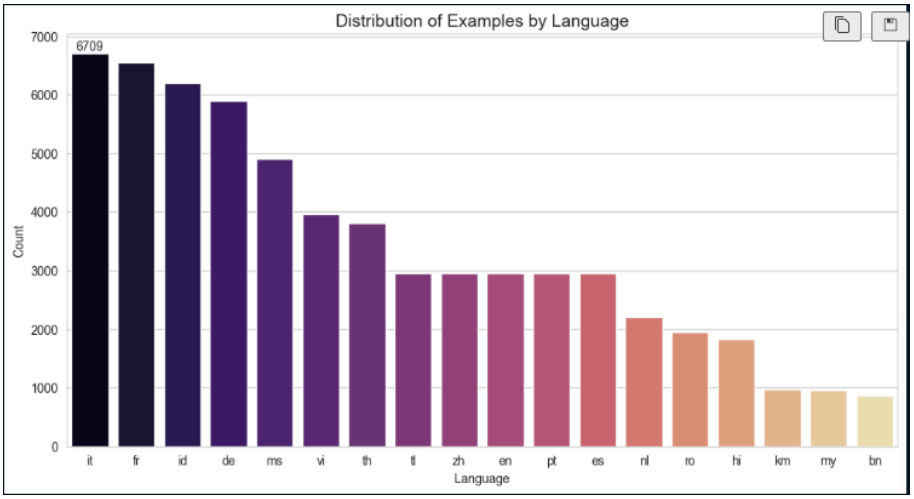
language	bn	de	fr	hi	id	it	km	ms	my	nl	ro	th	v
emotion													
anger	120	425	382	274	382	472	115	326	177	150	97	244	17(
anticipation	211	1475	1788	231	1841	1910	158	1344	130	788	560	938	11:
fear	7	8	22	8	32	20	23	34	9	10	8	21	39
joy	249	3388	3222	830	3077	3656	469	2566	412	981	923	2202	19:
sadness	282	606	1143	480	869	651	212	638	225	272	352	398	62:

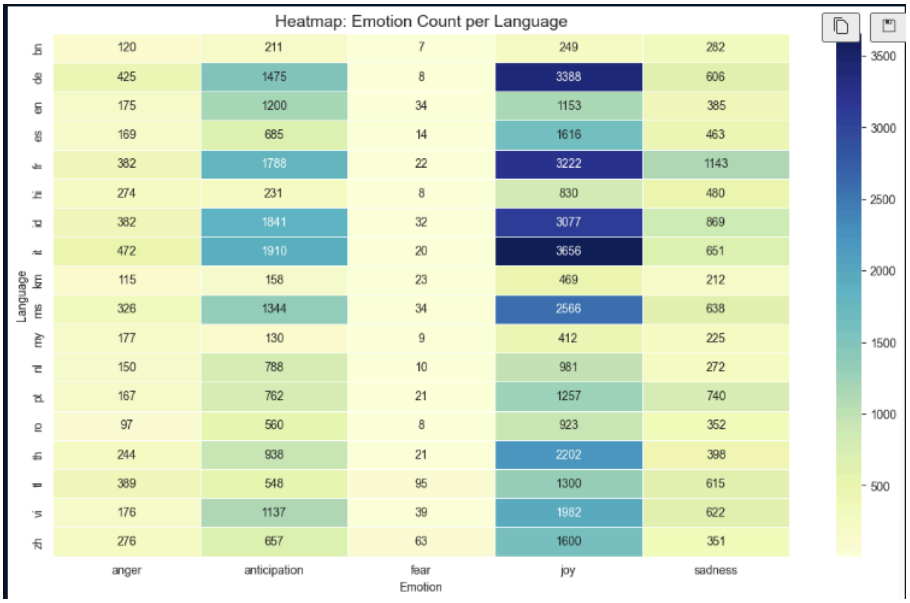
Small (2,947 instances per language)

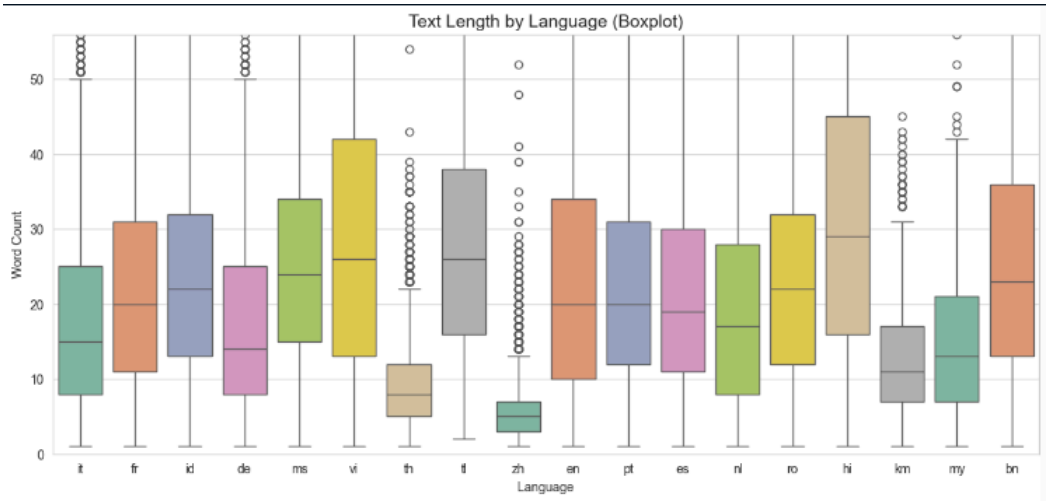
language	en	es	pt	tl	zh
emotion					
anger	175	169	167	389	276
anticipation	1200	685	762	548	657
fear	34	14	21	95	63
joy	1153	1616	1257	1300	1600
sadness	385	463	740	615	351

The dataset has shape (61534, 3), spanning multiple languages (English, Spanish, French, Hindi, etc.).
Columns: text, label, language
Labels: 4 emotion classes: joy, sadness, anger, fear, anticipation.
Relevance: This dataset represents a real-world scenario of multilingual sentiment analysis from social media, characterized by informal language, noise, and significant class imbalance, which makes it a challenging and practical NLP task.


Distribution of Examples by Class (Emotion)
30883
30000
25000
20000
15000
10000
5000
0
Count
joy
anticipation
sadness
anger
fear
Emotion


Distribution of Examples by Language
6709
7000
6000
5000
4000
3000
2000
1000
0
Count
it
fr
id
de
ms
vi
th
tl
zh
en
pt
es
nl
ro
hi
km
my
bn
Language

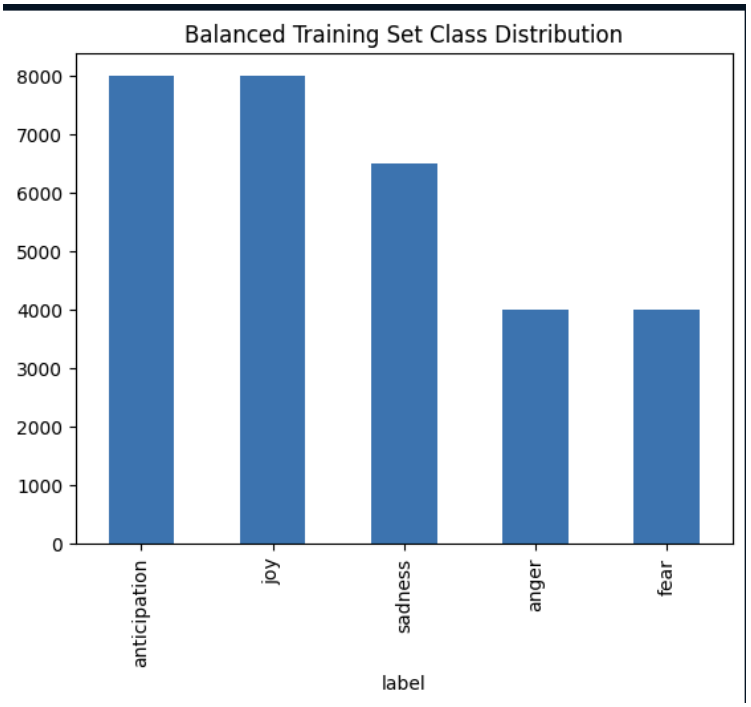

Heatmap: Emotion Count per Language
bn 120 211 7 249 282
de 425 1475 8 3388 606
en 175 1200 34 1153 385
es 169 685 14 1616 463
fr 382 1788 22 3222 1143
hi 274 231 8 830 480
id 382 1841 32 3077 869
it 472 1910 20 3656 651
km 115 158 23 469 212
ms 326 1344 34 2566 638
my 177 130 9 412 225
nl 150 788 10 981 272
pt 167 762 21 1257 740
ro 97 560 8 923 352
th 244 938 21 2202 398
tl 389 548 95 1300 615
vi 176 1137 39 1982 622
zh 276 657 63 1600 351
Language
anger
anticipation
fear
joy
sadness
Emotion
3500
3000
2500
2000
1500
1000
500

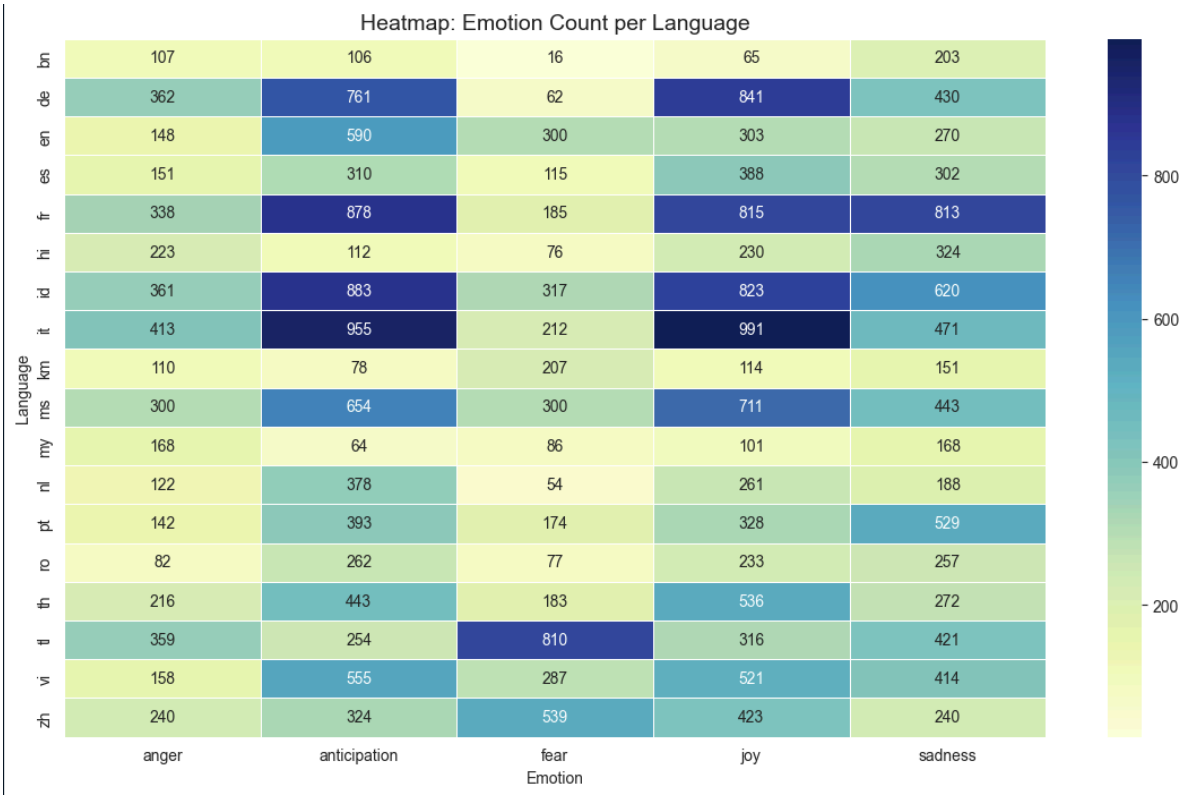


Missing Values: Was removed, also text column included inappropriate values such as [PHOTO],[PERSON], etc. also were removed.

Data Augmentation: Applied Back-Translation (via M2M100) specifically for underrepresented (Language, Label) pairs prior to training. Applied synonymous generation via BERT multilingual.

After augmentation:





Related Work (Literature Review)

Before designing the solution, three key papers were reviewed to guide the methodology:

1. He, P., Gao, J., & Chen, W. (2021). "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training".
    - Summary: This paper introduces DeBERTaV3, which replaces Masked Language Modeling (MLM) with Replaced Token Detection (RTD). It demonstrates that mDeBERTaV3 significantly outperforms XLM-R in cross-lingual tasks.
    - Relevance: This influenced the decision to choose microsoft/mdeberta-v3-base over the standard bert-base-multilingual-cased or xlm-roberta. The hypothesis was that mDeBERTa's disentangled attention would handle the diverse syntax of low-resource languages (like Hindi) better.
2. Sennrich, R., et al. (2016). "Improving Neural Machine Translation Models with Monolingual Data".
    - Summary: The foundational paper on Back-Translation. It shows that translating target data to source and back generates high-quality synthetic paraphrases.
    - Relevance: To address the scarcity of fear samples, this method was adopted for Data Augmentation. Instead of simple synonym replacement (which often breaks context in tweets), Back-Translation using M2M100 was chosen to generate diverse, context-aware training examples for minority classes.
3. Cui, Y., et al. (2019). "Class-Balanced Loss Based on Effective Number of Samples". (Standard reference for Class Weighting)

- Summary: Discusses how standard Cross-Entropy Loss fails on long-tail datasets. It proposes re-weighting loss based on class frequency.
- Relevance: This justified the implementation of a WeightedTrainer. Simply training on the raw dataset would bias the model towards joy; applying inverse class weights allows the model to "pay more attention" to the rare fear signals during backpropagation.

Problem Statement & Hypothesis

Problem: Multilingual emotion detection suffers from "resource poverty" for specific languages (e.g., Burmese, Hindi) and specific emotions (fear), leading to poor generalization.
Hypothesis: Fine-tuning a state-of-the-art cross-lingual model (mDeBERTa) combined with hybrid balancing strategies (Class Weights + Targeted Back-Translation) will yield a Macro-F1 score > 0.50, significantly outperforming a baseline majority classifier, even on low-resource languages.

Part 2: Model Training and Evaluation

1. Model Choice

Selected Model: microsoft/mdeberta-v3-base
Justification:

- Cross-lingual capability: Unlike English-only BERT, it supports 100+ languages.
- Performance: Benchmarks (XTREME) show it outperforms XLM-RoBERTa base.
- Efficiency: The "base" version fits within GPU memory constraints (Colab T4) while maintaining high accuracy.

2. Model Architecture

- Backbone: 12-layer Transformer Encoder (Hidden size 768).
- Head: A linear classification layer on top of the [CLS] token (768 -> num_labels).
- Tokenizer: SentencePiece-based tokenizer (DeBERTaV3 tokenizer).

3. Training Details

- Framework: Hugging Face Trainer API.
- Strategy:
  - Class Weights: Computed using sklearn.utils.class_weight.compute_class_weight and passed to a custom WeightedTrainer to penalize errors on minority classes (fear) more heavily.
  - Data Augmentation: Applied Back-Translation (via M2M100) specifically for underrepresented (Language, Label) pairs prior to training.
- Hyperparameters:
  - Learning Rate: 1e-5 (small for fine-tuning).
  - Batch Size: 16.

- Epochs: 5 (with Early Stopping).
- Dropout: 0.3 (model easily overfits)
- Weight Decay: 0.1

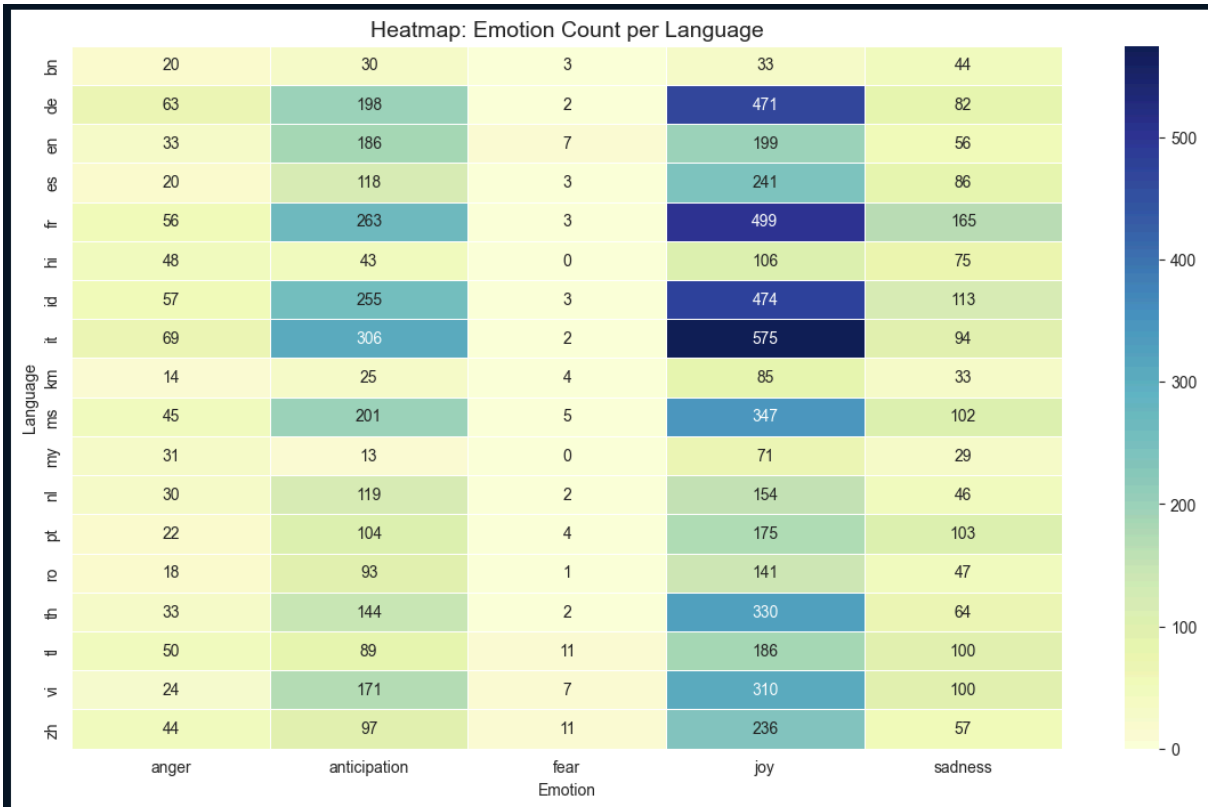
4. Evaluation

Metrics:

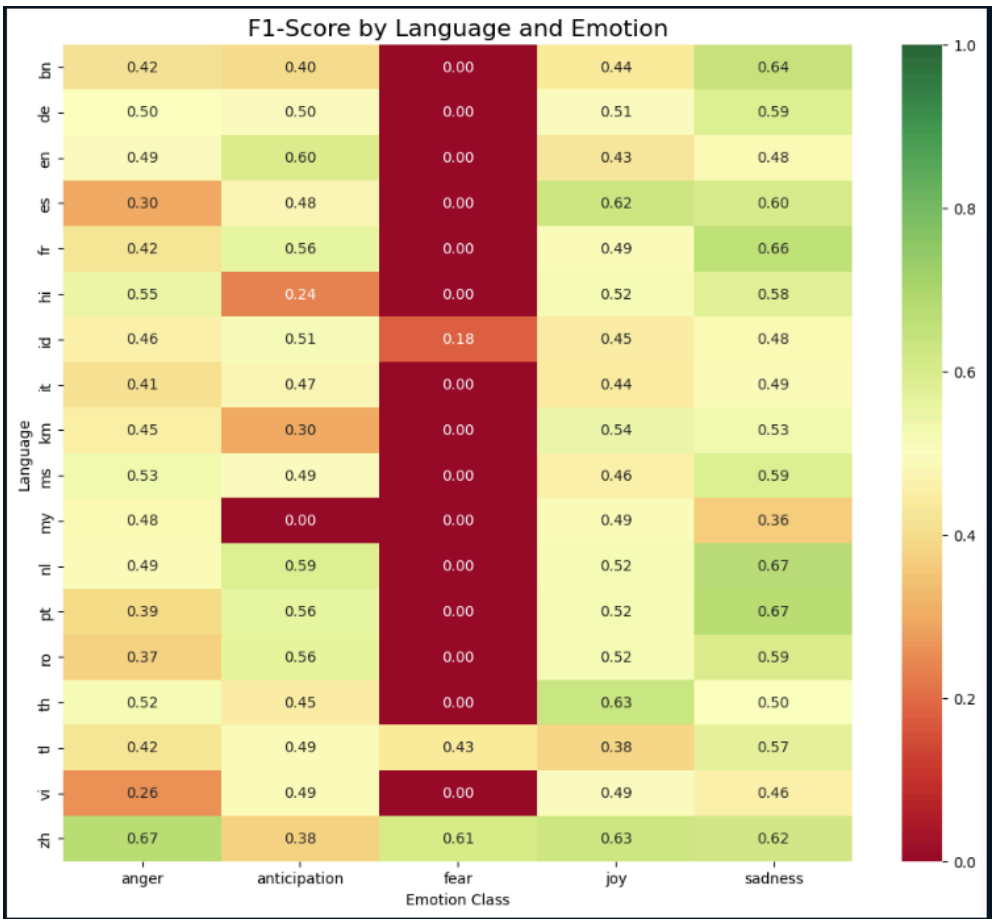
- Macro F1-Score: Chosen as the primary metric because Accuracy is misleading in imbalanced datasets.
- Confusion Matrix: To visualize misclassifications.

Results (Two test datasets were tested, one with huge class disbalance, second with generated via back-translation 430 samples for fear):

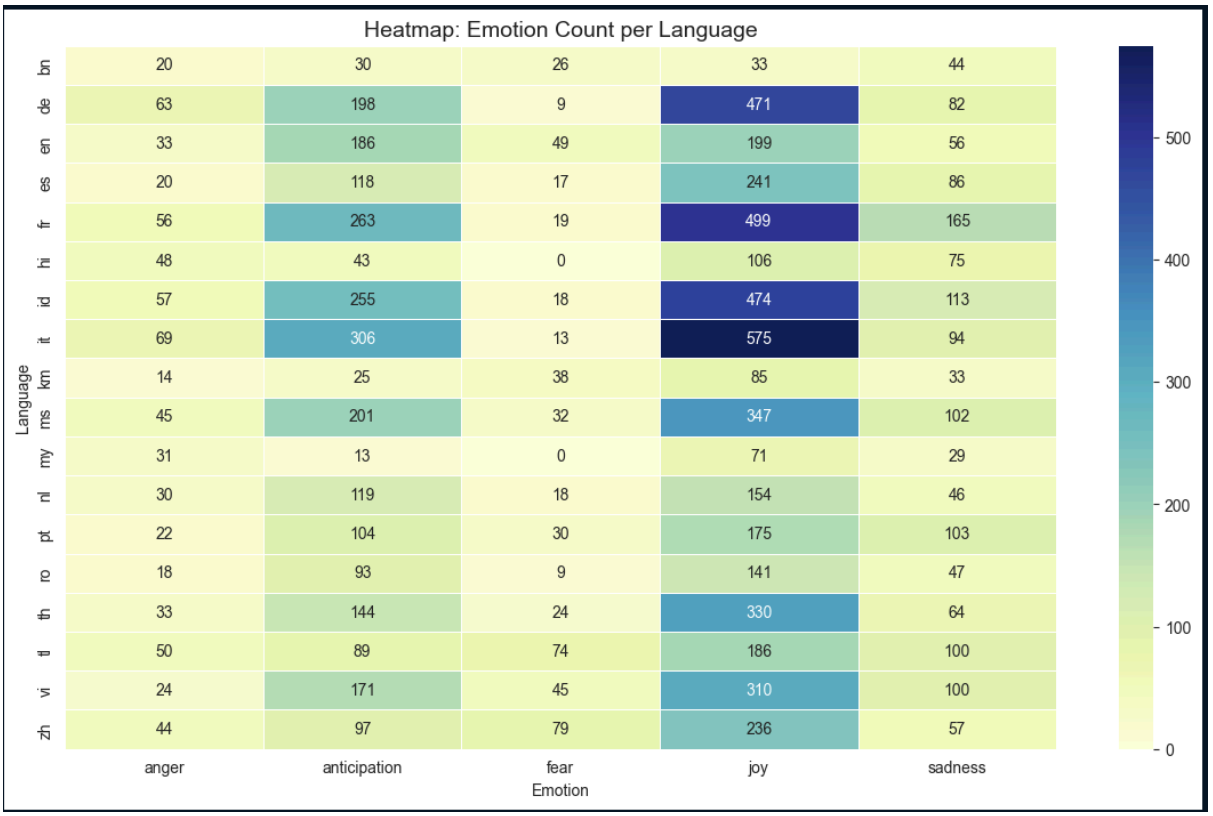
Experiment 1)



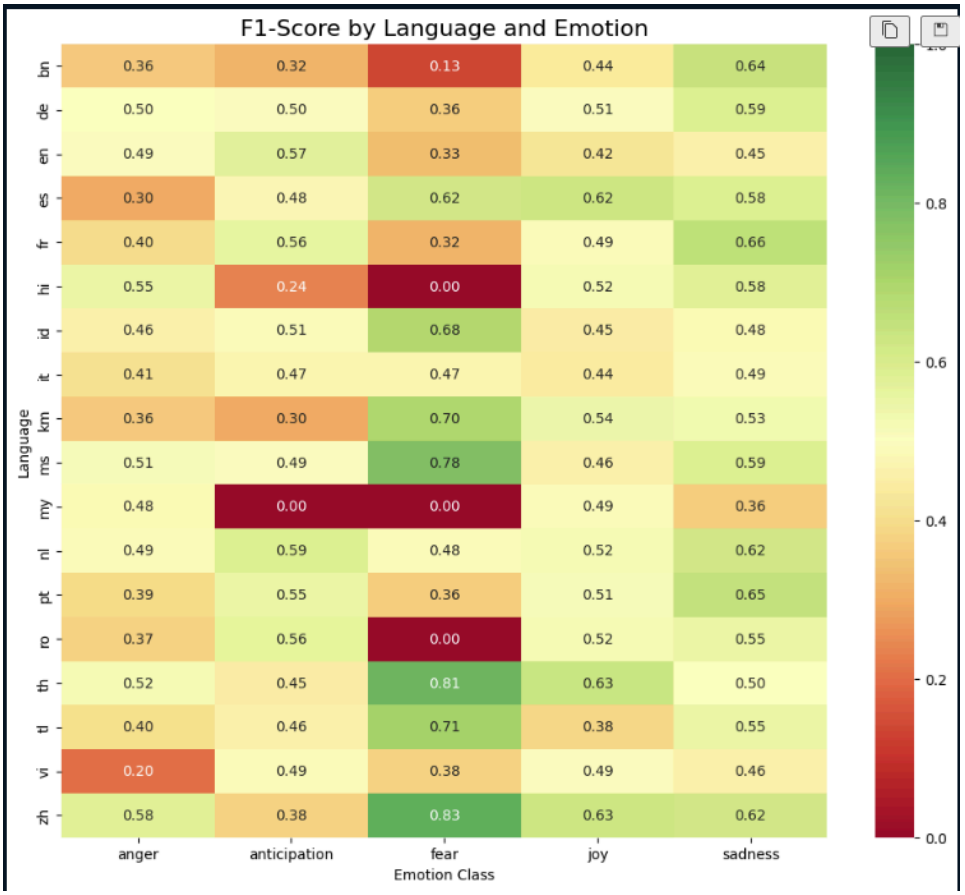
	precision	recall	f1-score	support
anger	0.36	0.63	0.46	677
anticipation	0.43	0.60	0.50	2452
fear	0.09	0.19	0.12	70
joy	0.70	0.39	0.51	4623
sadness	0.49	0.67	0.57	1392
accuracy			0.51	9214
macro avg	0.42	0.50	0.43	9214
weighted avg	0.57	0.51	0.51	9214



Experiment 2) +430 samples for fear

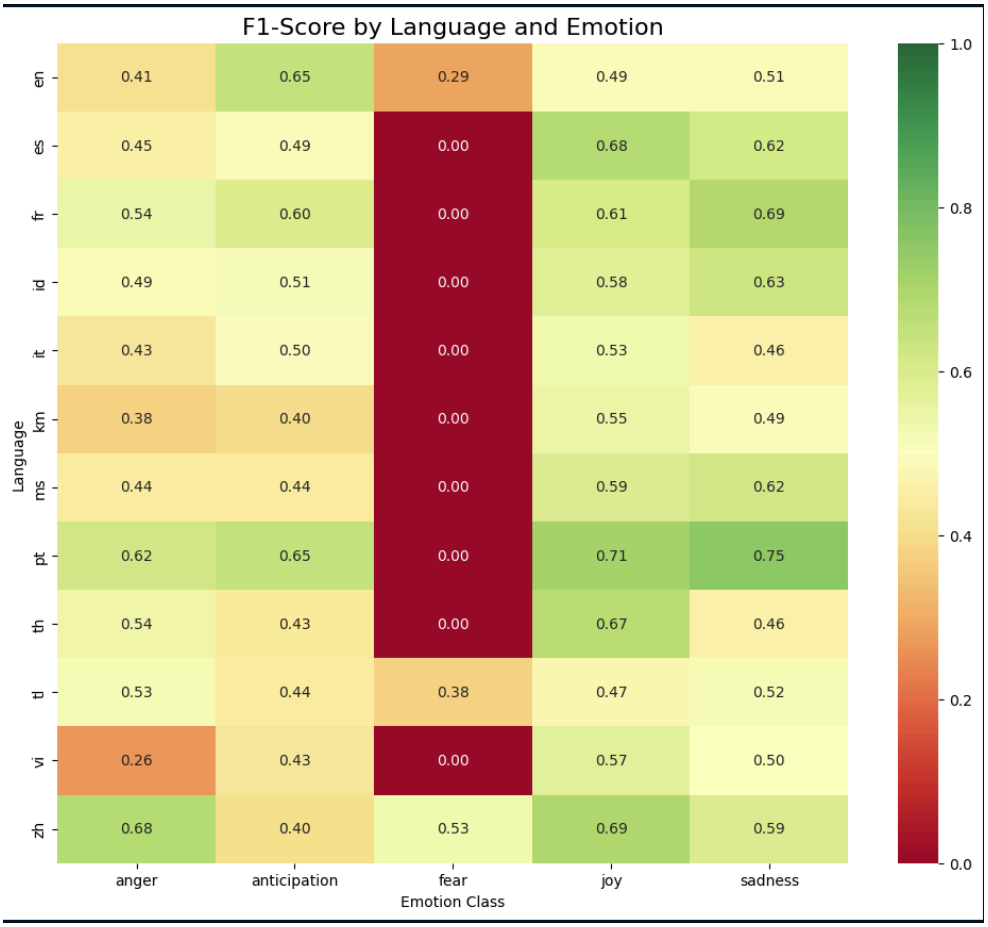


	precision	recall	f1-score	support
anger	0.34	0.63	0.44	677
anticipation	0.43	0.60	0.50	2452
fear	0.67	0.52	0.58	500
joy	0.70	0.39	0.50	4623
sadness	0.48	0.67	0.56	1392
accuracy			0.51	9644
macro avg	0.52	0.56	0.52	9644
weighted avg	0.57	0.51	0.51	9644



Experiment 3) Balanced Dataset (minimum of 500 samples per each class for each of 11 languages)

	precision	recall	f1-score	support
anger	0.44	0.54	0.48	325
anticipation	0.45	0.59	0.51	1296
fear	0.15	0.19	0.17	42
joy	0.70	0.52	0.60	2406
sadness	0.54	0.65	0.59	710
accuracy			0.56	4779
macro avg	0.46	0.50	0.47	4779
weighted avg	0.59	0.56	0.56	4779



5. Analysis & Error Analysis

Success Cases:
The model performs decently on high-resource languages and dominant classes (sadness). It captures semantic sentiment even in short texts.

Failure Cases (The "Pain Points"):

- Language-Specific Failures: As seen in the generated Heatmap, the model fails completely (F1=0.0) for fear in Hindi (hi) and Burmese (my) for augmented test dataset. For the disbalanced one model fails to classify almost all languages for fear.
- Reasoning:
    1. Extreme Data Scarcity: These specific pairs (e.g., Hindi+Fear) had <10 samples in the original training set.
    2. Translation Artifacts: The augmentation might have introduced noise, or the pre-trained mDeBERTa representation for these specific low-resource languages is weaker.

Improvements (Future Work):

- Targeted Augmentation: Instead of random augmentation, generate 500+ synthetic examples specifically for the "red zones" (Hindi/Burmese Fear) identified in the heatmap.

- Focal Loss: Replace Weighted Cross-Entropy with Focal Loss to force the model to focus on these hard-to-classify examples.

6. LLM Analysis (Llama 3.1 and Llama 3.3)

Prompt Engineering:

```python
def get_system_prompt(config_type):
    valid_labels = "joy, sadness, anger, fear, anticipation"

    if config_type == "zero-shot":
        return (
            f"You are an expert linguist specialized in emotion detection. "
            f"Classify the emotion of the text into exactly one of these labels: {valid_labels}. "
            f"Return ONLY JSON in the format {{\"emotion\": \"label\"}}."
        )

    elif config_type == "few-shot":
        return (
            "You are an expert in multilingual emotion classification.\n"
            "Possible labels: joy, sadness, anger, fear, anticipation.\n"
            "Use ONLY these labels.\n\n"

            "Examples:\n"
            "Text: I finally got the job I wanted!\nEmotion: joy\n\n"
            "Text: I lost my wallet yesterday.\nEmotion: sadness\n\n"
            "Text: This is so unfair, I'm furious.\nEmotion: anger\n\n"
            "Text: I don't know what will happen tomorrow.\nEmotion: anticipation\n\n"
            "Text: Mujhe andhere se dar lagta hai.\nEmotion: fear\n\n"

            "Now classify the following text.\n"
            "Return ONLY JSON in the format {\"emotion\": \"label\"}"
        )

    elif config_type == "chain-of-thought":
        return (
            "You are an expert multilingual linguist and psychologist. "
            "Your task is to analyze the emotional state of a social media post.\n\n"
            "Follow these steps:\n"
            "1. Identify the language of the text.\n"
            "2. Briefly analyze key emotional keywords and context.\n"
            "3. Choose the most fitting emotion from this list: [joy, sadness, anger, fear, anticipation].\n\n"
            "Output your reasoning and the final label in JSON format like this:\n"
            "{\n"
            "  \"reasoning\": \"Step-by-step analysis...\",\n"
            "  \"emotion\": \"label\"\n"
            "}"
        )
```

Test dataset: one example per each class for each of 18 languages

Results:

Few-shot

```
Detailed Classification Report (llama-3.1-8b-instant)
              precision    recall  f1-score   support

         joy       0.34      0.67      0.45        18
     sadness       0.45      0.56      0.50        18
       anger       0.62      0.28      0.38        18
        fear       0.60      0.53      0.56        17
anticipation       0.50      0.22      0.31        18

   micro avg       0.45      0.45      0.45        89
   macro avg       0.50      0.45      0.44        89
weighted avg       0.50      0.45      0.44        89

Detailed Classification Report (llama-3.3-70b-versatile)
              precision    recall  f1-score   support

         joy       0.45      0.78      0.57        18
     sadness       0.64      0.78      0.70        18
       anger       0.60      0.33      0.43        18
        fear       1.00      0.47      0.64        17
anticipation       0.50      0.50      0.50        18

    accuracy                           0.57        89
   macro avg       0.64      0.57      0.57        89
weighted avg       0.63      0.57      0.57        89
```

Zero-Shot

```
Detailed Classification Report (llama-3.1-8b-instant)
              precision    recall  f1-score   support

         joy       0.35      0.72      0.47        18
     sadness       0.39      0.67      0.49        18
       anger       0.67      0.33      0.44        18
        fear       0.78      0.41      0.54        17
anticipation       1.00      0.11      0.20        18

   micro avg       0.45      0.45      0.45        89
   macro avg       0.64      0.45      0.43        89
weighted avg       0.63      0.45      0.43        89

Detailed Classification Report (llama-3.3-70b-versatile)
              precision    recall  f1-score   support

         joy       0.45      0.78      0.57        18
     sadness       0.56      0.78      0.65        18
       anger       0.50      0.28      0.36        18
        fear       1.00      0.29      0.45        17
anticipation       0.50      0.50      0.50        18

    accuracy                           0.53        89
   macro avg       0.60      0.53      0.51        89
weighted avg       0.60      0.53      0.51        89
```

Chain-of-Thought

Detailed Classification Report (llama-3.1-8b-instant)

	precision	recall	f1-score	support
joy	0.44	0.67	0.53	18
sadness	0.40	0.56	0.47	18
anger	0.75	0.50	0.60	18
fear	0.88	0.41	0.56	17
anticipation	0.53	0.44	0.48	18
micro avg	0.53	0.52	0.52	89
macro avg	0.60	0.52	0.53	89
weighted avg	0.60	0.52	0.53	89

Detailed Classification Report (llama-3.3-70b-versatile)

	precision	recall	f1-score	support
joy	0.44	0.83	0.58	18
sadness	0.55	0.67	0.60	18
anger	0.64	0.39	0.48	18
fear	1.00	0.41	0.58	17
anticipation	0.69	0.50	0.58	18
micro avg	0.57	0.56	0.57	89
macro avg	0.66	0.56	0.56	89
weighted avg	0.66	0.56	0.56	89

Pretrained mDeBERTa

Classification Report for Fine-Tuned Model

	precision	recall	f1-score	support
anger	0.69	0.50	0.58	18
anticipation	0.50	0.50	0.50	18
fear	1.00	0.24	0.38	17
joy	0.50	0.67	0.57	18
sadness	0.47	0.78	0.58	18
accuracy			0.54	89
macro avg	0.63	0.54	0.52	89
weighted avg	0.63	0.54	0.52	89