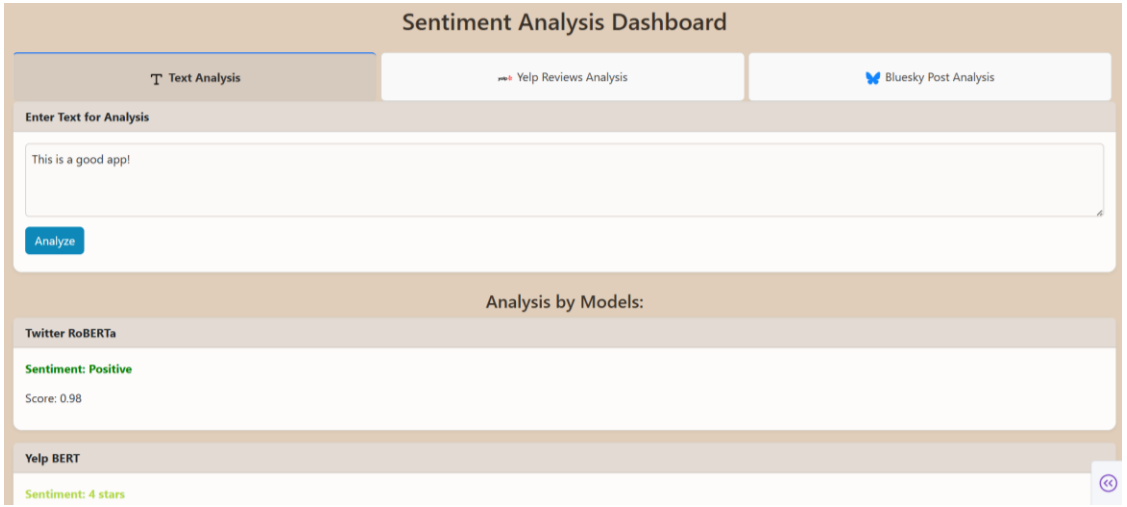
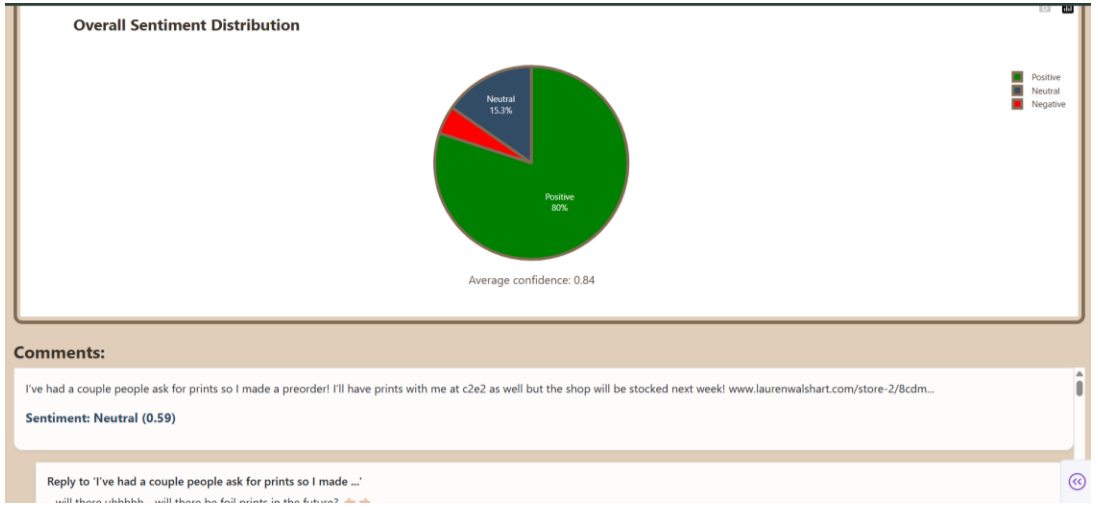
Team 7
Text Sentiment Analysis

Team Members: Ross Kugler, Jason Eigeman, Justin Lawrence,
Audrey Lindstrom, Linh (Jason) Nguyen, Huy (Harry) Ky

Project Overview

- Problem: How to analyze and classify the overall sentiment of text comments/reviews?
- Solution: Web app that utilizes AI models trained on text sentiment analysis to provide insights





Datasets

For Training:

- GoEmotions
  - human-annotated emotion data
  - dominant heuristic uses emotion-related Twitter tags

For Testing/Using with our Models:

- Yelp Academic Dataset
- Bluesky API

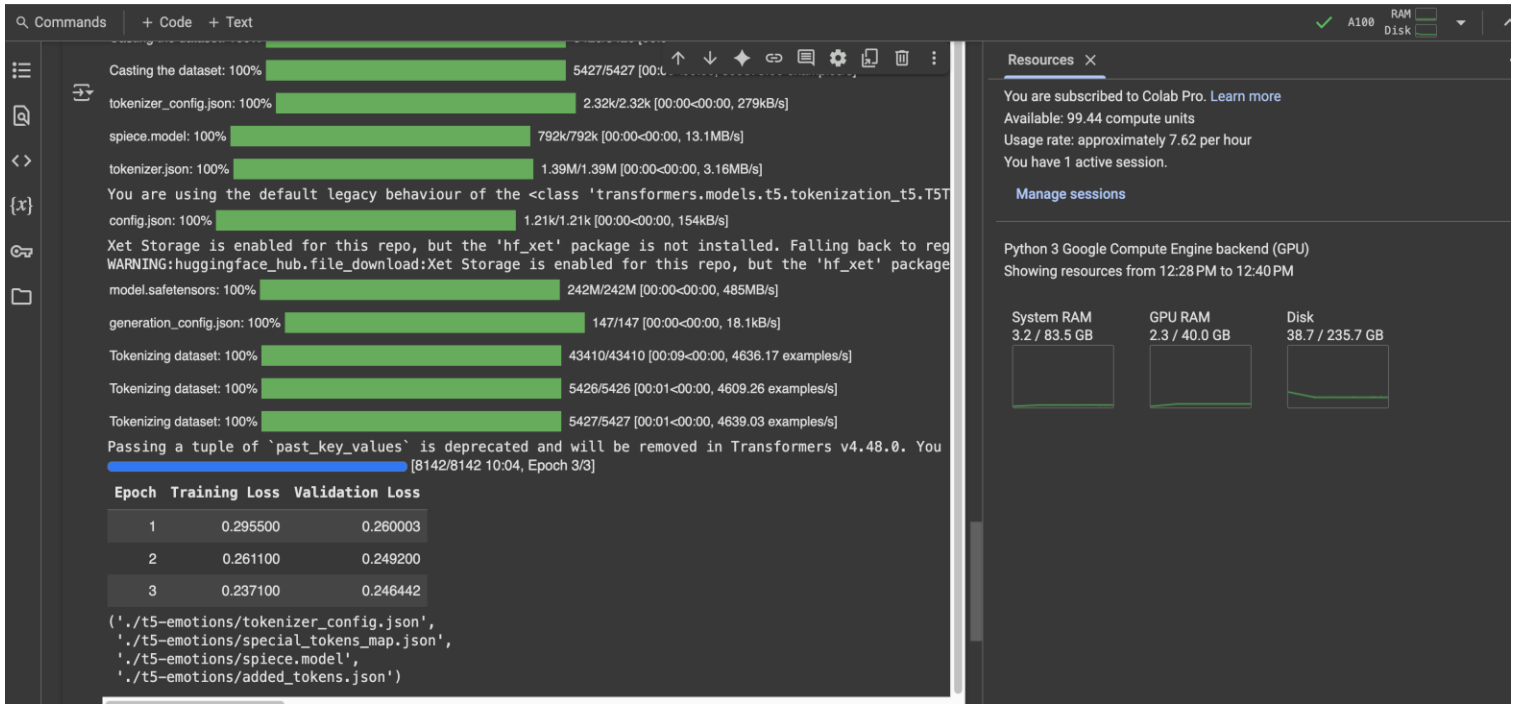
Models

- T5Emotions (custom model, fine-tuned on GoEmotions dataset)
- Twitter RoBERTa (pretrained, from HuggingFace)
- Yelp BERT (pretrained, from HuggingFace)

Model	Base Architecture	Dataset Source	Task Type	**Output Format**
GoEmotions	BERT-base	Reddit comments	Multi-label classification	List of emotions
T5Emotions	T5 (text-to-text)	GoEmotions	Text generation (emotions)	Generated label text
Twitter RoBERTa	RoBERTa-base	Twitter data	Classification	Single/multi emotion
Yelp BERT	BERT-base	Yelp reviews	Sentiment classification	Star rating or emotion

Fine-Tuning T5 with GoEmotions Dataset

- We utilized Google Colab Pro to access A100 to effectively train the model.
- With each epoch, both the training and validation loss consistently decreased, indicating that the model was effectively learning from the training data and improving its performance over time.



Fine-Tuned T5 Metrics

- Utilized GoEmotions Test data to run metrics on how well the model learned from the training data.
- Micro F1 Score: Measures the model's overall ability to correctly predict all emotions. Higher is better.
- Macro F1 Score: Averages the model's performance on each individual emotion. Higher is better.
- Subset Accuracy: Calculates how often the model predicts all the correct emotions for a sentence with no mistakes. Higher is better.
- Hamming Loss: Shows the percentage of individual label predictions the model got wrong. Lower is better

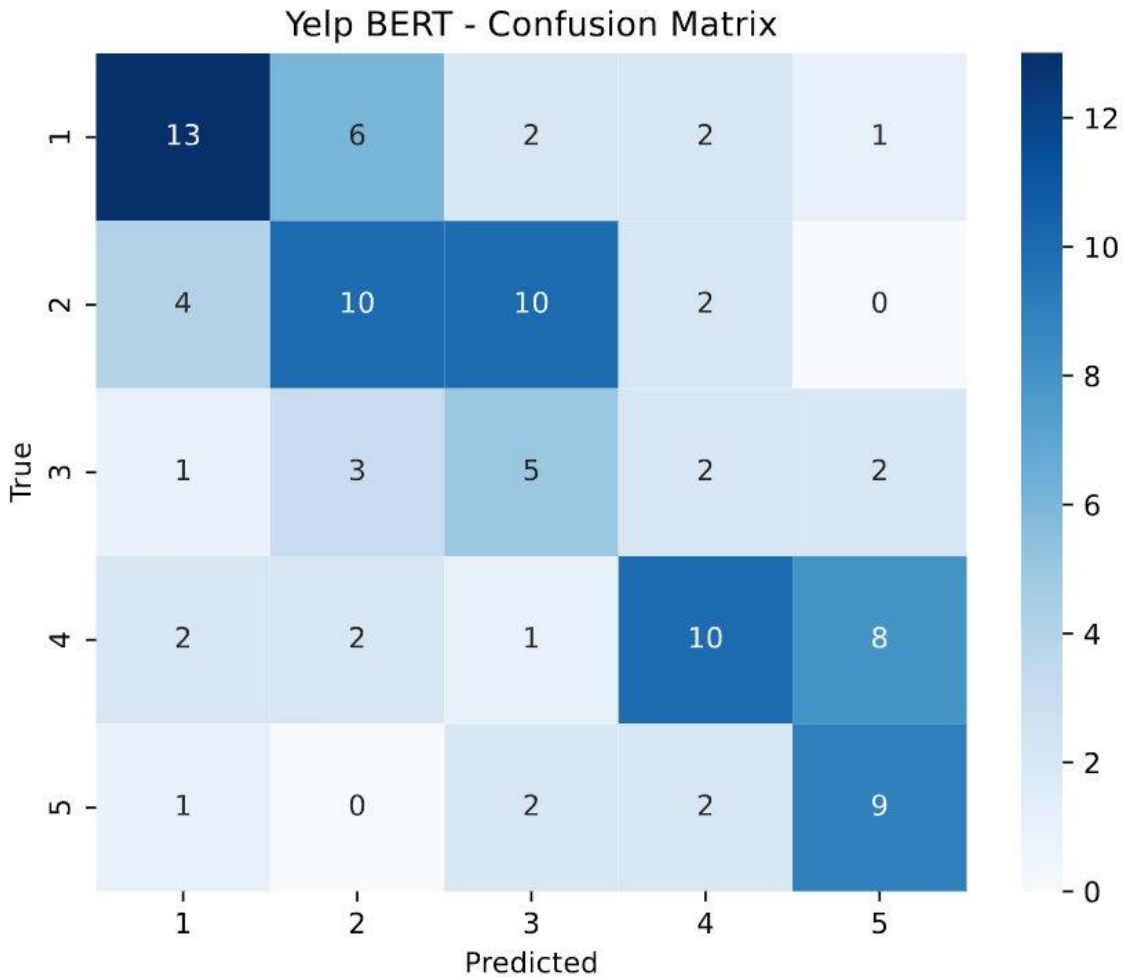
```
----------Evaluation Metrics on Google_Emotions----------

Test set was used for metrics.
The model was fine tuned using the training set
Micro F1 Score: 0.5940
Micro F1: The model was accurate 59.40% of the time across all emotion predictions.
Macro F1 Score: 0.4775
Macro F1: On average, the model was 47.75% accurate across each individual emotion.
Subset Accuracy: 0.5222
Subset Accuracy: The model made perfect emotion predictions for 52.22% of sentences.
Hamming Loss: 0.0320
Hamming Loss: On average, the model made mistakes on 3.20% of possible emotion labels.
```

Evaluation

	Model	Micro F1	Macro F1	Accuracy	Hamming Loss	MAE	Exact Match	Within ±1
0	T5Emotions	0.6061	0.3711	0.51	0.0325	–	–	–
1	GoEmotions	0.6239	0.3901	0.54	0.0293	–	–	–
2	Twitter RoBERTa	0.4	0.3315	0.4	0.6	–	–	–
3	Yelp BERT	0.47	0.4647	0.47	0.53	0.78	0.47	0.83

Evaluation


Yelp BERT - Confusion Matrix
True
Predicted
1
2
3
4
5
13 6 2 2 1
4 10 10 2 0
1 3 5 2 2
2 2 1 10 8
1 0 2 2 9
12
10
8
6
4
2
0

Project Difficulties

- Data Quality
  - Typos, Negations, and Sarcasm
- Combining and Normalizing different datasets
- Time-Constraints (would improve UI, more time to train models)