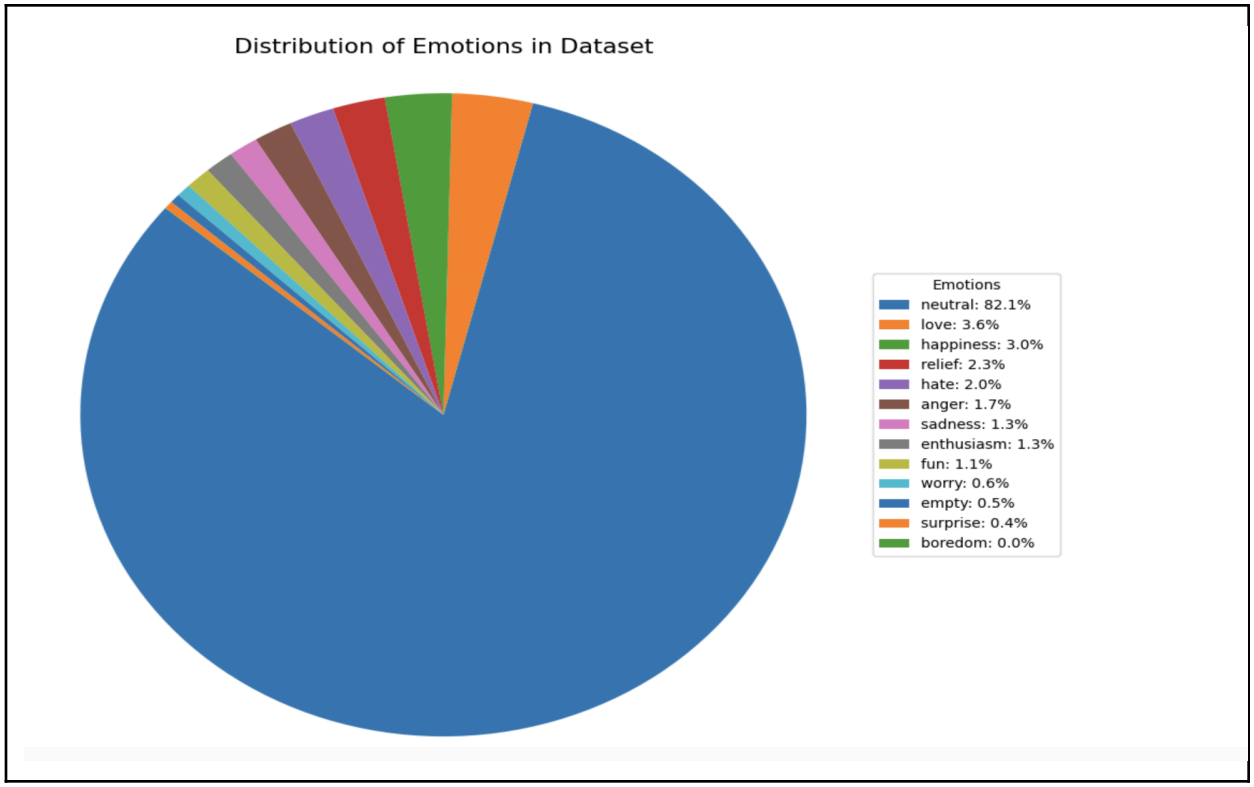
# Project 3 - NLP

**Student Name:** Vivian Umansky
**Submission Date:** 03.06.25

---

## Part 1 - Text Classification

**1.1**




**1.2**

The dataset is imbalanced, with over 80% of the samples labeled as "neutral". A model could achieve high accuracy by always predicting "neutral", even if it fails to detect other emotions. Accuracy does not reflect how well the model handles the minority classes.

Precision and recall, on the other hand, focus on how well the model identifies the non-neutral emotions.Therefore, precision and recall are more meaningful metrics for evaluating performance in this case.

**1.3**

The model performs very well on both the training and testing datasets. After training for 5 epochs on 10,000 samples using the DistilBERT classifier, the performance metrics were:

- Training Precision: 1.0000
- Training Recall: 0.9986
- Testing Precision: 0.9915
- Testing Recall: 0.9560

The confusion matrix on the test set shows very low error rates:

- Only 3 false positives (neutral misclassified as non-neutral)
- Only 16 false negatives (non-neutral misclassified as neutral)

These results indicate that the model is highly accurate and generalizes well beyond the training data. The small number of misclassifications demonstrates a good balance between precision and recall, particularly for the minority (non-neutral) class, which is often harder to detect in imbalanced settings.

Looking at the validation loss per epoch:

- It decreases significantly from epoch 1 to 2
- It flattens slightly between epochs 2 and 3
- It slightly increases in epochs 4 and 5

This slight increase in validation loss does not come with a drop in performance metrics (precision and recall remain high), so there is no clear sign of overfitting.
The model performs strongly and shows no significant overfitting. **Early stopping is not necessary in this case**, though it could be considered in future runs to avoid unnecessary computation after the 3rd epoch, where performance already stabilizes

**1.4**

To adjust both datasets to neutral vs. non-neutral classification, I reduced the ground truth labels to binary values:

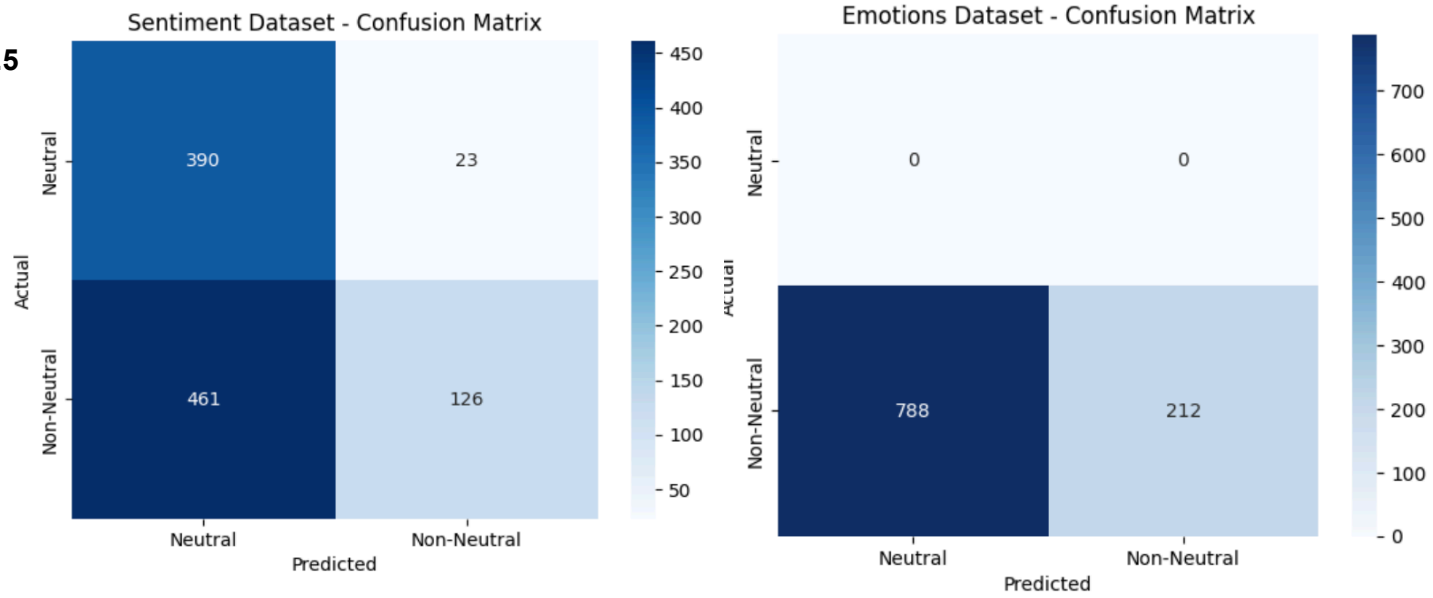
- 0 for neutral
- 1 for non-neutral

In the Sentiment Analysis dataset, I mapped the label neutral to 0, and both positive and negative to 1, since they express emotional polarity.

In the Emotions dataset, all labels represent emotional states, so I mapped all of them to 1(non neutral). This dataset does not contain a neutral class.

```
Mapping Table:
| Dataset   | Original Label           | Mapped Label    |
|:----------|:-------------------------|:----------------|
| Sentiment | neutral                  | 0 (neutral)     |
| Sentiment | positive/negative        | 1 (non-neutral) |
| Emotions  | all (sadness, joy, etc.) | 1 (non-neutral) |
```

**1.5**




a. Evaluated the trained model on two external datasets (1K samples each), without retraining.

- Sentiment Dataset: Precision: 0.8456, Recall: 0.2147
- Emotions Dataset: Precision: 1.0000, Recall: 0.2120

Both confusion matrices show that the model predicts "Neutral" too often, missing many non-neutral cases.

b. Yes, there is a noticeable drop in recall across both datasets.
The model tends to over predict the "Neutral" class, missing many non-neutral examples. This suggests it is biased toward the majority class it saw during training.

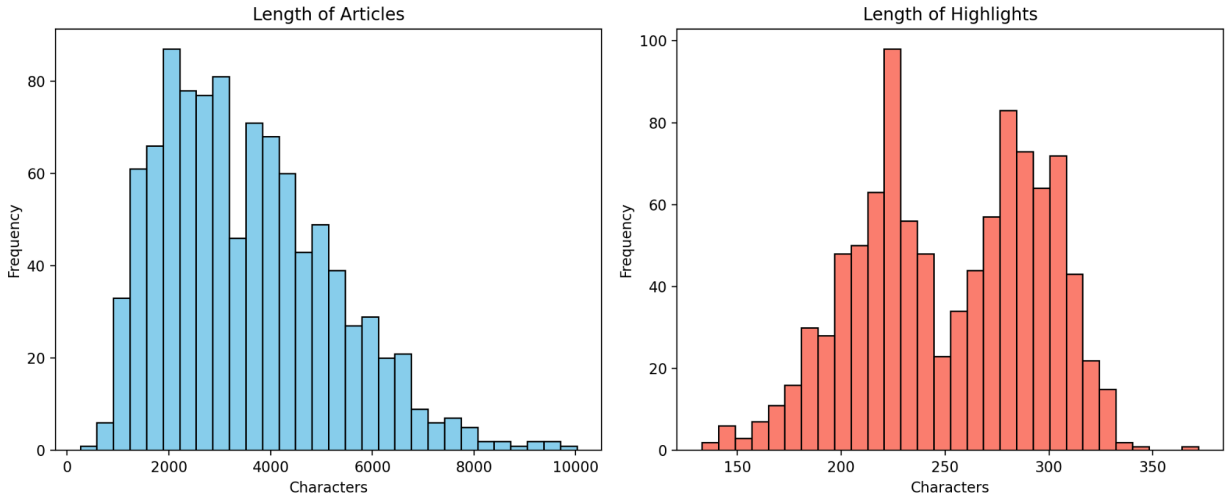
Two main factors contribute to this:

1. **Class imbalance** in the original training data made the model cautious, favoring "Neutral" predictions.
2. **Domain shift** – the external datasets include emotional language or phrasing that the model did not encounter during training, reducing its ability to generalize.

As a result, while precision remains high, recall drops significantly indicating the model is too cautious in predicting non-neutral examples.

# Part 2 - Text summarization

**2.2**




**2.3**

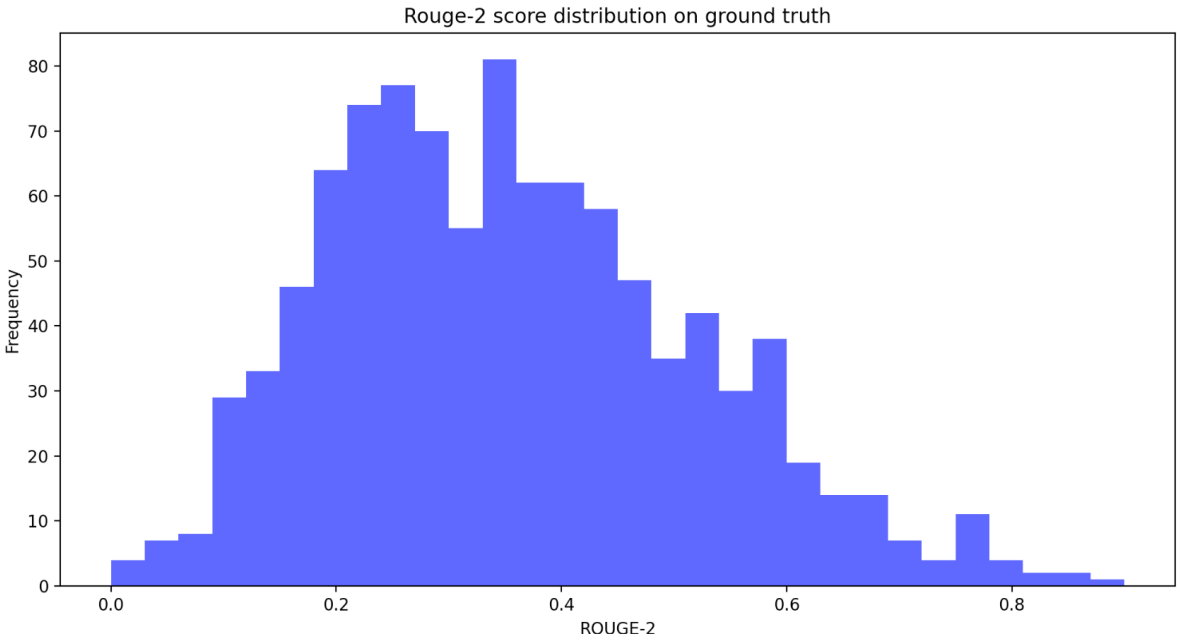
I wrote a custom function to compute ROUGE-N recall:

- ROUGE-1: unigram overlap
- ROUGE-2: bigram overlap
- I applied it on 1000 rows from the CNN/DailyMail dataset.

| Metric | Highest Score | Lowest Score |
|---|---|---|
| ROUGE -1 | 0.90 | 0.00 |
| ROUGE -2 | 0.87 | 0.00 |

Example with Lowest ROUGE-2:

- Article: About Qualcomm's business success and patents
- Highlight: Summarizes Qualcomm’s growth, patent portfolio, and company ranking
- Why score = 0.00?
  Even though the highlight is a valid summary of the article, it doesn’t share any bigrams with the article text. ROUGE-2 only checks for exact word pair matches, so valid paraphrases are penalized.
- ROUGE-N is a surface-level metric: it does not capture meaning or paraphrasing only literal n-gram overlap.


Rouge-2 score distribution on ground truth
Frequency
ROUGE-2
0
10
20
30
40
50
60
70
80
0.0
0.2
0.4
0.6
0.8

**2.4**

My Process:

- Loaded the T5-small model using HuggingFace pipeline.
- Generated summaries for the first 10 articles.
- Used my custom rouge_n function to compute ROUGE-2 scores for each generated summary vs. reference highlight.

| Example | Rouge- 2 Score |
|---|---|
| 1 | 0.1667 |
| 2 | 0.0000 |
| 3 | 0.0870 |
| 4 | 0.0500 |
| 5 | 0.0000 |
| 6 | 0.0000 |
| 7 | 0.0323 |
| 8 | 0.1667 |
| 9 | 0.0526 |
| 10 | 0.0000 |

Many model-generated summaries received a lower ROUGE-2 score than the human-written summaries achieved in section 2.3.

Example: ROUGE-2 score: 0.0000

Generated summary (T5 model):inmates with most severe mental illnesses are incarcerated until they're ready to appear

Reference summary (ground truth): “ Mentally ill inmates in Miami are housed on the forgotten floor Judge Steven Leifman says most are there as a result of avoidable felonies
While CNN tours facility, patient shouts: I am the son of the president
Leifman says the system is unjust and he's fighting for change”

Although the generated summary captures the general topic (mentally ill inmates in prison), it does not share any exact bigrams with the reference summary. As a result, ROUGE-2 gives it a score of 0, despite being a reasonable paraphrase.

## Advanced Section - Extra Points

**2.5**
ROUGE-N is a useful objective metric, but it has important limitations. It only checks for overlapping words or word pairs between the generated summary and the reference. It doesn't recognize valid paraphrasing, doesn't measure how well a summary flows, and cannot detect factual errors. Therefore, it's important to include subjective evaluation by human readers.

Evaluation Strategy with 100 People:

We ask 100 people to rate each generated summary based on 3 aspects:

- Fluency: Is the summary grammatically correct and easy to read?
- Relevance: Does the summary stay on-topic and cover the main idea of the article?
- Informativeness: Does the summary contain key information or is it missing key points?

Each criterion is rated on a 1 to 5 scale, where:

- 1 = very poor
- 5 = excellent

Each person gives 3 scores per summary.

For each summary, we compute the average score per person:

Person's score = (Fluency + Relevance + Informativeness) / 3

Then, we average over all 100 people:

Final score = (Sum of all 100 individual scores) / 100

This gives a single value between 1.0 and 5.0 for each summary.

This method gives a more complete picture of summary quality than ROUGE, because it includes human judgment of language quality and meaning, not just surface similarity.

## Part 3 - Information Retrieval

**3.2**

| Query | Most relevant article | Similarity Score |
|---|---|---|
| Leonardo DiCaprio | Elizabeth was portrayed in a variety of media by many notable artists, including painters Pietro Annigoni, Peter Blake, Chinwe Chukwuogo-Roy, Terence Cuneo, Lucian Freud, Rolf Harris, Damien Hirst, Juliet Pannett and Tai-Shan Schierenberg | 0.53681934 |
| France | In May 2022, FIFA announced the list of 36 referees, 69 assistant referees, and 24 video assistant referees for the tournament. Of the 36 referees, FIFA included two each from Argentina, Brazil, England, and France | 0.36249435 |
| Python | SharePoint, a web collaboration platform codenamed as Office Server, has integration and compatibility with Office 2003 | 0.5572225 |
| Deep Learning | SharePoint, a web collaboration platform codenamed as Office Server, has integration and compatibility with Office 2003 | 0.5605742 |