ERROR ANALYSIS – GROUP 24

(Vladislav Stoimenov, Mario Velichkov, Peter Paskalev, Raya-Neda Borisova)

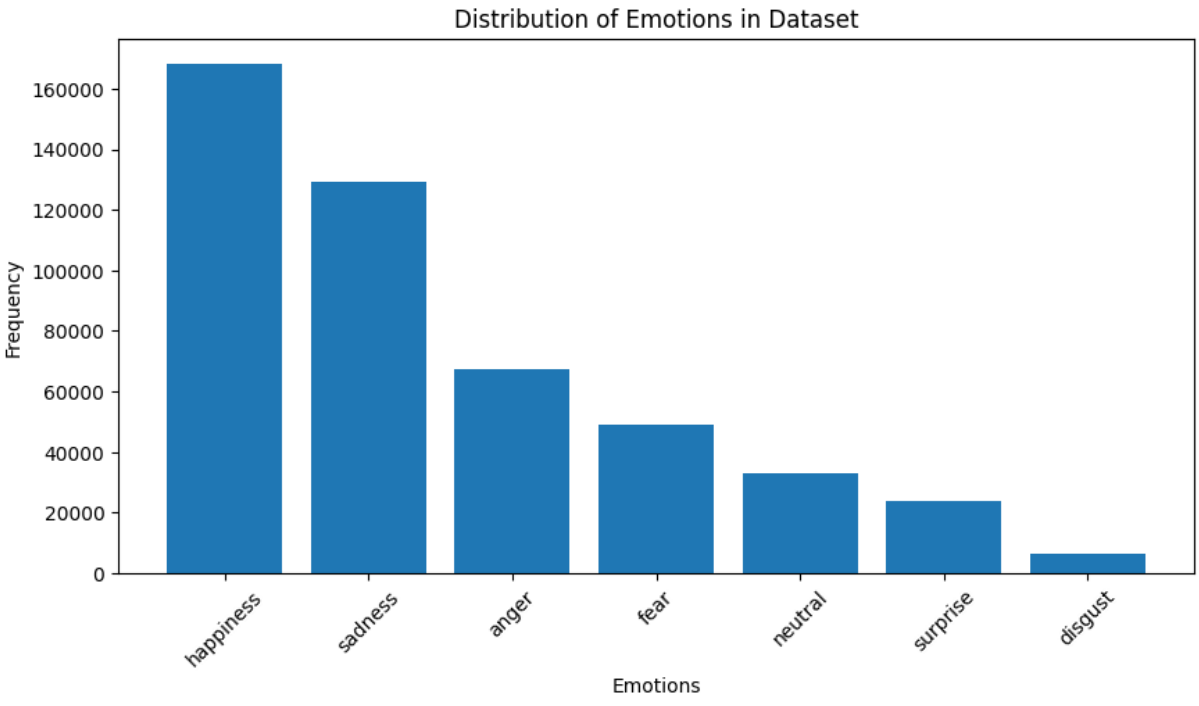
1. Introduction

In the project, we as a group trained a Transformer-based sentiment classification model to detect one of the seven core emotions in short textual inputs: happiness, sadness, anger, fear, disgust, surprise, and neutral. While the model achieved reasonable overall performance, this report focuses on a detailed error analysis. The goal is to identify patterns, evaluate model weaknesses, and suggest actionable improvements based on linguistic insights.

2. Dataset Distribution, Class Imbalances, Limitation, Biases

The training dataset has a big difference in how many examples each emotion has. Happiness and sadness have the most examples, while disgust, surprise, and neutral have much fewer. This kind of imbalance can cause the model to focus too much on the common emotions and ignore the rare ones.

To fix this, class weights are used during training. This means the model gave more importance to the rare emotions while learning. As a result, the model was still able to handle all emotions fairly well, even though the data was unbalanced.



3. Error before relabelling

During the training iterations of the Bert transformer the validation F1-score and Accuracy were increasing among the epochs but the predictions on the CIA test set were really inaccurate. The transformer model got an F1 score of 0.71 during the training process, but on the test set it decreased to around 0.50. The possible issues were due to overfitting of the model or wrong labels in the test set. As a group we decided to check the labels of the test set and to check whether they were accurate or not. We found out that a lot of them were not labelled properly. There were a lot of misclassified sentences in the Content Intelligence Agency's test set, so we managed to fix the issues with them. There were two typical errors in the test set, the first was that the emotion on many sentences were not accurate, and the second one is due to incorrect transcription or translation.

In a lot of sentences, the corresponding emotion that we got from CIA were not relevant to the corresponding sentences. Examples:

- According to CIA, the sentence below expresses happiness, but It does not express any emotion, so we set it 'neutral'.

144	For my daughter.

- CIA got this sentence as 'happiness' but while not looking at the context and the previous or the next sentences, it is just informative and does not express any emotion, so it is supposed to be 'neutral'.

151	Today is the premiere of the movie,

- The sentence below is classified as 'neutral' but since it expresses kind of nostalgia, which is more about 'sadness'.

184	We were better with my father.

There is a lot of incomplete sentences which have missing words and the translation on some of them were really inaccurate. Examples:

- These two lines represent one sentence, but they are separated as two distinct sentences, which makes it harder to determine the emotion of each of them since they are incomplete.

121	I also thought it might be
122	For domestic violence.

128	Ok if there is any problem
129	You can tell me.

- In this sentence the transcription or the translation itself is not complete. The phrase 'behaves with' is unclear and it is hard to understand what exactly means in the context. 'Behaves' usually needs more context, that we don't have in this sentence, so it's difficult to define the emotion of it.

97	Lily does not behave like a parent, but behaves with one

- Below is another example of unclear sentence structure where it is hard to get the intended meaning. The error here is probably caused due to incorrect transcription. Some of phrases of the show we have chosen to transcribe were really unclear, so the Whisper did not catch all of the words and then after the translation we got this awkward sentence, which meaning and emotion is hard to be defined.

125	You must not be that this injury

After fixing the test set, we looked at how the labels changed. We noticed that many sentences that were first labelled as 'anger', 'happiness', or 'fear' were changed to 'neutral.' This shows that some of the original labels made the sentences seem more emotional than they really were. On the other hand, a lot of 'neutral' sentences were actually expressing emotions like 'anger', 'sadness', or 'happiness', so we updated those too. For example, 34 'neutral' sentences were changed to 'anger', and 24 to 'sadness'. It also happened that some 'sadness' labels were better described as 'anger', and a few 'fear' sentences were closer to 'sadness'. These changes show that the original labels often didn't match the actual emotion in the sentence—either because the emotion was too subtle, or the sentence was hard to understand without context. After the corrections, the test set was much more accurate and gave us a better way to measure how well the model was really working. The plot below shows the Corrected (columns) and the Original (rows) values.

Corrected	anger	happiness	neutral	sadness
Original				
anger	0	0	24	1
disgust	3	0	1	0
fear	3	0	3	2
happiness	1	0	16	2
neutral	34	17	0	24
sadness	17	0	4	0
surprise	0	2	5	2

4. Confusion Matrix Analysis

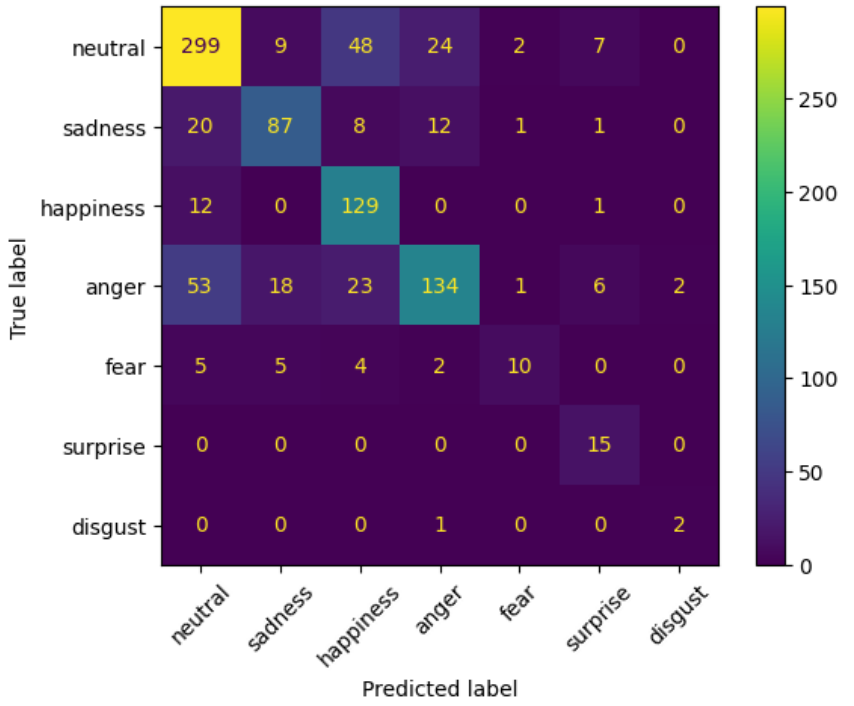
The confusion matrix offers a clear overview of where the model struggles the most. The most frequent misclassifications were:

- Anger – Neutral (53 times)
- Neutral – Happiness (48 times)
- Anger – Happiness (23 times)
- Sadness – Neutral (20 times)

The confusion matrix helped to see where the model made the most mistakes. One big problem was that the model often predicted 'neutral' when the correct emotion was something else. For example, a lot of 'anger' sentences were predicted as 'neutral'. This probably happened because the 'angry' sentences didn't always use strong or clear words—sometimes the emotion was more hidden or said in a calm way.

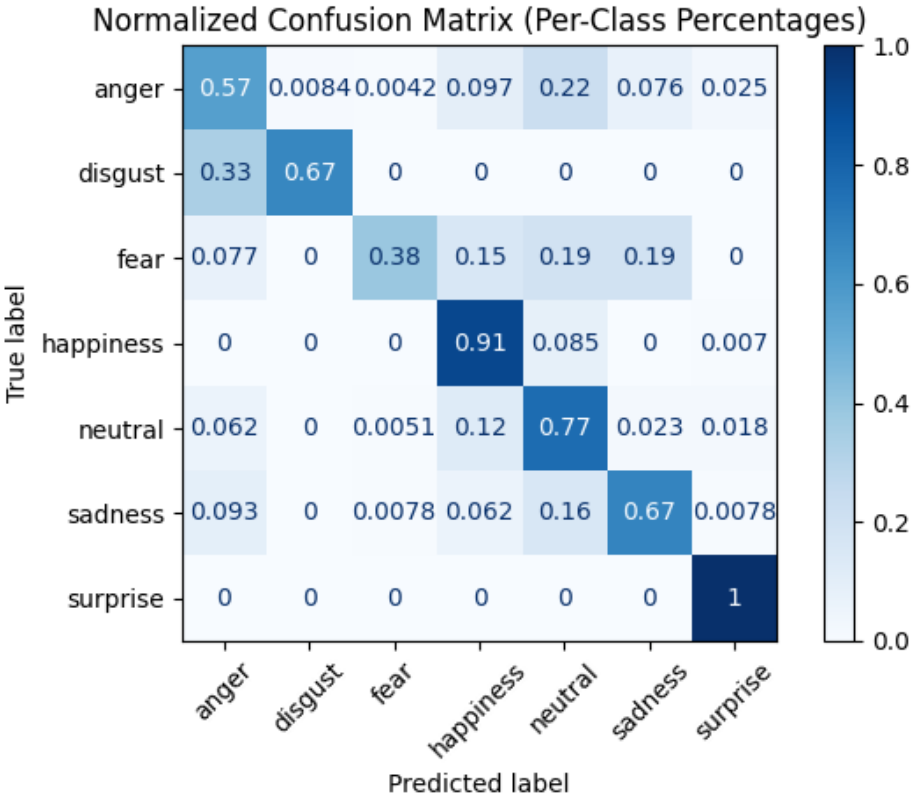
Another common mistake was the model predicting 'happiness' for sentences that were actually 'neutral'. This might be because the sentence used positive sounding words, even if

it wasn't really expressing 'happiness'. The model also confused 'sadness' with 'neutral', and sometimes even 'anger' with 'happiness'.

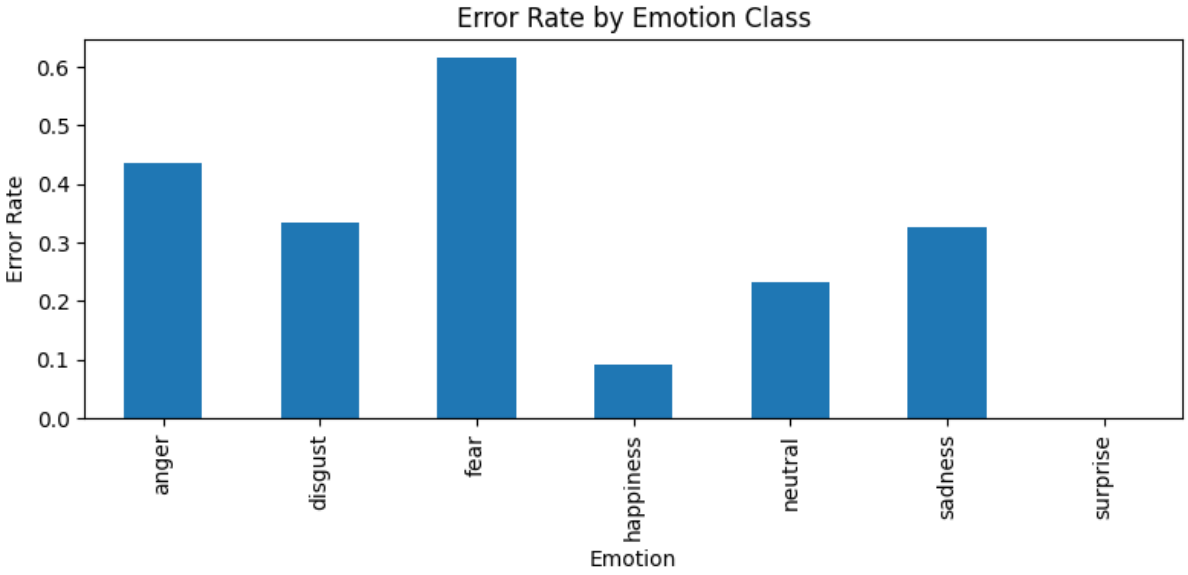


5. Normalized Confusion Matrix and Error Rate per Class

We also looked at the normalized confusion matrix, which shows the percentage of correct and incorrect predictions for each emotion. This helped us understand how well the model performs per class, even when the number of examples is different for each emotion. The results showed that the model was most accurate with 'happiness' and 'neutral', meaning it got a high percentage of those right. On the other hand, emotions like 'fear', 'disgust', and anger had much lower accuracy. These were often confused with other emotions or labelled as 'neutral'.

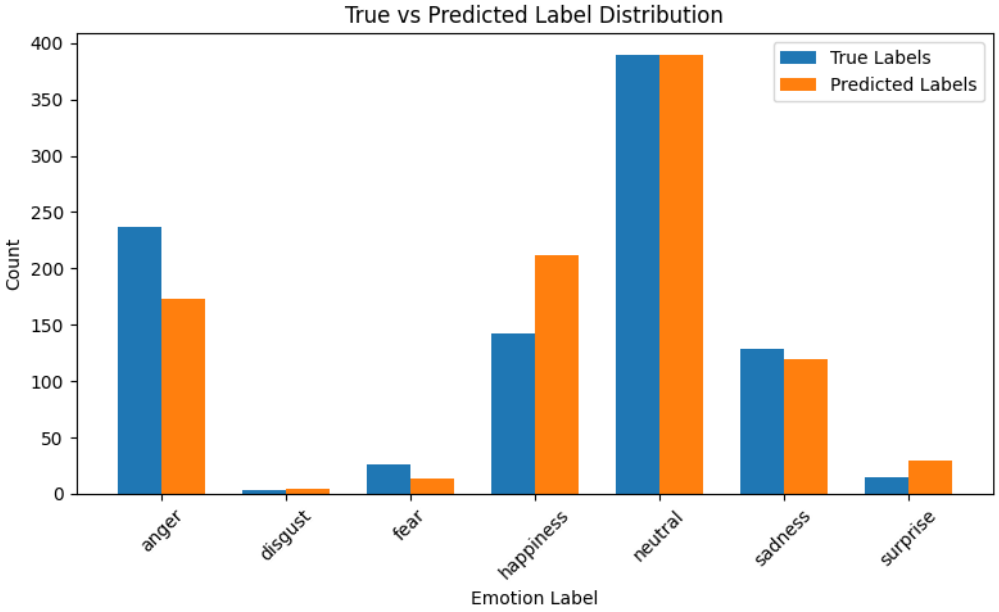


We also plotted the error rate for each class, which showed how often the model got each emotion wrong. Again, ‘fear’ and ‘anger’ had the highest error rates. This tells us that the model struggles the most with these less common or subtler emotions. These insights confirm what we saw earlier in the confusion matrix and show that even with class weighting, some emotions are still much harder for the model to handle correctly.



6. Class Distribution Bias

A class distribution plot comparing true and predicted labels showed clear over prediction of ‘neutral’ and ‘happiness’, with ‘disgust’, ‘surprise’, and ‘fear’ being underrepresented in predictions. This reflects class imbalance in the dataset and indicates that the model may be biased toward dominant classes, leading to poor generalization for less frequent emotions.
In such setting, the model likely favours ‘safe’ predictions {like ‘neutral’) unless it finds strong signals of negative or rare emotions. This limits its effectiveness in detecting subtle but important emotional states.



7. Sentence-Level Error Examples

To better understand where the model struggles, we manually reviewed a set of incorrectly predicted sentences for different emotions. Below are the most common patterns we noticed, based on emotion-specific examples.

1. Anger – Often Predicted as 'Neutral' or 'Happiness'

The model often failed to detect 'anger' when it was expressed in a calm or indirect way. Many of these sentences lacked strong emotional words, which may have led the model to label them as 'neutral' or even positive.

Examples:

- "She talked Mom not to buy me clothes" – Predicted: Neutral
- "Velina just doesn't know the measure" – Predicted: Neutral
- "She must take responsibility for her actions." – Predicted: Happiness
- "Come on, we're leaving" – Predicted: Neutral

Short commands or emotionally subtle sentences were frequently labelled as 'neutral'. In some cases, sarcastic or frustrated tones were completely missed.

2. Neutral – Often Predicted as Happiness or Surprise

The model sometimes interpreted polite or hopeful sentences as positive emotions. This happened even when the sentence was factual or emotionally flat.

Examples:

- "However, she is a mother." – Predicted: Happiness
- "All teenagers are like that" – Predicted: Happiness
- "The next morning" – Predicted: Anger
- "But I hope things will work out in the future" – Predicted: Happiness

This suggests the model may rely too much on surface-level words like "hope," mistaking neutral phrasing for emotional expression.

3. Fear – Spread Across Happiness, Sadness, Neutral, Anger

Fear was one of the hardest emotions for the model to get right. It was confused with many other emotions, often depending on sentence phrasing and urgency

Examples:

- "They kill the children. Please" – Predicted: Anger
- "What my passion, not to lose Theo" – Predicted: Happiness
- "Well now if something bad happened to my daughter" – Predicted: Sadness
- "You have to do something. It has become indispensable" – Predicted: Happiness

Fear is often expressed in indirect or vague ways, which may be difficult for the model to pick up, especially without strong fear-related keywords.

4. **Sadness – Sometimes Predicted as Neutral or Happiness**
   Sadness was often missed when it was expressed through regret or subtle disappointment instead explicit emotional terms.
   Examples:
   - "If I had money…" – Predicted: Neutral
   - "Not everything can be bought with money and gifts." – Predicted: Happiness
   - "Ever since Velina's father abandoned her family 6 years ago, the girl has been constantly pampered by her mother." – Predicted: Happiness

   The model may have focused more on content or keywords, rather than recognizing emotional tone from overall sentence structure.

5. **Happiness – Sometimes Predicted as Neutral**
   Happiness was usually misclassified when the sentence was short, factual, or didn't include emotional language, even if the overall tone was positive.
   Examples:
   - "Velina falls into real euphoria" – Predicted: Neutral
   - "You have some idea to be together" – Predicted: Neutral
   - "Velina and Nicholas no longer quarrel" – Predicted: Neutral

   The model may struggle to recognize low-intensity or implied happiness when it's not clearly stated.

8. Recommendations & Reflections

Based on the results of this error analysis, we identified several ways the model could be improved in future versions.

First, although we used class weights to handle imbalanced data, the model still struggled with rare emotions like 'fear', 'disgust', and 'surprise'. Adding more labelled examples for these emotions or using data augmentation techniques could help the model learn to recognize them better. Another idea is to include contextual features such as sentence position in a dialogue or surrounding sentences, since many emotions are expressed indirectly and depend on what was said before or after.

The model also often missed emotions when they were subtle or not directly stated. This suggests that relying only on sentence-level embeddings may not be enough. Using additional features like sentiment intensity scores and negation handling could make the model more sensitive to the meaning behind the words.

From this process, we also learned how important good labelling is. The original test set had many incorrect or unclear labels, which led to a big gap between validation and test performance. After fixing the test set, the results became much more reliable and useful. This showed us how much the quality of the data affects the outcome.

Overall, this error analysis helped us understand not just what the model got wrong, but also why. It gave us clearer ideas on what to improve next and how to make emotion classification more accurate and fairer across all emotion categories.

9. Conclusion

This error analysis helped us better understand the strengths and weaknesses of our Transformer-based emotion classification model. While the model performed well on common emotions like 'happiness' and 'neutral', it struggled with less frequent and subtler emotions such as 'fear', 'disgust', and 'sadness'. By reviewing misclassified examples and

analysing confusion patterns, we saw that the model often failed when the emotion was not clearly expressed or when the language was vague or polite.

Correcting the test set labels also made a big difference. It improved the quality of the evaluation and gave us more trust in the results. Through this process, we learned how important data quality, context, and balance are for training reliable models. The insights we gained will help guide improvements in future versions, both in model design and in dataset preparation.