Error Analysis

Notebook

1. Introduction

Error analysis is an important step in any NLP project. It helps identify the limitations of models and offers insights for improvement. Before I begin the analysis, I noticed some mislabelling in total of 48 in certain columns of the test set. This is expected since the output from the agency is used as the ground truth, and it would be nearly impossible to develop a model that labels all emotions with 100% accuracy. Moreover, working with Arabic requires careful analysis and a deep understanding of its linguistic nuances, making it crucial to analyse errors.

2. Data Preparation and Prediction Generation

• Dataset Loading & Label Mapping:

We loaded the test dataset from the file “Test_label_final_preprocessed.csv” and mapped the provided emotion labels (neutral, anger, sadness, happiness, surprise, fear, disgust) to the correct numerical values.

• Preprocessing & Tokenization:

The text was pre-processed to ensure consistency (e.g., renaming columns and handling missing values) and then tokenized using the same tokenizer employed during training. This consistency ensures that the model interprets the input text in the same manner as during the training phase.

• Model Predictions:

The trained model was used to predict on the pre-processed test data the (agency dataset) . The output logits were converted to class labels using an argmax operation, providing both the ground truth labels and the model’s predictions for subsequent analysis.

3. Quantitative Error Analysis

• Classification Report:

We computed key performance metrics, including precision, recall, weighted F1-score, and accuracy for each emotion class.

The classification report showed us an overall accuracy of 75.76%, with the neutral class performing very well (precision 0.8210, recall 0.9042, F1-score 0.8606). However, classes such as anger (F1 0.3000), happiness (F1 0.2462), fear (F1 0.2667), and disgust (F1 0.0000) showed poor performance, indicating that the model struggles with distinguishing these emotions accurately.

```
Classification Report:
              precision    recall  f1-score   support

     neutral     0.8210    0.9042    0.8606       355
       anger     0.3333    0.2727    0.3000        22
     sadness     0.5625    0.3750    0.4500        24
```

```
   happiness     0.2581    0.2353    0.2462        34
    surprise     1.0000    0.3636    0.5333        11
        fear     1.0000    0.1538    0.2667        13
     disgust     0.0000    0.0000    0.0000         3

    accuracy                         0.7576       462
   macro avg     0.5678    0.3292    0.3795       462
weighted avg     0.7469    0.7576    0.7373       462
```

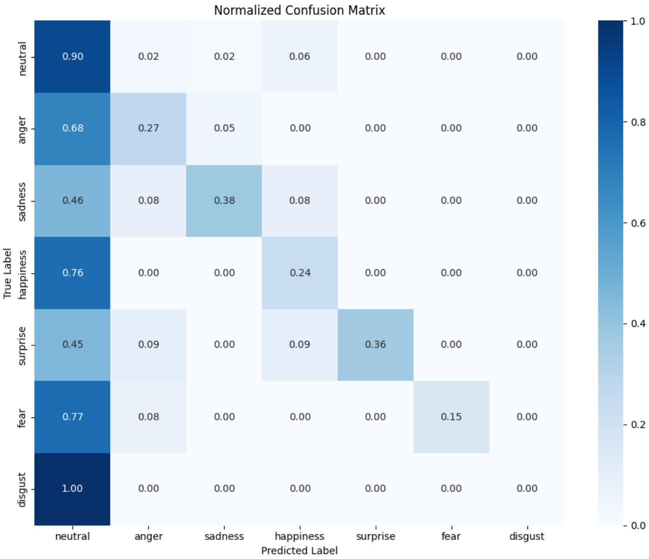
Confusion Matrices:

We generated both raw and normalized confusion matrices.

The raw confusion matrix shows the absolute number of misclassifications per class, while the normalized version provides the relative error rate per true class.

These matrices helped us identify specific overlaps, such as many happiness samples being predicted as neutral, which highlights areas where the model's decision boundaries might be blurred.

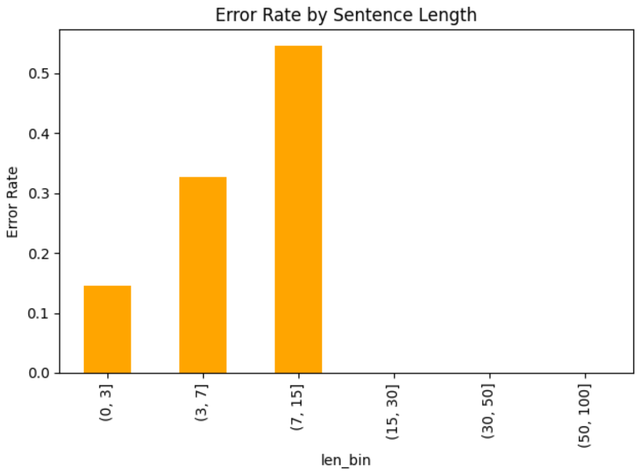
From what we noticed, is that the second predicted class for the emotion aligns with the true value, which is a strong indication of to which level is the model biased here



	neutral	anger	sadness	happiness	surprise	fear	disgust
neutral	0.90	0.02	0.02	0.06	0.00	0.00	0.00
anger	0.68	0.27	0.05	0.00	0.00	0.00	0.00
sadness	0.46	0.08	0.38	0.08	0.00	0.00	0.00
happiness	0.76	0.00	0.00	0.24	0.00	0.00	0.00
surprise	0.45	0.09	0.00	0.09	0.36	0.00	0.00
fear	0.77	0.08	0.00	0.00	0.00	0.15	0.00
disgust	1.00	0.00	0.00	0.00	0.00	0.00	0.00

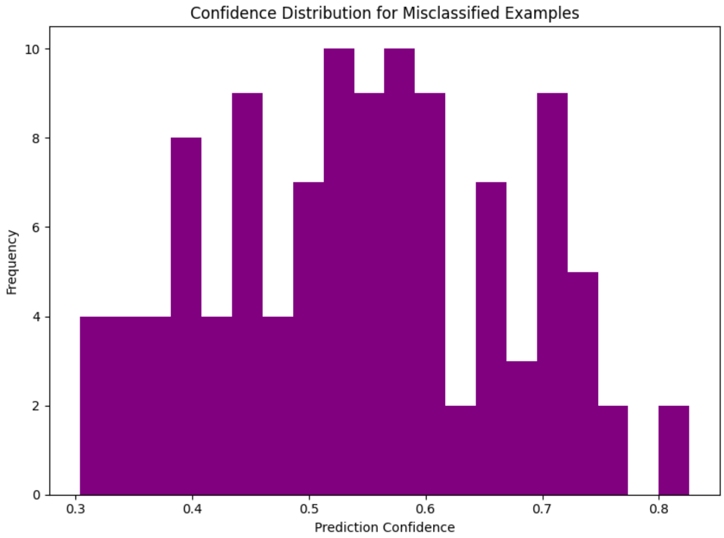
Error rate by sentence length:

Another fun fact we noticed is that the longer a sentence is, the more likely it is to be misclassified. Based on our analysis, sentences with 0–3 tokens have the lowest error rate at 0.15%, sentences with 3–7 tokens have an error rate of 0.35%, and sentences between 7 and 15 tokens reach an error rate of 0.55%.



Misclassified Examples:

Another interesting observation is that many of the model's misclassifications were made with **moderate to high confidence**. Based on the chart, most wrong predictions had confidence levels between **0.5 and 0.7**, and a few even went beyond **0.7**. This shows that the model was often quite sure about its incorrect guesses, which can be risky. It suggests that the model might be **overconfident**, especially when dealing with certain dominant classes like "neutral." In contrast, there were only a few misclassifications when the model was less confident (below 0.4), meaning the model usually knows when it's unsure — but when it's wrong, it's often *very* sure.



Advanced N-gram Analysis in Misclassified Examples:

Another interesting insight we found is from the n-gram analysis of the misclassified samples. Some phrases like **"من كده"**, **"أكبر من كده"**, and **"العفو عند المغدرة"** showed up multiple times in wrong predictions. This probably means the model is getting confused by **common or emotional expressions** that could mean different things depending on the context.

For example, **"أكبر من كده"** could be used in happy, sad, or even sarcastic ways. So even though the model might recognize the words, it struggles to understand **what the person is really feeling**. Also, phrases like **"بعلاقة مع زوجها"** or **"زي ما"** are too vague or need more context, which makes them tricky.

Based on the results we had, we decided to perform qualitative error analysis and to understand more in depth the reason behind some misclassified classes.

4. Qualitative Error Analysis
   • Misclassified Samples Extraction:

   We extracted all misclassified samples from the test set (112 in total) for manual inspection. Reviewing these examples enabled us to detect patterns in un-clear language and subtle emotional cues that might be leading to errors.
   • Word Frequency Analysis:

   An analysis of the misclassified samples revealed that common words such as "ما", "من", "يعني", and "على" frequently appear. These filler words may dilute the emotional signal, especially in shorter texts, making it harder for the model to correctly infer the sentiment.
   • Advanced N-gram Analysis:

   Bigrams and trigrams were generated from the misclassified texts. The most common bigrams included "مع زوجها", "من كده," and "العفو عند," while the top trigrams included "العفو عند المغدرة" and "أكبر من كده."

   These recurring phrases suggest that certain linguistic patterns might be contributing to systematic misclassifications, potentially confusing the model between similar emotional states.
   • Per-Class Error Breakdown and Detailed Commentary:

   We grouped misclassifications by true label to quantify the errors per class.

   For example, for the anger class, 16 errors were observed, with 15 of them being misclassified as neutral and one as sadness. A sample misclassified text for anger was:
   "هون مشيرة شكت بصديقتها ويمكن اتهمتها ضملا وأعلنا أنه مشاعر على العلاقة مع زوجها"
   Here, the subtle cues for anger might be overshadowed by generic negative sentiment, leading the model to default to neutral.

   In the happiness class, all 26 errors were misclassified as neutral, indicating that subtle expressions of joy without explicit positive markers (e.g., "تعالي والله" vs. a clear positive statement) are not being captured effectively.

Similar analyses for fear, disgust, sadness, and surprise revealed that most errors involve a default prediction of neutral, which may suggest that when the model is uncertain or the text lacks strong emotional indicators, it leans toward a neutral prediction.

These findings, along with the error distributions, highlight overlapping cues and ambiguous language as key factors contributing to misclassifications.

5. 5. Our Interpretation and Recommendations

After looking closely at the misclassified examples and going over all the confusion matrices and prediction patterns, a few things became very clear.

• Neutral Takes Over

One of the most obvious things we noticed is how much the model leans toward predicting neutral. Even in examples where the emotion was clearly something else like anger or happiness, the model would often still go with neutral. What's interesting is that when you check the logits, the true label was often the second-highest prediction. This shows that the model actually sees the emotion, but it plays it safe and chooses neutral, probably because it has seen so many neutral samples during training.

• Short Sentences, Fewer Mistakes

A really fun thing we found was the relationship between sentence length and error rate. Sentences with 0–3 tokens had the lowest error (just 0.15%), while those with 7–15 tokens had the highest (up to 0.55%). This makes sense because short sentences often contain very clear expressions, while longer ones might have more filler or mixed emotions that confuse the model.

• Overlapping Emotions and Vague Language

Many of the errors came from sentences that were either vague or contained emotional expressions that could belong to more than one class. For example, phrases like “أكبر من كده” or “العفو عند المغدرة” showed up in anger, sadness, and even in neutral samples. The model isn't wrong for getting confused—it's just that these expressions are too dependent on context. This explains why we often saw anger being predicted as sadness or happiness as neutral.

• Confidence Doesn't Always Mean Correct

Another surprising insight is that the model is often confident even when it's wrong. A lot of misclassifications had logits between 0.5 and 0.7—and some even higher. That means the model is confidently making the wrong decision. But when the model is unsure, the confidence is much lower (under 0.4), which shows it knows when it's guessing. That's a good sign overall, but the high-confidence mistakes are something to watch out for.

6. Conclusion

This comprehensive error analysis provides both quantitative metrics and qualitative insights into our emotion classification model. While the model demonstrates strong performance on the neutral class, it struggles with accurately identifying emotions such as anger, happiness, and fear due to overlapping linguistic cues and un-clear expressions. The analysis highlights specific areas for improvement, such as data augmentation and feature engineering, and suggests that a balanced F1-score is critical for this task, knowing that we used here the weighted F1. These insights will be integrated into our model card and serve as the basis for our recommendations to the client, ensuring that future enhancements are both targeted and effective.