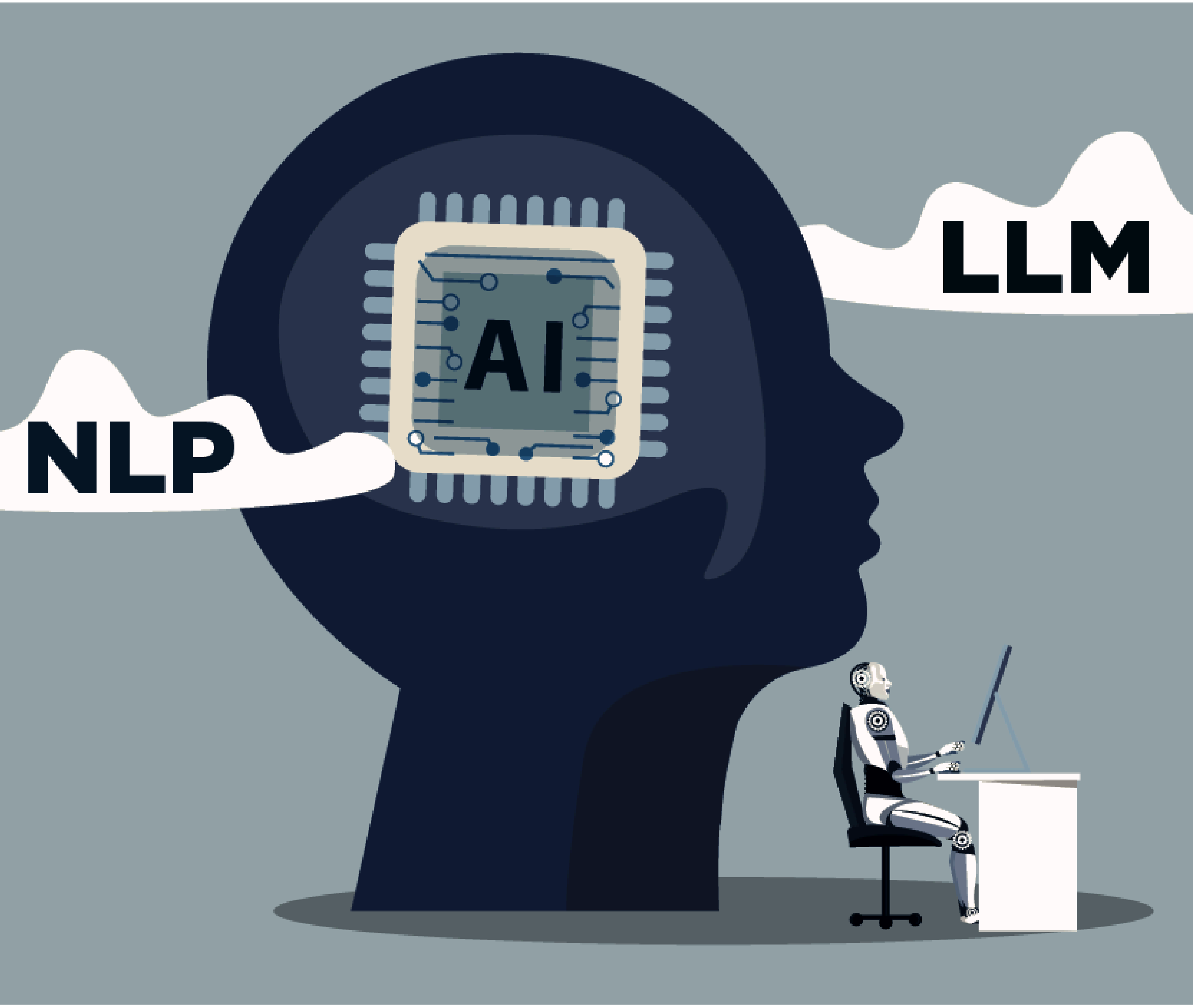
ERROR ANALYSIS

BREDA UNIVERSITY OF APPLIED SCIENCES



Presented by:

Jonas Vos and Ayumi Chotoe

Introduction

In natural language processing (NLP), error analysis is essential for uncovering the limitations of a model and identifying areas for future improvement. For emotion classification tasks, where inputs are linguistic and labels are subjective emotional states, understanding prediction errors can reveal valuable insights into how a model processes language and why it may fail to correctly classify certain emotional tones.

This report presents an error analysis of our best-performing emotion classification model. The analysis includes both quantitative evaluations (e.g., confusion matrix, F1 scores) and qualitative linguistic insights to better understand systematic errors. Our findings will inform the final recommendation to the client on how to best use and potentially improve the model.

Model Overview

BERTje, a Dutch pre-trained transformer model developed by the University of Groningen, was used for the classification task. Its performance was evaluated by comparing the model's predictions, stored in the "predicted_emotion" column of the CSV file, to the "true" emotion labels provided in the "main_emotion" column. The true labels were originally created by the Content Intelligence Agency and carefully reviewed by students for accuracy. This error analysis explores where and why the model made incorrect predictions and offers recommendations for future improvements.

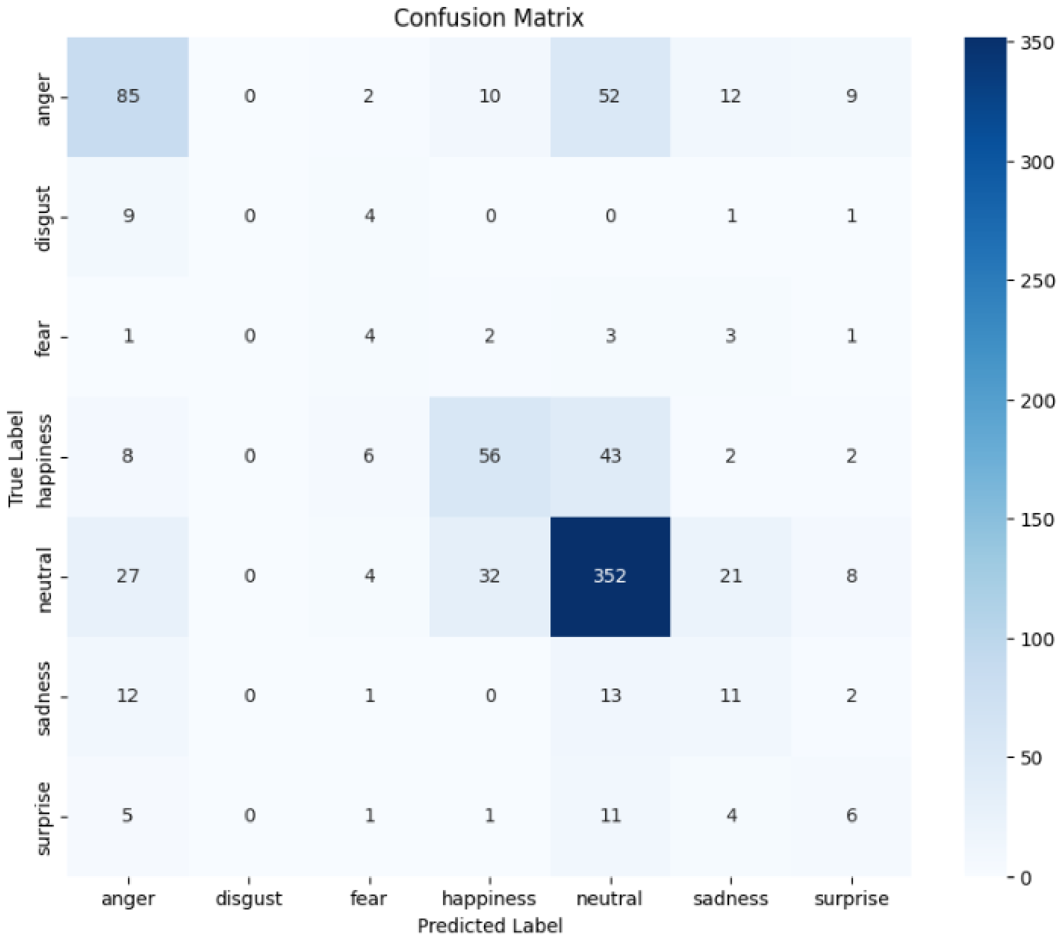
The CSV file contains each sentence with its corresponding ground truth emotion and the predicted emotion. In addition to the confusion matrices generated from this data, the analysis also delved into specific linguistic characteristics of the sentences, including sentence length, specific word types (such as adjectives, adverbs, and figurative language), and the presence of borderline and ambiguous examples. Borderline examples refer to sentences that sit between two or more emotional categories, making them difficult to classify even for humans. Ambiguous examples include sentences with vague wording, mixed emotional cues, or irony and sarcasm that may confuse the model.

The evaluated model achieved an overall accuracy of **62.2%**, with relatively strong performance on the dominant class (*neutral*) and significantly weaker performance on minority classes (*fear*, *disgust*, and *surprise*).

Confusion Matrix Analysis

The confusion matrix (Figure 1) provides an overview of where the model tends to misclassify certain emotions. The matrix is heavily influenced toward the *neutral* class, with many non-neutral emotions being misclassified as *neutral*. Below are key findings:

Figure 1: Confusion Matrix of Predicted vs. True Emotions



Neutral (444 samples): The model performs well here (F1-score = 0.77), correctly classifying the majority of *neutral* sentences. However, it also over-predicts this class, suggesting a bias toward neutrality.

Anger (170 samples): Frequently misclassified as *neutral,* with a recall of 50&, indicating medium recognition of negative emotions.

Happiness (117 samples): Often confused with *neutral*, possibly due to subtle differences in tone not captured by the model.

Fear, Disgust, Surprise: These classes suffer from both low precision and low recall, suggesting the model fails to recognize them consistently.

Disgust (15 samples): Received no correct predictions at all, with an F1-score of 0.00.

Quantitative Metrics Summary

Emotion	Precision	Recall	F1-Score	Support
Neutral	0.74	0.79	0.77	444
Happiness	0.55	0.48	0.51	117
Anger	0.58	0.50	0.54	170
Sadness	0.20	0.28	0.24	39
Surprise	0.21	0.21	0.21	28
Fear	0.18	0.29	0.22	14
Disgust	0.00	0.00	0.00	15

Qualitative Error Analysis

Sentence Length

Longer sentences with multiple clauses were more likely to be misclassified. This suggests the model struggles with contextual dependencies across longer inputs, especially when emotional cues are buried within more neutral statements.

Example:

- *True: Anger*, *Predicted: Neutral*
- *"Als ze van de trap afvalt, zou ik het niet erg vinden."* The model misses the sarcastic and hostile undertone, classifying it as neutral due to the sentence structure.

Lexical Ambiguity and Sarcasm

Words that change meaning based on tone or context often lead to misclassification. Sarcastic or ironic expressions, which are inherently hard for models to detect, were frequently mislabeled as *neutral*.

Emotionally Charged Verbs

Sentences containing verbs that express emotion (e.g., *haten*, *houden van*, *verliezen*) tend to be classified more accurately, suggesting that the model relies heavily on explicit emotional indicators.

Word Types and Intensity

Sentences labeled *anger* or *disgust* often included exclamations, interjections, or informal speech patterns. These were typically classified as *neutral* unless explicit hate speech or strong negative adjectives were present.

Example:

- *True: Disgust*, *Predicted: Neutral*
- *"Wat een smerige vertoning!"* Despite being an expressive statement, the lack of training examples for this class likely led to a neutral prediction.

Class-Specific Analysis

- **Anger**: Many *anger* instances were toned down and expressed indirectly. The model often missed these subtleties and classified them as *neutral* or *happiness*.
- **Happiness**: Misclassifications typically happened when positive sentiments were embedded in otherwise factual or neutral statements.
- **Disgust**: The model appears entirely incapable of detecting this class. A likely cause is the very low number of training examples (15), making generalization nearly impossible.

Evaluation Priorities: Accuracy vs F1

Given the **imbalance in class distribution** and the importance of correctly identifying less frequent but emotionally intense classes (e.g., *anger*, *fear*, *disgust*), **F1-score** is a more meaningful metric than accuracy. Optimizing for F1 would help mitigate the model's bias toward the *neutral* class and promote balanced performance.

Recommendations to the Client

1. **Augment Underrepresented Classes**: Increase the number of labeled examples for *disgust*, *fear*, and *surprise* to help the model learn their features.
2. **Handle Sarcasm and Indirect Emotion**: Consider using sentiment-aware or context-enhanced models (e.g., BERT with attention to discourse markers).
3. **Refine Data Preprocessing**: Incorporate syntactic features or discourse cues that might aid in detecting emotions spread across longer text.
4. **Post-processing Heuristics**: Implement a confidence threshold mechanism to flag uncertain predictions for human review, especially in sensitive applications.

Conclusion

Our model shows competent performance in detecting *neutral* and to a lesser extent *happiness* and *anger*. However, it struggles significantly with underrepresented and context-heavy emotions like *disgust* and *fear*. Through both quantitative metrics and linguistic analysis, we found that model bias, lexical ambiguity, sentence length, and training data limitations all contribute to these shortcomings. By addressing these weaknesses with targeted data improvements and architecture adjustments, the model can be refined to better serve its intended purpose in real-world emotion classification tasks.