1. Introduction

Understanding how emotion classification models make decisions is crucial for building trust and improving performance. This report explores explainability techniques applied to a transformer-based emotion classifier, focusing on Gradient × Input, Layer-wise Relevance Propagation (LRP), and Perturbation Testing. These methods help identify which words influence the model's predictions and reveal how the model processes both obvious and subtle emotional cues. By analyzing token relevance and model confidence, we gain insights into its decision-making behavior. The goal is to ensure that the model not only performs well but also reasons in a way that is interpretable and robust across different inputs.

2. Gradient × Input

- This method was used to determine the relevance of each token by calculating the gradient of the output with respect to each input token.
- The resulting relevance scores were used to generate a bar graph where the x-axis represents tokens and the y-axis shows their relevance score.
- As seen in the examples, the model focuses on highly emotional words that serve as strong indicators of the emotion being conveyed in the sentence.
- These examples highlight how transformer models can detect emotional meaning even in complex subword combinations, such as *unloved* and *obnoxious*.

3. Layer-wise Relevance Propagation (LRP)

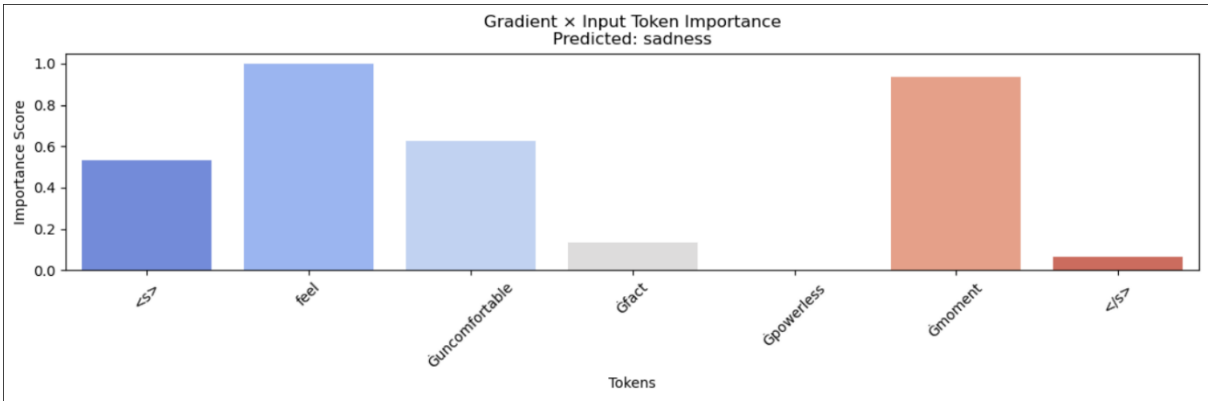
- LRP helps us understand how much each word contributes to the model's decision regarding the emotion in a sentence.
- Unlike Gradient × Input, which measures how small changes in input affect the output, LRP traces the model's final decision backward, distributing relevance scores across all tokens in a balanced way.
- In one example, words like *feel*, *like*, *vital*, and *game* were the most important for predicting the emotion "happiness", indicating that the model focused its attention on them.
- Overall, LRP provides more consistent and interpretable explanations than gradient-based methods, which can sometimes highlight irrelevant or unclear parts of a sentence.

4. Perturbation Test

- To evaluate the reliability of the model's predictions, we applied a perturbation test by removing words one at a time and observing how the model's confidence changed.
- This test helps us determine whether the model relies too heavily on a few emotionally obvious words or whether it takes the overall sentence context into account.
- A gradual drop in confidence suggests that the model uses distributed information across multiple tokens, while a sharp drop indicates an over-reliance on specific keywords.
- This insight is valuable in understanding whether the model is capturing subtle emotional cues or simply keyword spotting.

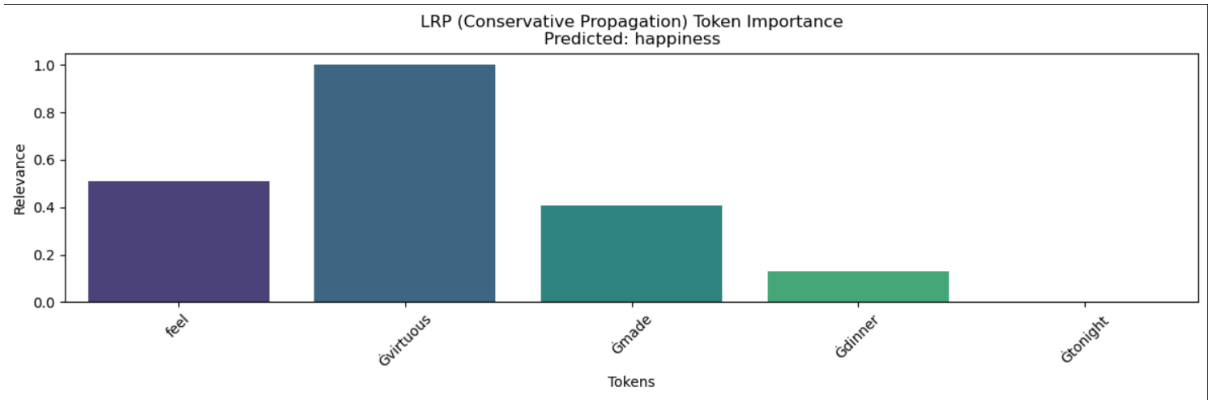
Gradient × Input: Results

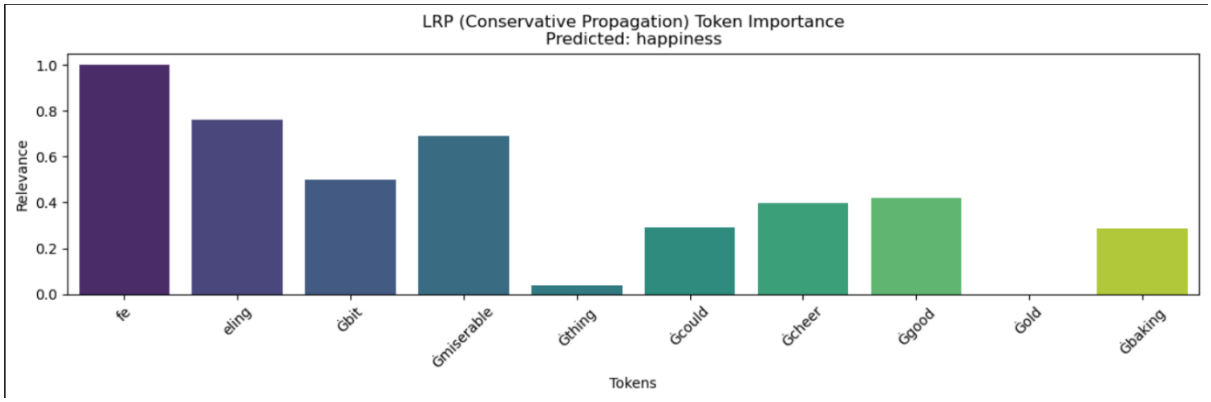
- This method revealed that the model assigns higher relevance to emotionally charged tokens in sentences.
- For example, words like *loved* or *Uncomfortable* received high relevance scores, indicating the model's strong response to key emotional indicators.
- This suggests that the model is capable of identifying meaningful features for emotion classification.



Layer-wise Relevance Propagation (LRP):results

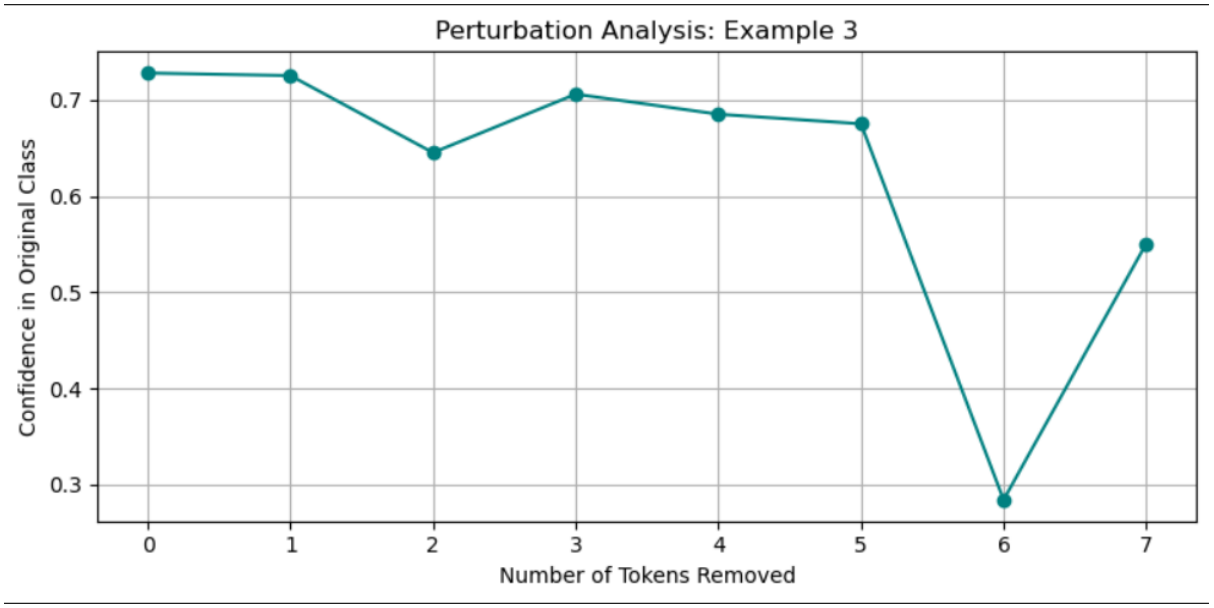
- LRP gave a more balanced view of the important tokens, often spreading relevance across several words instead of focusing solely on one.
- This indicates a more context-aware approach by the model.
- However, in longer sentences, interpretation can become more difficult, as many tokens receive moderate to high relevance scores. This may be due to the model capturing subtle contextual cues that are not always obvious to human readers.





Perturbation Test: Results

- In shorter, emotionally explicit sentences, the model heavily relies on a few key words. Removing these causes a significant drop in prediction confidence.
- Such sentences are deterministic—specific tokens dominate the output.
- In contrast, more subtle sentences like *"It's a good day today"* or *"I feel okay"* show a gradual decline in confidence as words are removed. These examples demonstrate that:
    - The model's decision relies on the **combination** of tokens rather than any single word.
    - Sentences are more **context-dependent**, requiring a deeper understanding of semantic relationships between words like *good* and *day*, or *feel* and *okay*.
- This gradual drop in confidence confirms that the model builds a distributed representation of emotion, which is more robust and generalizable.



Conclusion

The model demonstrates robust behavior, with a gradual decrease in confidence as tokens are removed. The explanations generated through Gradient × Input and Layer-wise Relevance Propagation (LRP) confirm that predictions are based on multiple tokens, especially emotionally significant ones.

The Perturbation Test further supports this, showing that the model does not overly depend on a single word. This robustness is desirable for real-world applications where inputs can vary due to noise or phrasing.

All these tests allow us to say the following of our transformer model The model shows reliable performance by using a distributed representation of tokens, avoiding over-reliance on single keywords. The model effectively identifies emotionally charged words and uses them to make accurate predictions.