XAI in Emotion Classification Using Transformer Models

1.Introduction

The rapid growth of Natural Language processing advancement is unnegotiable, yet when it comes to understanding the reasoning behind its predictions, it feels like a magic black box with billions of parameters that feel impossible to understand why the decision has been made, which makes them hardly to be interpreted.

Explainable artificial intelligence (XAI):

To address this challenge, Explainable artificial intelligence (XAI) provides us with great techniques to make it more understandable which part of the input influenced the model prediction. XAI helps in bridging the gap between model abilities and human understanding.

Our task (emotions classification):

In this project, we applied XAI to our fine-tuned model Transformer model (Marbert), for our goal emotion classification. The goal of the model it to predict the correct emotion based on short sentences, the labels used are as follows: Happiness, Sadness, Anger, Surprise, Fear and finally Disgust, we excluded Neutral from the XAI to focus solely on the emotional labels, knowing that Neutral was used in the training of the model.

Methods:

To achieve this, we applied three explainability methods:

- Gradient × Input
- Layer-wise Relevance propagation (LRP) via conservative Propagation
- Input perturbation

We applied these methods on 18 labelled examples (3 per emotion)

2. part 1- Gradient × Input

What is Gradient × Input?

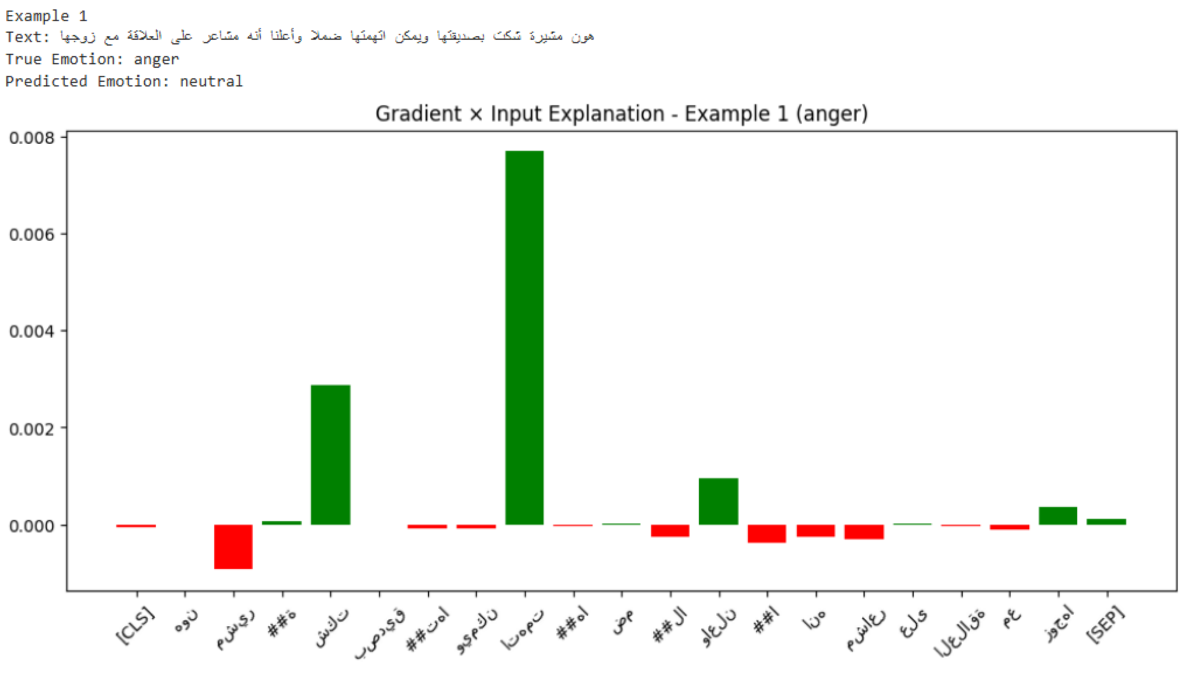
Gradient × Input is one of the simplest forward explainability techniques for interpreting deep learning models. It measures how much each input token influences the model's prediction and combines that with the token’s internal representation which is known as embeddings to highlight the most important parts. The results is a relevance score for each token, showing how much it contributed to the final prediction. A positive score means the token shared the model’s decision, while a negative score means it was against it. This technique shows us what model focuses on

How we Applied it

 We applied it to our fine-tuned model Transformer model (Marbert). The analysis was made on 18 selected test sentences, 3 for each of the 6 emotions, the emotions used are as follows happiness, sadness, anger, surprise, fear, and disgust. The inputs were tokenized from the model we saved then pass it through   the model, and for each example, we calculated the gradients and visualized the relevance scores, with a color-coded bar chart. In each chart, green represented positive contributions (tokens that helped the model to predict an emotion), while red bars indicated negative contribution.

Analysis & Findings

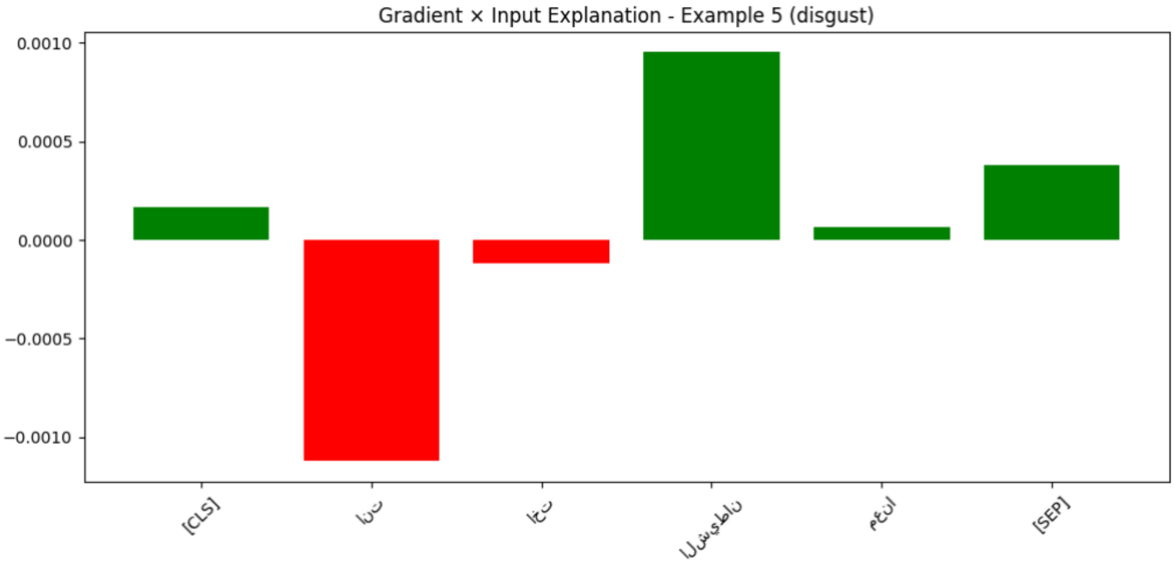
Starting with the Anger label we noticed that all the examples had the same issue strong word indication which mean that these contributed to the predictions, despite that it picked the correct tokens and the one the supported the model it sometimes misclassified for example these “شكت” and “اتهمت” should be a strong indication for anger it predicted as neutral



```
Example 1
Text: هون مشيرة شكت بصديقتها ويمكن اتهمتها ضملا وأعلنا أنه مشاعر على العلاقة مع زوجها
True Emotion: anger
Predicted Emotion: neutral
```

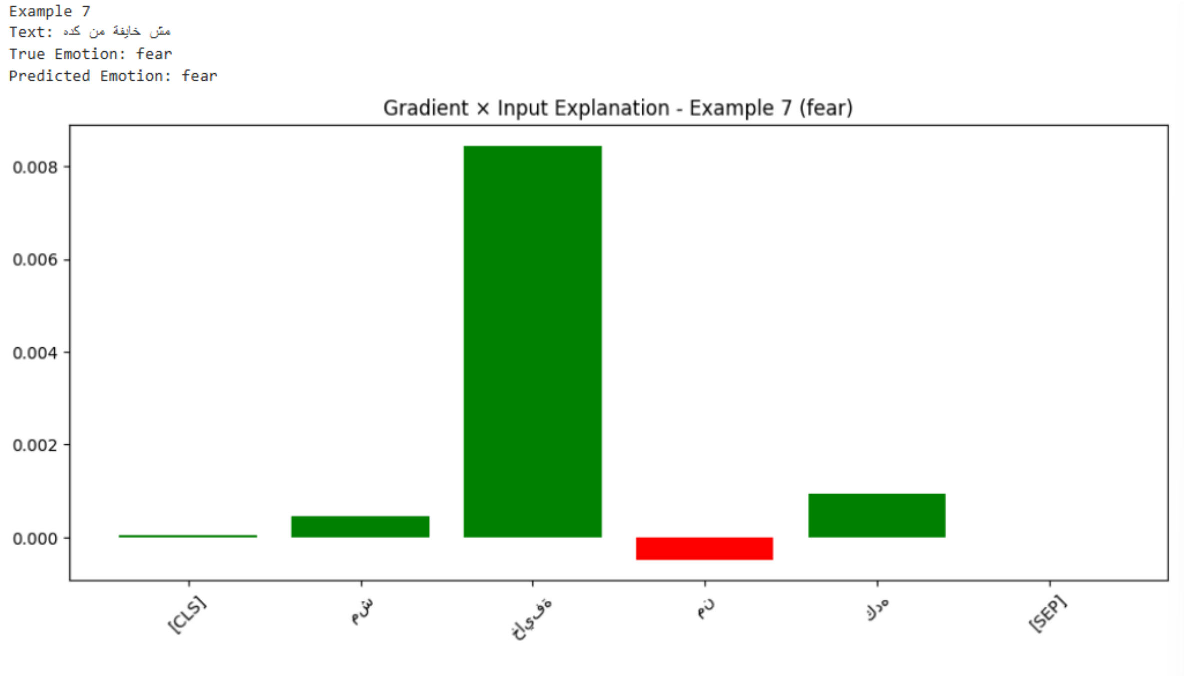
Disgust

Had an interesting look starting with our first example with “ زبالة” which means trash here it was working the model prediction which the made choose neutral as it did not help another example was Text: " أنت أخت الشيطان معنا True Emotion: disgust Predicted Emotion: neutral despite that some of the contribute towed the correct emotion other tokes affect like أنت you and أخت which translated you and sister respectively



Fear

For the fear we had a challenge due to the short sentence the tokens were not learning and subjective enough to give correct prediction and this is what we got from the one that had a somehow long sentence

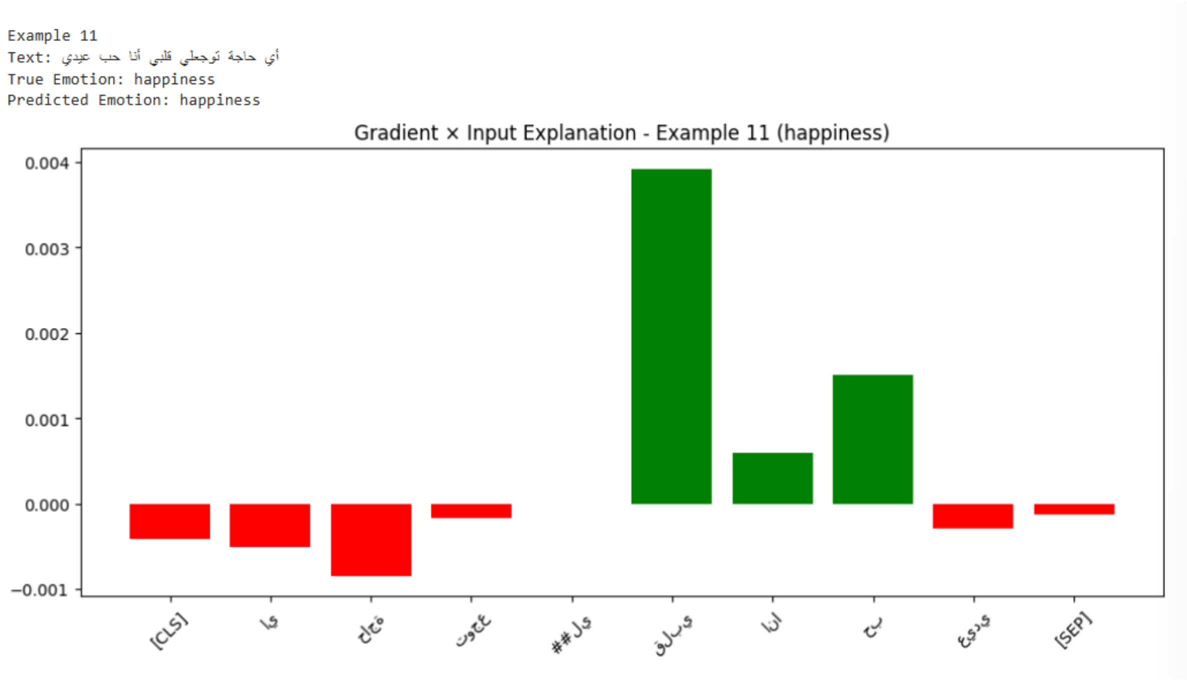


```
Example 7
Text: مش خايفة من كده
True Emotion: fear
Predicted Emotion: fear
```

Which indicated that correct the translation for each token from the left to the right is as follows (I’m not, scared, from, that) and we can see that it picked the second one the most scared

Happiness

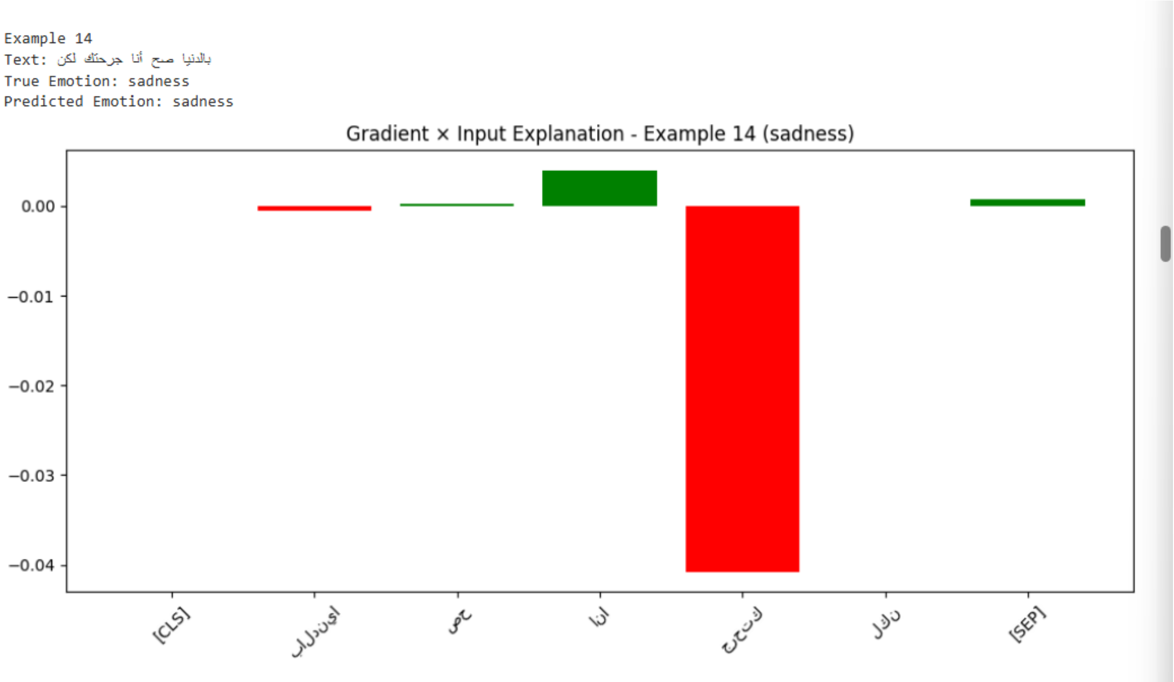
For happiness, the model’s performance depended on the clarity the emotional markers. For example, in Example 11 (‘أي حاجة توجعلي قلبي أنا حب عيدي’), the positive cues in ‘حب عيدي’ clearly drove the model to predict happiness. In contrast, in hard to distinguish cases like the religious quotation in Example 10 (‘زي ما قال الرسول عليه الصلاة والسلام’) or the brief phrase ‘خلاص أنت’ (Example 12) lacked strong positive emotional words, making it harder for the model to confidently assign a happiness label. This shows us that the model successfully detects happiness when it encounters direct signals of joy or affection but struggles when the emotional tone is not well-defined or context dependent.”



```
Example 11
Text: أي حاجة توجعلي قلبي أنا حب عيدي
True Emotion: happiness
Predicted Emotion: happiness
```

Sadness

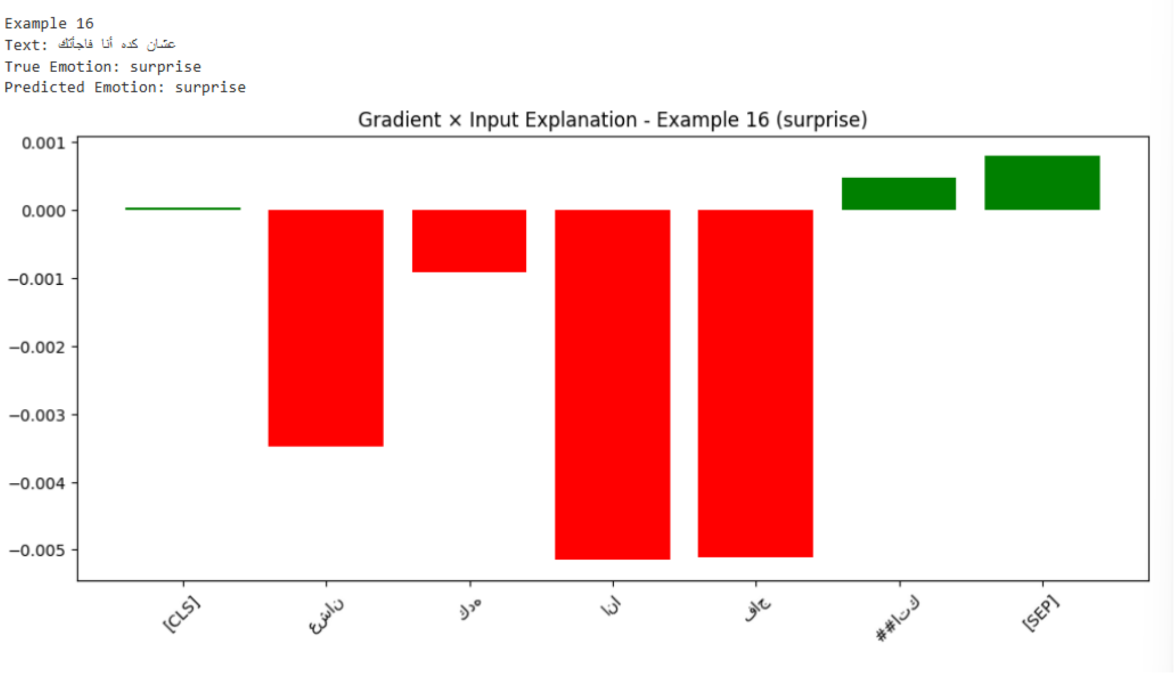
The model detects sadness well when it sees direct or clear emotional signals. For example, in Example 14 (‘بالدنيا صح أنا جرحتك لكن’), despite some unexpected token contributions, the presence of a regretful, apologetic tone allowed the model to correctly predict sadness. However, in cases like Example 15, where the context is more unclear (especially when a key word like ‘مشاعر’ turns out to be a name rather than an emotional cue), the model tends to default to a neutral prediction. This shows that while the model can pick up on overt sadness, it struggles with more nuanced or context-dependent expressions of sorrow.”



```
Example 14
Text: بالدنيا صح أنا جرحتك لكن
True Emotion: sadness
Predicted Emotion: sadness
```

Surprise

The model seems to do well when the sentence includes clear and direct signals. In Example 16 (“عشان كده أنا فاجأتك”), even though there were some token-level oddities, the word “فاجأتك” clearly guided the model to predict surprise. But in Example 17 (“يعني خبر خلافكم وصل على السودان”) and Example 18 (“اللي عملنا لك ياها”), the model predicted neutral. That’s likely because these inputs don’t contain strong emotional words or punctuation — instead, they rely on implied meaning, which the model often misses. So overall, it handles surprise well when the emotion is obvious but struggles when it's subtle or depends on context.



```
Example 16
Text: عشان كده أنا فاجأتك
True Emotion: surprise
Predicted Emotion: surprise
```

Part 2 – Improved Explanation with Conservative Propagation (LRP)

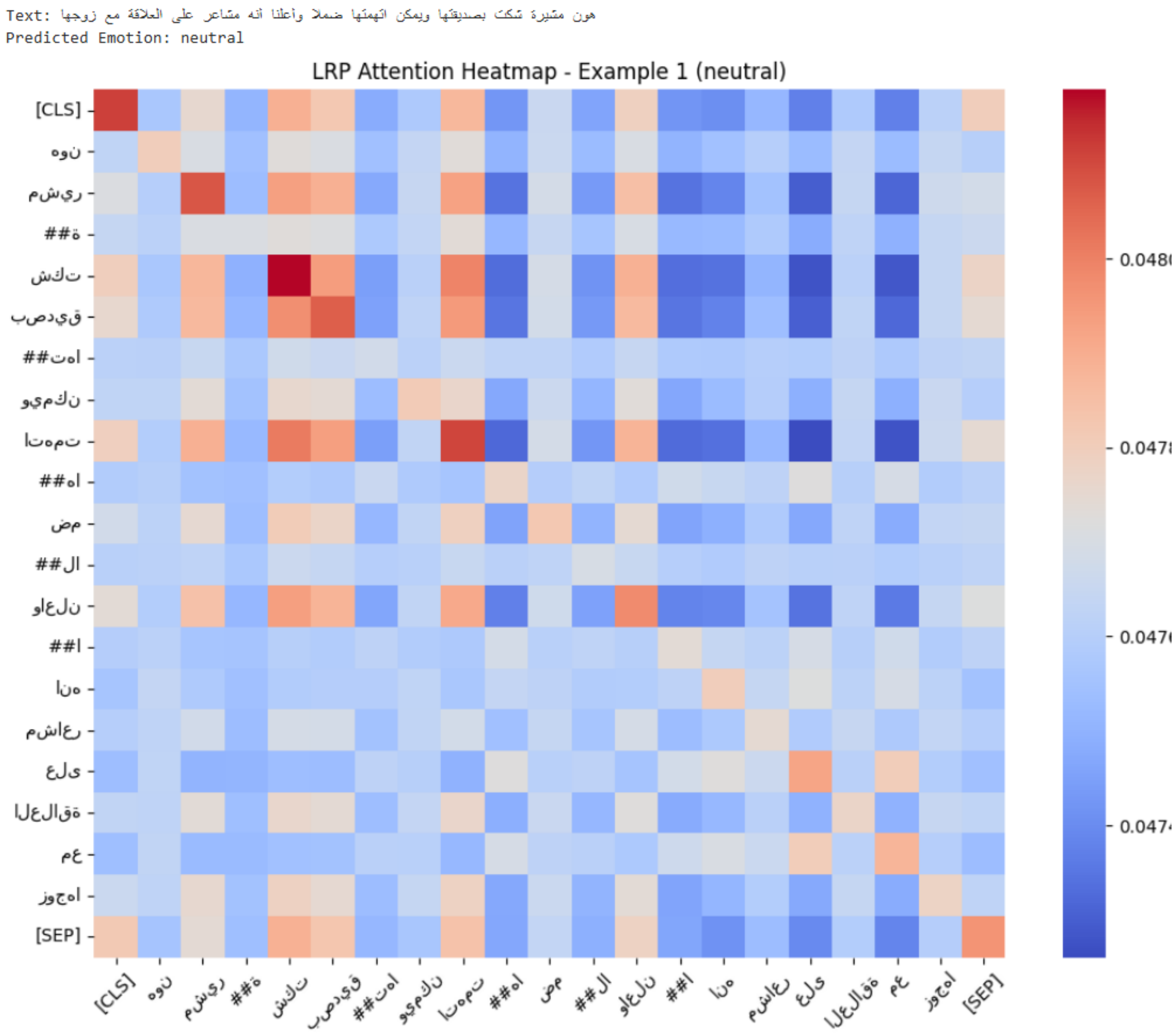
What is Conservative Propagation?

Conservative Propagation, also called Layer-wise Relevance Propagation (LRP), is a backward explainability method that builds on what Gradient × Input lacks, especially in deep models like Transformers. It works by pushing relevance scores backwards from the output, through layers like attention and normalization, while keeping the model's structure intact. In the end, it gives a clearer and more organized score for each input token, showing which parts truly influenced the model's decision.

How we Applied it

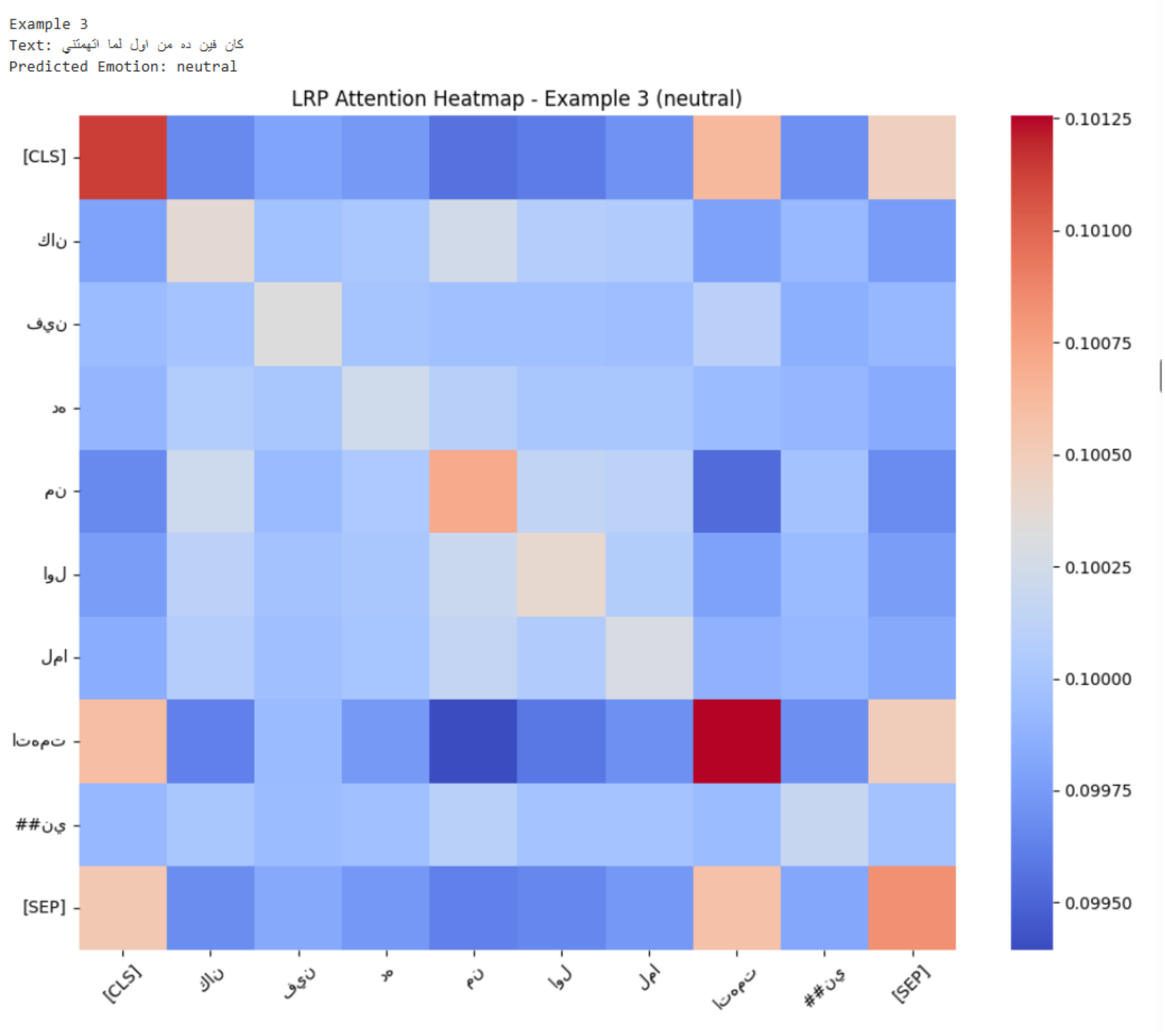
we modified the attention and layer normalization components, to support Conservative Propagation. we applied this to the same selected sentences and visualized the output using attention heatmaps, showing how relevance is distributed across tokens.

Analysis & Findings



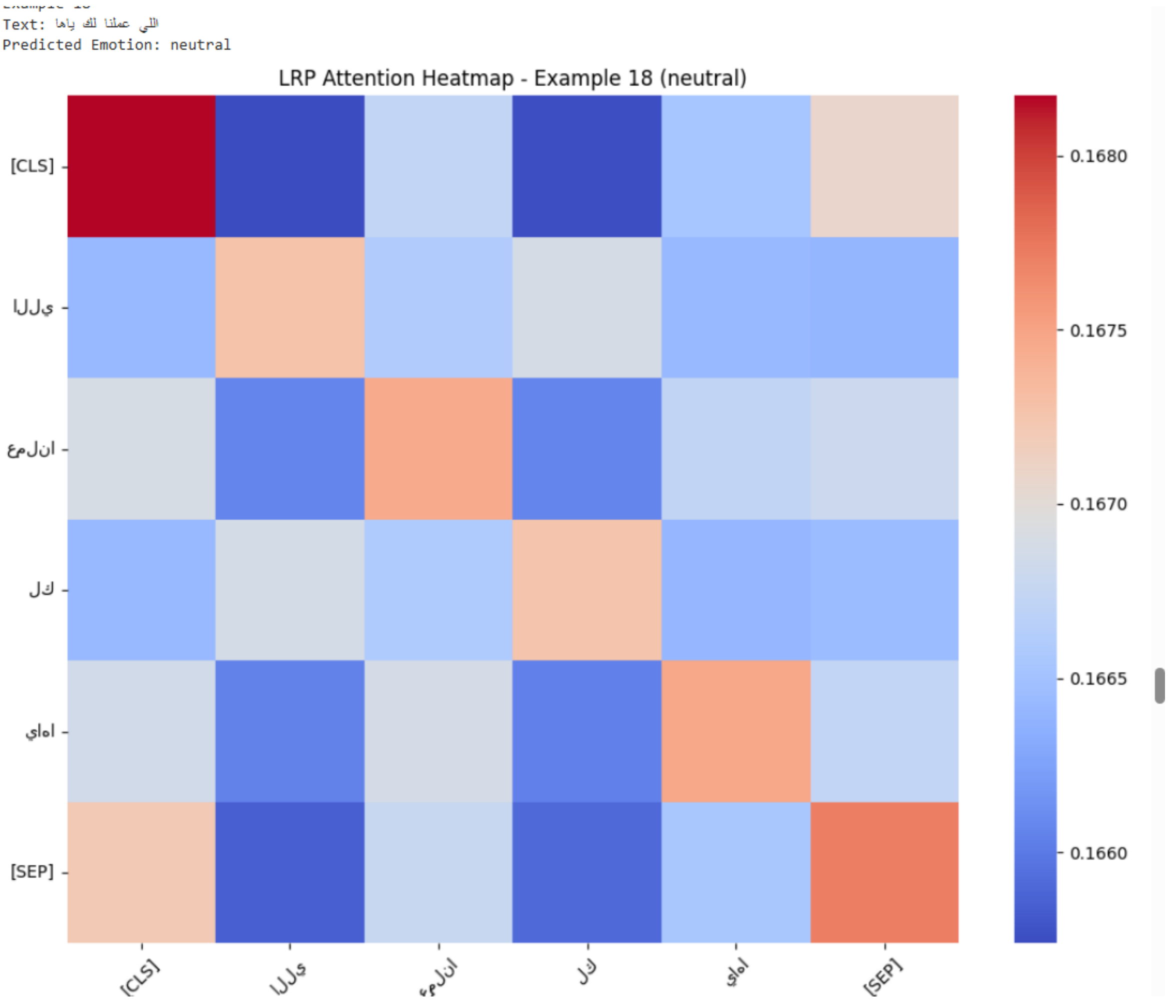
```
Text: هون مشيرة شكت بصديقتها ويمكن اتهمتها ضملا وأعلنا أنه مشاعر على العلاقة مع زوجها
Predicted Emotion: neutral
```

In this example, the LRP heatmap shows that words like “شكت” and “اتهمت” were clearly active in the sentence and interacted strongly with other tokens. But the attention is spread out across too many parts, which weakens the overall focus. That might be why the model went for neutral instead of anger even though the emotional signals were there, they just weren’t highlighted enough.



```
Example 3
Text: كان فين ده من اول لما اتهمتني
Predicted Emotion: neutral
```

In Example 3, the LRP heatmap highlights “اتهمت” as the most relevant token, meaning the model saw it as very important. But the rest of the attention was scattered, and the final prediction was still neutral. So even though the model caught an anger-related word, it didn’t give it enough weight to shift the whole decision — showing a gap between what the model notices locally and what it predicts overall.



```
Example 18
Text: اللي عملنا لك ياها
Predicted Emotion: neutral
```

In Example 18, the phrase “اللي عملنا لك ياها” could suggest surprise depending on how it's said, but the LRP heatmap shows that none of the tokens stood out as emotionally important. Most of them got similar low scores, and because there wasn’t a strong emotional signal, the model ended up predicting neutral. This shows that the model still struggles with subtle or implied emotions—especially when things like sarcasm or tone matter.

Compared to Gradient × Input, LRP offers a **more structured and realistic view** of how the model processes emotional content.
It does a better job at highlighting key emotional triggers (like “اتهمتني” and “فاجأتك”) while distributing less weight to irrelevant or noisy tokens.
This makes LRP a more reliable method for understanding **why** the model makes emotion predictions—and where it may still fall short.

Input Perturbation

What is Input Perturbation?

Input Perturbation is an explainability technique that tests how much the model relies on specific tokens. It works by gradually removing tokens from a sentence — usually the ones with the lowest relevance — and observing how the model's confidence in its prediction changes.

What we did:

In my case, we used the relevance scores from the LRP method to determine the least relevant tokens. For each of the 18 selected sentences, I removed tokens step by step and recorded the change in the model’s prediction confidence. I visualized the results using line graphs, where the X-axis shows how many tokens were removed, and the Y-axis shows the model's confidence.

What the results showed:

Example 1 – "هون مشيرة شكت بصديقتها ويمكن اتهمتها ضملا وأعلنا أنه مشاعر على العلاقة مع زوجها":
For this example, the perturbation graph indicates a relatively sharp decline in confidence once a few key tokens are removed. For instance, when tokens such as **"مشيرة"** or **"شكت"** are perturbed, the model’s confidence drops steeply. This behaviour suggests that the model is highly dependent on these specific words to signal the emotional content—if these tokens are missing, the prediction confidence falls dramatically, implying an over-reliance on a small subset of emotionally relevant cues.

Example 9 – "أن أخسرك":
Given the brevity of this sentence, the perturbation graph shows an immediate and pronounced drop in confidence upon removal of the main token. Because the sentence is very short and essentially composed of a single, critical concept, the removal of its central word causes the confidence to collapse nearly entirely. This is a clear indication that for such concise expressions, the model’s decision is entirely driven by that specific token.

Example 16 – "عشان كده أنا فاجأتك" (Predicted Emotion: surprise):
In this case, the line graph demonstrates moderate behaviour. As tokens are removed, the confidence decreases at a steady, intermediate rate—not as abruptly as in Example 9 nor as distributed as in some longer sentences. This indicates that while tokens (for example, **"فاجأتك"**) play a significant role, the prediction is also supported by other words in the sentence. Thus, the reliance is somewhat distributed, reflecting a more balanced decision-making process.

Example 10 – "زي ما قال الرسول عليه الصلاة والسلام":
The graph for this example shows a more gradual decline in confidence with incremental token removal. This gradual drop suggests that the model’s prediction does not depend solely on one or two words. Instead, the decision appears to be formed by a broader context covering the

entire sentence. A distributed reliance generally points to a more robust model behaviour, as the loss of any single token causes only a modest decrease in overall confidence.

Summary

Based on the analysis we did using the XAI techniques, we acknowledge that transformers models offer impressive results, and understanding why a predictions was made is not 100 guaranteed, yet we got a clear indication of how the predictions was made, we noticed that with clear and strong indication of an emotion like as in like “حب” which means love predicted accurately to love, the model faced challenges in more un-clear emotion with lead then to default of neutral.

In some cases, it picked up on strong emotions word like “شكت” or “اتهمت” for anger, but because of the model did not give more weight to these despite it picked and giving it more to the other tokens and predicted neutral.

What was interesting to us, is that the how some irrelevant tokens had high influence, and at the same time, some important emotional tokens were underweighted, which raised a key question to us:

Do the emotionally significant words that we (as humans) would highlight also receive high relevance by the model?

In some examples, yes but not always. There was clear mismatch whee model missed the main emotional triggers and focused on unimportant parts.

This made us as well to think about this question:

Would we assign the same emotion based on these tokens?

Yes. Based on the content and tone, I would them this as they are. The tokens support that, so the model's final prediction doesn’t match what we would expect — even though it looked at the right words.

Final word XAI helped us understand where the model focuses, it also helped us reveal its blind spots, the techniques we applied helped us evaluate how reliable it is, and how it reasons for some predations which makes us proud of.