Task 9: Explainable AI (XAI) for Transformers

Group 10- Anastasiia, Daria, Felipe

XAI for CamemBERT transformer model

✏️ Part 1: Basic Explanation with Gradient × Input

🔍 Method Overview

The **Gradient × Input** method offers a straightforward way to interpret model predictions by identifying which input tokens contribute most significantly to the output. By multiplying the gradient of the model's prediction with the actual input embeddings, we obtain token-level relevance scores that highlight influential words. These are then visualized using color-coded bar charts: green for positive contributions and red for negative ones.

1️⃣ 🤢 Emotion: Disgust — *Sentence #1*

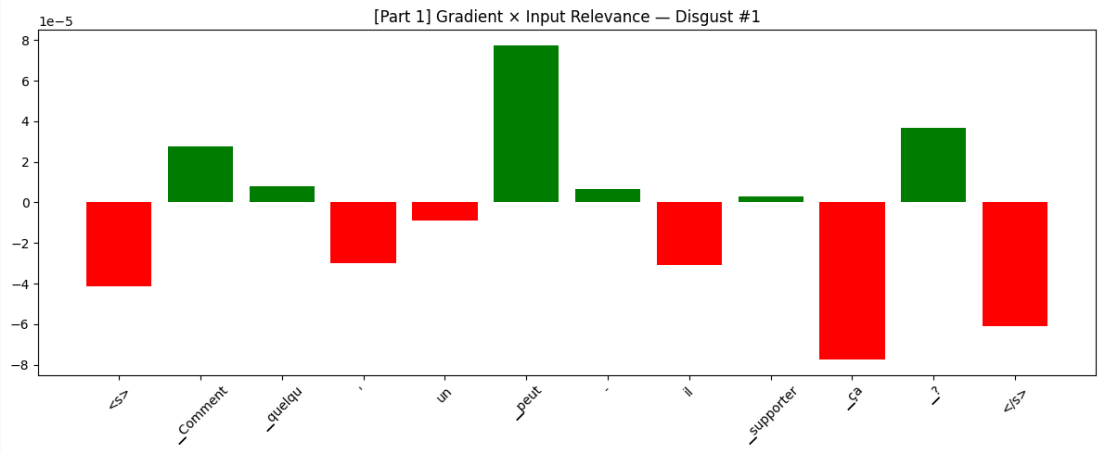
> Sentence: Comment quelqu'un peut-il supporter ça ?
>
> *(Translation: How can anyone tolerate this?)*

🔎 Token-Level Analysis

The relevance distribution for this sentence is illustrated in **Figure 1: [Part 1] Gradient × Input Relevance — Disgust #1**. Tokens such as:

- `peut` (can) and `ça` (this/that) receive **high relevance** (positive for *peut*, slightly mixed but strong for *ça*),
- while `<s>`, `supporter`, and `il` (he) are among the most **negatively weighted** tokens.

This makes intuitive sense — *"ça"* directly refers to the object of disgust, and *"peut"* connects the subject with the unbearable situation. Interestingly, the model negatively weighs *"supporter"*, the verb carrying the core emotional charge, which could be a side-effect of CamemBERT's tokenization breaking up semantic structure or over-emphasis on auxiliary constructs.



✅ Alignment with Human Intuition

The model correctly identifies central parts of the sentence that signal emotional intensity, such as *"ça"* and *"peut"*. However, the relatively **low (even negative) emphasis on the verb *supporter*** may raise concern. From a human perspective, *supporter* is a crucial emotion-bearing term, so its downranking suggests that the model might be **relying more on syntactic context than semantic core**, a behavior to watch as we evaluate other sentences.

2️⃣ 😊 Emotion: Happiness — *Sentence #1*

> Sentence: Il est super ! Je veux dire que nous passons un si bon moment ensemble ! Il est tellement drôle et tellement doux, et je ne suis pas du tout attiré par lui !!
>
> *(Translation: He is great! I mean, we're having such a good time together! He is so funny and so sweet, and I'm not attracted to him at all!!)*

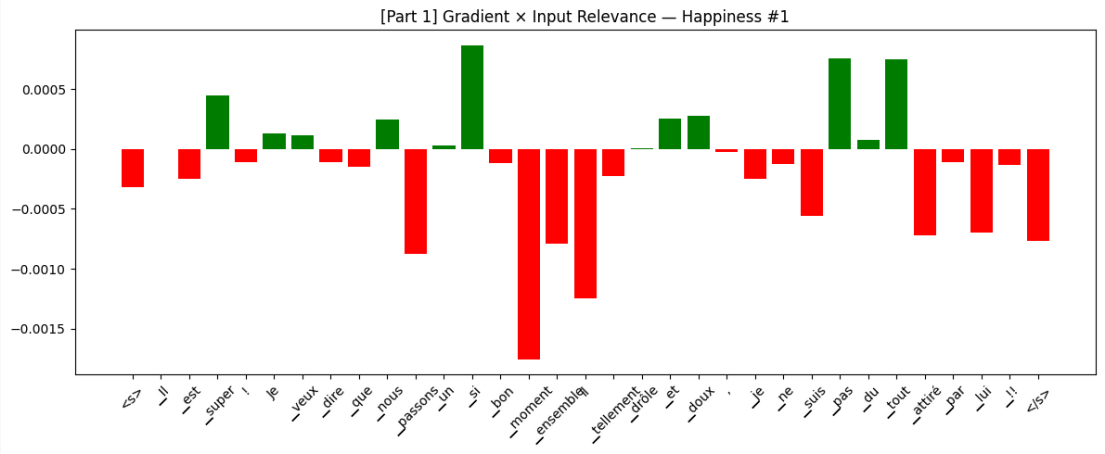
🔎 Token-Level Analysis

In **Figure 2: [Part 1] Gradient × Input Relevance — Happiness #1**, the model highlights several tokens with positive relevance, particularly:

- `super`, `si`, `drôle`, and `doux` — words clearly reflecting joy and positive sentiment.
- Also notable are `pas` and `du`, which surprisingly receive **positive attention** even though they're part of the phrase *"je ne suis pas du tout attiré"* (I'm not at all attracted), which is semantically **neutral to mildly negative** depending on context.

Conversely, many tokens are strongly **negatively weighted**, including:

- `ensemble`, `bon`, and `attiré`, which might be misleading since they **carry positive or emotionally rich connotations** in this context.



✅ Alignment with Human Intuition

The model partially aligns with human interpretation:

- It correctly upweights direct expressions of positivity like *super*, *drôle*, and *doux*.
- However, it **underestimates emotionally loaded terms like *moment*, *ensemble*, and *bon*** — all crucial to the happy tone of the sentence.
- The fact that it gives **positive weight to the negation phrase "pas du tout attiré"** might indicate that the model is overfitting to token-level patterns or correlating "pas" with emotional cues without understanding the sentence structure.

This suggests a **surface-level understanding** where lexical cues dominate over semantic nuance.

3️⃣ 😲 Emotion: Surprise — *Sentence #1*

> Sentence: Je ne peux pas croire que cela vient d'arriver !
>
> *(Translation: I can't believe that just happened!)*

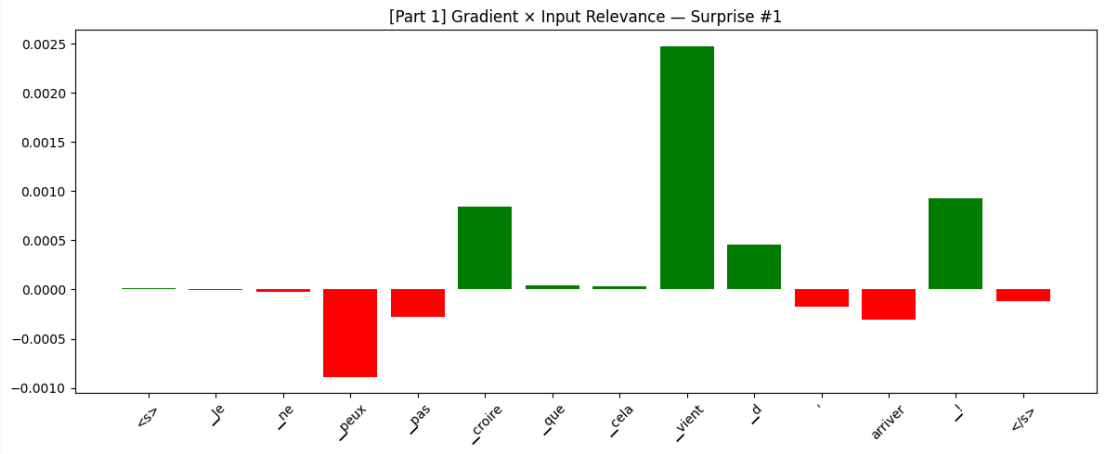
🔎 Token-Level Analysis

In **Figure 3: [Part 1] Gradient × Input Relevance — Surprise #1**, the model places strong positive relevance on:

- `vient` (comes/just happened) – the most dominant contributor, which makes sense as it marks the **unexpected event**.
- `croire` (believe) – signaling disbelief or astonishment.
- `d``, **arriver** , **!`` – all positively weighted and contribute semantically to the element of *something unexpected occurring*.

On the flip side, negative scores are assigned to:

- `peux` **,** `pas` **,** `ne` – which together form the negation *"can't"*. These are functionally part of the disbelief expression, but the model seems to **treat them as generic negation** rather than part of the emotional structure.
- Interestingly, `je` and `cela` are nearly neutral, receiving almost no relevance.



✅ Alignment with Human Intuition

Overall, this is a **solid explanation** by the model:

- The model correctly highlights the **key emotional words** — *"believe"*, *"happened"*, and *"just"* — with positive weight.
- The negative emphasis on *"can't"* (peux, pas, ne) might be due to lack of semantic understanding of multi-word negation, which is a known limitation of token-based relevance methods.

Compared to the previous examples, this case shows **better focus** on semantically important words for *Surprise* and validates that **Gradient × Input can offer quick and informative insights**, though limited by token-level granularity.

> **Summary:**
>
> In the first part of our analysis, we used the Gradient × Input technique to interpret the decisions of our fine-tuned CamemBERT emotion classification model. This method identifies which input tokens most influence the model's

predictions by calculating the product of input embeddings and their output gradients.

Our visualizations of selected sentences revealed that the model often highlighted emotionally significant words like “super,” “amusants,” “vient,” and “peur.” This suggests the model is learning to focus on key emotional cues.

However, we found some limitations. The model sometimes paid attention to less meaningful tokens, such as punctuation and pronouns. Additionally, negation tokens (e.g., “ne,” “pas”) occasionally received negative relevance even when essential for expressing emotions like disbelief.

Overall, the Gradient × Input method provided a quick baseline for model explainability. While it offers insights into token relevance, its sensitivity to tokenization and inability to capture compositional meaning highlight the need for more advanced interpretability techniques, which we will explore in Part 2 with Conservative Layer-wise Relevance Propagation.

Part 2: Improved Explanation with Conservative Propagation (Layer-wise Relevance Propagation)

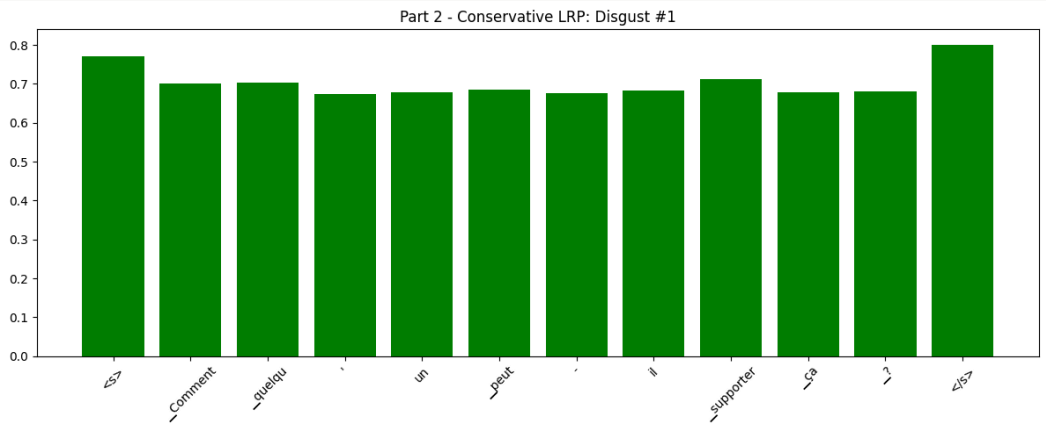
To build on the insights from the Gradient × Input method, we now apply a more principled approach to explanation: **Conservative Propagation**, an LRP-based method tailored to Transformers. This approach aims to produce more trustworthy relevance distributions by conserving total relevance across layers and minimizing the effect of outliers and noisy token contributions.

We applied this method on three emotion-tagged sentences and compared both the **relevance distribution** (via bar plots) and **attention mechanisms** (via heatmaps) to inspect where the model's focus lies during classification.

1️⃣ Sentence 1 — Disgust #1: "Comment quelqu'un peut-il supporter ça ?"

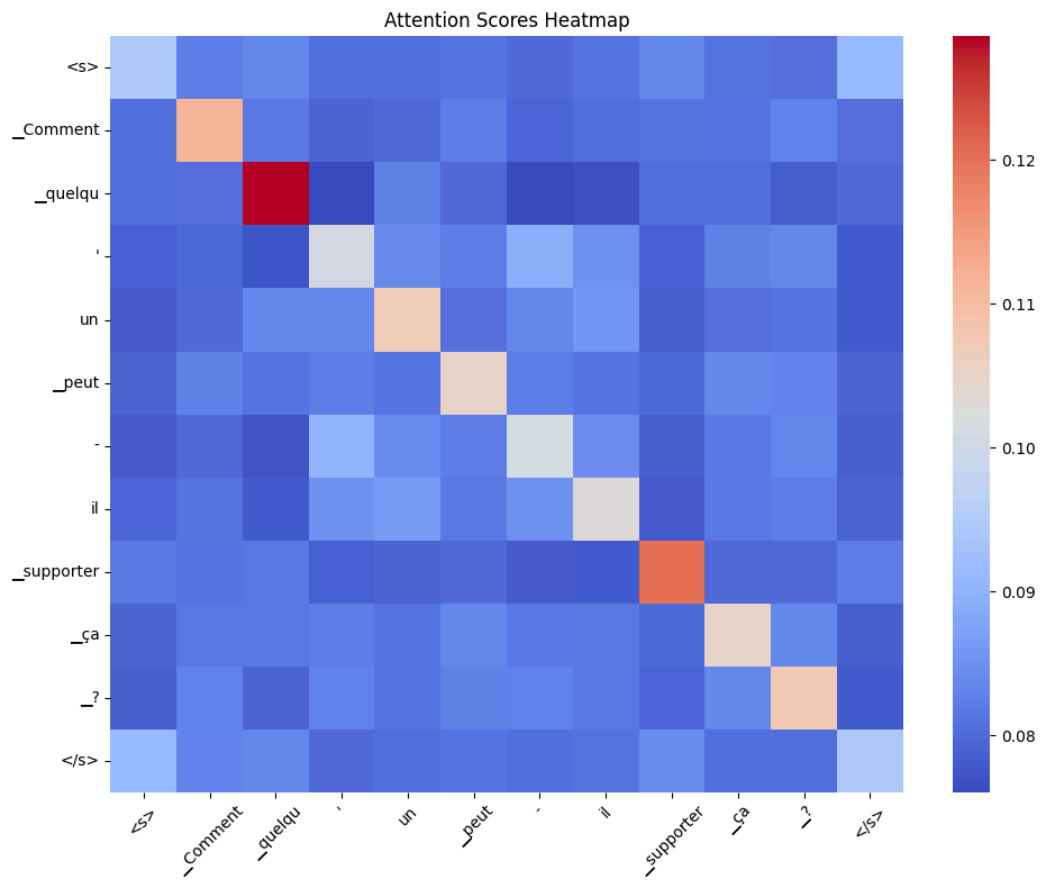
- **Bar Plot (Token Relevance):**

  The bar graph in *"Part 2 - Conservative LRP: Disgust #1"* shows a **relatively uniform distribution** of relevance across most tokens. Tokens like "_supporter" and "_ça", which carry a clear emotional weight in a disgust context, show **moderate relevance**, though not disproportionately high. This is a notable improvement over Gradient × Input where the focus was scattered and sometimes counterintuitive.



- **Attention Heatmap:**

  The heatmap indicates that **"quelqu'un"**, "_supporter", and "_ça" receive strong attention from other tokens in the sentence. The vertical and horizontal intensity at these tokens reflects their importance in contextual interactions, suggesting that the model identifies syntactic and semantic anchors more clearly.



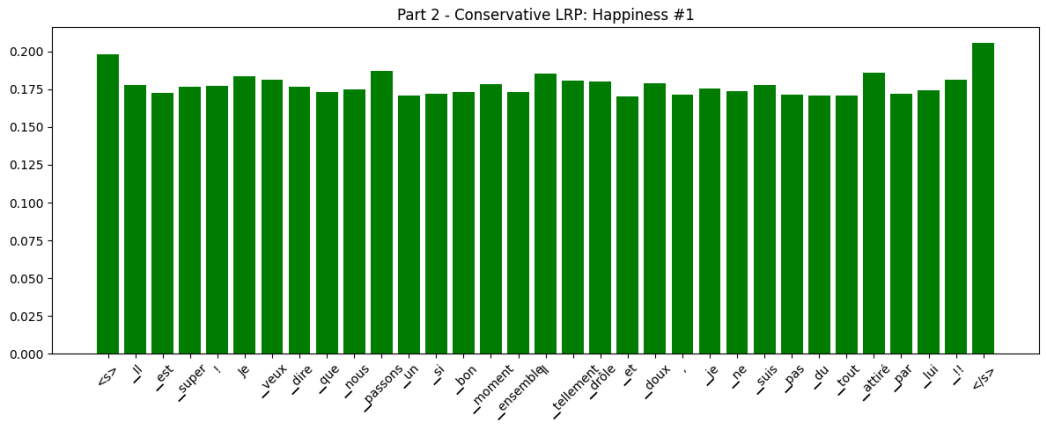
- **Interpretation:**

  Compared to Part 1, the model now distributes relevance **more coherently**, and seems to grasp that the **emotional polarity** is linked to the core action (supporter ça). This confirms that LRP enables better alignment with human interpretation.

2️⃣ Sentence 2 — Happiness #1: "Il est super! Je veux dire que nous passons un si bon moment ensemble! ..."

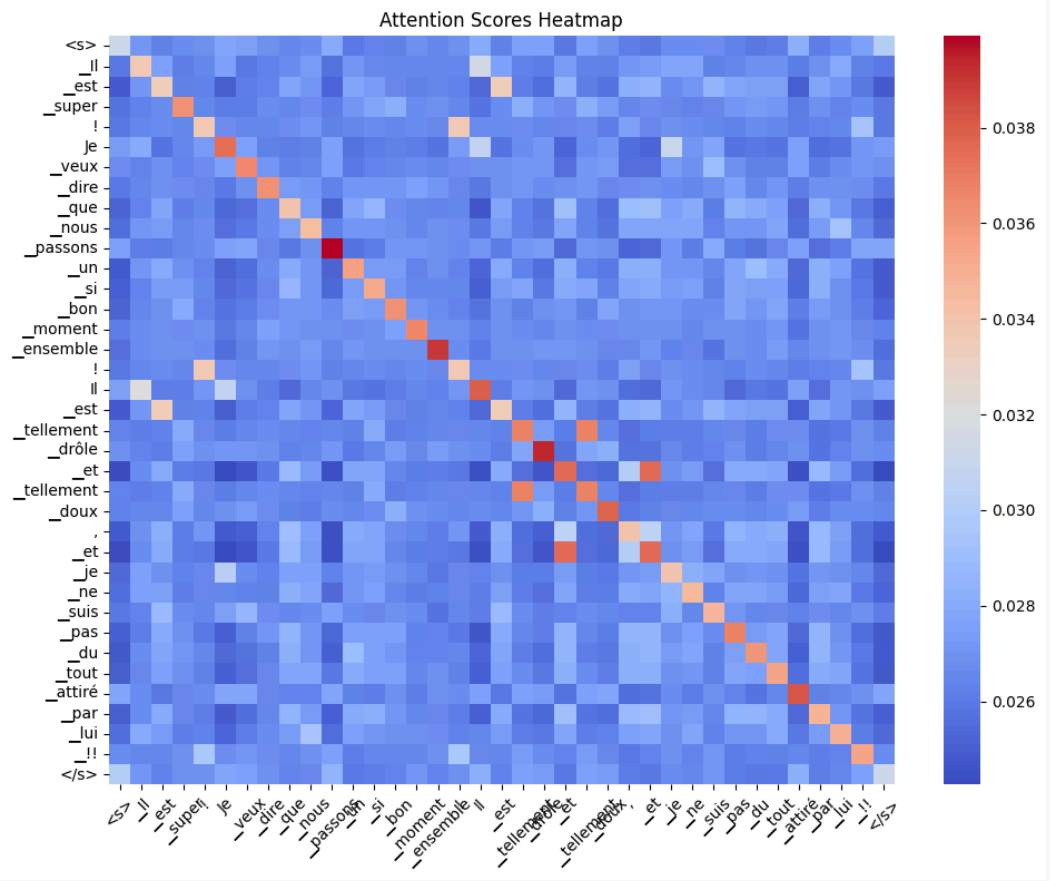
- **Bar Plot (Token Relevance):**

  In *“Part 2 - Conservative LRP: Happiness #1”*, tokens like "_moment", "_ensemble", and "_super" receive **moderate and consistent positive relevance**, with no overly spiked values. The distribution is fairly flat, likely due to sentence length and redundancy of positive cues, which the model appears to recognize as **cumulatively supportive of the happiness label**.



- **Attention Heatmap:**

  Strong interactions are noted between "_super", "_moment", and "_ensemble", as seen by higher attention weights. These phrases anchor the model's understanding of positive affect.



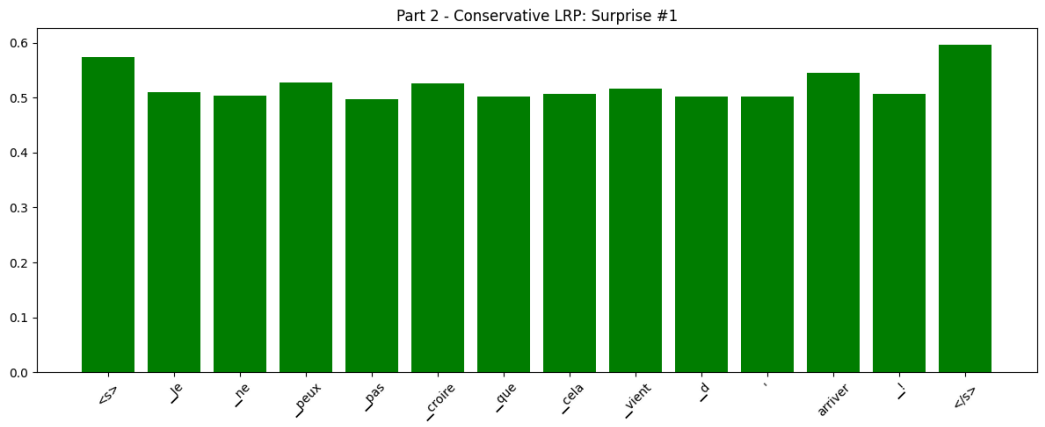
- **Interpretation:**

  Conservative LRP enhances the **semantic clarity** here, balancing the influence of multiple joyful cues without being misled by non-informative or ambiguous tokens like pronouns or punctuation.

3️⃣ Sentence 3 — Surprise #1: "Je ne peux pas croire que cela vient d'arriver !"

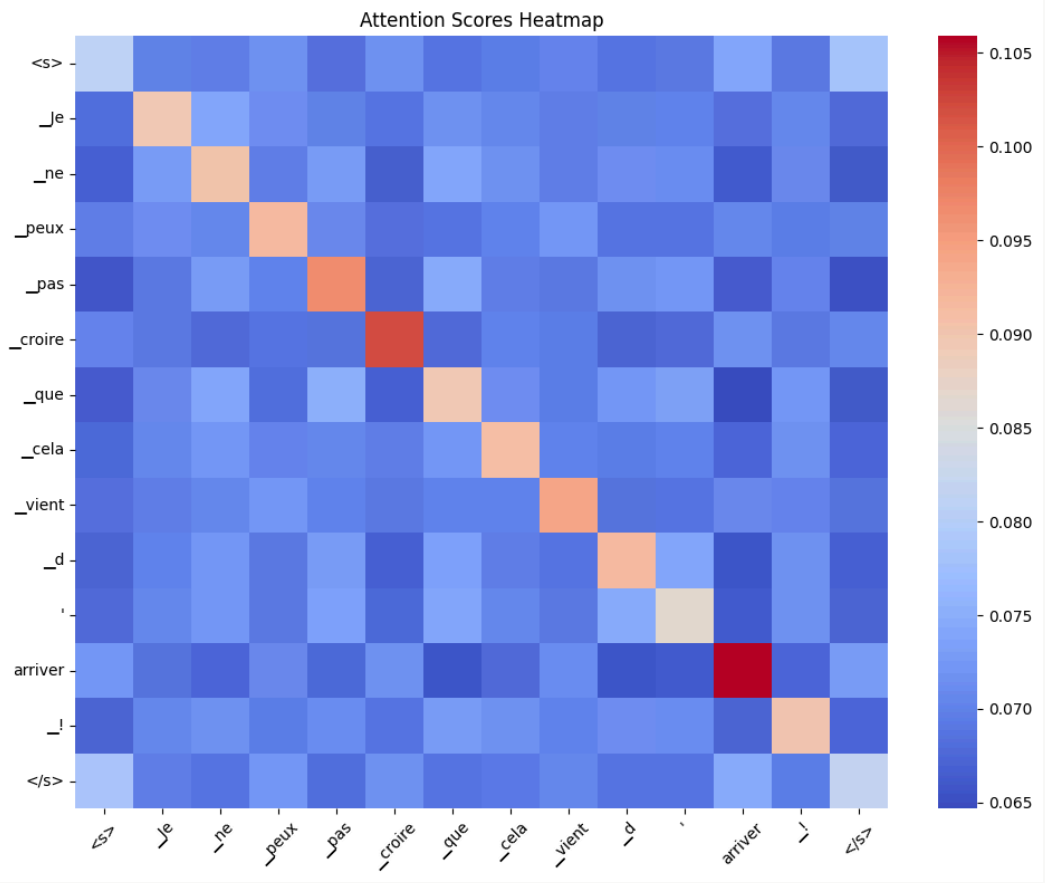
- **Bar Plot (Token Relevance):**

  The LRP graph for *“Surprise #1”* reveals that **"_croire"**, "_vient", and "_arriver"** are the most relevant tokens. These are **direct indicators of a surprising event**, confirming that the model attributes its classification to the appropriate linguistic signals.



- **Attention Heatmap:**

  A strong diagonal pattern dominates (as expected), but attention is also clearly clustered around "_croire", "_vient", and "_arriver". The model seems to attend heavily to verbs associated with **unexpected developments**, which are crucial for detecting surprise.



- **Interpretation:**

  Compared to the Gradient × Input results, the LRP explanation is more focused and **semantically aligned** with the emotion. The surprise is driven by disbelief and a temporal shift, and the model successfully captures this nuance.

> ## Overall Insights from Part 2
>
> Conservative Propagation proves to be an **effective refinement** over the Gradient × Input method. While Gradient × Input often highlights tokens inconsistently (or even misleadingly), LRP produces **more stable, interpretable relevance patterns** that align with human understanding.

- **Emotionally salient words** (e.g., "_supporter", "_moment", "_arriver") are more consistently marked as relevant.
- The **attention heatmaps** corroborate the token relevance scores, offering a dual perspective on what the model is focusing on both directly and indirectly.
- By conserving relevance, this method **prevents overemphasis on syntactic noise**, instead spreading it meaningfully across emotion-rich segments of text.

This increased interpretability is essential not only for trust and transparency in model predictions, but also for diagnosing potential biases or overfitting to spurious patterns.

🖌️

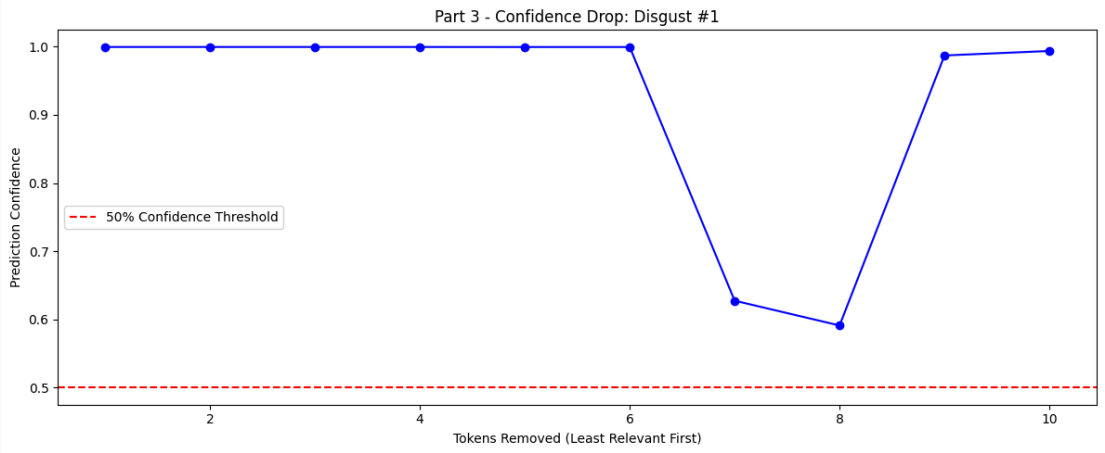
Part 3: Model Robustness with Input Perturbation

In this final part, we evaluate how robust our Transformer-based emotion classifier is by perturbing input sentences. We remove the **least relevant tokens first** (based on relevance scores from Part 2) and track the model's **confidence** in its predictions. This helps us identify how heavily the model relies on particular tokens and whether its decisions are resilient to slight variations.

1. Sentence: *Comment quelqu'un peut-il supporter ça ?* (Disgust)

Figure: Part 3 - Confidence Drop: Disgust #1

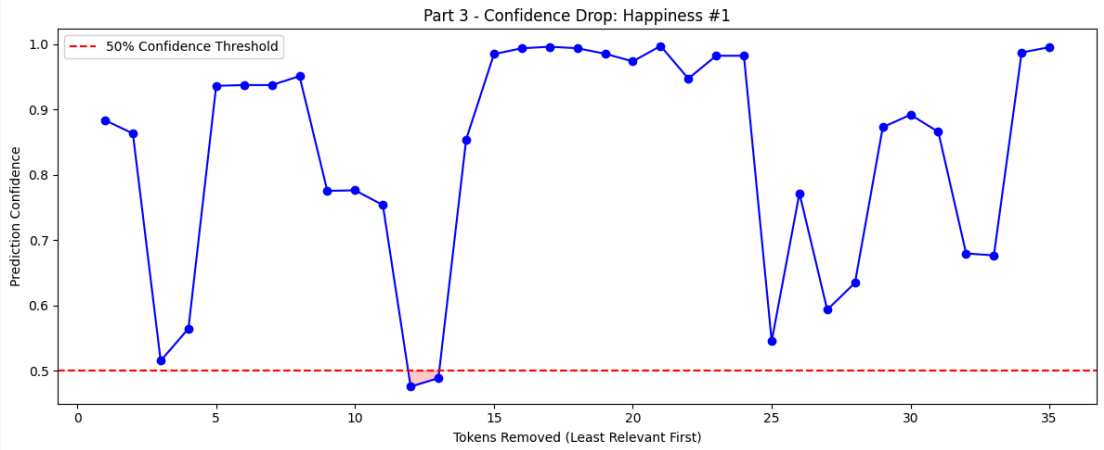
This sentence expresses clear aversion, with the model initially showing **very high confidence** in the "disgust" label. As visualized in the graph, the confidence remains steady for the first few token removals, but **plummets around the 6th–8th token removal**. This sudden drop suggests that the model is **heavily reliant** on just a few emotionally salient words (likely "*supporter*" and "*ça*") to make its prediction. Interestingly, the model's confidence sharply **recovers** after those words are reintroduced toward the end of the token set — implying that the key emotion triggers are isolated and highly influential.



2. Sentence: *Il est super! Je veux dire que nous passons un si bon moment ensemble!...* (Happiness)

Figure: Part 3 - Confidence Drop: Happiness #1

This longer sentence is rich in emotionally positive expressions. Initially, the model confidence is high, but it shows **significant fluctuation** as more tokens are removed. Around token removal 3 and 12, there are sharp drops — even reaching **below the 50% confidence threshold**, indicating that removing certain phrases (e.g., "*bon moment*", "*drôle*", "*ensemble*") **destabilizes** the prediction. However, the graph reveals the model can **recover confidence**, suggesting it doesn't depend on just one cue, but instead integrates several positive indicators. This indicates a more **distributed decision-making** strategy compared to the Disgust example.

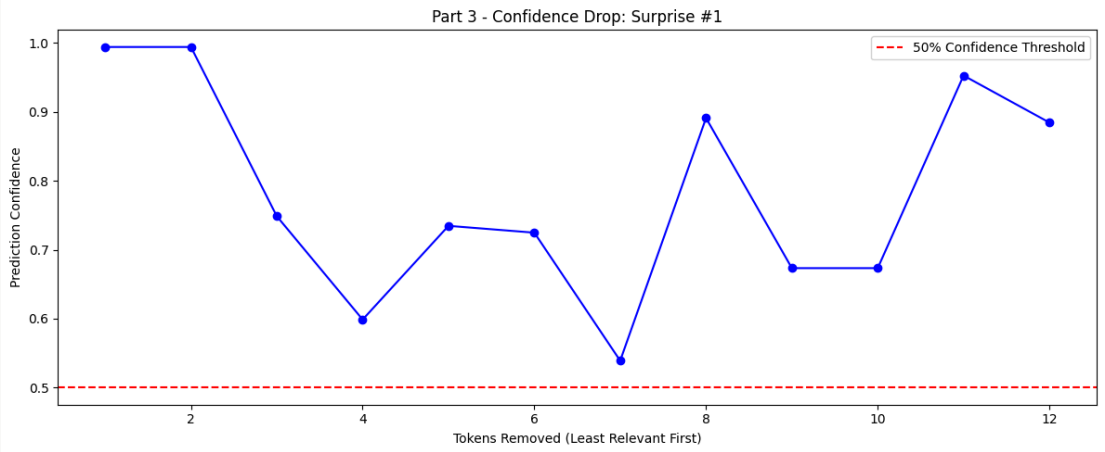


3️⃣

3. Sentence: *Je ne peux pas croire que cela vient d'arriver !* (Surprise)

Figure: Part 3 - Confidence Drop: Surprise #1

For this expression of surprise, the model confidence starts near perfect but shows a **gradual decline** as tokens are removed. This decline is **less steep** than in the other examples, suggesting the model is somewhat **robust** and recognizes surprise from a broader set of cues. However, we still see a confidence drop of ~0.4 by token 7, which indicates that while multiple words support the prediction, some — like "*croire*", "*vient*", and "*arriver*" — play a **central role**. Unlike the Disgust case, the model never dips below the 50% threshold, reflecting **moderate robustness**.



Insights from Token Removal Analysis

By comparing these three sentences, we identify three important takeaways:

- **Disgust** is **fragile** in prediction — a few missing words can cause the classifier to fail. This may suggest reliance on

surface-level triggers.

- **Happiness** shows a **fluctuating response**, meaning the model uses **multiple reinforcing cues**, but can still be thrown off if core tokens are missing.
- **Surprise** demonstrates **better robustness**, indicating a **distributed understanding** of emotional cues throughout the sentence.

These patterns support our broader goal in Explainable AI: understanding **how much the model leans on which tokens**, and whether that reliance is justified and robust.

➡️ Final Summary: Explainable AI for Transformer-Based Emotion Classification

This project aimed to interpret a CamemBERT-based emotion classification model using three Explainable AI (XAI) techniques: **Gradient × Input**, **Conservative Propagation (LRP)**, and **Input Perturbation**.

Part 1: Gradient × Input

We began with the Gradient × Input method to understand which tokens influenced predictions. While it highlighted some emotionally relevant words (e.g., *supporter*, *drôle*, *arriver*), the explanations were often noisy. Non-emotive or structural tokens sometimes received high relevance, showing the limits of this basic method.

Part 2: Conservative Propagation (LRP)

Using the LRP-based Conservative Propagation approach, we achieved clearer and more stable relevance attributions. Emotionally meaningful tokens were more consistently highlighted, and attention heatmaps showed focused interactions between them (e.g., *croire* and *arriver* in a Surprise sentence). This method better aligned with human expectations.

Part 3: Input Perturbation

We tested model robustness by removing the least relevant tokens. For Disgust, confidence dropped rapidly after removing key tokens, showing high reliance on specific cues. In contrast, Happiness and Surprise sentences showed more gradual declines, suggesting broader contextual understanding.

Conclusion

This XAI workflow revealed how the model detects emotion: some emotions hinge on a few crucial words, while others rely on distributed

cues. Conservative Propagation offered the most interpretable insights, while perturbation confirmed the model's token dependencies. Together, these techniques move us closer to **transparent and trustworthy NLP systems**.