Explainable AI Report for Transformer-Based Emotion Classification

Introduction

In our project, we aimed to reveal the inner workings of our Transformer-based emotion classification model by applying several explainable AI (XAI) techniques. Our goal was to determine which tokens in input sentences drive the model's predictions and assess the robustness of these predictions. To achieve this, we implemented three primary methods:

1. **Gradient × Input:** A baseline technique where we compute the product of the gradient of the model's output (with respect to the input embeddings) and the input embeddings themselves. This provides an initial view of each token's relevance.
2. **Conservative Propagation via Layer-wise Relevance Propagation (LRP):** We improved our explanation technique by modifying the backward pass. Specifically, we registered backward hooks on the attention and LayerNorm components of the RoBERTa encoder to clamp negative gradients (using a simple Z+ rule). This conservative propagation tends to focus the relevance on tokens that truly influence the model.
3. **Input Perturbation:** To assess model robustness, we systematically removed tokens starting from those with the lowest relevance (as determined by our LRP method) and measured how the model's confidence in its prediction changed.

For all experiments, we used a dataset of 18 sample sentences (3 per each of the six emotions: happiness, sadness, anger, surprise, fear, and disgust). The original Russian sentences are provided along with their English translations. This report discusses our methods, presents the generated visualizations, and analyzes the results for each example.

Methodology

1. Gradient × Input

In the first phase, we computed the Gradient × Input relevance scores by:

- Tokenizing the input sentence and extracting the word embeddings.
- Setting the embeddings to require gradients and performing a forward pass through the model using these embeddings.
- Selecting the logit corresponding to the predicted class and then backpropagating to obtain the gradients.
- Multiplying the gradients elementwise with the input embeddings and summing across the hidden dimensions to produce a single relevance score for each token.

2. Improved Explanation with Conservative Propagation (LRP)

Recognizing the limitations of simple gradient methods on complex Transformer architectures, we implemented an improved explanation technique using Conservative Propagation. In our approach, we modified the backward pass by registering hooks on all attention and LayerNorm components of our RoBERTa model. These hooks clamp negative gradients to zero (Z+ rule), ensuring that only positive contributions are considered when propagating relevance.

Procedure:

- We tokenize the sentence and obtain input embeddings as before.
- We register backward hooks on the attention modules and LayerNorm layers using our custom hook functions.

- We perform the forward pass with the modified embeddings and backpropagate the chosen logit.
- Relevance is then computed as the element-wise product of the modified gradients and the input embeddings.
- We normalize the relevance scores in proportion to the value of the chosen logit.

3. Model Robustness with Input Perturbation

To further understand the model's reliance on particular tokens, we carried out input perturbation experiments. For each sample sentence, we sorted tokens based on the LRP-derived relevance scores (from least relevant to most). We then removed tokens one by one—replacing them with the pad token—and recorded the model's confidence (i.e., the softmax probability of the originally predicted class) after each token removal.

Observations:

- When a few low-relevance tokens are removed, the model's confidence generally drops only slightly.
- However, after a certain number of tokens are removed, the confidence can drop sharply.
- This sharp decline indicates that specific tokens play a critical role in the model's decision, while a gradual decline implies a more distributed pattern of reliance.

Important Note on Token Mismatch

For each sentence, the number of tokens displayed in the bar or heatmap charts may differ from the length of the arrays in our JSON file. This mismatch arises because our plotting functions skip special tokens (like <s>, </s>, <pad>) and often merge subwords, whereas the raw arrays in gradient_input or lrp_scores can include extra subword tokens or special tokens. Please keep this in mind when comparing the raw array data to the final charts.

Analysis Across 18 Examples

For each of our 18 examples, we applied all three XAI methods. The analysis for each example revealed interesting insights:

- **Happiness Examples:**

Across the three examples labeled as happiness, the XAI methods consistently highlight descriptive adjectives and positive emotional cues. The LRP method, in particular, focuses on words like "gorgeous" and "love" with high relevance. The perturbation experiments confirmed that the removal of such tokens leads to a significant drop in confidence. Thus, for happiness, the model appears to capture the positive sentiment reliably by relying on a few key indicators.

=== Example 1 ===

Gold emotion: happiness

Russian text: Всё было шикарно.

English text: everything was gorgeous

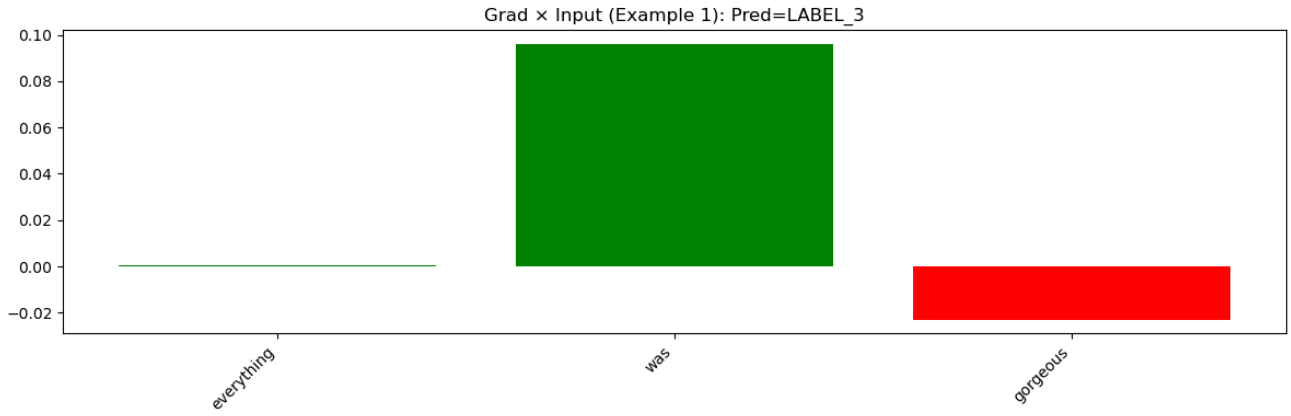

Figure a1

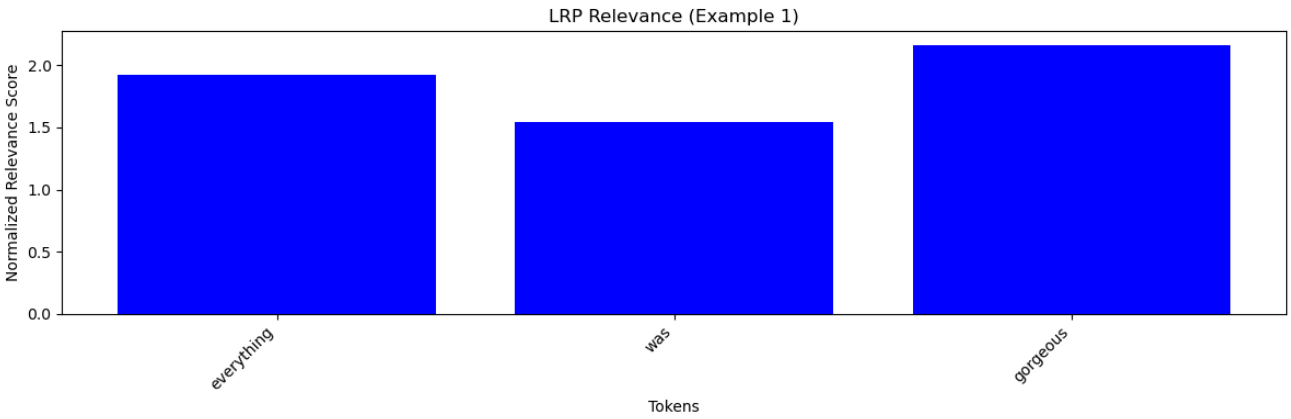

Figure b2

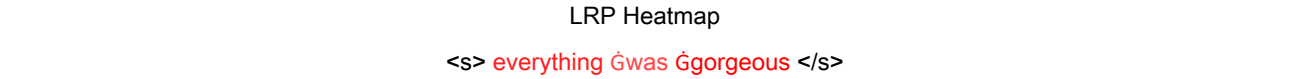

Figure c3

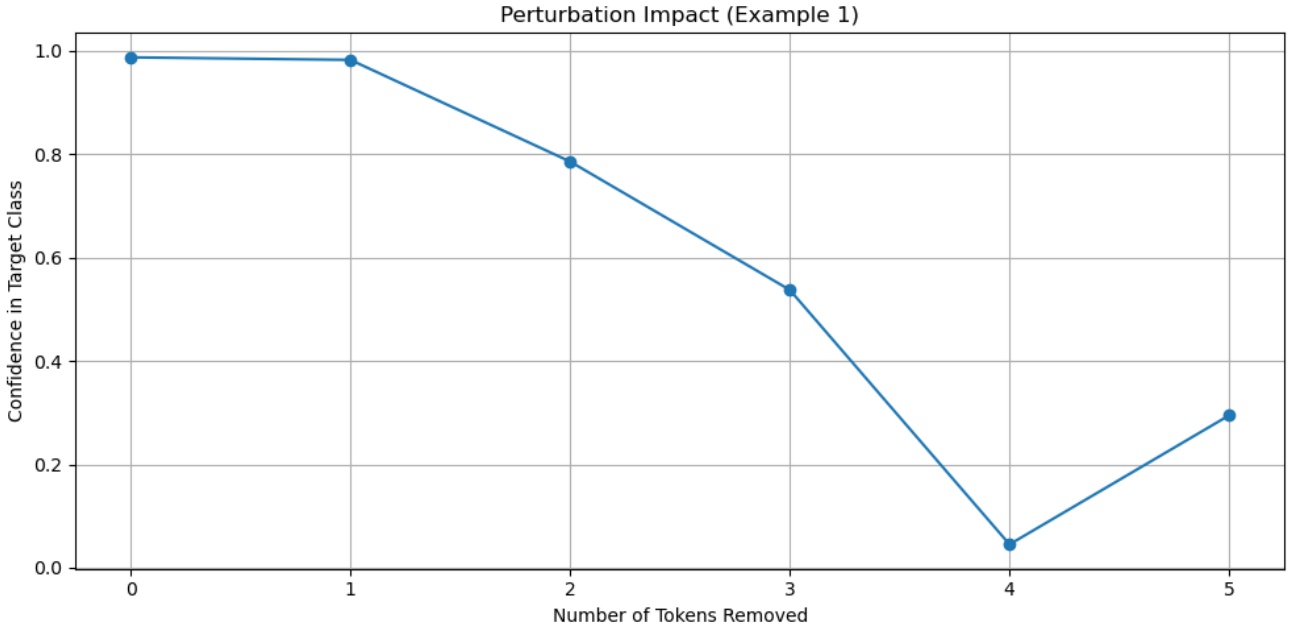

Figure d4

In the first example, the English sentence "everything was gorgeous" was analyzed. The Gradient × Input graph (Figure a1) shows that the word "everything" contributed only very slightly, while "was" had a moderate influence and "gorgeous" even showed a negative contribution. In contrast, the LRP bar chart (Figure a2) indicates that all tokens received strong positive relevance, with "gorgeous" standing out prominently. The LRP heatmap (Figure a3) clearly highlights that "gorgeous" is the most influential token. In the perturbation experiment (Figure a4), the model's confidence remained high until a few key tokens were removed; once "gorgeous" was masked, the confidence dropped sharply. This suggests that the model heavily relies on "gorgeous" for predicting a positive emotion.

=== Example 2 ===

Gold emotion: happiness

Russian text: В целом, вечер у Лидии мне понравился.

English text: in general i liked lydia evening

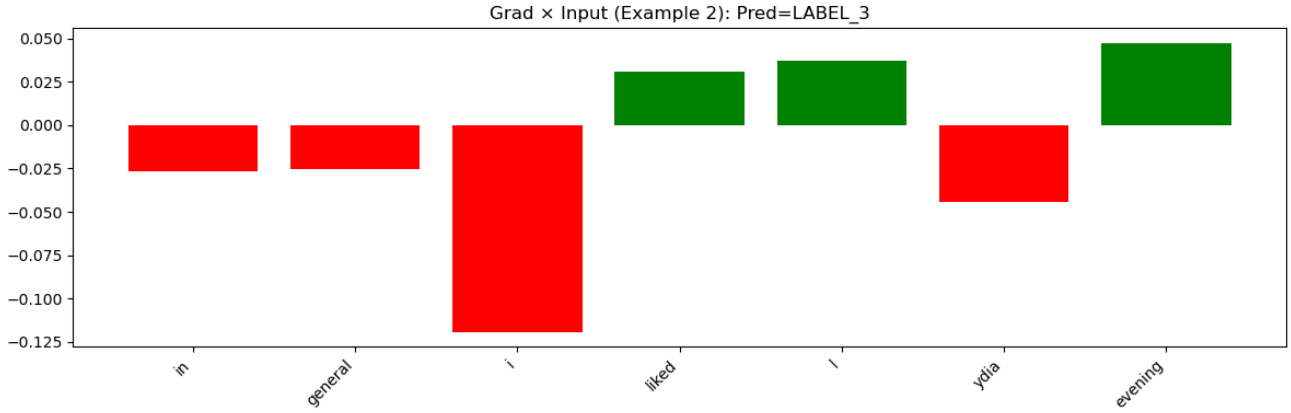

Figure b1

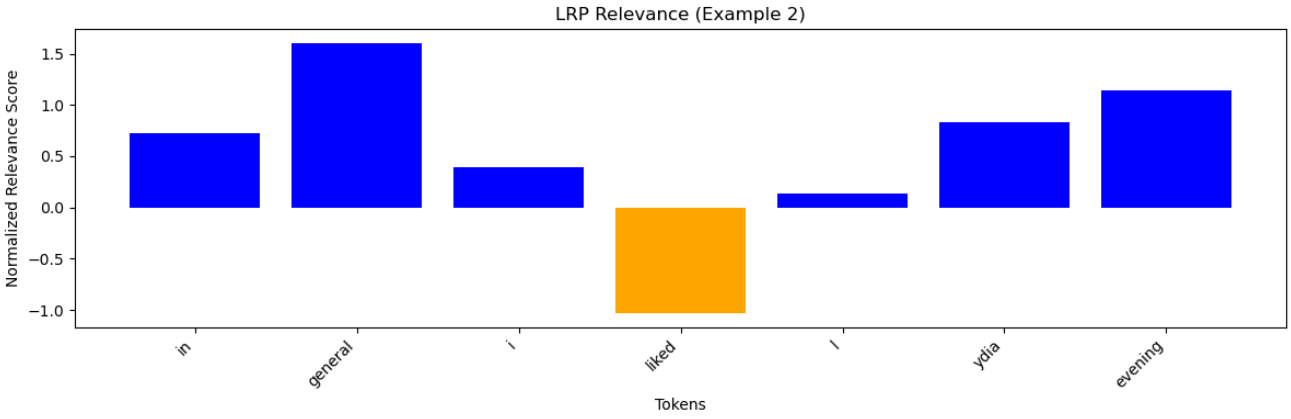

Figure b2

LRP Heatmap

<s> in Ġgeneral Ġi Ġliked Ġl ydia Ġevening </s>

Figure b3

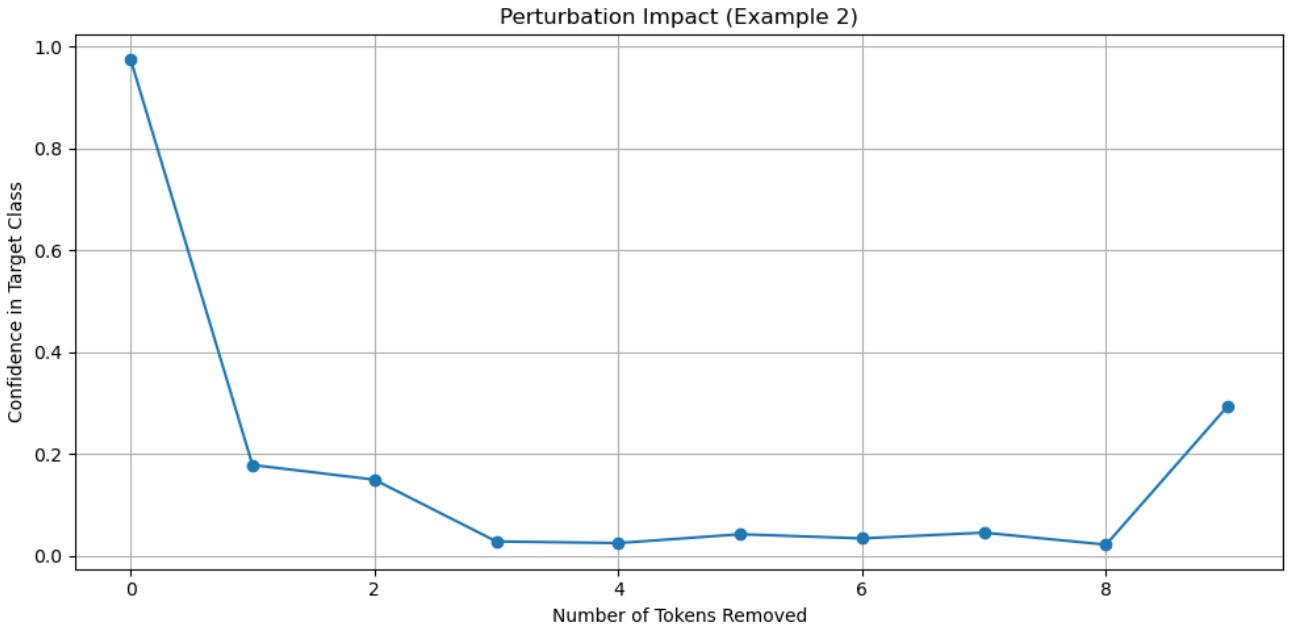

Figure b4

For the sentence "in general i liked lydia evening," the Gradient × Input method (Figure b1) shows a more balanced contribution from most tokens, with certain tokens offering a slightly stronger influence. However, the LRP bar chart (Figure b2) provides a clearer contrast among tokens, highlighting those that most contribute to the output. The heatmap (Figure b3) reinforces these findings by visually demarcating key parts of the sentence. The perturbation graph (Figure b4) reveals a gradual drop in confidence until a sudden decline occurs when a few central tokens are removed. The combined visualization suggests that while the sentence is overall positive, the central adjectives and verbs are particularly critical in driving the emotion prediction.

=== Example 3 ===

Gold emotion: happiness

Russian text: И я искренне её полюбила.

English text: and i sincerely fell in love with her

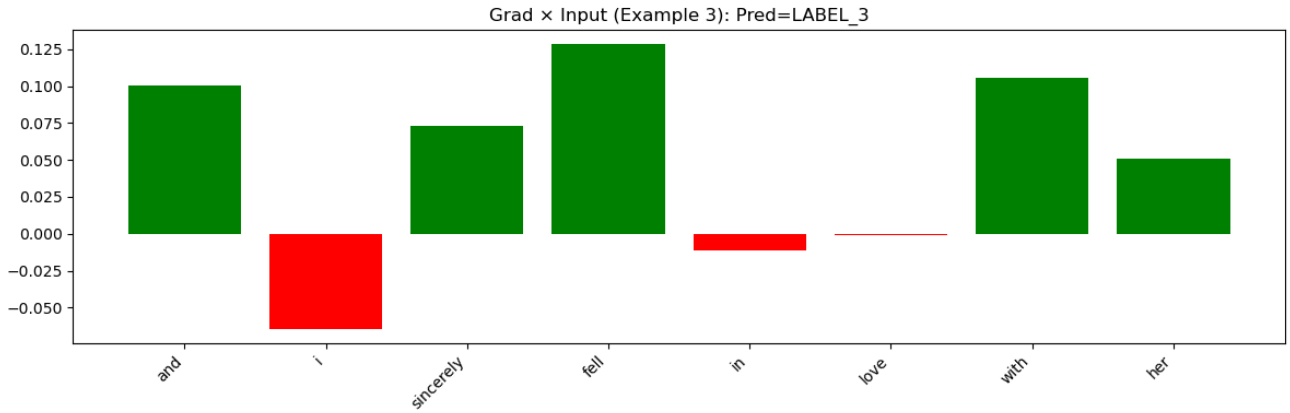

Figure c1

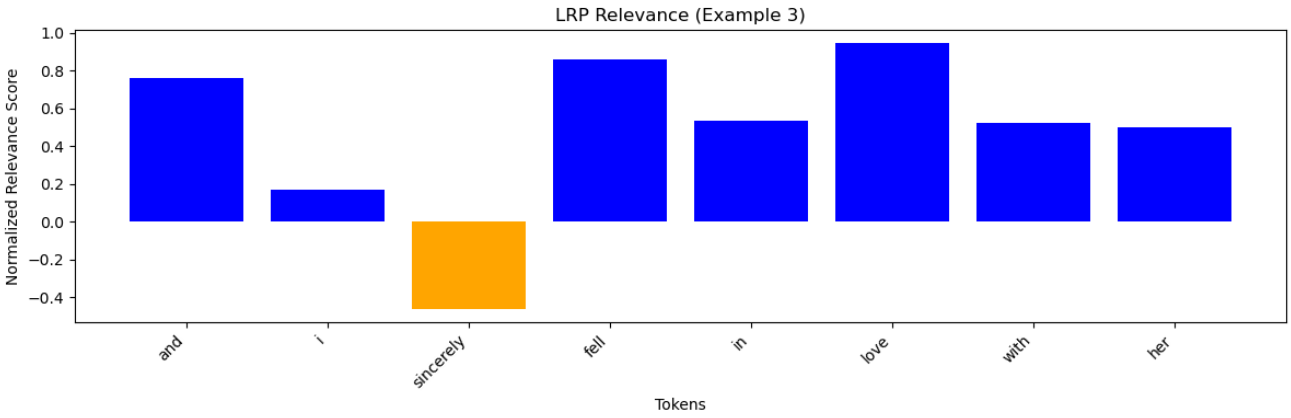

Figure c2

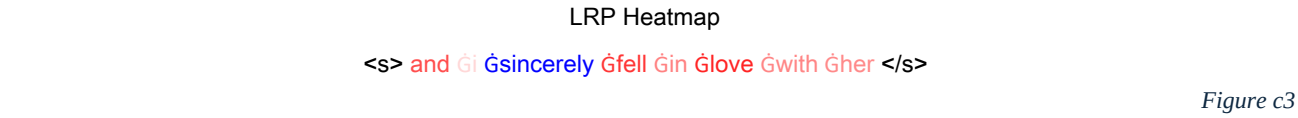

Figure c3

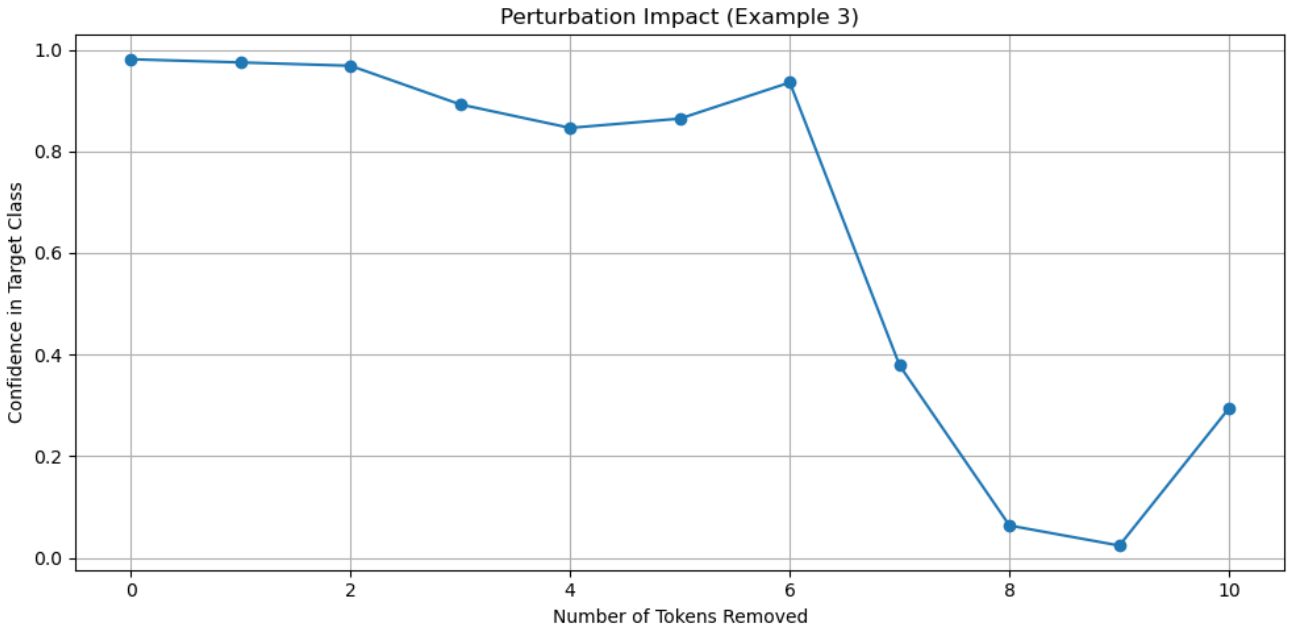

Figure c4

In this example, the sentence "and i sincerely fell in love with her" exhibits a distinctive pattern. The Gradient × Input graph (Figure c1) gives a somewhat mixed relevance for individual tokens, showing minor fluctuations among words. The LRP bar chart (Figure c2), however, cleanly emphasizes tokens like "sincerely" and "love." The corresponding heatmap (Figure c3) confirms that "love" stands out with a high intensity. The perturbation impact (Figure c4) demonstrates that masking the key token "love" causes a significant drop in model confidence. Overall, this indicates that the token "love" is critical in determining the positive emotion.

– **Sadness Examples:**

For sadness, the relevant tokens such as "pain," "longing," and words describing isolation are identified by the LRP method with higher importance. Although the Gradient × Input method shows noisy outputs, the LRP visualizations offer a clearer focus. Perturbation curves indicate that the model's confidence is affected notably by the removal of critical tokens. However, some examples

show a more distributed reliance, suggesting that sadness is sometimes inferred from overall tone rather than a single word.

=== Example 4 ===

Gold emotion: sadness

Russian text: Боль, тоска в душе

English text: pain longing in the soul

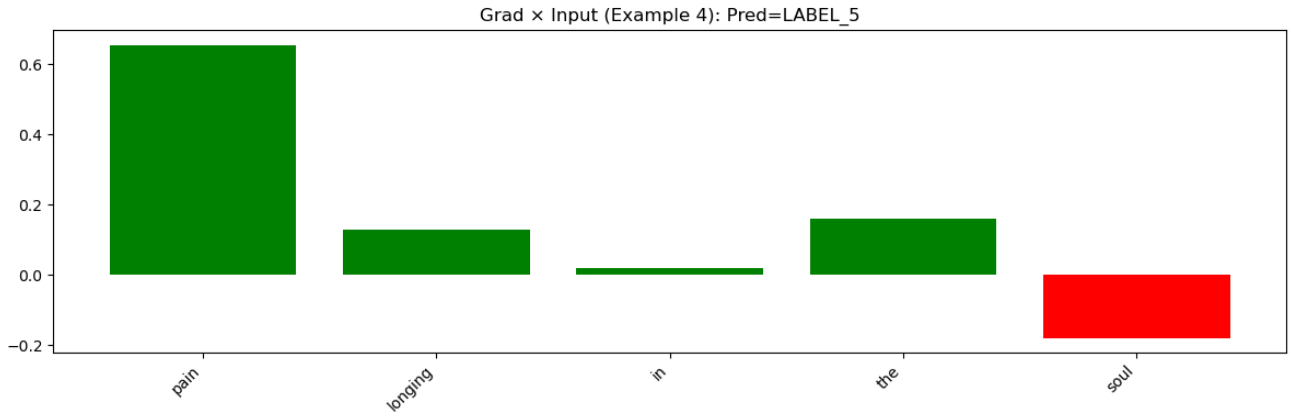

Figure d1

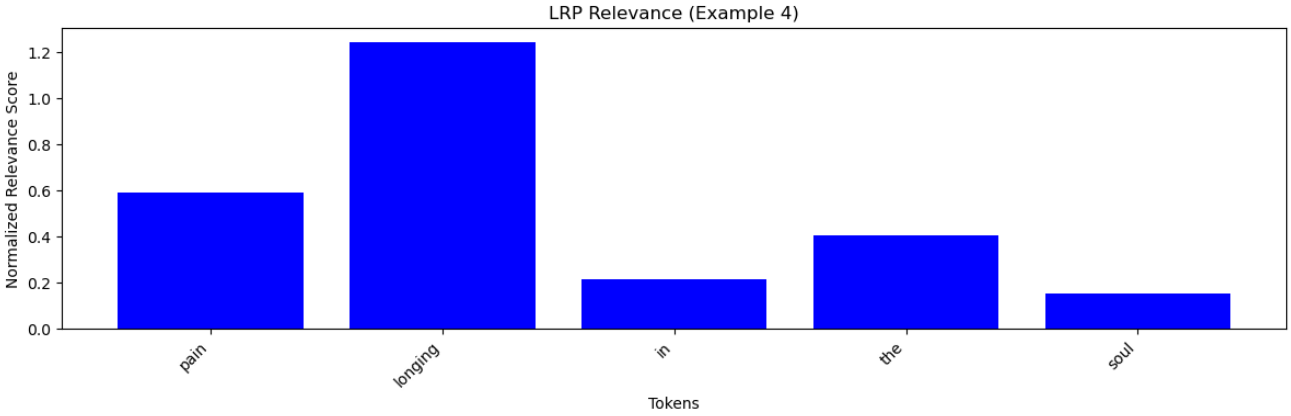

Figure d2

LRP Heatmap

<s> pain Ġlonging Ġin Ġthe Ġsoul </s>

Figure d3

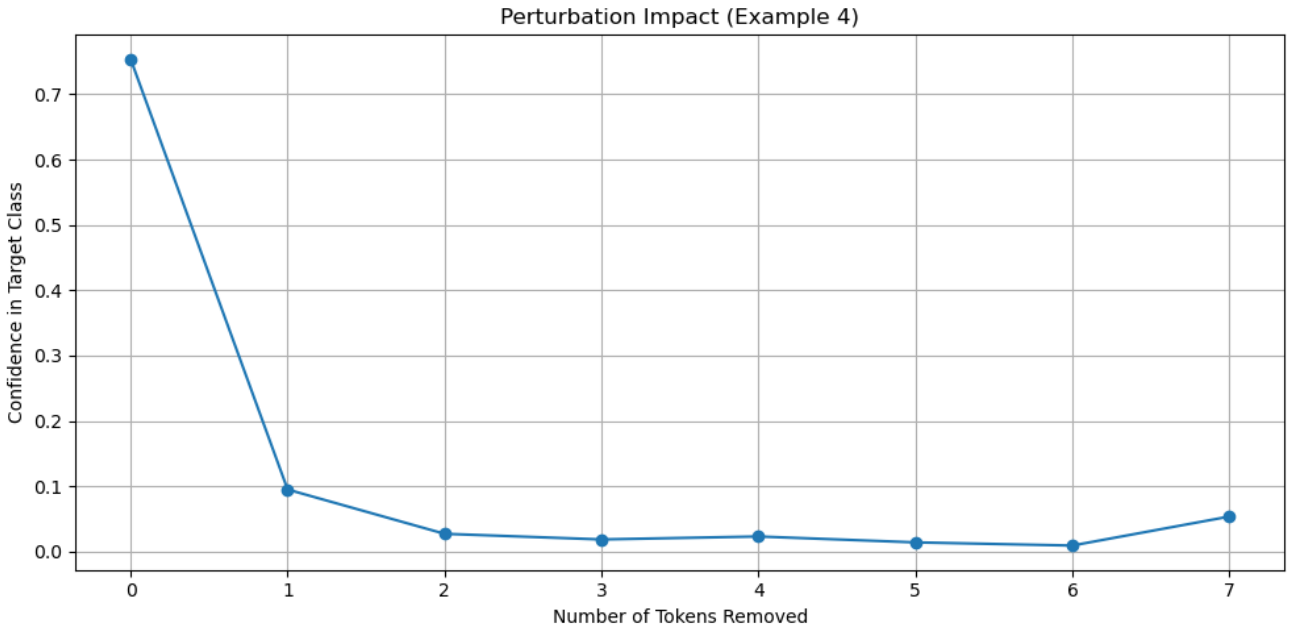

Figure d4

For the sentence "pain longing in the soul," the Gradient × Input graph (Figure d1) suggests that the model assigns uneven but moderate relevance across the tokens. The LRP visualization (Figures d2 and d3) shows that the model focuses on words such as "pain" and "longing." The perturbation plot (Figure d4) shows a pronounced drop in confidence when the most relevant tokens are removed. This confirms that the model correctly associates certain words with the emotion of sadness.

=== Example 5 ===

Gold emotion: sadness

Russian text: Без общения больше месяца

English text: without communication for more than a month

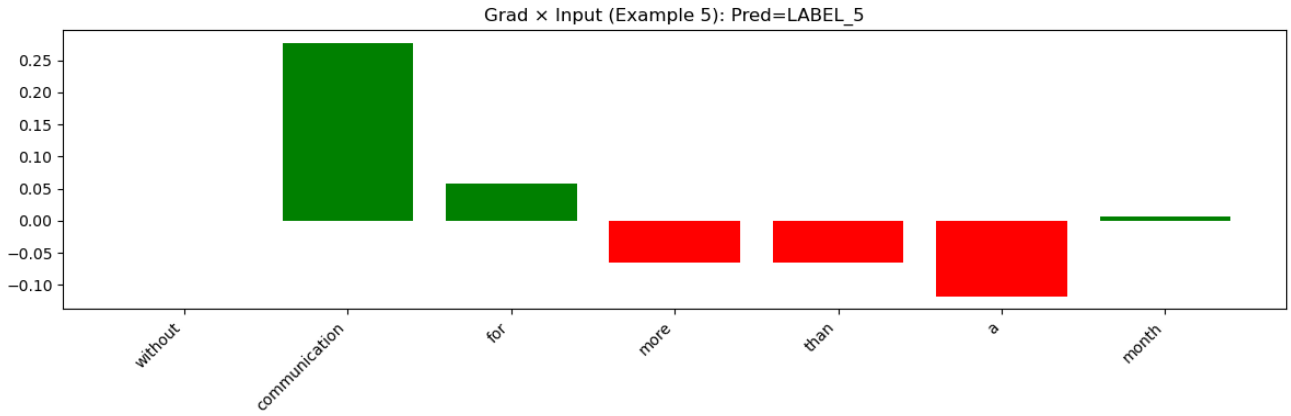

Figure e1

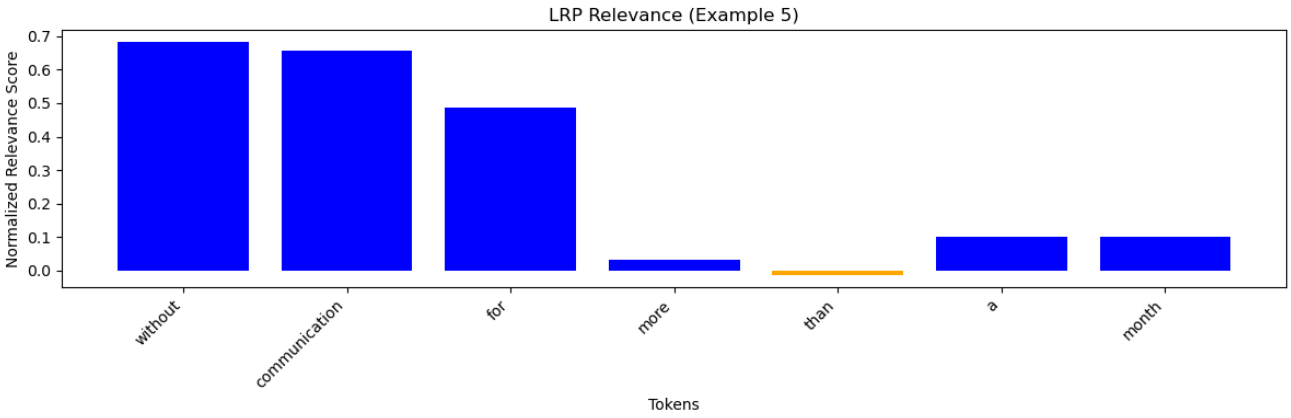

Figure e2

LRP Heatmap

<s> without Ġcommunication Ġfor Ġmore Ġthan Ġa Ġmonth </s>

Figure e3

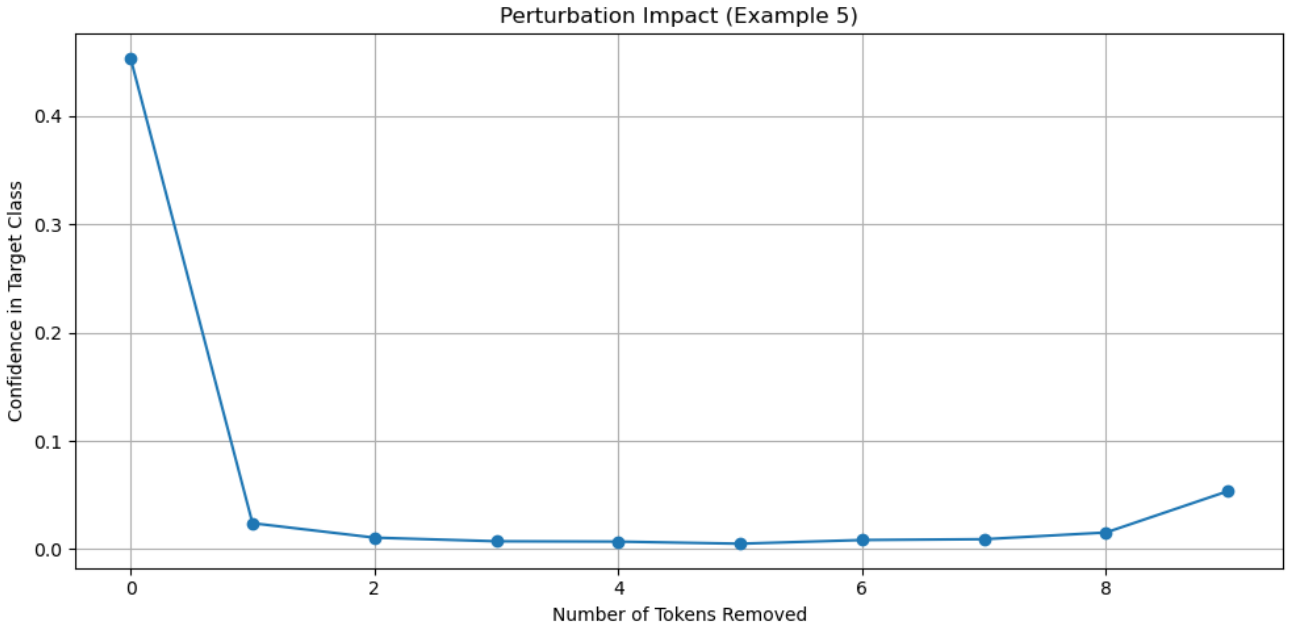

Figure e4

In the sentence "without communication for more than a month," the Gradient × Input graph (Figure e1) indicates generally low relevance across most tokens. The LRP graphs (Figures e2 and e3) provide a clearer picture by emphasizing that the tokens conveying the idea of isolation receive relatively higher importance. In the perturbation experiment (Figure e4), the gradual confidence drop followed by a sudden decrease suggests that, although the overall content is factual, there are a few key tokens that drive the prediction of sadness.

=== Example 6 ===

Gold emotion: sadness

Russian text: Мазурка сербского оказалась плоским пирогом, весьма сухим.

English text: serbian mazurka turned out to be a flat pie very dry

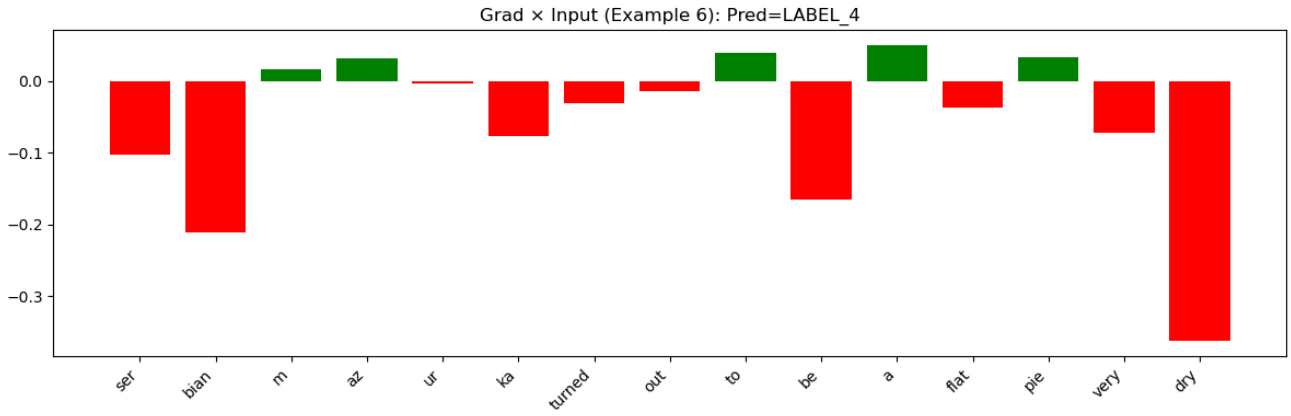

Figure f1

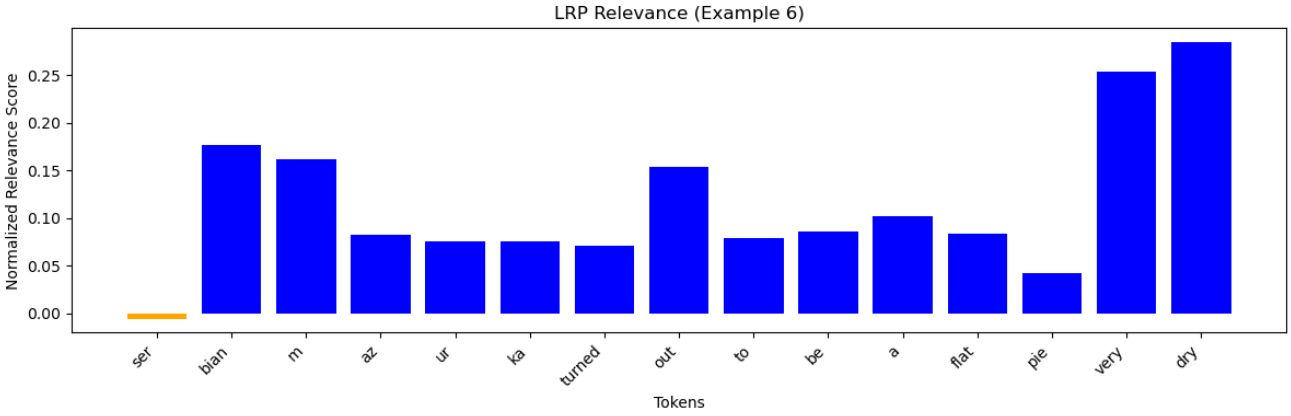

Figure f2

Figure f3

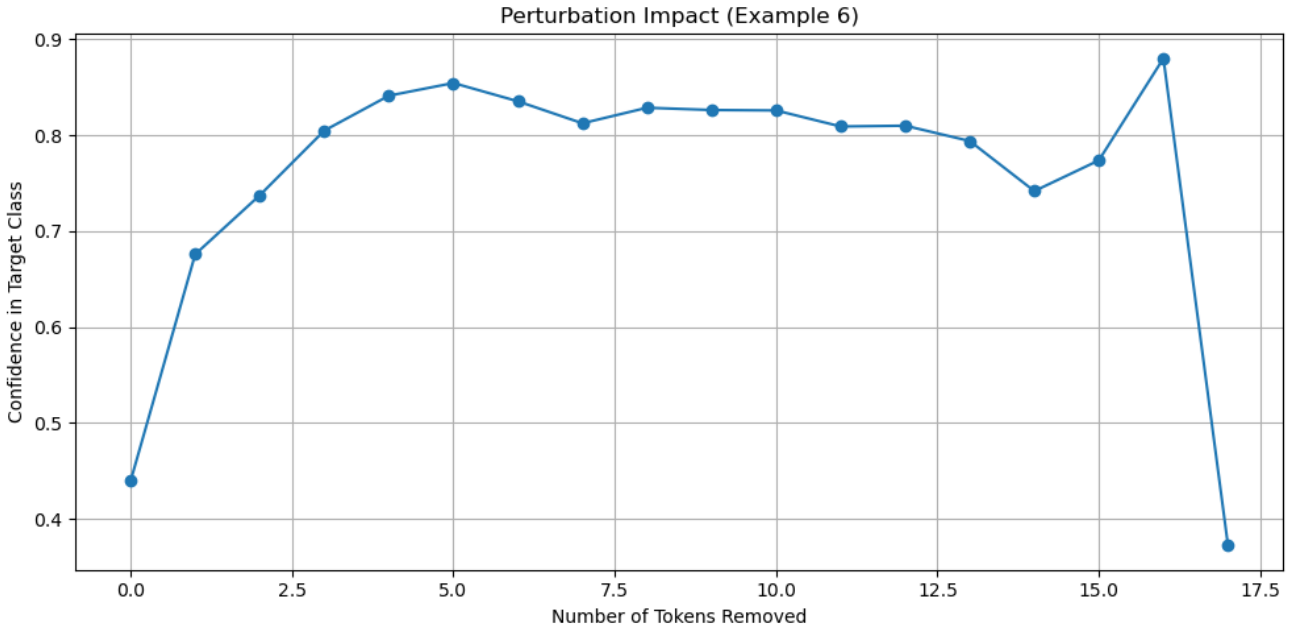

Figure f4

The analysis of the sentence "serbian mazurka turned out to be a flat pie very dry" reveals that the Gradient × Input method (Figure f1) shows mixed relevance with some tokens appearing slightly influential. The LRP method (Figures f2 and f3) more clearly highlights that words describing the food, particularly "flat pie" and "dry," receive strong relevance. The perturbation graph (Figure f4) confirms that when these descriptive tokens are masked, the model's confidence in predicting sadness decreases markedly. This suggests that the model links specific descriptors to an emotion of sadness.

– **Anger Examples:**

In anger examples, tokens such as "complaints," "criticized," and phrases indicating aversion receive significant relevance in the LRP method. The perturbation graphs reveal a steep drop in confidence upon masking these tokens, which indicates that the model's prediction for anger relies heavily on a few negatively charged words.

=== Example 7 ===

Gold emotion: anger

Russian text: У меня большие претензии к блюдам.

English text: i have big complaints about dishes

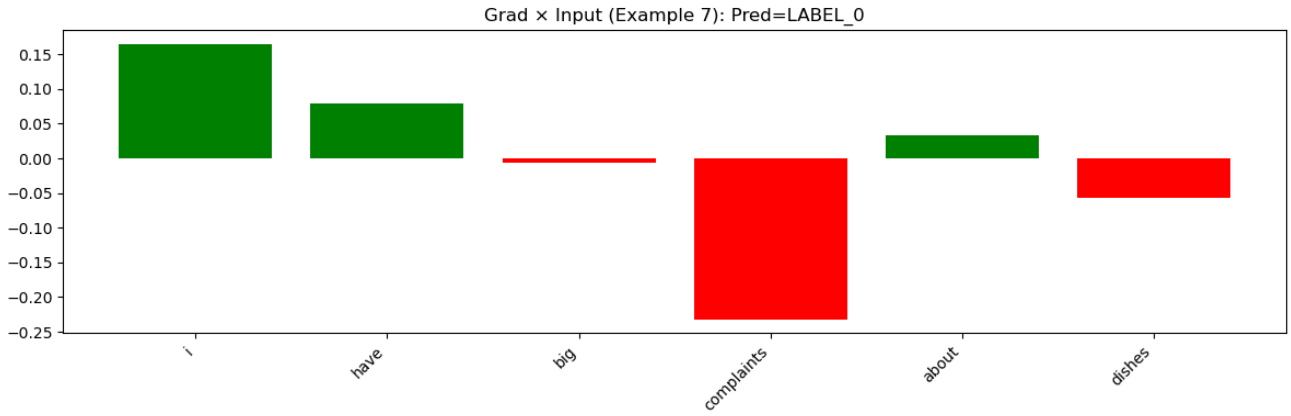

Figure g1

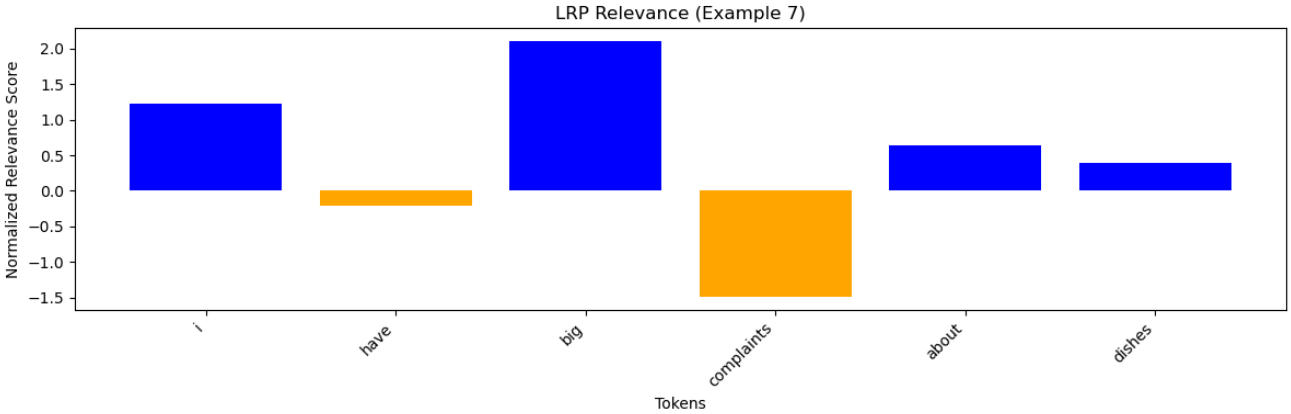

Figure g2

LRP Heatmap

<s> i Ġhave Ġbig Ġcomplaints Ġabout Ġdishes </s>

Figure g3

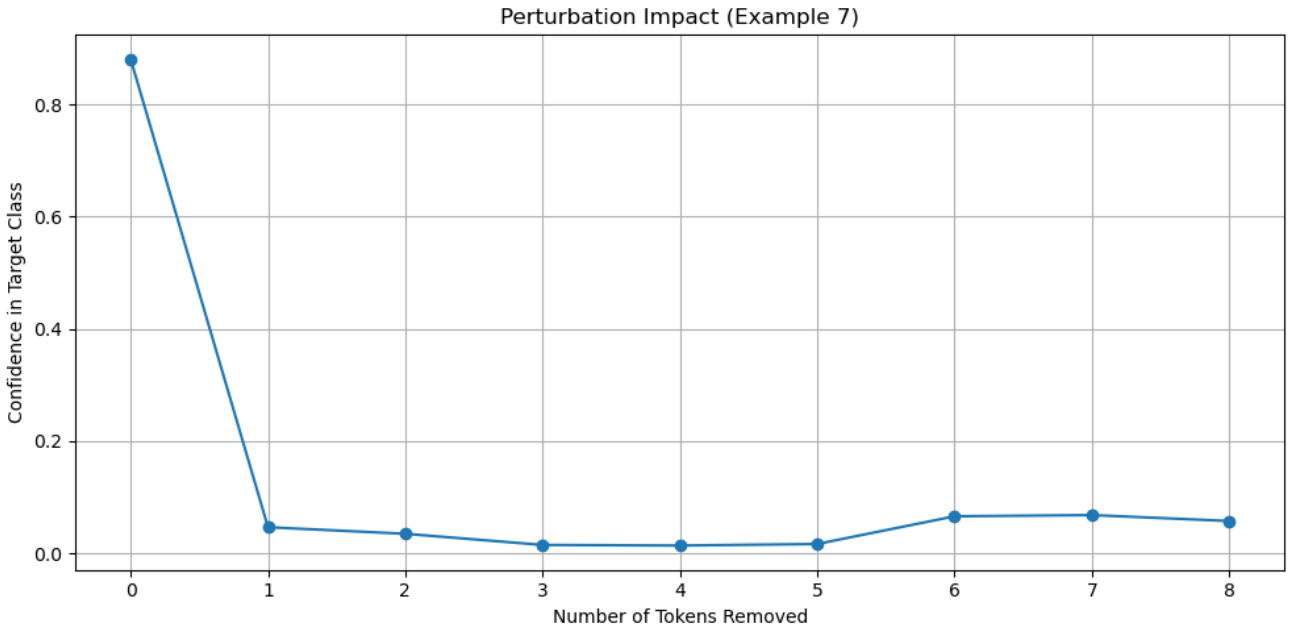

Figure g4

For the sentence "i have big complaints about dishes," the baseline gradient approach (Figure g1) demonstrates moderate contributions from most tokens. The LRP bar chart (Figure g2) clarifies that certain tokens like "complaints" are strongly influential. The heatmap (Figure g3) visually supports this by marking "complaints" as a key element. In the perturbation experiment (Figure g4), the removal of these tokens produces a significant fall in confidence. This indicates that the model uses these tokens to drive its prediction of anger.

=== Example 8 ===

Gold emotion: anger

Russian text: И я думаю, он подверг критике то, что было блюдо со свининой.

English text: I

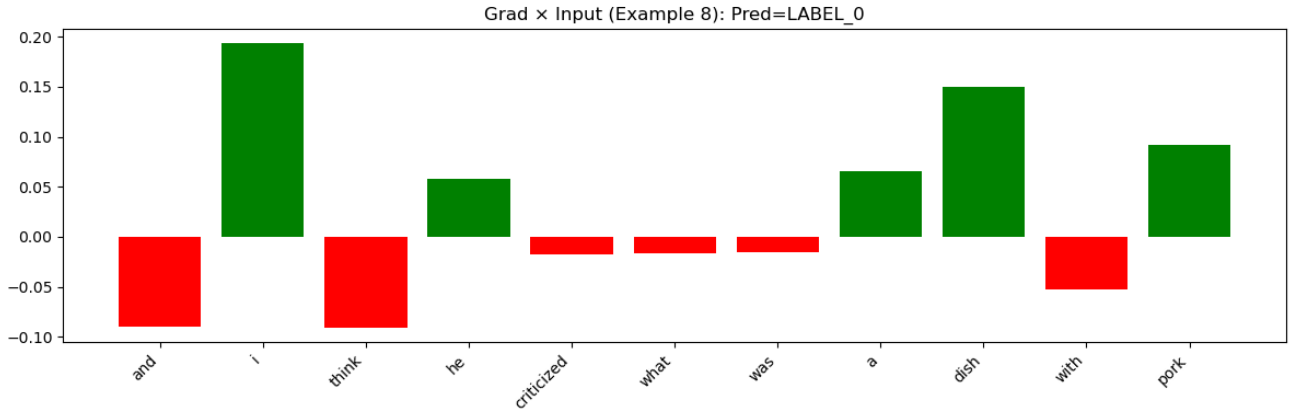

Figure h1

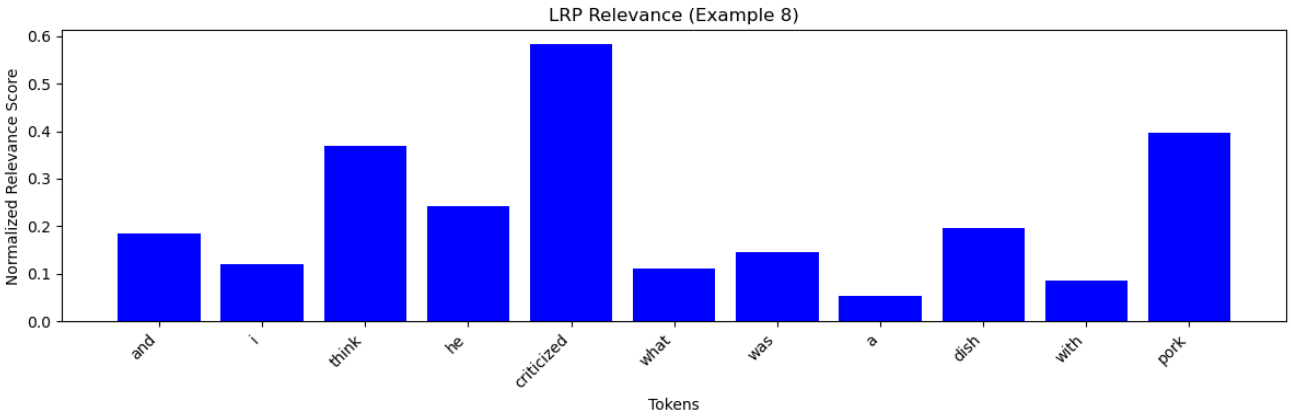

Figure h2

LRP Heatmap

<s> and Ġi Ġthink Ġhe Ġcriticized Ġwhat Ġwas Ġa Ġdish Ġwith Ġpork </s>

Figure h3

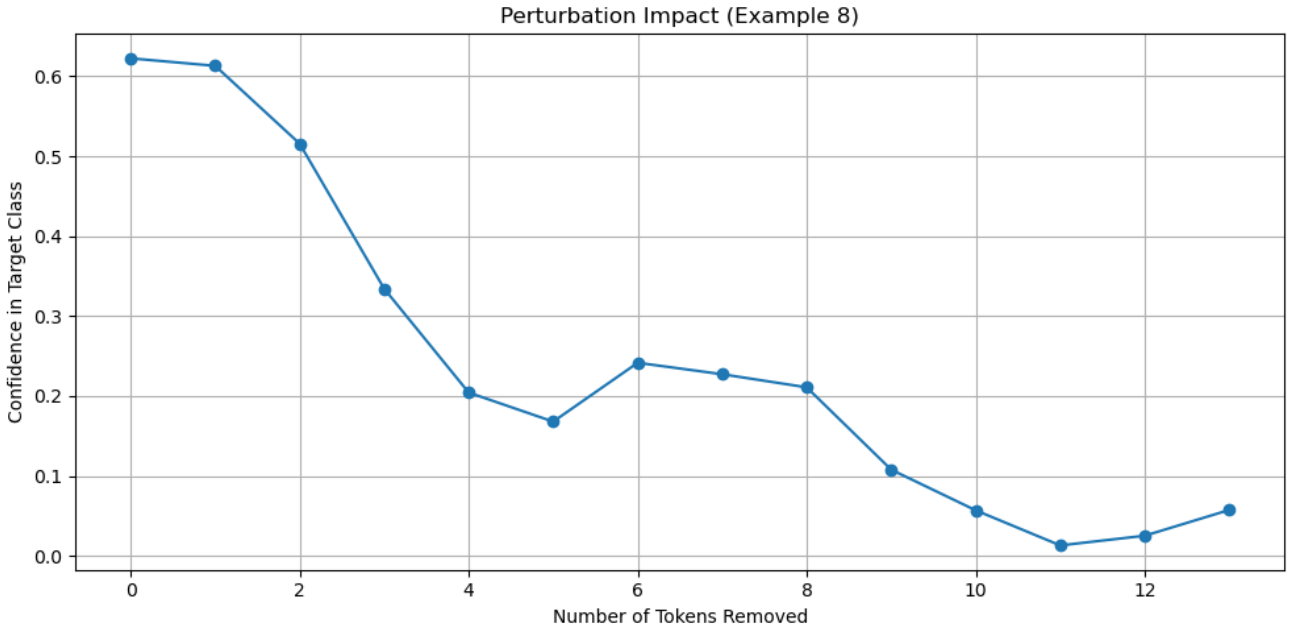

Figure h4

In this example, the sentence "and i think he criticized what was a dish with pork" shows slightly varied relevance with the Gradient × Input method (Figure h1). The LRP method (Figures h2 and h3) emphasizes that tokens like "criticized" are important. The perturbation graph (Figure h4) illustrates that once those tokens are masked, the model's confidence declines noticeably. This further confirms the critical role of key tokens in predicting anger.

=== Example 9 ===

Gold emotion: anger

Russian text: Но я, если честно, руками есть не люблю.

English text: but to be honest i don t like my hands

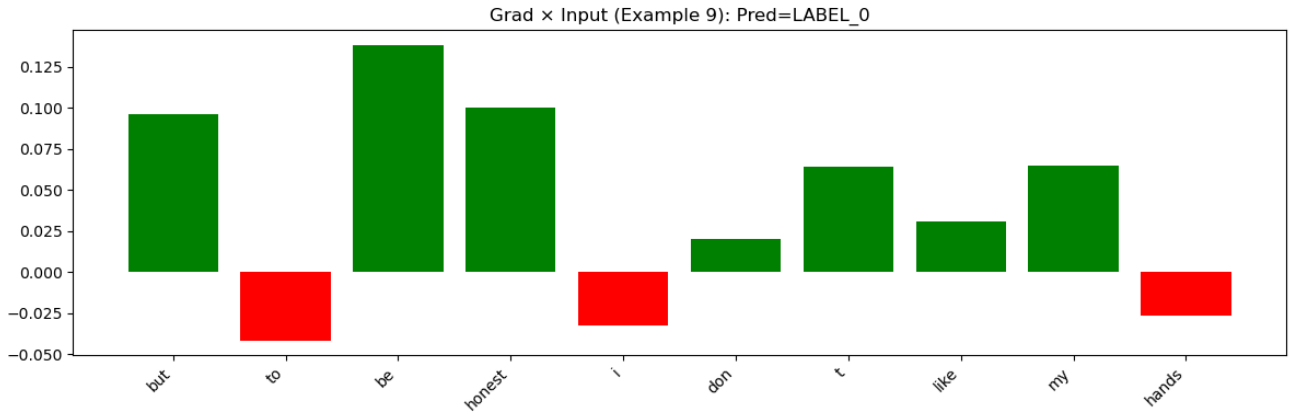

Figure i1

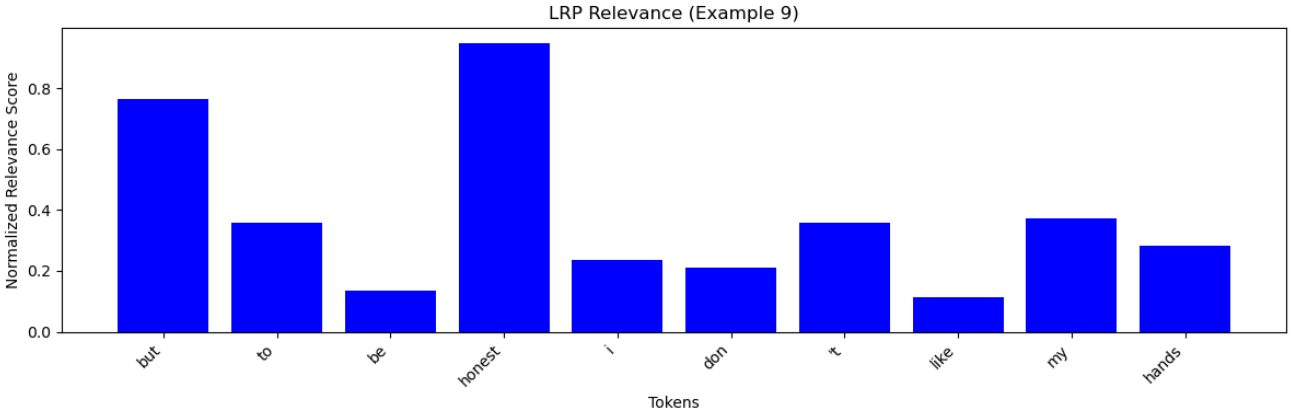

Figure i2

Figure i3

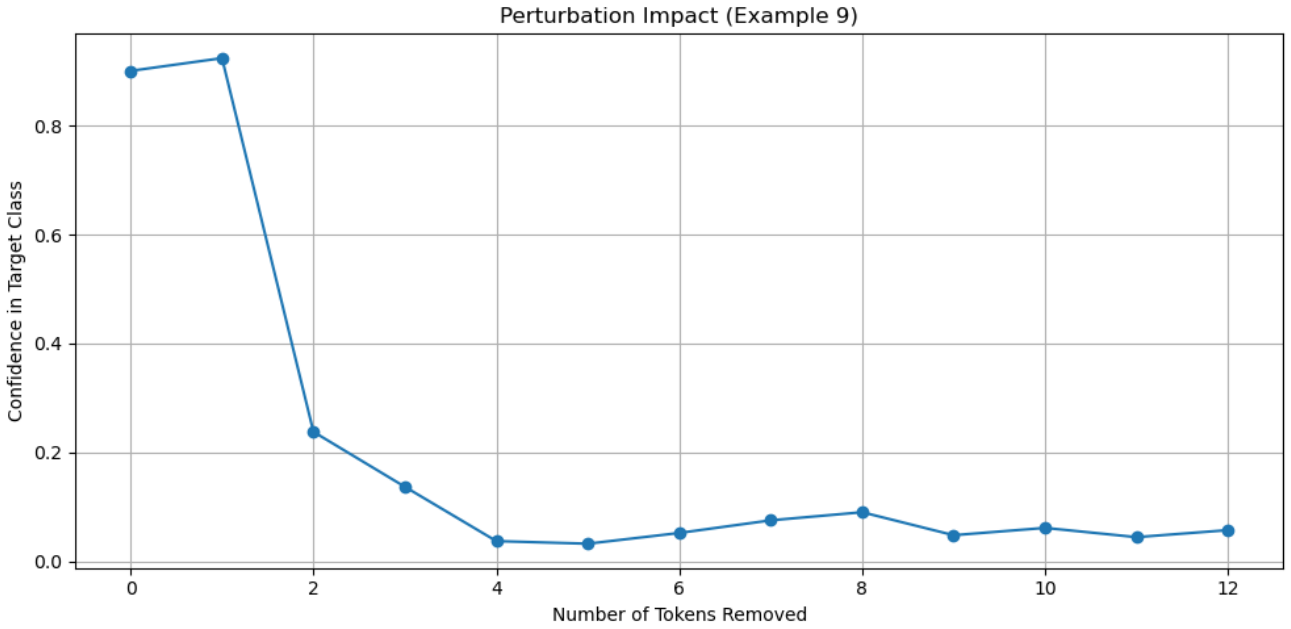

Figure i4

The sentence "but to be honest i don't like my hands" reveals that the Gradient × Input visualization (Figure i1) indicates moderate importance in most tokens. The LRP analysis (Figures i2 and i3) focuses particularly on the phrases "don't like" and "hands." The perturbation experiment (Figure i4) shows that removal of these tokens results in a clear drop in confidence, indicating that the model utilizes these tokens to inform its prediction of anger.

- **Surprise Examples:**

Surprise examples often display a lower overall relevance in both Gradient × Input and LRP methods. While certain tokens (like "gender" or "mutual sympathy") are highlighted, the overall signal is less pronounced. The perturbation experiments show a gradual decline in confidence, suggesting that the model does not depend on any single token to predict surprise. This may indicate that the model's surprise prediction is more distributed or ambiguous.

=== Example 10 ===

Gold emotion: surprise

Russian text: Я его половую принадлежность даже не могу понять.

English text: i can not even understand his gender

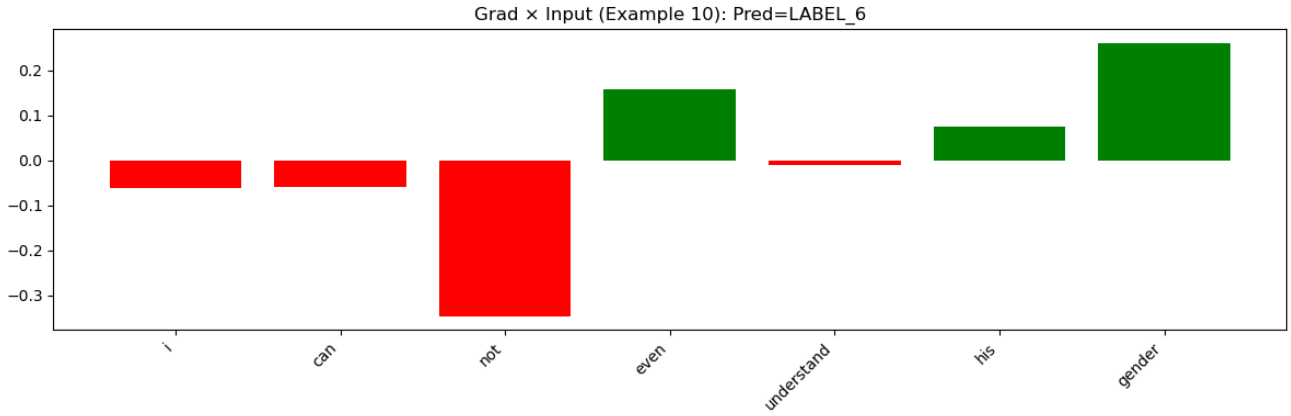

Figure j1

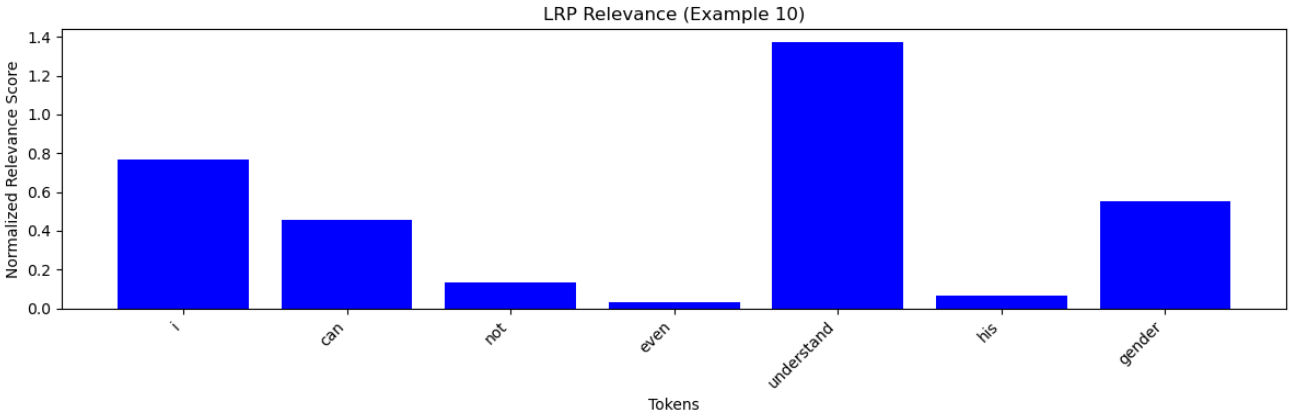

Figure j2

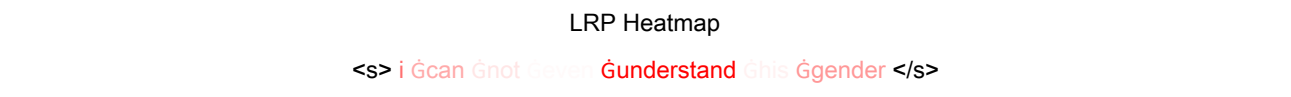

Figure j3

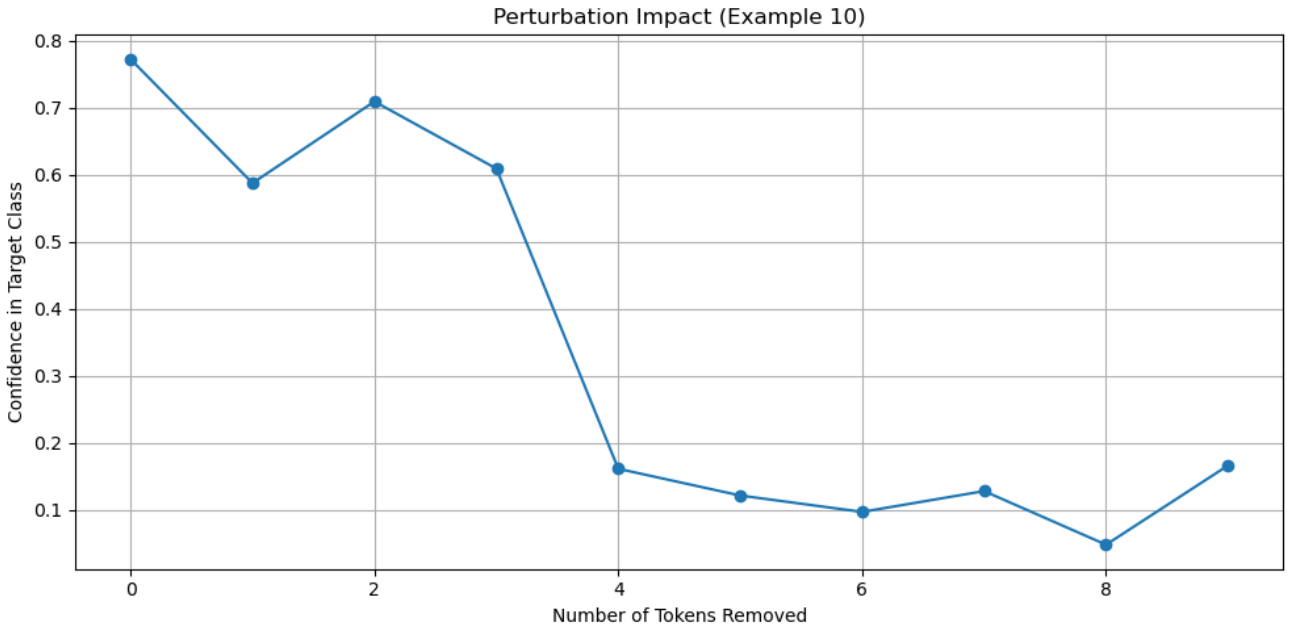

Figure j4

In the sentence "i can not even understand his gender," the Gradient × Input method (Figure f1) shows low to moderate relevance across the tokens. The LRP graphs (Figures f2 and f3) consistently highlight that the word "gender" is significant. The perturbation experiment (Figure f4) indicates that masking "gender" causes a notable decline in the confidence of the predicted class for surprise. Thus, the model appears to rely on this token to infer surprise.

=== Example 11 ===

Gold emotion: surprise

Russian text: что между ними возникла взаимная симпатия.

English text: that mutual sympathy arose between them

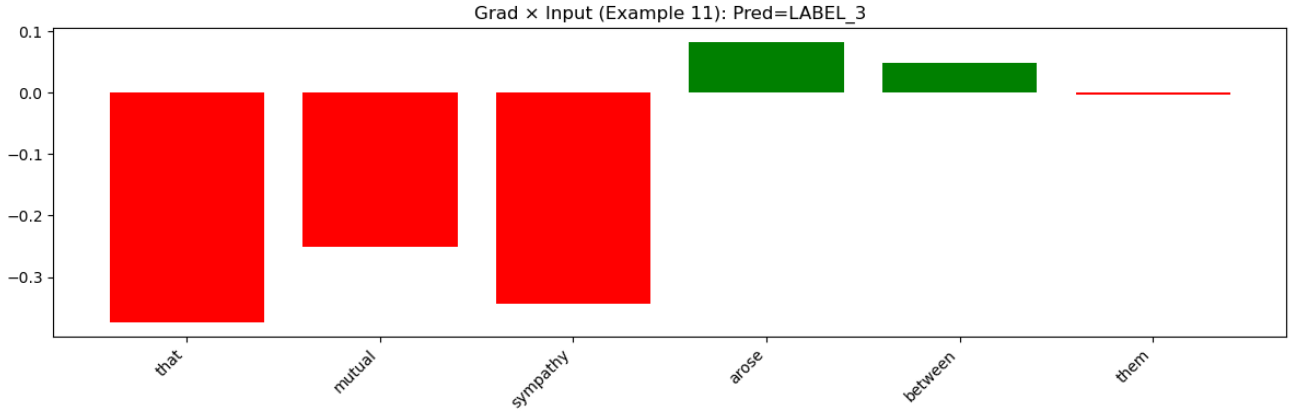

Figure k1

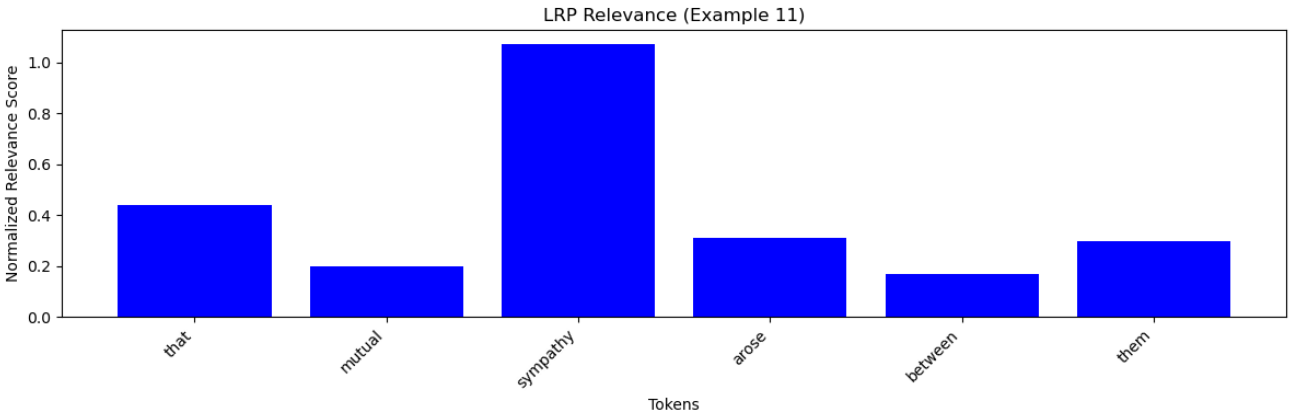

Figure k2

LRP Heatmap

<s> that Ġmutual Ġsympathy Ġarose Ġbetween Ġthem </s>

Figure k3

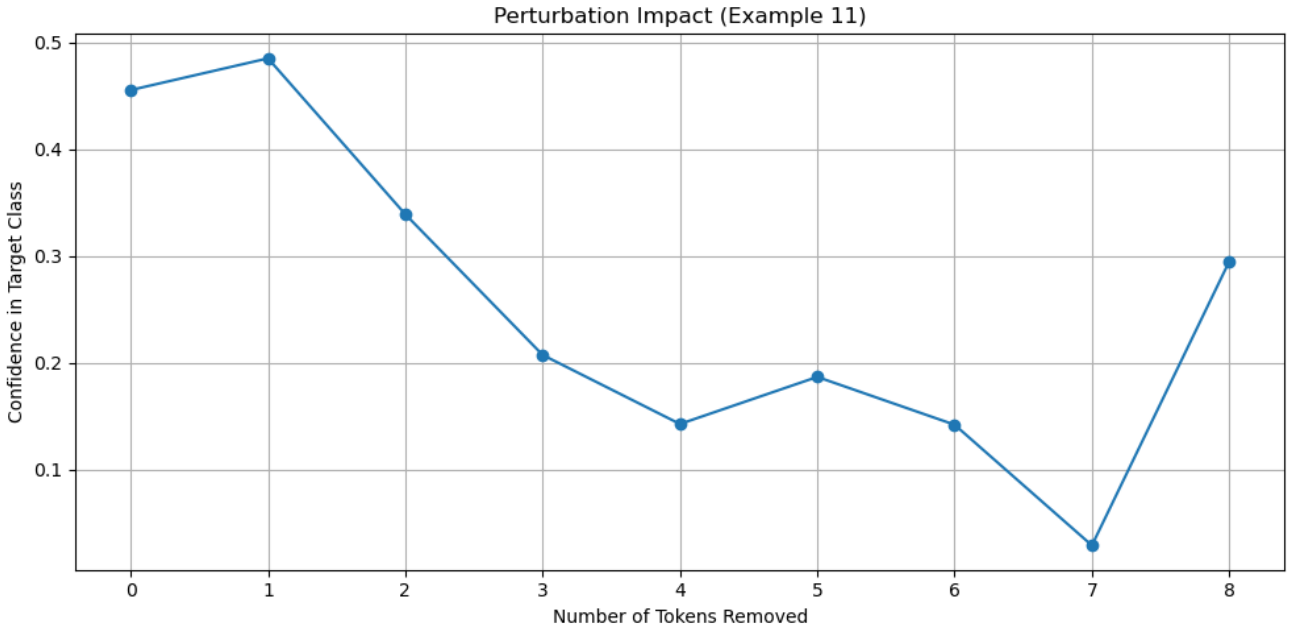

Figure k4

For "that mutual sympathy arose between them," the Gradient × Input graph (Figure k1) reveals mild signals, while the LRP method (Figures k2 and k3) concentrates relevance on tokens such as "mutual" and "sympathy." The perturbation curve (Figure k4) displays a moderate drop in confidence, suggesting that although the overall tone is ambiguous, the key token "sympathy" influences the prediction for surprise.

=== Example 12 ===

Gold emotion: surprise

Russian text: Вы знаете, я немножечко о каждом уже кое-что знаю.

English text: you know i already know something a little about everyone

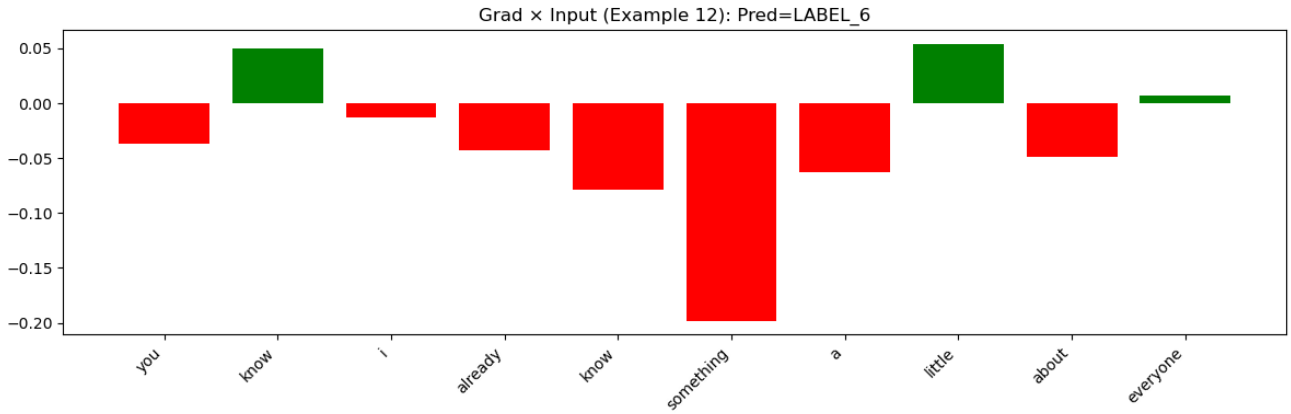

Figure l1

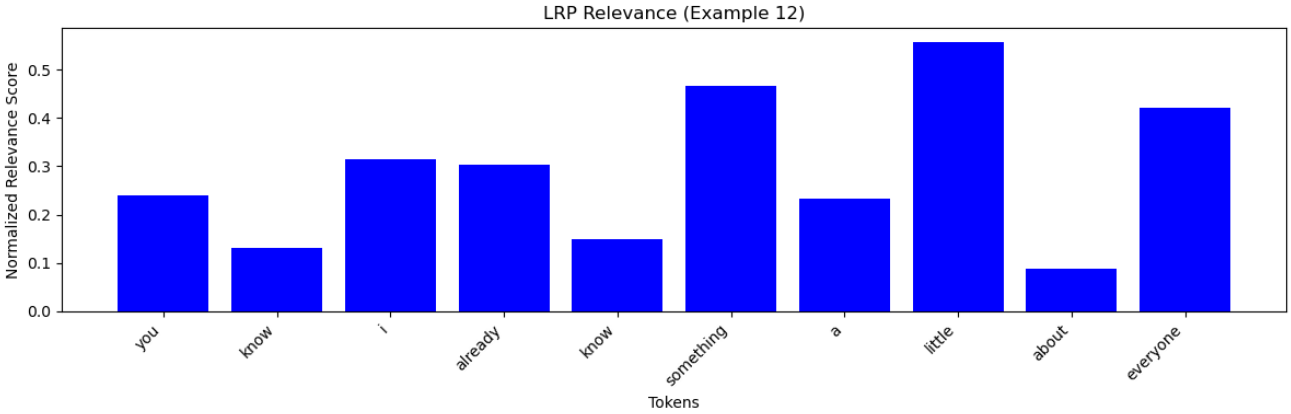

Figure l2

LRP Heatmap

<s> you Ġknow Ġi Ġalready Ġknow Ġsomething Ġa Ġlittle Ġabout Ġeveryone </s>

Figure l3

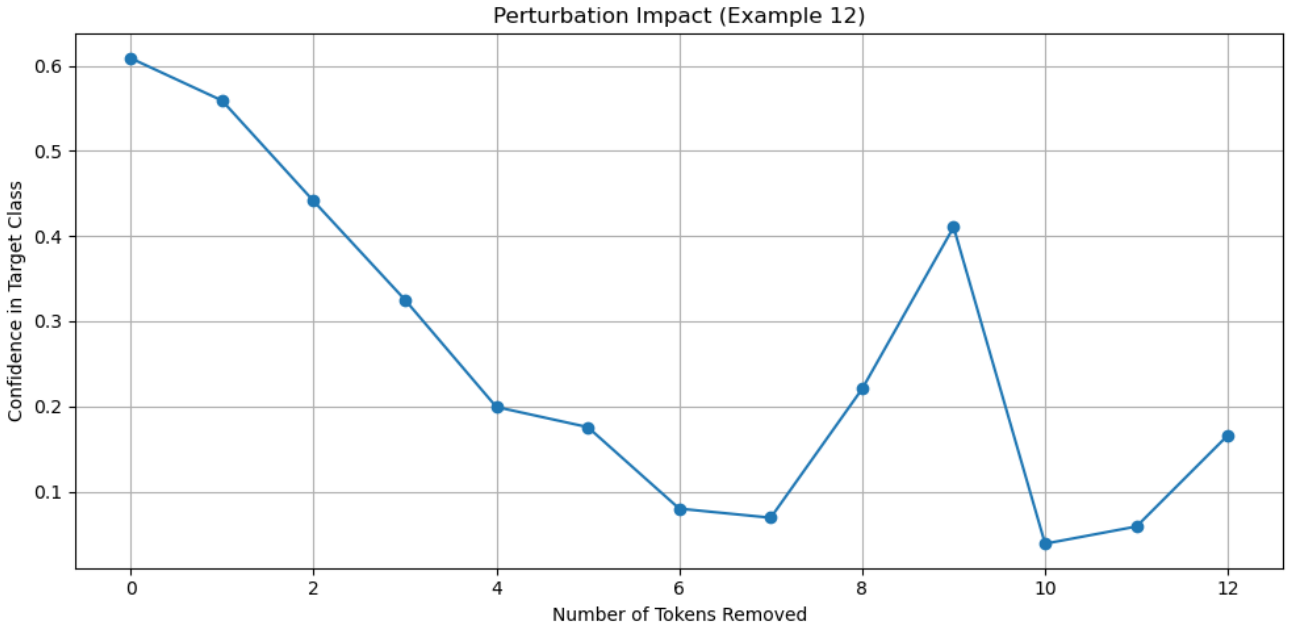

Figure l4

In the sentence "you know i already know something a little about everyone," the Gradient × Input method (Figure l1) shows weak to moderate relevance in each token. The LRP visualizations (Figures l2 and l3) generally suggest a uniform distribution, without any single token dominating. The perturbation experiment (Figure l4) indicates only a gradual drop in confidence as tokens are removed, which is expected for a sentence that is more factual in tone. This example aligns with a prediction of surprise, where no strong emotional cue is present.

– **Fear Examples:**

For fear, the XAI methods identify that tokens such as "afraid" or phrases like "did not know who would come" play a crucial role. The LRP heatmap clearly shows these tokens as having high relevance, and the perturbation curves exhibit a sharp decline in confidence when they are removed. This consistent pattern across fear examples confirms that the model heavily relies on a few critical indicators for fear.

=== Example 13 ===

Gold emotion: fear

Russian text: Скажу откровенно, я не знала ведь, кто ко мне придет,

English text: frankly i did not know who would come to me

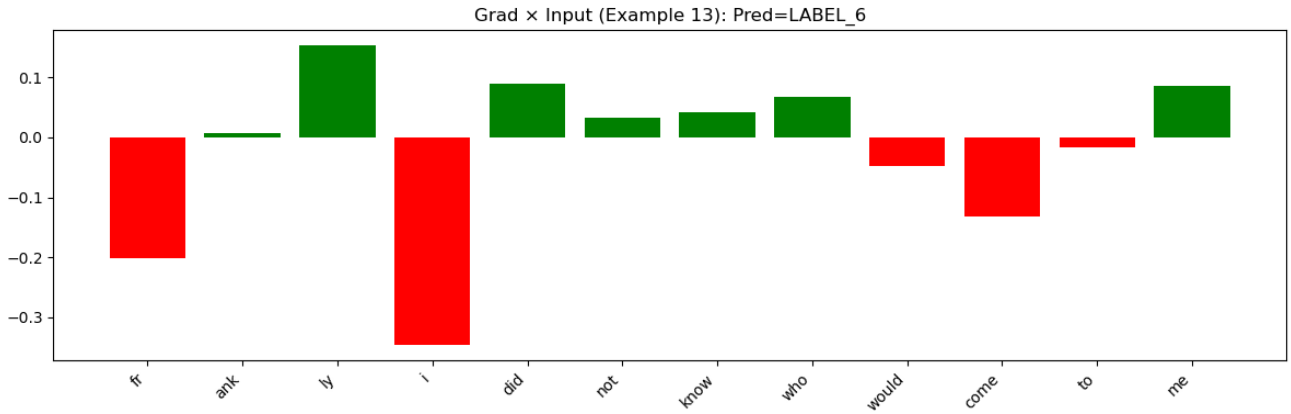

Figure m1

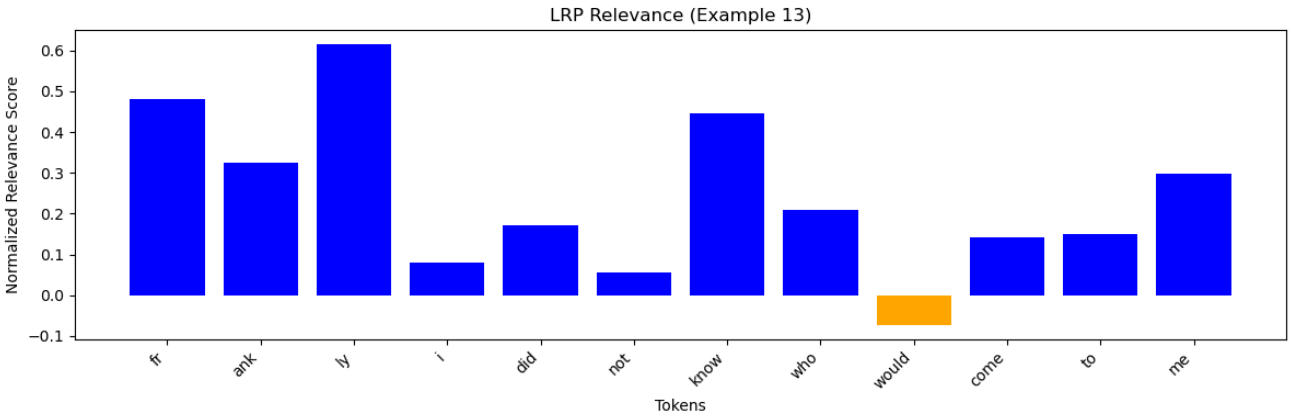

Figure m2

Figure m3

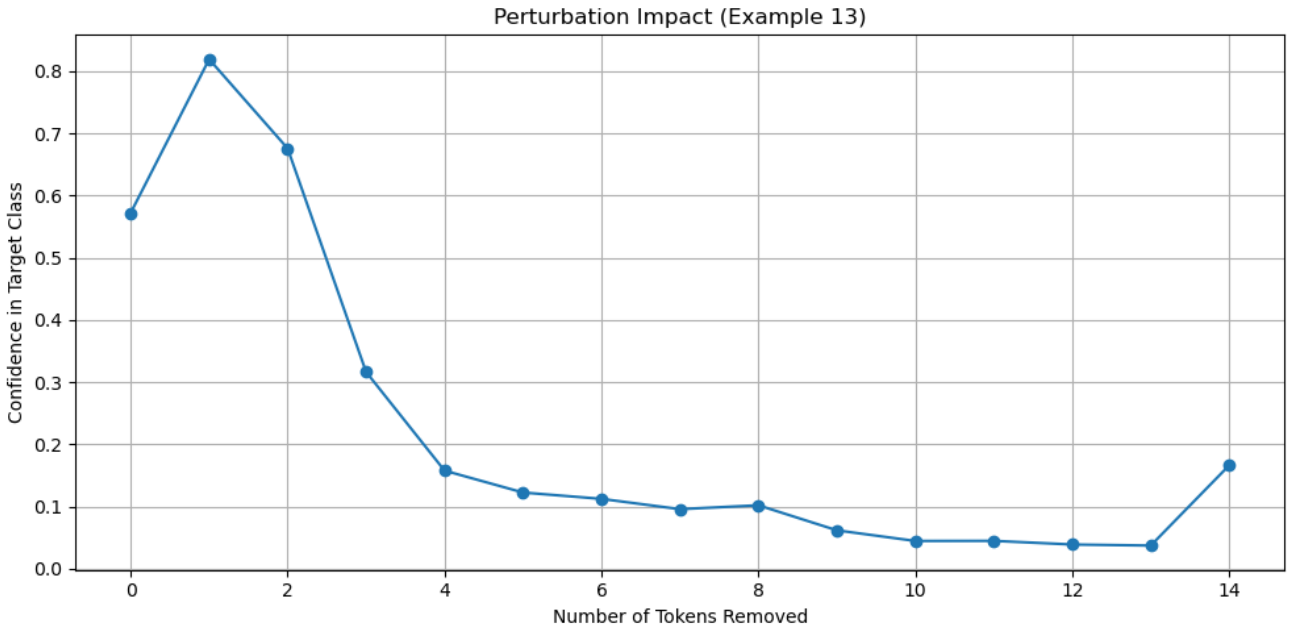

Figure m4

In the sentence "frankly i did not know who would come to me," the Gradient × Input method (Figure m1) shows mixed signals across tokens. The LRP method (Figures m2 and m3) reveals that tokens like "did not know" and "come to me" have higher relevance. The perturbation graph (Figure m4) demonstrates a sharp confidence drop when those key tokens are removed. This indicates that the model relies on those critical components to predict fear.

=== Example 14 ===

Gold emotion: fear

Russian text: Я боюсь к нему идти.

English text: i m afraid to go to him

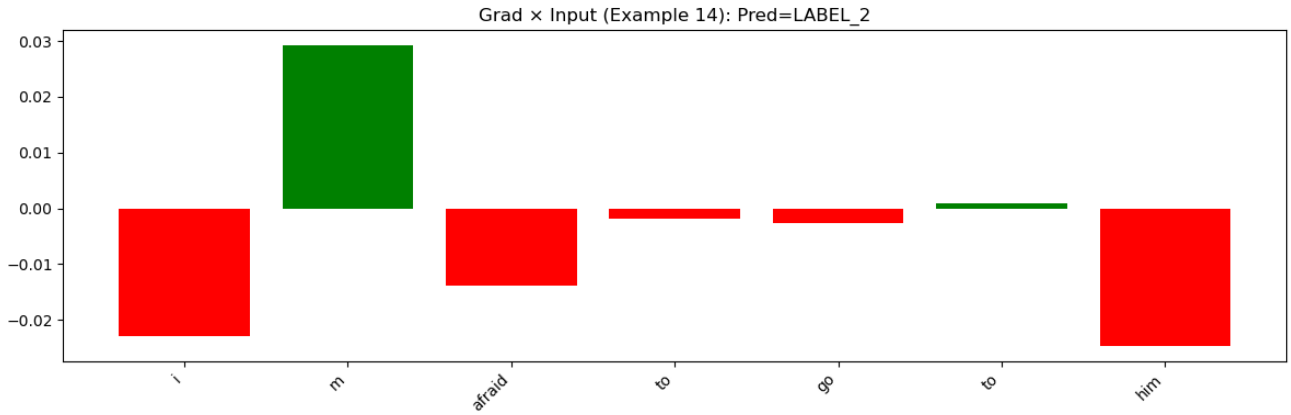

Figure n1

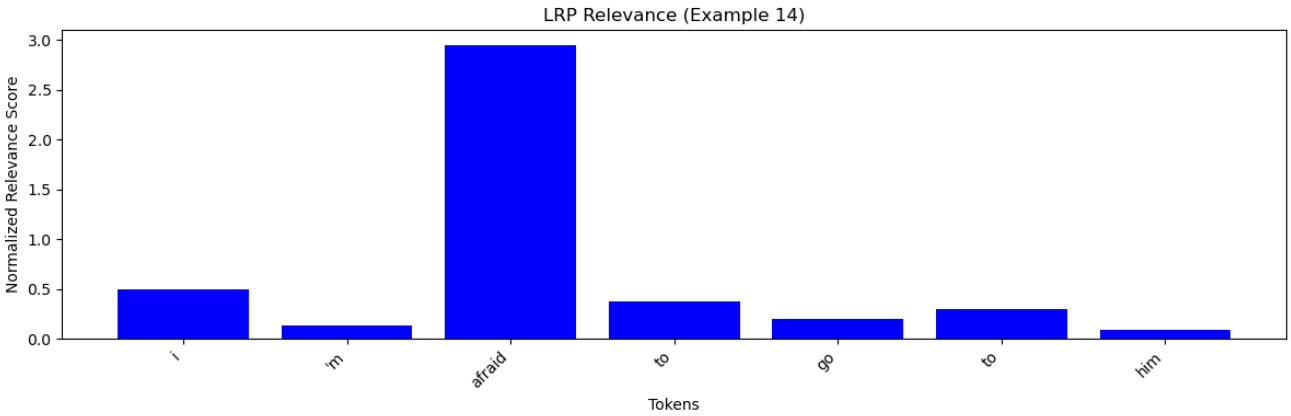

Figure n2

Figure n3

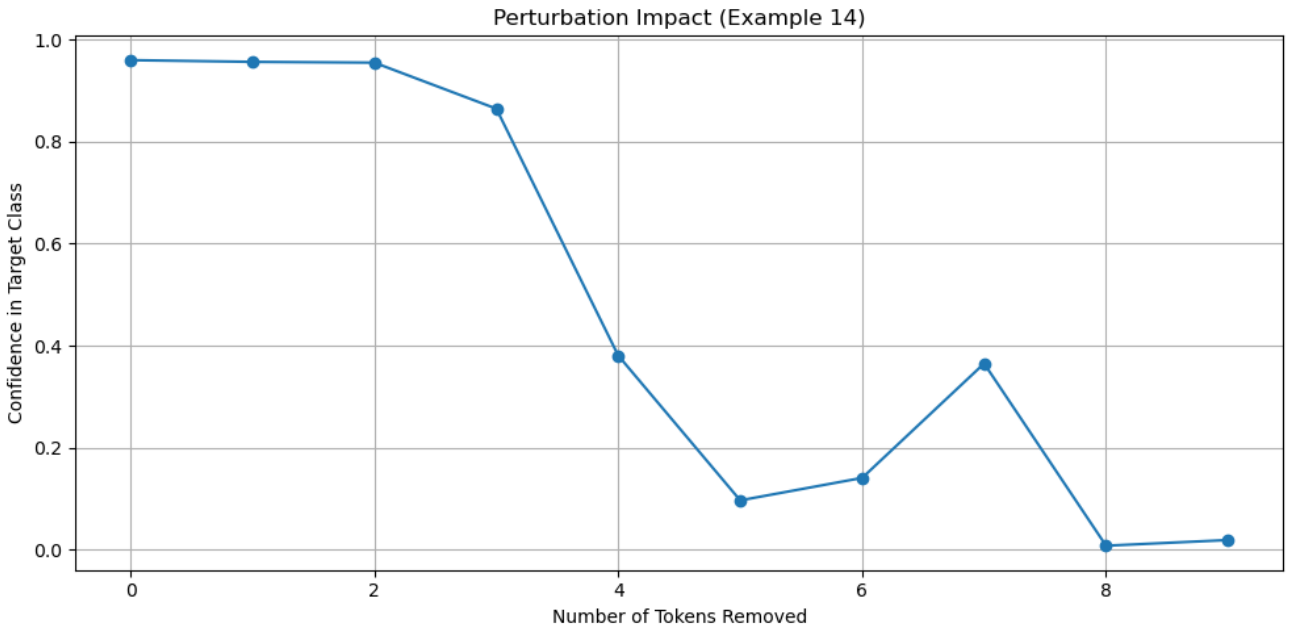

Figure n4

For "i'm afraid to go to him," the Gradient × Input method (Figure n1) exhibits a relatively consistent relevance distribution, while the LRP graphs (Figures n2 and n3) emphasize the token "afraid" prominently. The perturbation experiment (Figure n4) shows a sudden drop in model confidence when "afraid" is masked. This confirms that the model uses the token "afraid" as a strong signal for fear.

=== Example 15 ===

Gold emotion: fear

Russian text: Очень вкусно, но я боюсь, что там

English text: very tasty but i m afraid that there

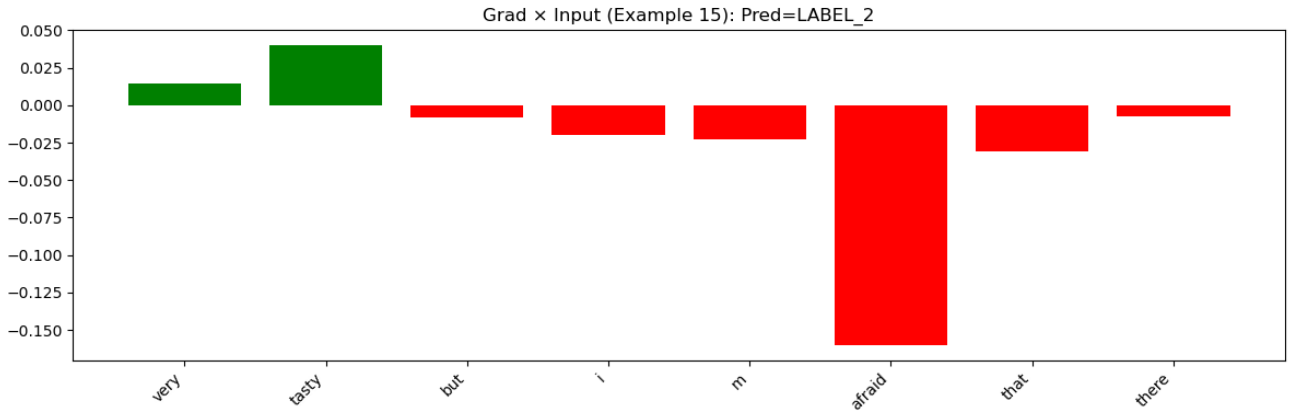

Figure o1

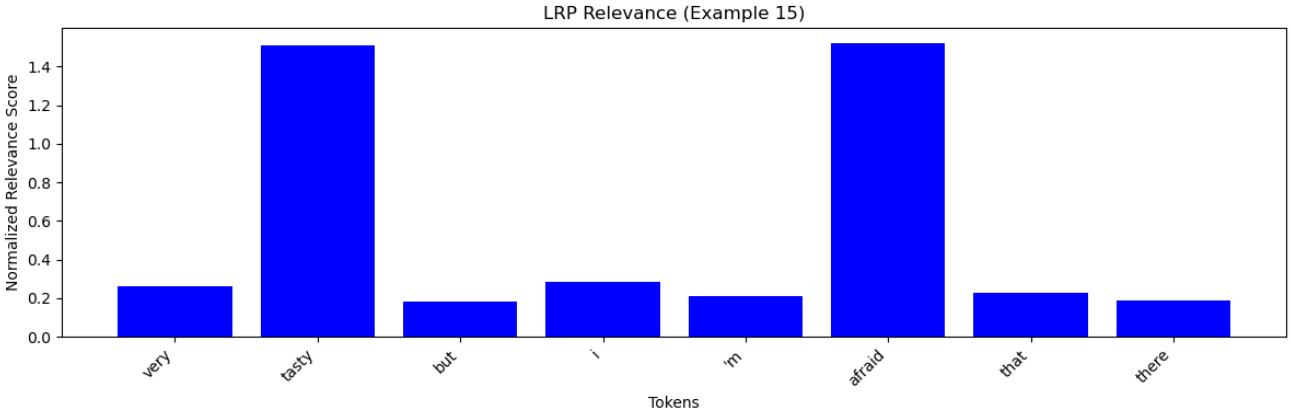

Figure o2

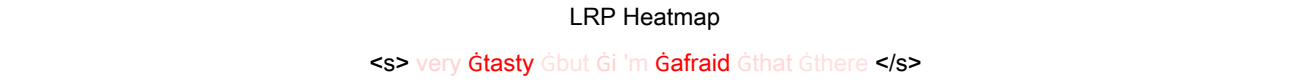

Figure o3

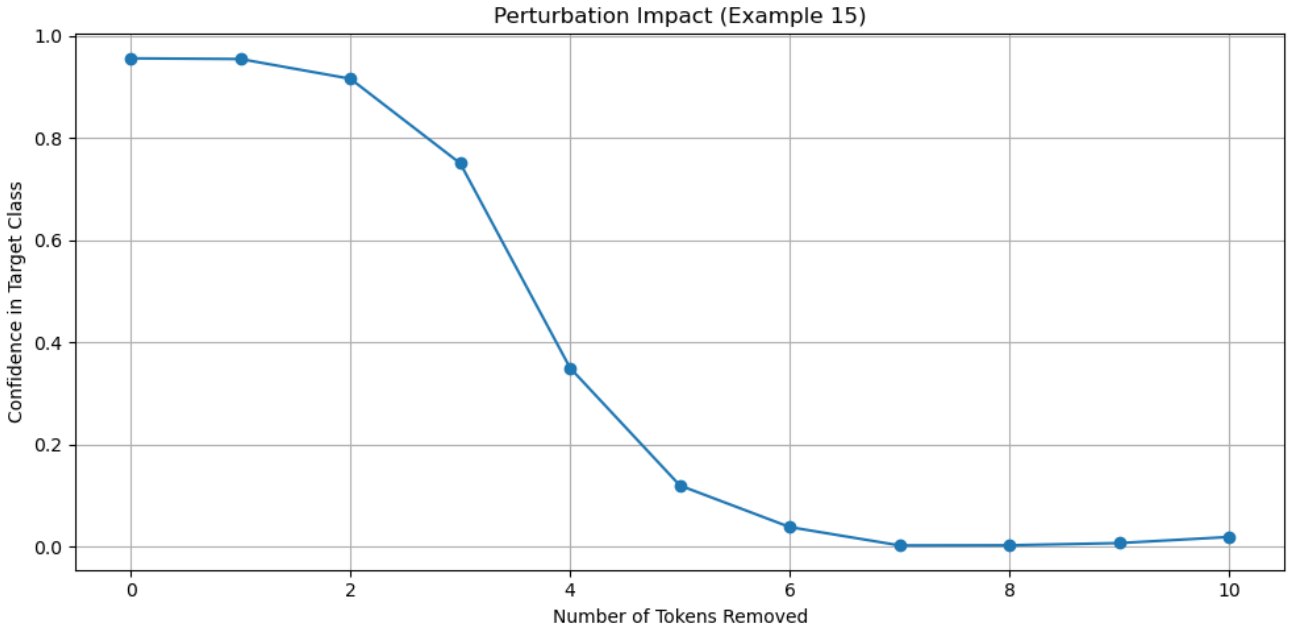

Figure o4

In this example, "very tasty but i'm afraid that there" yields mixed results in the Gradient × Input graph (Figure o1), where the descriptive parts and the word "afraid" vary in importance. The LRP visualizations (Figures o2 and o3) indicate that while the phrase "very tasty" contributes positively, the word "afraid" is particularly emphasized as critical. The perturbation plot (Figure o4) shows a clear decline in confidence upon removal of "afraid," suggesting that despite the positive adjective, the fear component drives the model's final prediction.

– **Disgust Examples:**

In disgust examples, the LRP and heatmap visualizations consistently emphasize negative descriptors (e.g., "clogged," "acute rejection"). The Gradient × Input signals in these cases are more variable, but the improved LRP method provides a clear focus. Perturbation experiments also show that removal of the key tokens results in a significant drop in confidence. Overall, the model's predictions for disgust are well-explained by the XAI methods.

=== Example 16 ===

Gold emotion: disgust

Russian text: Желудок-то забитый у вас у всех.

English text: the stomach is clogged with you all

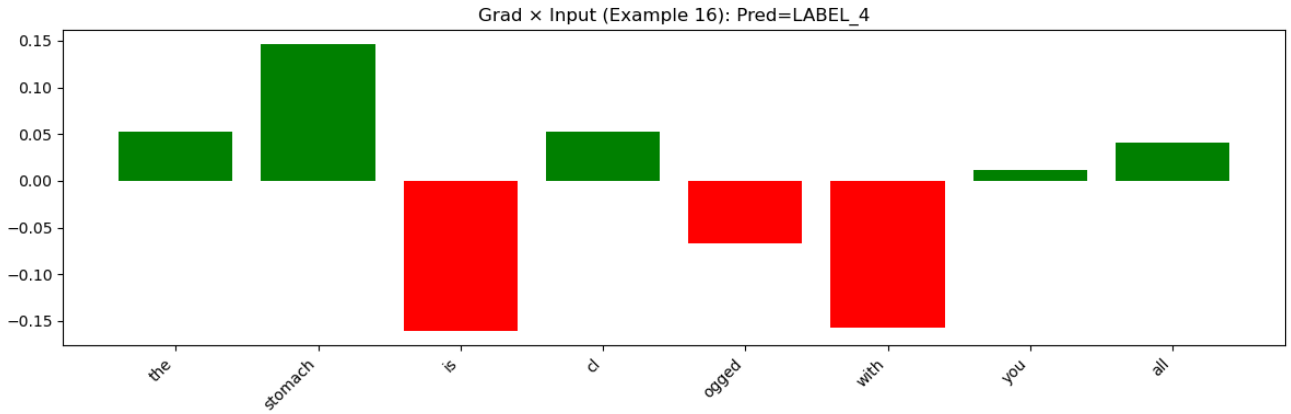

Figure p1

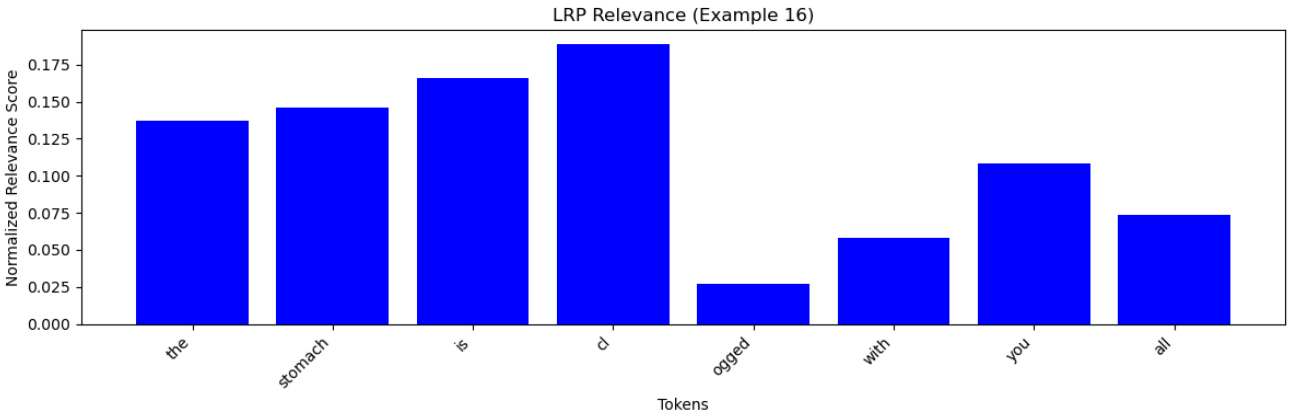

Figure p2

LRP Heatmap

<s> the Ġstomach Ġis Ġcl ogged Ġwith Ġyou Ġall </s>

Figure p3

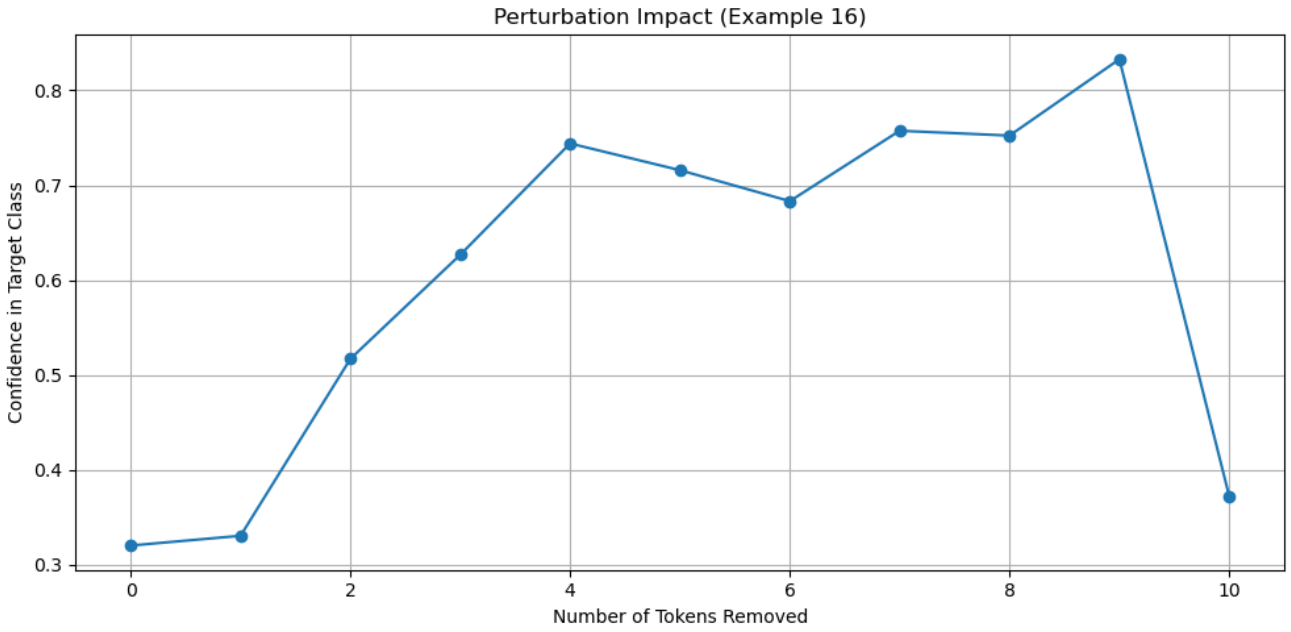

Figure p4

For "the stomach is clogged with you all," the Gradient × Input method (Figure p1) suggests that most tokens have a low or moderate contribution. The LRP method (Figures p2 and p3) highlights that the token corresponding to "clogged" holds notable relevance. The input perturbation graph (Figure p4) confirms that the removal of this key descriptive word causes the model's confidence to drop considerably. This observation supports the prediction of disgust.

=== Example 17 ===

Gold emotion: disgust

Russian text: острое неприятие

English text: acute rejection

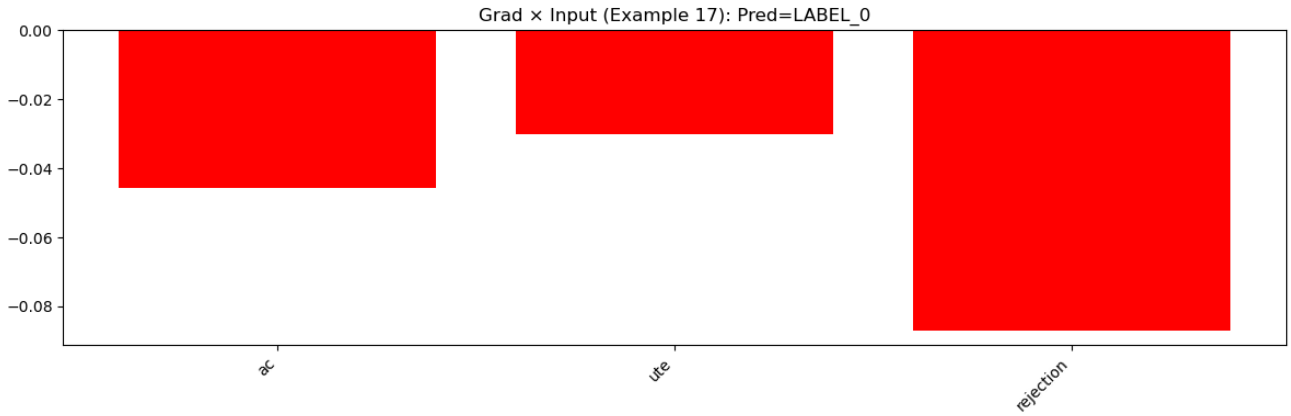

Figure q1

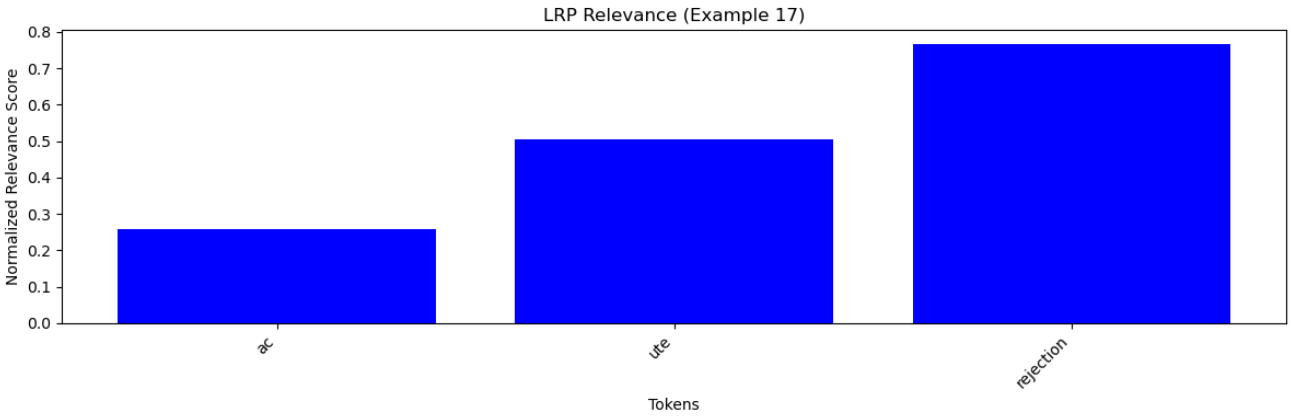

Figure q2

Figure q3

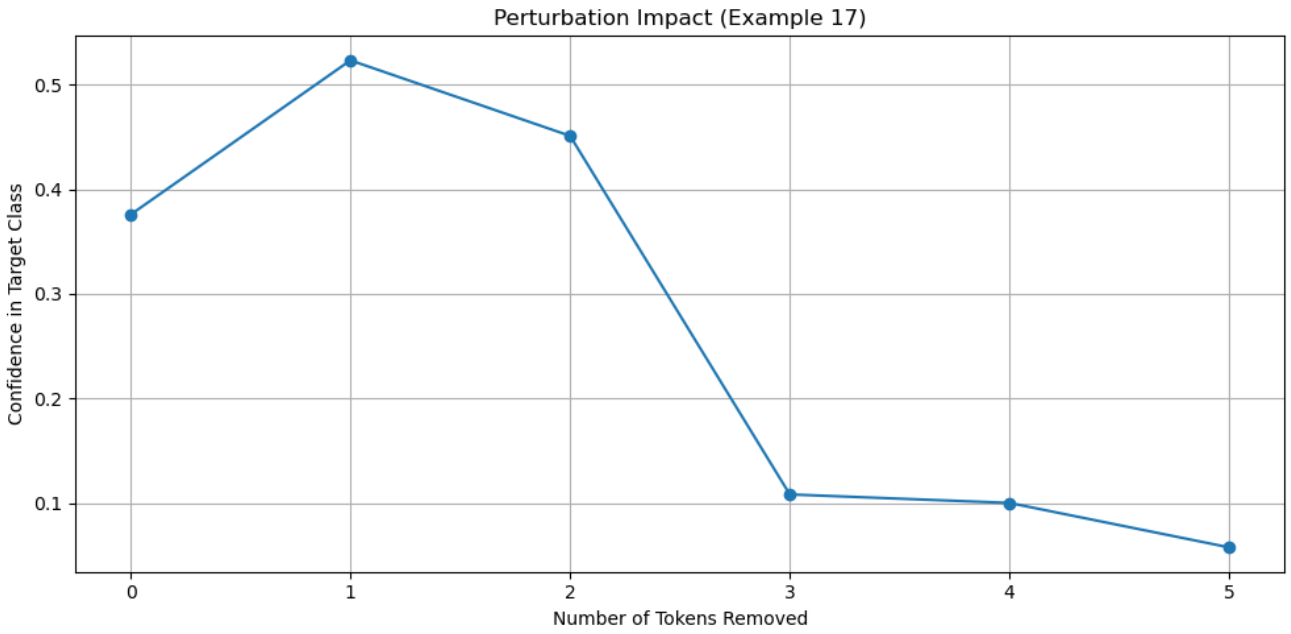

Figure q4

In the sentence "acute rejection," the Gradient × Input graph (Figure q1) shows a minimal relevance on average with some variation. The LRP results (Figures q2 and q3) clearly indicate that both tokens are given strong positive relevance. The perturbation analysis (Figure q4) reveals that masking either token leads to a swift drop in confidence, indicating that the model's prediction of disgust is highly dependent on these specific emotionally charged words.

=== Example 18 ===

Gold emotion: disgust

Russian text: того блюда, которое тебе подали.

English text: the dishes that you were served

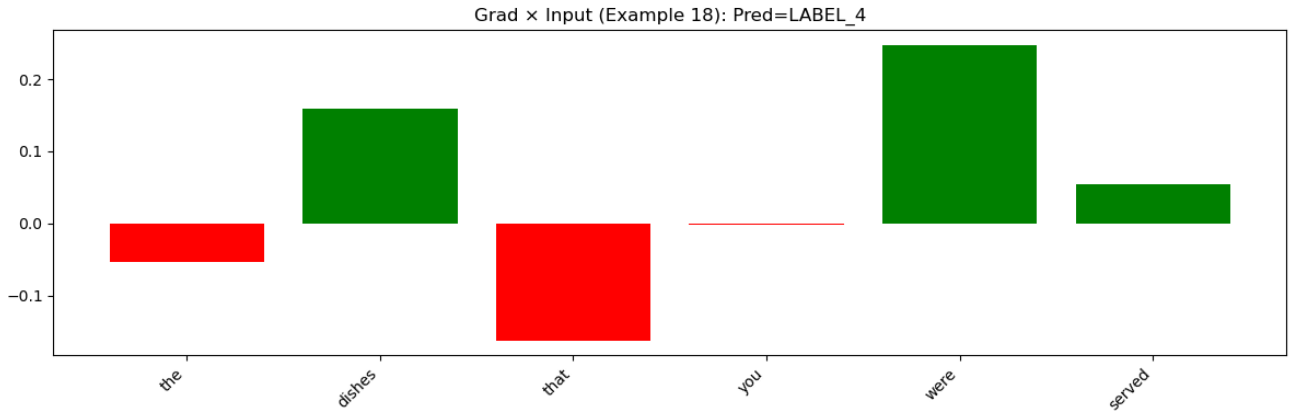

Figure r1

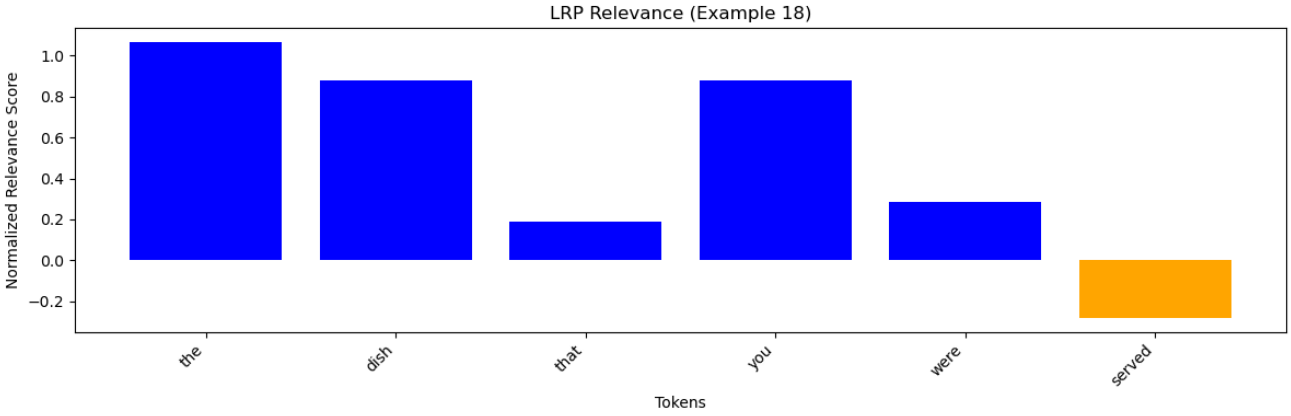

Figure r2

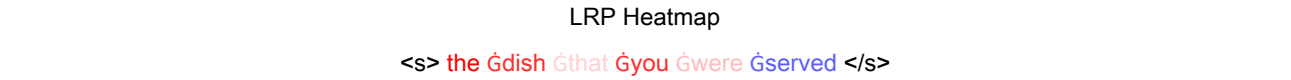

Figure r3

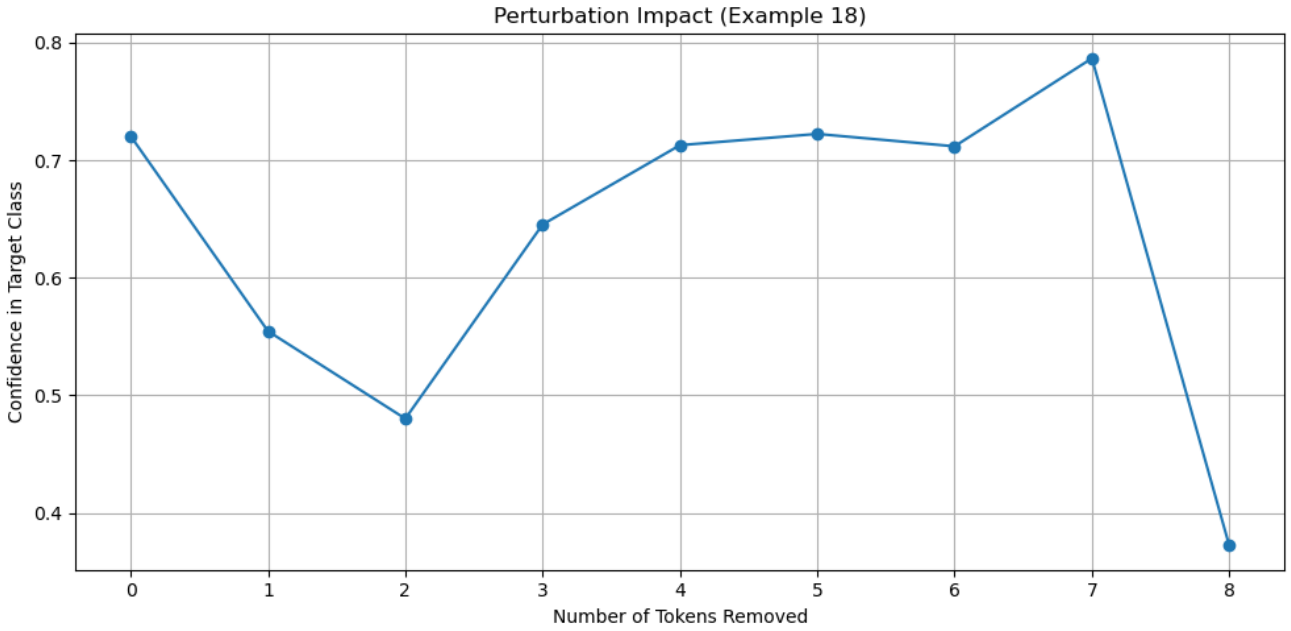

Figure r4

Finally, for "the dish that you were served," the Gradient × Input method (Figure r1) indicates a diverse contribution from the tokens, but the LRP method (Figures r2 and r3) sharply emphasizes the descriptive tokens in the sentence. The perturbation experiment (Figure r4) demonstrates that when these key tokens are removed, the model's confidence in predicting disgust falls sharply. This again shows that certain descriptors drive the model's prediction in this case.

Conclusion

Overall, the XAI experiments provide a valuable insight into how our Transformer-based emotion classification model makes its decisions. Both the Gradient × Input and the improved LRP methods highlight that the model tends to rely on a small set of emotionally charged tokens—especially in cases of happiness, anger, fear, and disgust. In contrast, for emotions like surprise, the reliance appears more distributed.

Furthermore, the perturbation analysis reveals that the model's confidence generally remains stable until a critical subset of tokens is removed, at which point the confidence drops sharply. This confirms that the model's decision-making process is sensitive to key tokens, which aligns with our expectations for an emotion classifier.

These findings suggest that our model captures the emotional content of text by focusing on a few pivotal words. The improved LRP method is particularly effective in filtering out noise and providing a clearer interpretation compared to the basic Gradient × Input method. As a result, these XAI techniques help us understand both the strengths and potential weaknesses of our model. In future work, these insights might guide us in refining the model further to achieve even better interpretability and performance.

In summary, the XAI methods applied in this study help demystify the internal behavior of our emotion classification model. They show that while the model's predictions are largely driven by a few significant tokens, some emotions such as surprise exhibit a more distributed reliance on multiple tokens. These insights are crucial for understanding the model and informing further improvements in both model design and training.

Overall, this study improves our understanding of how individual tokens shape the decisions of our Transformer model in emotion classification and provides a roadmap for further improvements in model robustness and fairness.