Explainable AI for Transformers, team 22

Monika Stangenberg (213648), Michał Bątkowski (232079), Kamil Łęga (232752)

Part 1

I chose 21 sentences from the test set from my group. There are 3 sentences per class.

Daniel Klusek, welcome to the Promenada studio. We're going.

correct class: neutral, predicted class: neutral, CORRECT

I mean I will swim. At the moment I am on a walk and we will swim at 2pm.

correct class: neutral, predicted class: happiness, WRONG

Good morning.

correct class: neutral, predicted class: neutral, CORRECT

2/3 sentences from correct class were predicted correctly. One was misclassified as happiness, probably because activity related words (swim walk).

Thank you for the interview.

correct class: happiness, predicted class: happiness, CORRECT

Thank you kindly.

correct class: happiness, predicted class: happiness, CORRECT

As tourists said, the sea is beautiful at any time of the year and so rough too. So let's look how beautiful it is. Or maybe you can still take a bath this summer.

correct class: happiness, predicted class: happiness, CORRECT

The model correctly predicted all 3 happiness sentences, showing strong performance on polite expressions and positive, descriptive language.

What to do when you can't sunbathe and bathe? We will ask tourists about this.

correct class: surprise, predicted class: neutral, WRONG

Where did you come from?

correct class: surprise, predicted class: neutral, WRONG

Does the end of summer holidays mean the end of the season?

correct class: surprise, predicted class: neutral, WRONG

The model misclassified all 3 surprise sentences as neutral. It failed to recognize question structures.

The Lord knows that this is a jerk.

correct class: disgust, predicted class: anger, WRONG

Well, he can wish you not to start.

correct class: disgust, predicted class: anger, WRONG

Well, they are like that, because I understand that he is not right and just

correct class: disgust, predicted class: anger, WRONG

The model got all 3 disgust sentences wrong, predicting anger instead. It seems to react to negative words but can’t tell the difference between anger and disgust.

It is a pity to bear it by mowing him in the ass.

correct class: anger, predicted class: sadness, WRONG

When I look at him that this is a bandit, not a minister.

correct class: anger, predicted class: anger, CORRECT

That he robs me because he gives me what they haven't worked after 35 years like me.

correct class: anger, predicted class: sadness, WRONG

The model predicted 1/3 sentences correctly. It confused anger with sadness, probably because of words like *pity* and *because* that sound emotional.

The Swedish, Soviet invasion will survive.

correct class: fear, predicted class: fear, CORRECT

But there is aggression in this man.

correct class: fear, predicted class: anger, WRONG

And today there is even nowhere to hide

correct class: fear, predicted class: surprise, WRONG

The model predicted 1/3 sentences correctly. It confused fear with anger and surprise, probably because it didn’t understand the context of words like aggression and hide.

Not beach is the weather, although the sun comes out, but not enough to sunbathing until today.

correct class: sadness, predicted class: neutral, WRONG

Well, but in the morning the rain was a bit.

correct class: sadness, predicted class: neutral, WRONG

When he was after a carol, he cried and said that just what was happening in the church now,

correct class: sadness, predicted class: sadness, CORRECT

The model predicted 1/3 sentences correctly. It confused sadness with neutral, possibly because the emotional tone was subtle.

Predicted classes:

anger

Tokens with big positive impact: /s, just, well, But

Tokens with big negative impact: Lord, jerk, can, to, start, they, is, band, “’”, minister, aggression

disgust

No predicted sentences for this class

fear (1 predicted sentence in this class)

Tokens with big positive impact: survive

Tokens with big negative impact: The, invasion, will

happiness

Tokens with big positive impact: tourists, look, maybe, summer, kindly, interview, “.”, am

Tokens with big negative impact: beautiful, rough, too, “.”, Thank, you, swim, walk, pm

neutral

Tokens with big positive impact: Kl, Prom, ada, studio, /s, “.”, What, ‘t, about, ?, Does, end, of, beach, until, rain

Tokens with big negative impact: Welcome, morning, sun, Where, from, ?, /s, not, is, weather, well, in, the

sadness

Tokens with big positive impact: It, pity, owing, ass, ro, bs, because, 35, car, cried, now

Tokens with big negative impact: bear, m, he, ‘t, after, When, church

surprise

Tokens with big positive impact: /s

Tokens with big negative impact: there, hide

Part 2

Daniel Klusek, welcome to the Promenada studio. We’re going.

correct class: neutral, predicted class: happiness, WRONG

confidence: 0.75

models pays a lot of attention to “welcome”

I mean I will swim. At the moment I am on a walk and we will swim at 2pm.

correct class: neutral, predicted class: happiness, WRONG

confidence: 0.74

models pays a lot of attention to “we will”

Good morning.

correct class: neutral, predicted class: neutral, CORRECT

confidence: 0.65

models pays a lot of attention to “.”

Thank you for the interview.

correct class: happiness, predicted class: happiness, CORRECT

confidence: 0.96

models pays a lot of attention to “Thank you for”

Thank you kindly.

correct class: happiness, predicted class: happiness, CORRECT

confidence: 0.96

models pays a lot of attention to “Thank you kindy”

As tourists said, the sea is beautiful at any time of the year and so rough too. So let's look how beautiful it is. Or maybe you can still take a bath this summer.

correct class: happiness, predicted class: happiness, CORRECT

confidence: 0.88

models pays a lot of attention to “beautiful”

What to do when you can't sunbathe and bathe? We will ask tourists about this.

correct class: surprise, predicted class: neutral, WRONG

confidence: 0.74

models pays a lot of attention to “ask” and “tourist”

Where did you come from?

correct class: surprise, predicted class: neutral, WRONG

confidence: 0.83

models pays a lot of attention to “you” and “?”

Does the end of summer holidays mean the end of the season?

correct class: surprise, predicted class: neutral, WRONG

confidence: 0.70

models pays a lot of attention to “Does”

The Lord knows that this is a jerk.

correct class: disgust, predicted class: anger, WRONG

confidence: 0.89

models pays a lot of attention to “jerk”

Well, he can wish you not to start.

correct class: disgust, predicted class: neutral, WRONG

confidence: 0.81

models pays a lot of attention to “start”

Well, they are like that, because I understand that he is not right and just.

correct class: disgust, predicted class: neutral, WRONG

confidence: 0.60

models pays a lot of attention to “understand”

It is a pity to bear it by mowing him in the ass.

correct class: anger, predicted class: neutral, WRONG

confidence: 0.36

models pays a lot of attention to “the ass”

When I look at him that this is a bandit, not a minister.

correct class: anger, predicted class: anger, CORRECT

confidence: 0.42

models pays a lot of attention to “band it”

That he robs me because he gives me what they haven't worked after 35 years like me.

correct class: anger, predicted class: anger, CORRECT

confidence: 0.81

models pays a lot of attention to “bs”

The Swedish, Soviet invasion will survive.

correct class: fear, predicted class: neutral, WRONG

confidence: 0.77

models pays a lot of attention to “.”

But there is aggression in this man.

correct class: fear, predicted class: anger, WRONG

confidence: 0.92

models pays a lot of attention to “aggression”

And today there is even nowhere to hide.

correct class: fear, predicted class: surprise, WRONG

confidence: 0.72

models pays a lot of attention to “nowhere”

Not beach is the weather, although the sun comes out, but not enough to sunbathing until today.

correct class: sadness, predicted class: neutral, WRONG

confidence: 0.74

models pays a lot of attention to “.”

Well, but in the morning the rain was a bit.

correct class: sadness, predicted class: neutral, WRONG

confidence: 0.86

models pays a lot of attention to “.”

When he was after a carol, he cried and said that just what was happening in the church now,

correct class: sadness, predicted class: neutral, WRONG

confidence: 0.49

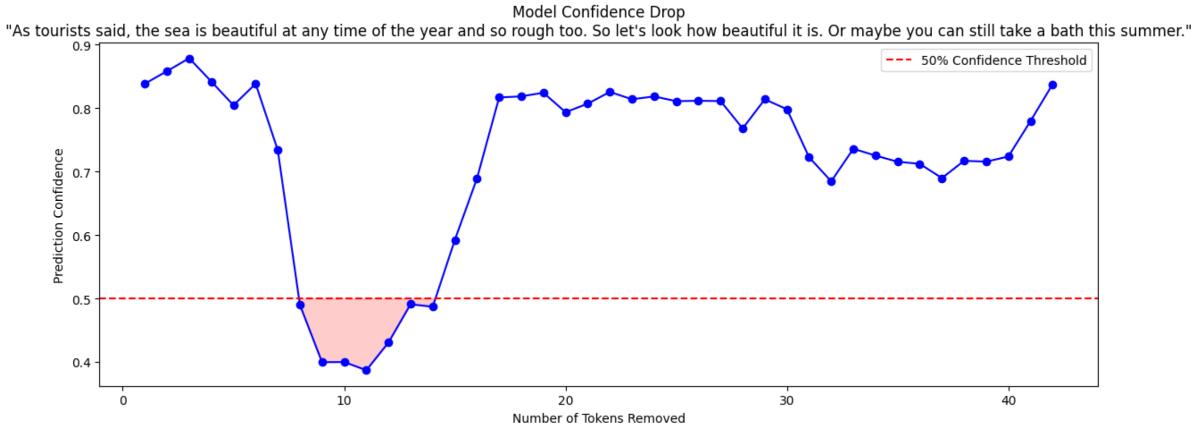
models pays a lot of attention to “,”

Conservative Propagation improved the interpretability of the model. Emotional words like “thank you”, “beautiful” and “aggression” received higher attention. However, the model still confused some emotions (disgust vs. anger, fear vs. neutral) and sometimes focused on punctuation.

Part 3

Results of Model Robustness with Input Perturbation are interesting.

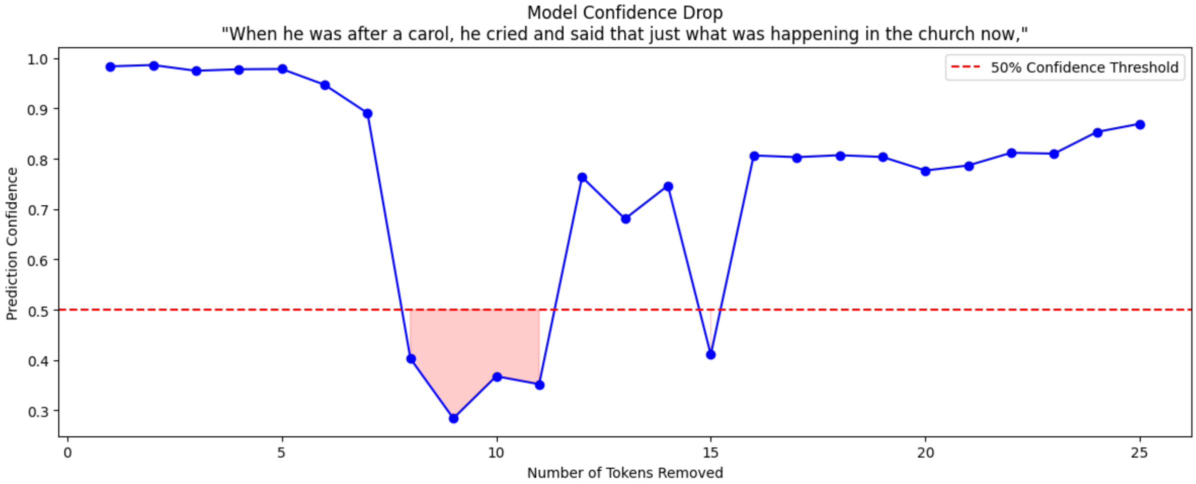
Example: (this is a correct prediction of happiness class)



The model loses confidence after removing 8 tokens but quickly recovers. This means it relies on key words but can still make accurate predictions using the rest of the sentence. It is robust because even when important words are missing, it can adapt.

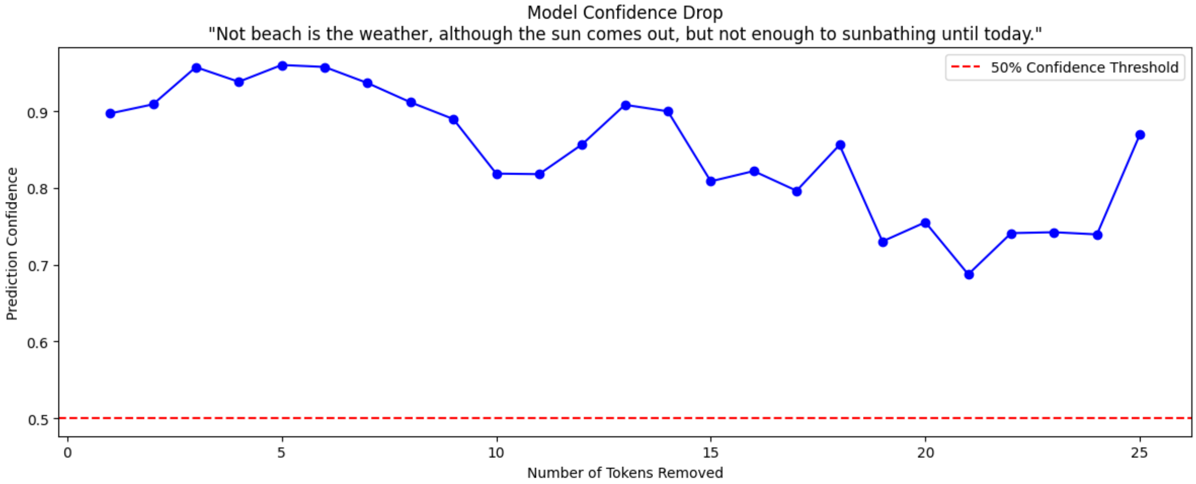
A similar pattern appears in most of the sentences. This shows that the model tends to rely on a few key words rather than distributing it’s attention equally. However, it also demonstrates the model's ability to recover quickly when those important parts are missing.

Example: (this is a incorrect prediction of sadness class)



Despite the fact that prediction is not correct, the model effectively recover after removing tokens which were very important for it's decision. The final confidence is lower that the beginning one, but the difference is around 10 percentage points.

Example: (this is a incorrect prediction of sadness class)



There are not a lot of sentences where we can clearly see that confidence is decreasing with every token missing. Yet, here we can see that the model stays confident even after removing many tokens. This shows it's not sensitive to individual words and makes predictions based on the full sentence.

Conclusion

The explainability methods showed that the model often focuses on a few key emotional words, which can help in correct predictions but they are also causing mistakes. Wrong predictions usually happen when the model misinterprets emotional tone or focuses too much on irrelevant tokens (punctuation). Input perturbation confirmed that while the model relies on specific words, it can still recover when they're removed.
Overall, the model is quite robust, but still struggles to fully understand subtle or similar emotions.