Explainability for Text Classifiers

Hello and welcome!

I'm Adrian Jaques Böck. This survey focuses on categorizing Counter speech and Non-Counter speech in social media, with the goal of making machine learning model predictions easier to understand for us humans.

The objective of this survey is to assess three methods that can help us understand the decision-making process of these text classifiers.

The survey will take you about 25 min.

Before each task, there will be a short introduction. **Please take a moment to read it carefully.**

Disclaimer:
This study includes exposure to hate speech for research purposes, aiming to combat online offensive language. The content may be distressing, so please consider your comfort level before participating. You may withdraw at any point without any consequences.
Your personal data is anonymized and treated with utmost confidentiality. By engaging in this activity, you acknowledge your comprehension and acceptance of the possibility of encountering hate speech and agree to share your data for research purposes.

Thank you for your time and contribution. Let's begin the survey.

There are 125 questions in this survey.

Infos

How old are you? *

Choose one of the following answers
Please choose **only one** of the following:

○ 17 or younger
○ 18 – 24
○ 25 – 34
○ 35 – 44
○ 45 – 54
○ 55 – 64
○ 65 – 74
○ 75 or older
○ Prefer not to answer

What is you gender? *

Choose one of the following answers
Please choose **only one** of the following:

○ Female
○ Male
○ Non-binary
○ A-gender
○ Other
○ Prefer not to say

Do you have a red-green color blindness? *

Choose one of the following answers
Please choose **only one** of the following:

○ Yes
○ No

Which of the following options best describes your current employment status? *

Choose one of the following answers
Please choose **only one** of the following:

○ Employed full-time (including freelance, self-employment etc.)
○ Employed part-time (including freelance, self-employment etc.)
○ Unemployed
○ Student
○ Prefer not to say

What is your highest level of education? *

Choose one of the following answers
Please choose **only one** of the following:

○ Secondary school
○ College or further education/ Bachelor’s Degree
○ Master’s Degree
○ Advanced Graduate work or PhD
○ Other
○ Prefer not to say

How would you rate your knowledge in AI from 1-5? (1= no knowledge, 5 = expert) *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

How would you rate your knoweledge in Explainable AI from 1-5 (1 = no knowledge, 5 = expert) *

Please choose **only one** of the following:

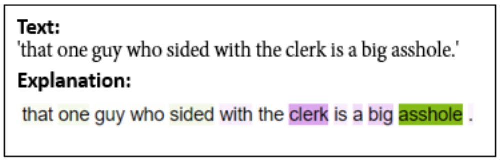
○ 1
○ 2
○ 3
○ 4
○ 5

Task 1 - Introduction

Task 1

You will see 5 texts and have to guess if each is classified by an AI as "Non-Counter speech " or "Counter speech".
Each text appears 4 times, with 3 accompanied by explanations and one without.

Here's an example:



Text:
'that one guy who sided with the clerk is a big asshole.'
Explanation:
that one guy who sided with the clerk is a big asshole .

The highlighted text units (“tokens”) represent which words or tokens the AI considered for it’s decision:

- **Green** indicates a positive attribution towards the class that the model predicts
- **Violet** indicates a negative attribution towards the class that the model predicts

The stronger/vibrant the color is, the more important the token is for the prediction.

Please note that some elements like "##" and spaces are added by some of the methods and can be ignored.
Also, some words might be split or removed by the method.

The **colors give no indication for the sentiment** of the text, but rather the **attribution towards the predicted class** of the model (which is unknown by you).

Your task is to guess the classification based on the standalone text and the text + explanation.

*

Definition of Counter speech:

In this study, **Counter speech** is defined as **direct response/comment that counters hateful or harmful speech**. The data consists of comments that where collected from youtube videos containing hateful content towards three target communities: Jews, African-Amerikan (Blacks) and LGBTQAI+.

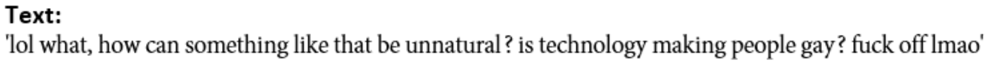
Some counter speech may contain hostile language where the counterspeaker uses abusive, hostile, or obscene comments in response to the original hate message

Definition of Non-Counter speech:

Non-Counter speech is refered as every comment that is **not directly responding/countering a (harmful) content**.

Task 1 - Baseline1

Without explanation



Text:
'lol what, how can something like that be unnatural ? is technology making people gay ? fuck off lmao'

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

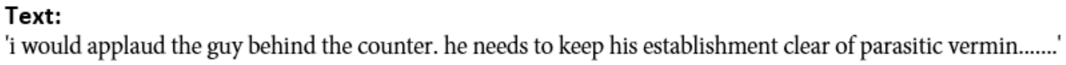
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

Task 1 - Baseline2

without explanation



Text:
'i would applaud the guy behind the counter. he needs to keep his establishment clear of parasitic vermin.......'

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

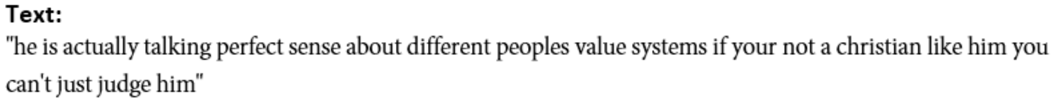
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

Task 1 - Baseline3

Without explanation



Text:
"he is actually talking perfect sense about different peoples value systems if your not a christian like him you can't just judge him"

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

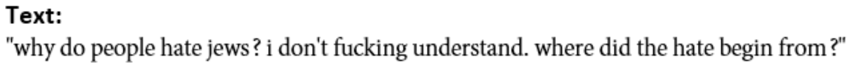
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

Task 1 - Baseline 4

Without Explanation



Text:
"why do people hate jews? i don't fucking understand. where did the hate begin from?"

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

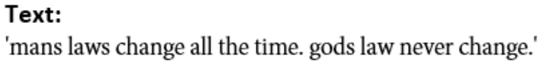
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

Task 1 - Baseline 5

Without explanation



Text:
'mans laws change all the time. gods law never change.'

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

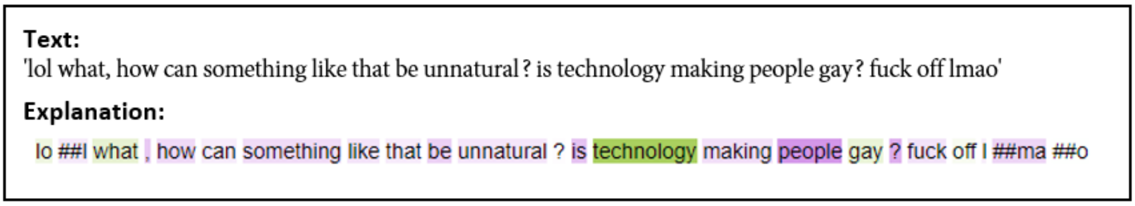
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

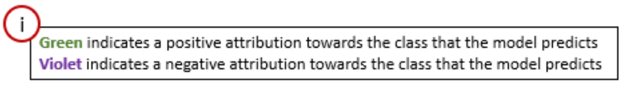
Task 1 - MA1

With explanation



Text:
'lol what, how can something like that be unnatural ? is technology making people gay ? fuck off lmao'

Explanation:
lo ##l what , how can something like that be unnatural ? is technology making people gay ? fuck off l ##ma ##o



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers

Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

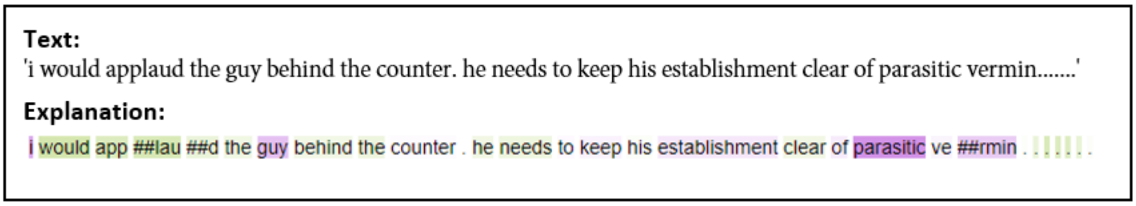
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

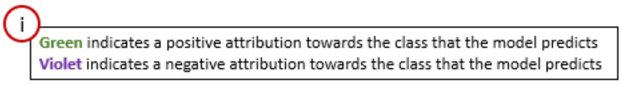
Task 1 - MA2

With explanation



Text:
'i would applaud the guy behind the counter. he needs to keep his establishment clear of parasitic vermin.......'

Explanation:
i would app ##lau ##d the guy behind the counter . he needs to keep his establishment clear of parasitic ve ##rmin . . . .



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers

Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

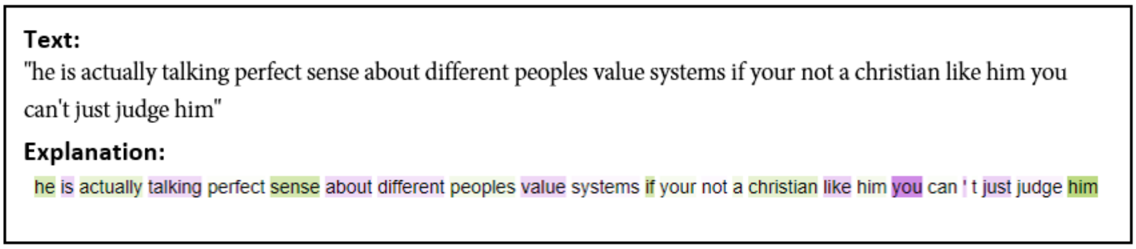
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

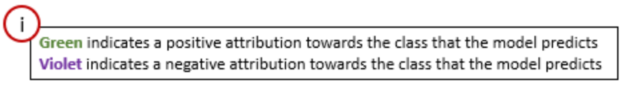
Task1 - MA3

With explanation



Text:
"he is actually talking perfect sense about different peoples value systems if your not a christian like him you can't just judge him"

Explanation:
he is actually talking perfect sense about different peoples value systems if your not a christian like him you can ' t just judge him



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

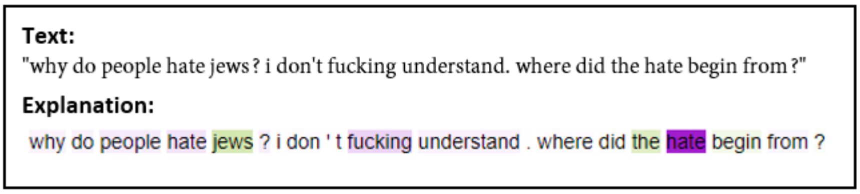
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

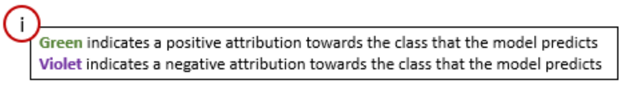
Task 1 - MA4

With explanation



Text:
"why do people hate jews? i don't fucking understand. where did the hate begin from?"

Explanation:
why do people hate jews ? i don ' t fucking understand . where did the hate begin from ?



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

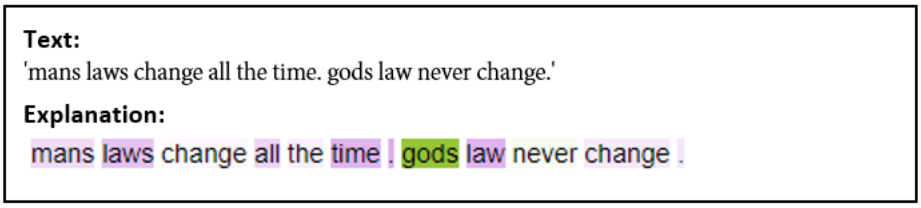
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

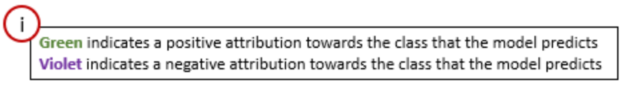
Task 1 - MA5

With explanation



Text:
'mans laws change all the time. gods law never change.'

Explanation:
mans laws change all the time . gods law never change .



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

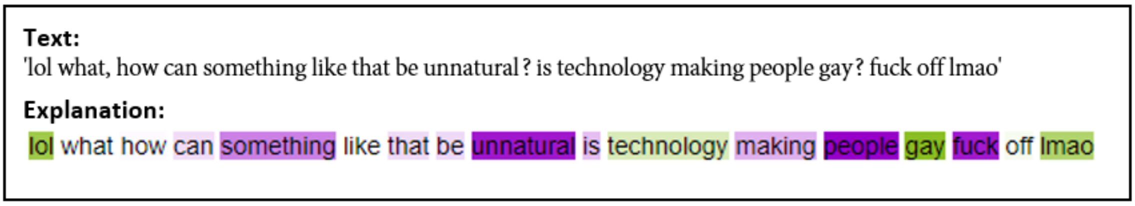
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

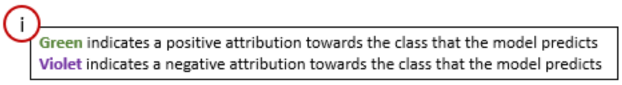
Task 1 - MB1

With explanation



Text:
'lol what, how can something like that be unnatural ? is technology making people gay ? fuck off lmao'

Explanation:
lol what how can something like that be unnatural is technology making people gay fuck off lmao



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

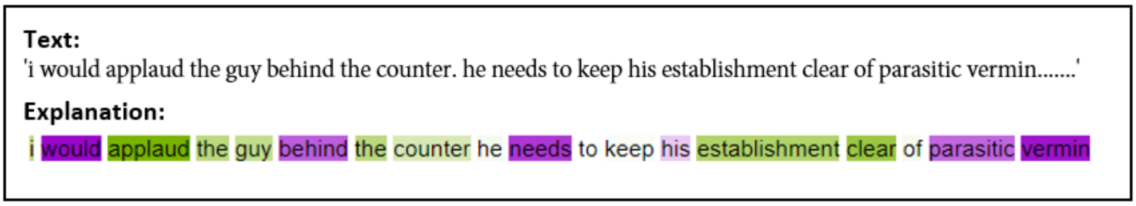
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

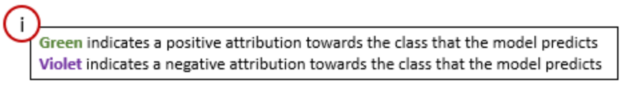
Task 1 - MB2

With explanation



Text:
'i would applaud the guy behind the counter. he needs to keep his establishment clear of parasitic vermin.......'

Explanation:
i would applaud the guy behind the counter he needs to keep his establishment clear of parasitic vermin



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

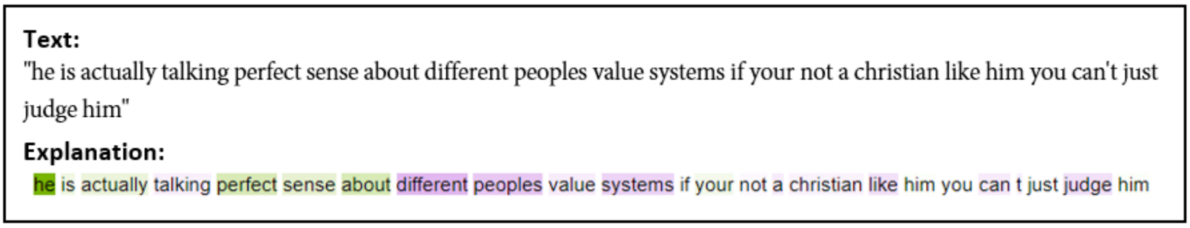
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

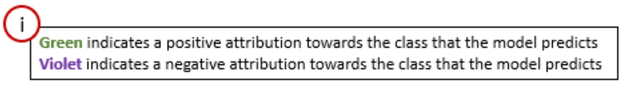
Task 1 - MB3

With explanation



Text:
"he is actually talking perfect sense about different peoples value systems if your not a christian like him you can't just judge him"

Explanation:
he is actually talking perfect sense about different peoples value systems if your not a christian like him you can t just judge him



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

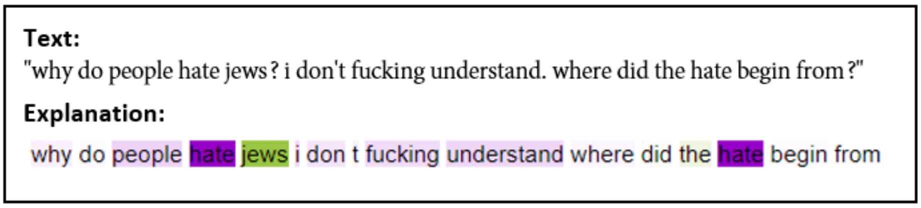
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

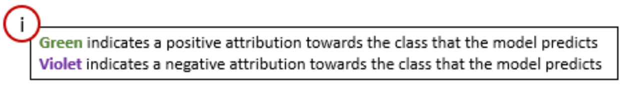
Task 1 - MB4

With explanation



Text:
"why do people hate jews? i don't fucking understand. where did the hate begin from?"

Explanation:
why do people hate jews i don t fucking understand where did the hate begin from



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers

Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

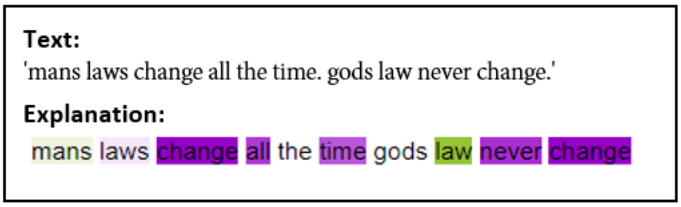
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

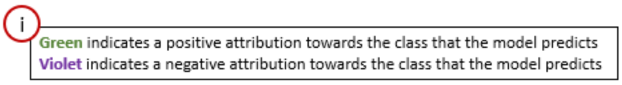
Task 1 - MB5

With explanation



Text:
'mans laws change all the time. gods law never change.'

Explanation:
mans laws change all the time gods law never change



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

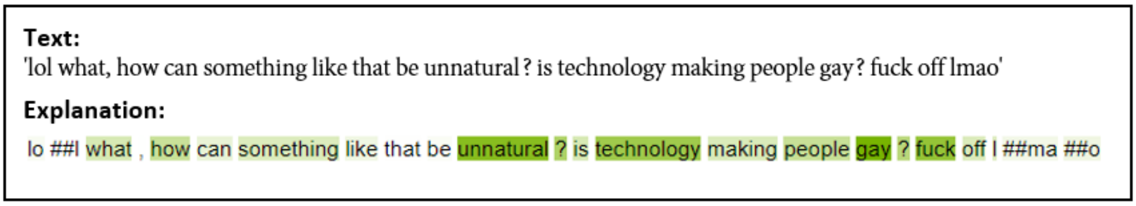
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

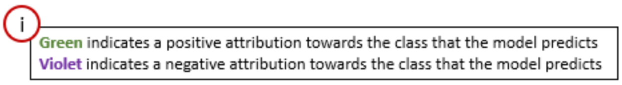
Task 1 - MC1

With explanation



Text:
'lol what, how can something like that be unnatural ? is technology making people gay ? fuck off lmao'

Explanation:
lo ##l what , how can something like that be unnatural ? is technology making people gay ? fuck off l ##ma ##o



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

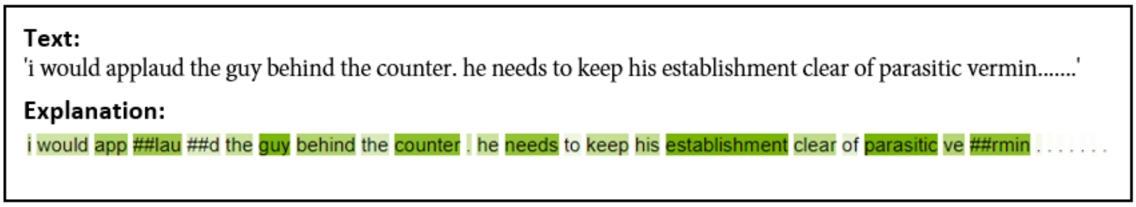
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

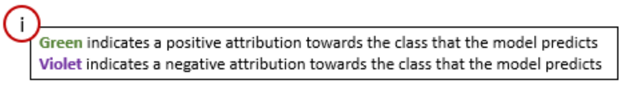
Task 1 - MC2

With explanation



Text:
'i would applaud the guy behind the counter. he needs to keep his establishment clear of parasitic vermin.......'

Explanation:
i would app ##lau ##d the guy behind the counter . he needs to keep his establishment clear of parasitic ve ##rmin . . . .



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

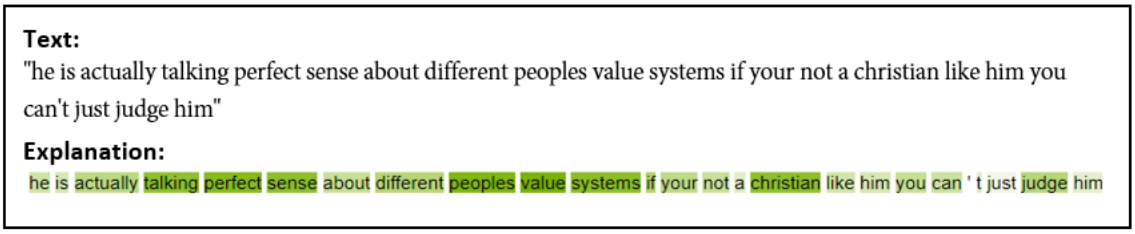
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

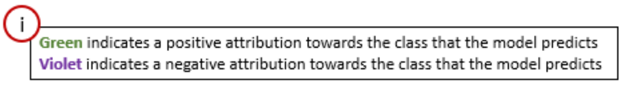
Task 1 - MC3

With explanation



Text:
"he is actually talking perfect sense about different peoples value systems if your not a christian like him you can't just judge him"

Explanation:
he is actually talking perfect sense about different peoples value systems if your not a christian like him you can ' t just judge him



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

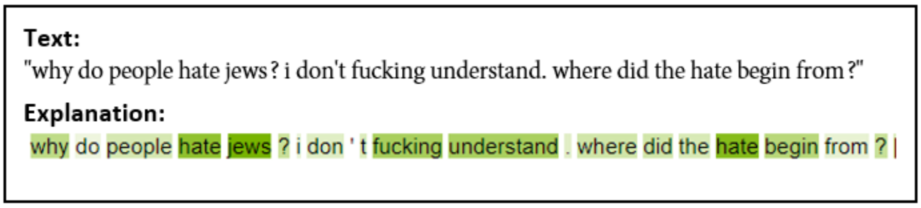
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

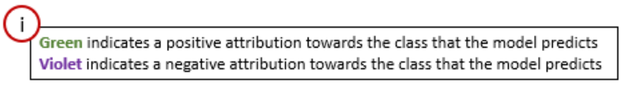
Task 1 - MC4

With explanation



Text:
"why do people hate jews? i don't fucking understand. where did the hate begin from?"

Explanation:
why do people hate jews ? i don ' t fucking understand . where did the hate begin from ? |



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers
Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

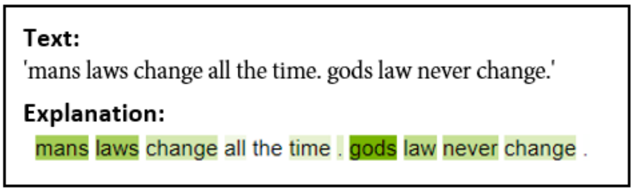
I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

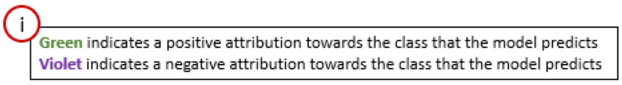
Task 1 - MC5

With explanation



Text:
'mans laws change all the time. gods law never change.'

Explanation:
mans laws change all the time . gods law never change .



Green indicates a positive attribution towards the class that the model predicts
Violet indicates a negative attribution towards the class that the model predicts

How will the model classify the text? *

Choose one of the following answers

Please choose **only one** of the following:

○ Counter speech
○ Non-Counter speech

I am confident with my answer *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

Task 2 Introduction

Task 2

In this part, we'll asses 3 methods (**A, B** and **C**) based on factors such as understandability, sufficiency, trustworthiness, satisfaction and helpfulness.
This time we use the original visualizations proposed by the creators of the method.

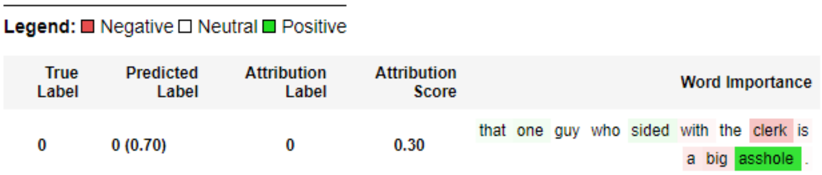
In Task 2, you will see 5 texts along with explanations of the 3 methods (A, B, and C) and the predicted labels (Non-Counter speech or Counter speech) from the classifier. You will be asked questions about the understandability, sufficiency, trustworthiness, satisfaction and helpfulness of these methods.

Please note that the label of Non-Counter speech = 0 and the label of Counter speech = 1

Method A und **B** are relatively similar to interpret:

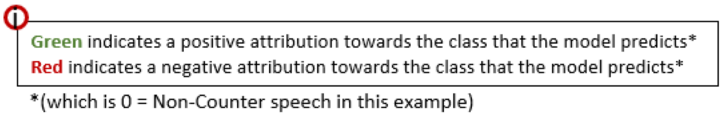
Method A:

True Label: 0 (Non-Counter speech)
Predicted Label: 0 (Non-Counter speech)



Legend: 🟥 Negative □ Neutral 🟩 Positive

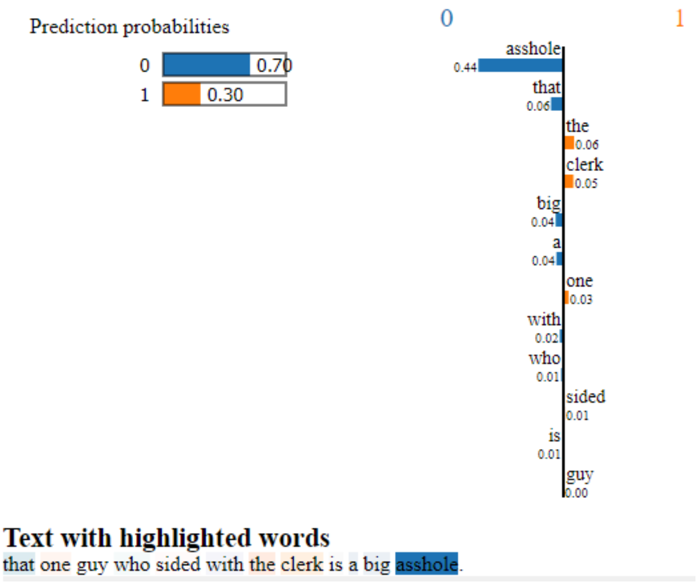
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.70)	0	0.30	that one guy who sided with the clerk is a big asshole .



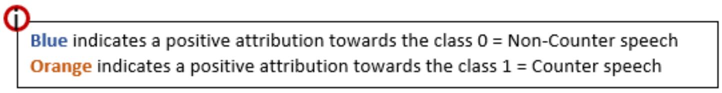
Green indicates a positive attribution towards the class that the model predicts*
Red indicates a negative attribution towards the class that the model predicts*
*(which is 0 = Non-Counter speech in this example)

Method B:

True Label: 0 (Non-Counter speech)
Predicted Label: 0 (Non-Counter speech)



Text with highlighted words
that one guy who sided with the clerk is a big asshole.

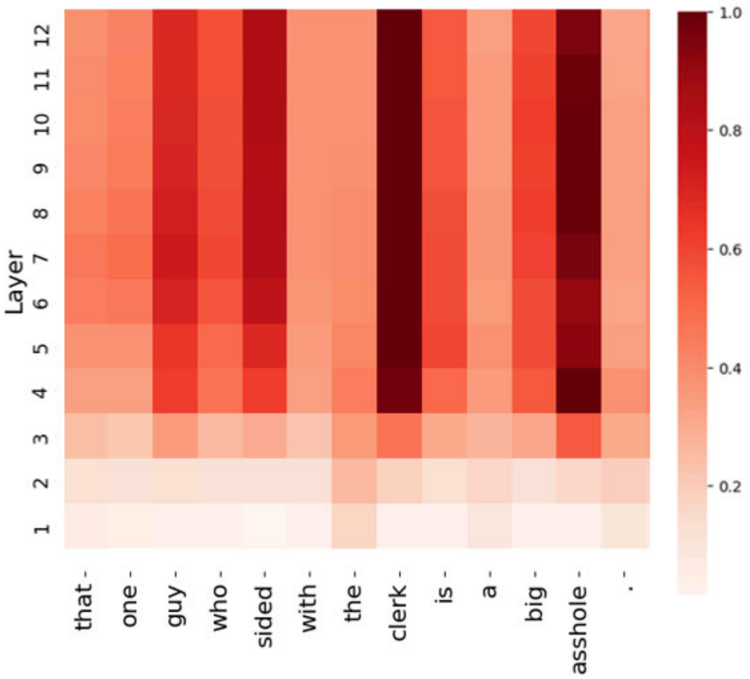


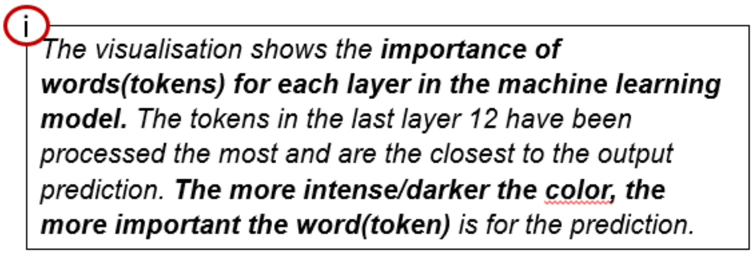
Blue indicates a positive attribution towards the class 0 = Non-Counter speech
Orange indicates a positive attribution towards the class 1 = Counter speech

Method C is more complex as it does not only visualizes the output text of the explanatory model but also displays the attributions of the words/tokens through the inner structure of the utilized model, which comprises multiple layers (12 in this case), in the form of a heatmap.

Method C:

True Label: 0 (Non-Counter speech)
Predicted Label: 0 (Non-Counter speech)

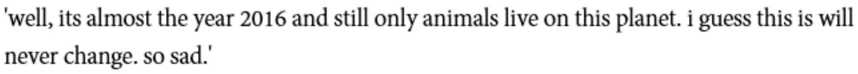




The visualisation shows the ***importance of words(tokens) for each layer in the machine learning model.*** *The tokens in the last layer 12 have been processed the most and are the closest to the output prediction.* ***The more intense/darker the color, the more important the word(token)*** *is for the prediction.*

Task 2 -MA 1

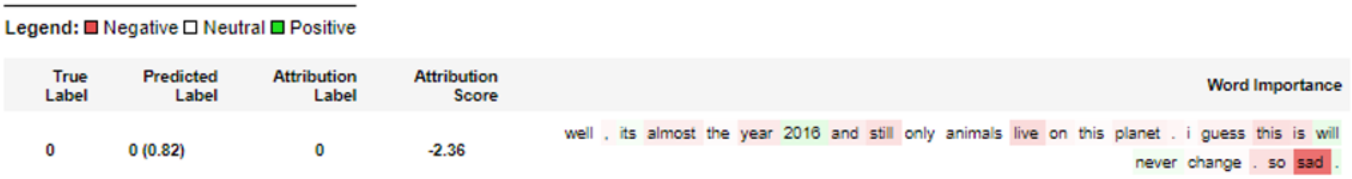
Text:



'well, its almost the year 2016 and still only animals live on this planet. i guess this is will never change. so sad.'

True Label: 0 (Non-Counter speech)
Predicted Label: 0 (Non-Counter speech)

Explanation:



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.82)	0	-2.36	well , its almost the year 2016 and still only animals live on this planet . i guess this is will never change . so sad .

1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

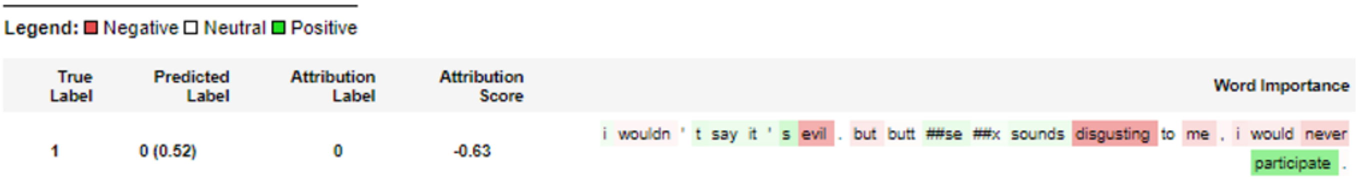
○ 1
○ 2
○ 3
○ 4
○ 5

Task 2 -MA 2

Text:
"i wouldn't say it's evil. but buttsex sounds disgusting to me, i would never participate."

True Label: 1 (counter speech)
Predicted Label: 0 (Non-Counter speech)

Explanation:



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	0 (0.52)	0	-0.63	i wouldn ' t say it ' s evil . but butt ##se ##x sounds disgusting to me , i would never participate .

1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

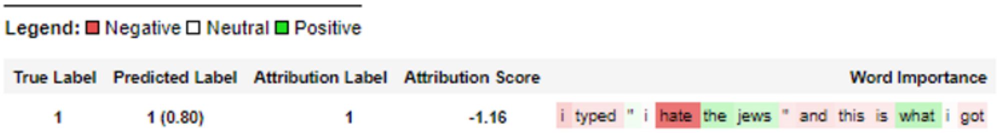
○ 1
○ 2
○ 3
○ 4
○ 5

Task 2 -MA 3

Text:
'i typed "i hate the jews" and this is what i got'

True Label: 1 (counter speech)
Predicted Label: 1 (counter speech)

Explanation:



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	1 (0.80)	1	-1.16	i typed " i hate the jews " and this is what i got

1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

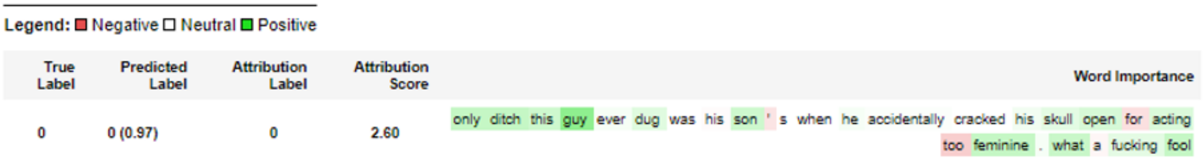
○ 1
○ 2
○ 3
○ 4
○ 5

Task 2 -MA 4

Text:
"only ditch this guy ever dug was his son's when he accidentally cracked his skull open for acting too feminine. what a fucking fool"

True Label: 0 (Non-Counter speech)
Predicted Label: 0 (Non-Counter speech)

Explanation:



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.97)	0	2.60	only ditch this guy ever dug was his son ' s when he accidentally cracked his skull open for acting too feminine . what a fucking fool

1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

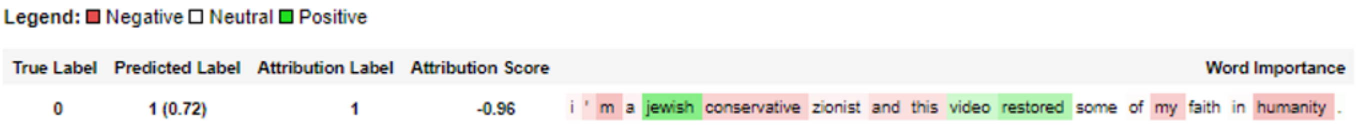
○ 1
○ 2
○ 3
○ 4
○ 5

Task 2 -MA 5

Text:
"i'm a jewish conservative zionist and this video restored some of my faith in humanity."

True Label: 0 (Non-Counter speech)
Predicted Label: 1 (Counter speech)

Explanation:



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	1 (0.72)	1	-0.96	i ' m a jewish conservative zionist and this video restored some of my faith in humanity .

1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

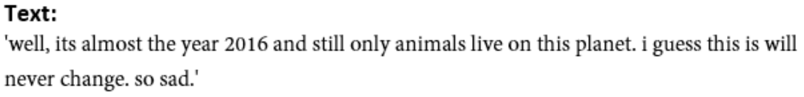
5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

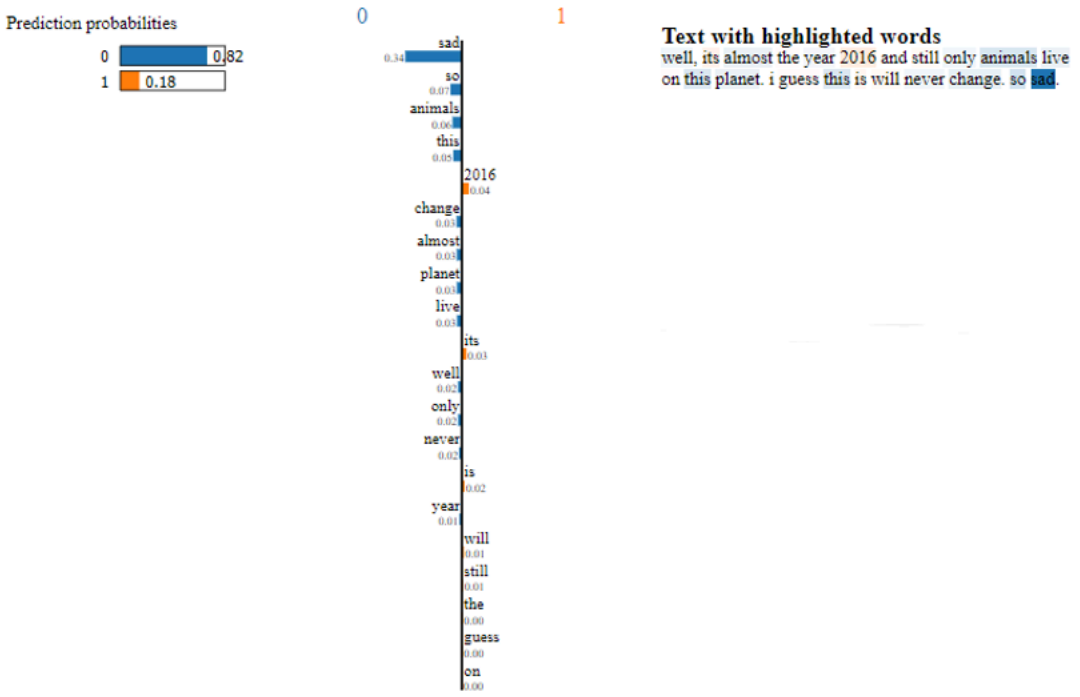
Task 2 -MB 1



Text:
'well, its almost the year 2016 and still only animals live on this planet. i guess this is will never change. so sad.'

True Label: 0 (Non-Counter speech)
Predicted Label: 0 (Non-Counter speech)

Explanation:



Text with highlighted words

well, its almost the year 2016 and still only animals live on this planet. i guess this is will never change. so sad.

1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

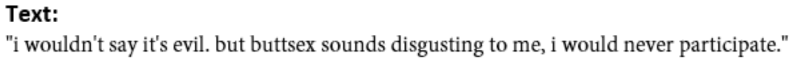
5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

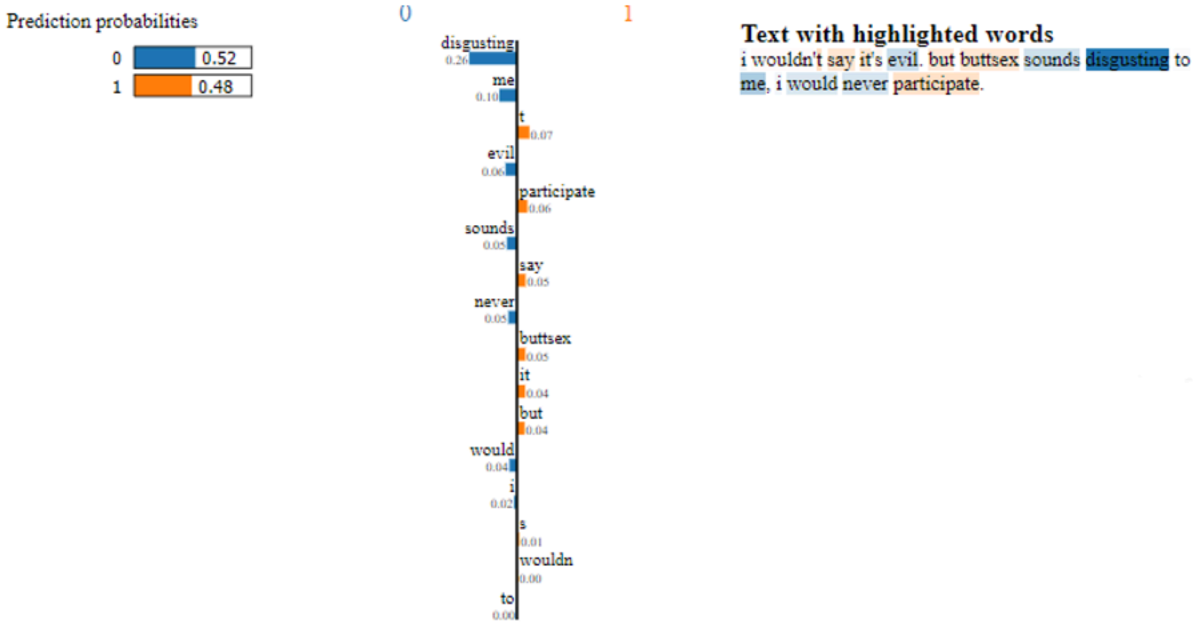
Task 2 -MB 2



Text:
"i wouldn't say it's evil. but buttsex sounds disgusting to me, i would never participate."

True Label: 1 (Counter speech)
Predicted Label: 0 (Non-Counter speech)

Explanation:



Text with highlighted words

i wouldn't say it's evil. but buttsex sounds disgusting to me, i would never participate.

1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

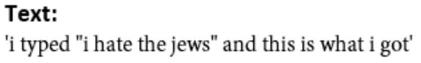
5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

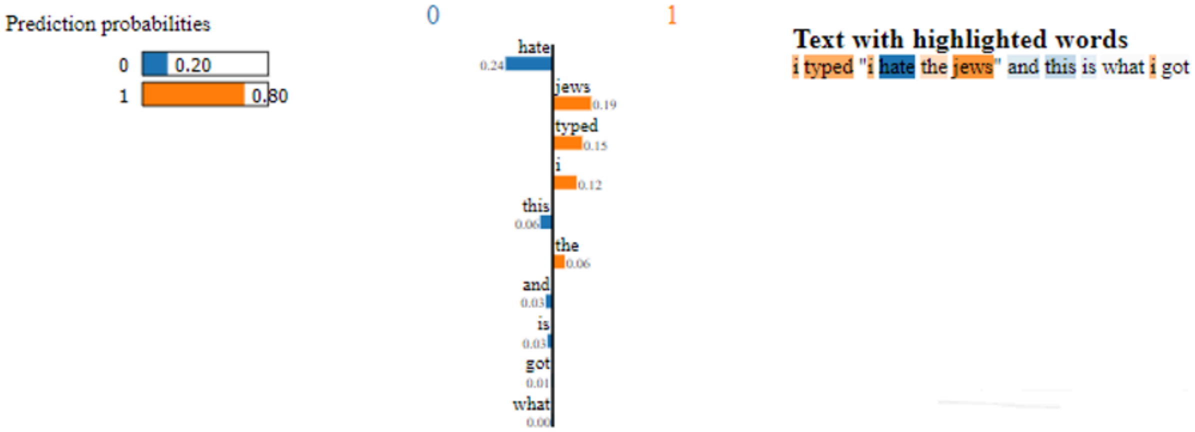
Task 2 -MB 3



Text:
'i typed "i hate the jews" and this is what i got'

True Label: 1 (Counter speech)
Predicted Label: 1 (Counter speech)

Explanation:



i typed "i hate the jews" and this is what i got

1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

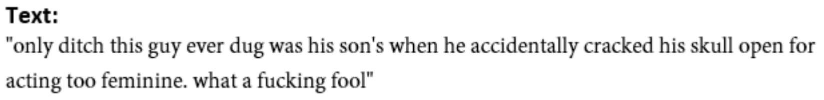
5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

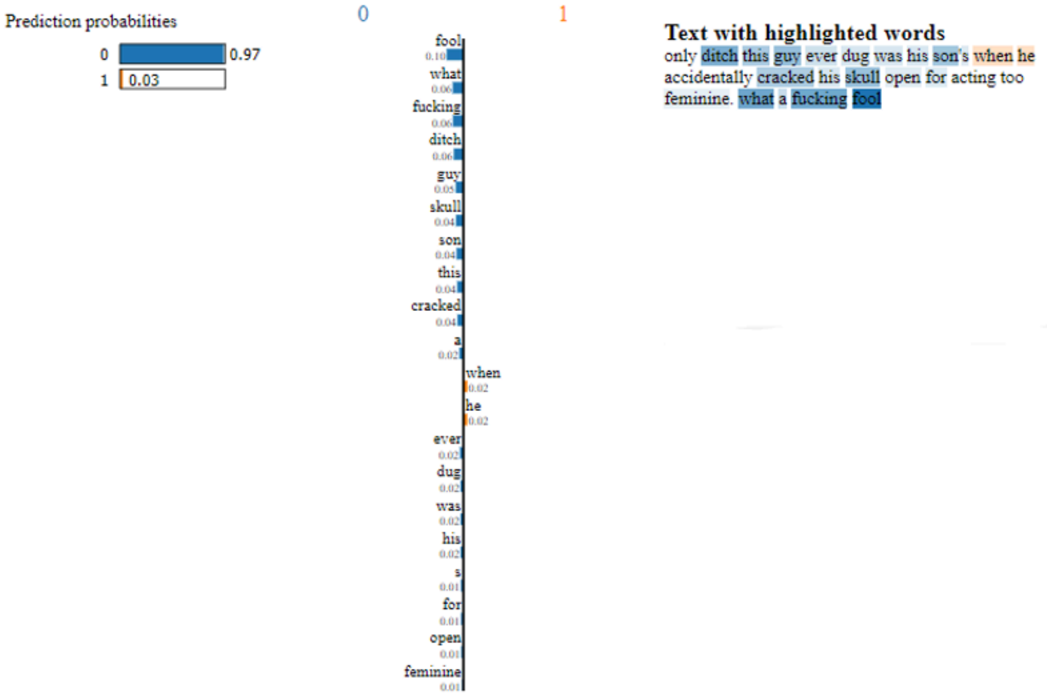
Task 2 -MB 4



Text:
"only ditch this guy ever dug was his son's when he accidentally cracked his skull open for acting too feminine. what a fucking fool"

True Label: 0 (Non-Counter speech)
Predicted Label: 0 (Non-Counter speech)

Explanation:



Text with highlighted words

only ditch this guy ever dug was his son's when he accidentally cracked his skull open for acting too feminine. what a fucking fool

1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

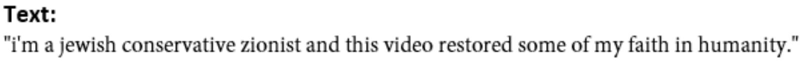
5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

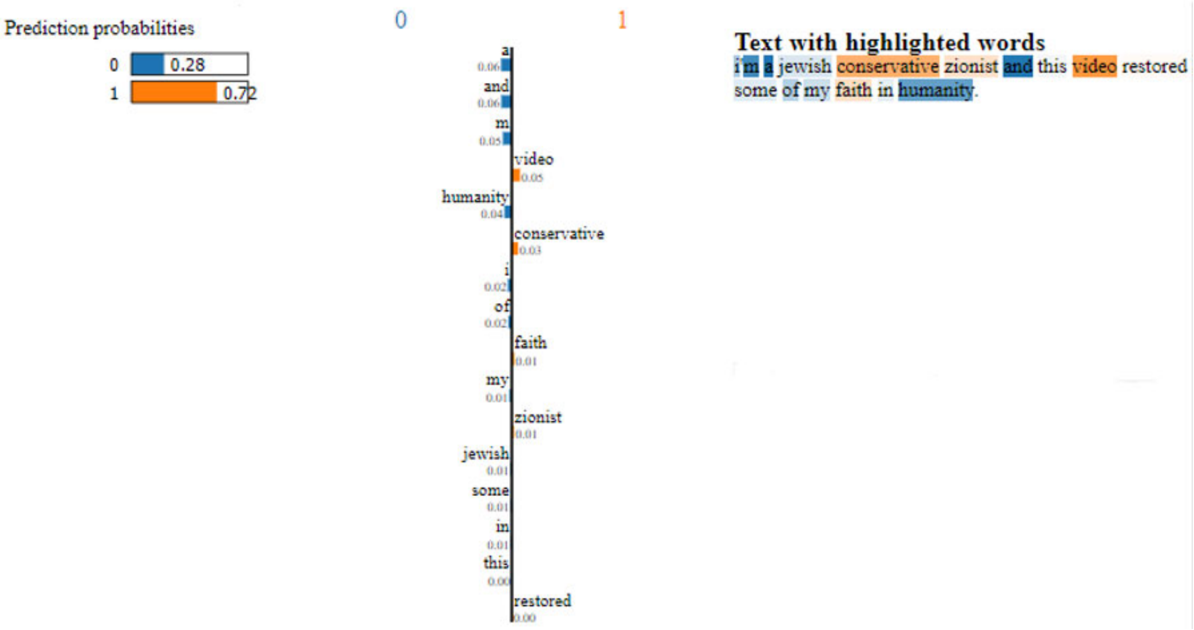
Task 2 -MB 5



Text:
"i'm a jewish conservative zionist and this video restored some of my faith in humanity."

True Label: 0 (Non-Counter speech)
Predicted Label: 1 (Counter speech)

Explanation:



Text with highlighted words

i'm a jewish conservative zionist and this video restored some of my faith in humanity.

1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

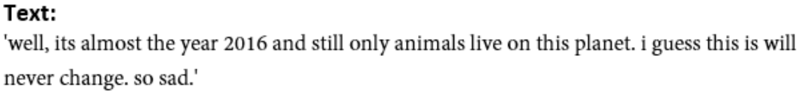
5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

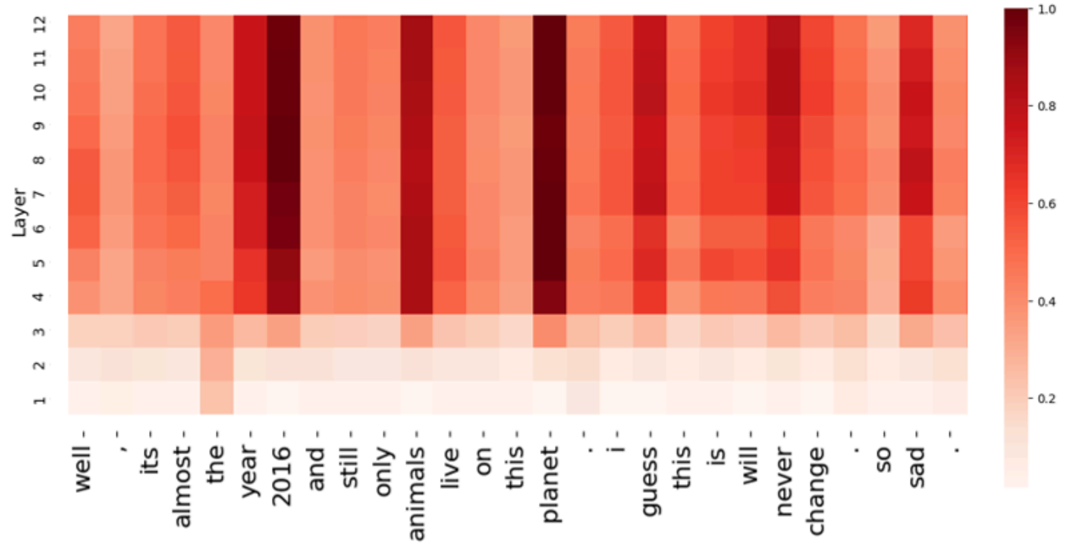
Task 2 -MC 1



Text:
'well, its almost the year 2016 and still only animals live on this planet. i guess this is will never change. so sad.'

True Label: 0 (Non-Counter speech)
Predicted Label: 0 (Non-Counter speech)

Explanation:



1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

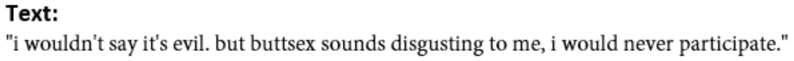
5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

Task 2 -MC 2

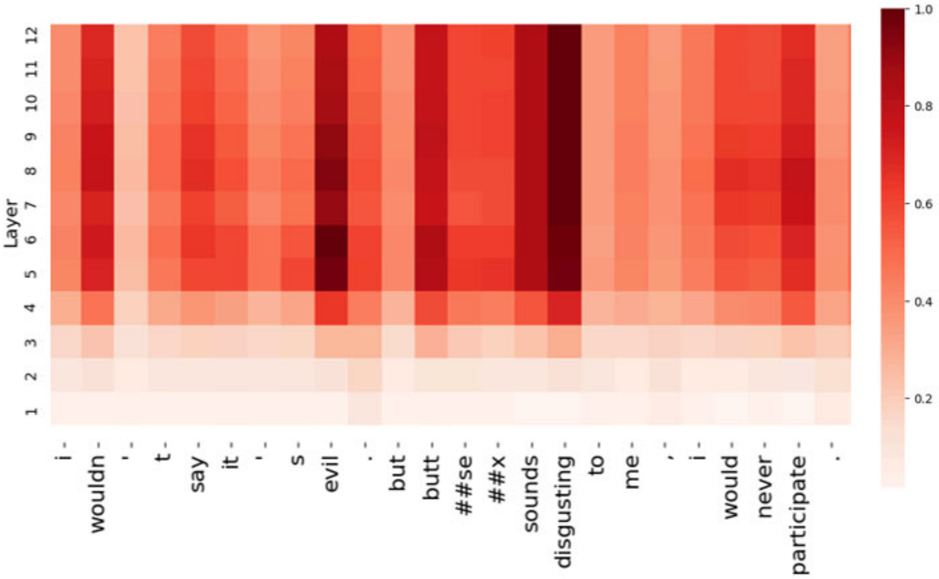


Text:

"i wouldn't say it's evil. but buttsex sounds disgusting to me, i would never participate."

True Label: 1 (Counter speech)
Predicted Label: 0 (Non-Counter speech)

Explanation:



1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree))

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

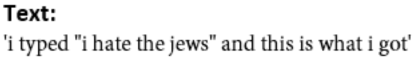
5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

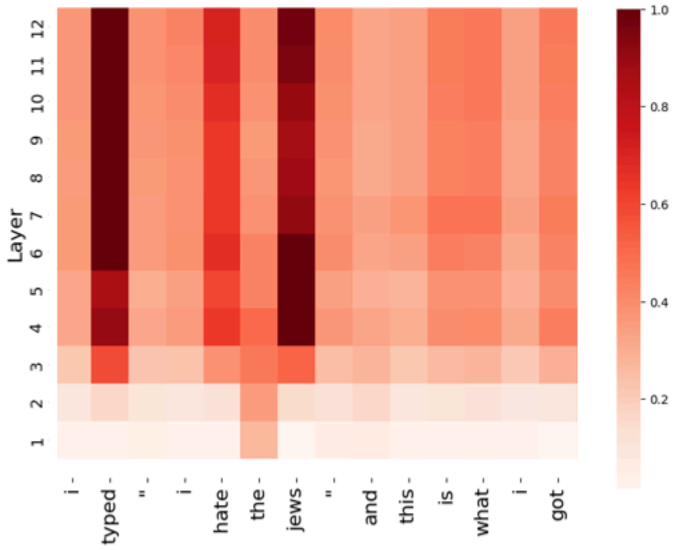
Task 2 -MC 3



Text:
'i typed "i hate the jews" and this is what i got'

True Label: 1 (Counter speech)
Predicted Label: 1 (Counter speech)

Explanation:



1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

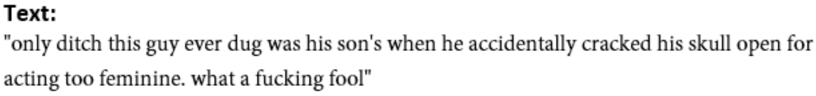
5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

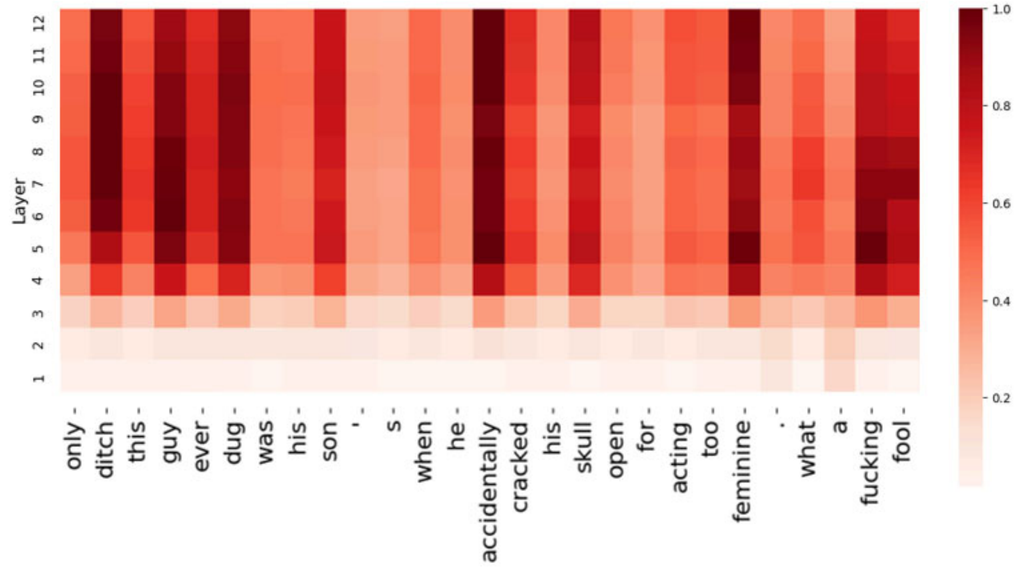
Task 2 -MC 4



Text:
"only ditch this guy ever dug was his son's when he accidentally cracked his skull open for acting too feminine. what a fucking fool"

True Label: 0 (Non-Counter speech)
Predicted Label: 0 (Non-Counter speech)

Explanation:



1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

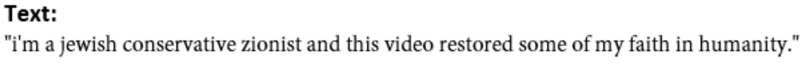
5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

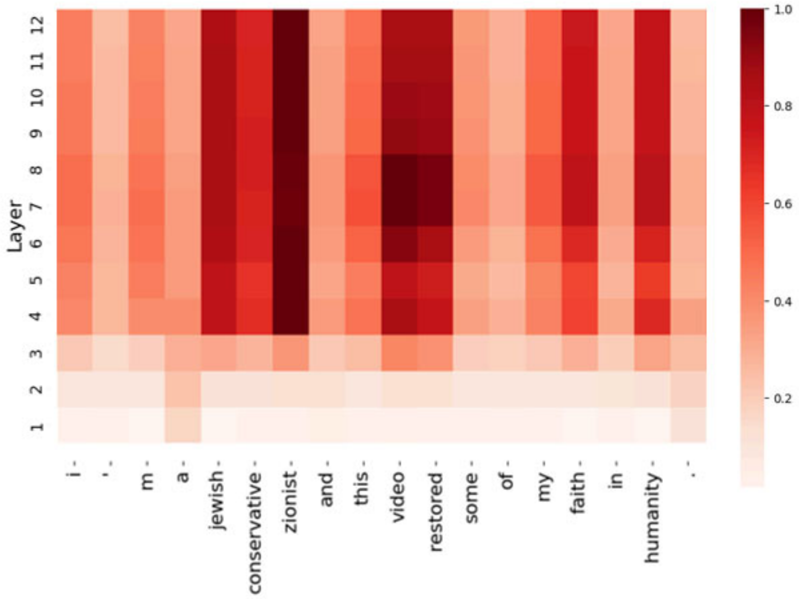
Task 2 -MC 5



Text:
"i'm a jewish conservative zionist and this video restored some of my faith in humanity."

True Label: 0 (Non-Counter speech)
Predicted Label: 1 (Counter speech)

Explanation:



1.I understand why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

2. I think the visualization is sufficient for explaining why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

3. I think the visual explanations increases my trust in the model (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

4. I am satisfied with the explanation why the model classified the text as 0 (Non-Counter speech)/ 1 (Counter speech) (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

5. I Find the explanations helpful (1 = Strongly disagree, 5 = Strongly agree)

*

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

You are done! Thank you so much for your time :)

for any questions please contact me at aboeck@fhstp.ac.at

Submit your survey.
Thank you for completing this survey.