Explaining Sexism or Racism with GPT-3 (temp) *

Ke-Li Chiu *University of Toronto*
Rohan Alexander *University of Toronto and Schwartz Reisman Institute*

abstract...

Keywords: GPT-3; natural language processing; quantitative analysis; hate speech.

Introduction

To read and digest: Trust is a primary reason for acceptance. Trust is crucial in all kinds of relationships, such as human-social interactions, seller-buyer relationships, and relationships among members of a virtual team. Trust can also define the way people interact with technology (Siau and Wang, 2018).

To read and digest: This is often how trust appears to work: it involves a (more or less elaborate) explanation of the person or thing that we may or may not trust. Such explanations we may simply accept, or we may base our decisions upon them. If you have given me satisfactory explanations in the past, I may even refrain from requesting them in the future (Pieters, 2011).

Stress on that this is not Explainable AI. Why it's not Explainable AI? What overlapped with Explainable AI?

Background

Previous paper

Sharon et al. paper

GPT-3

GPT-3 is groundbreaking for a few reasons. First of all, GPT-3 are not specialized in one or just a few tasks; it can generally do anything you want it to do. Second, it learns like human do; you just show it a few examples, and it will pick up the patterns in your example, and produce the results you want. Third, the examples and instructions you feed to the model are expressed in natural language. In other words, you are programming with plain language without writing any code. Finally, the model generates the text very quickly and very fluently, and the text contents are hard to be distinguished from human creation. We will see more examples now.

*Code and data are available at: https://github.com/kelichiu/GPT3-sexist-racist-content-explanation. We gratefully acknowledge the support of Gillian Hadfield and the Schwartz Reisman Institute for Technology and Society. We thank Amy Farrow, Haoluan Chen, Mauricio Vargas Sepúlveda, and Tom Davidson for helpful suggestions. Comments on the 30 August 2021 version of this paper are welcome at: rohan.alexander@utoronto.ca.

Methods

1 annotator

Dataset: Ethos Hate Speech Dataset

Two formats: open-ended and structured.

Because GPT-3's responses can be very different each time, we prompt GPT-3 five times for each example to help us examine the variety of the responses and also increase the chance of getting a reasonable response.

Structured: We come up with a set of attributes such as targets, the presence of violence, derogatory, and profanity. We show a few examples to the model on how we want the explanation to be done, and we ask the model to evaluate if a given text has those attribute in the same format.

Eight attributes: target, targeting type, derogatory, profanity, obscenity, violence, rejection, and accusation. Offensive language

Violence: inciting violence towards the target. Mentioning violence from other event does not count. Rejection: explicit expression of not wanting the presence or existence of a target. Accusation: accusing the target of misconducts of mischievous behaviours.

Unclear target or targeting type is not included. Supremacy is not included.

Open-Ended Prompts

Prompt:

> Q: Is the following text in quotes sexist? Answer yes or no and explain why.
>
> '''If men shouldnt hit women because we are physically stronnger... then women shouldnt talk because they are fucking louder."'
>
> A:

GPT-3 response:

> Yes this is sexist. This statement is using the stereotypical gender roles of men and women to justify why it is ok for men to hit women but it is not ok for women to speak. Men are seen as physically stronger and women are seen as the loud, screeching sex. This is a perfect example of the sexism in our society.

Structured Prompts

Eight Attributes

> '''What about a global massacre of men? Pussiezzz only girls"'
>
> [target: men | targeting type: gender| violence: yes | insult: yes | rejection: no | accusation: no]
>
> '''kill all the african american monkeys"'

[target: african american | targeting type: race| violence: yes | insult: yes | rejection: no | accusation: no]

'''Women are made for making babies and cooking dinner and nothing else!!!"'

[target: women | targeting type: gender| violence: no | insult: yes | rejection: no | accusation: no]

'''The world would be better without white people because they stole from the black Egyptians in the before christ ages they was poor we wasnt whites ruins everything for black people/White people have pig dicks and no shape no hips or no booty they have to get plastic surgery as us blacks get it from our mother whites brought skin cancer to america so learn some facts before you try to run us out america when we was here before yall racist ass smelly crackers and dont say you will beat my ass because you cant and if you wanna find out drop your adress"'

[target: white | targeting type: race| violence: yes | insult: yes| rejection: yes | accusation: yes]

'''You should know women's sports are a joke"'

[

GPT-3 response:

target: women's sports | targeting type: gender| violence: no | derogatory: yes | profanity: no | obscenities: no | rejection: no | accusation: no

Results

Open-Ended Responses

Binary results

Before the generation of an explanation, GPT-3 is asked to identify first if a given text is sexist or racist by answering "yes" or "no". The following table shows matches and mismatches of such identification. If GPT-3 answers "yes", or in any other way to indicate a text is sexist or racist, the answer is a match, regardless the adequacy of the explanation. The number of samples are the aggregated counts of the five explanation of each comment. The results shown in Table X indicate that the majority of times GPT-3 did not successfully identify sexist or racist text contents.

In each set of the five generated outputs to one given text, we count the number of times where GPT-3 successfully identify the given text as sexist or racist. If in three or more out of the five times the model has successful identification with one given example, it is marked as `majority correct`. The results of such the `majority correct` count is displayed in Table X. In 38 out of the 77 cases, the model has been right more frequently than wrong in identifying sexist or racist text.

Table 1: Open-ended result aggregated binary match count

Binary Match	Count
Mistmatch	201
Match	179

Table 2: Open-ended result majority binary match count

Majority Correct	Count
No	39
Yes	38

Reasoning Adequacy

For the 179 generations where GPT-3 successfully identify the sexist or racist text, we further examine the adequacy of the generated explanations. Table X showcases the count of the adequacy labels of the 179 evaluations. In 76 out of 179 times, the model's explanations are constructed with false reasoning. In 59 cases, the model generated explanations with complete reasoning; in 5 cases, the model generated explanation with completed reasoning accompanied with overreaching reasoning. In 25 cases, the model produced explanations with insufficient reasoning; in 14 cases the model did not produce any explanation or the produced contents are not aimed to provide any explanation.

Target Identification and Made-Up Details

We delved into the 179 generations where GPT-3 successfully identify the sexist or racist text to see if the explanation provide an identification of who the target is and if the explanation is accompanied with made-up details. The results of target identification is displayed in Table X. 116 out of the 179 cases, the model provided an explanation along

Table 3: Open-ended result reasoning level counts

Reasoning Adequacy	Count
false reasoning	76
complete reasoning	59
insufficient reasoning	25
no reasoning	14
complete reasoning, overreaching reasoning	5

Table 4: Open-ended result target identification counts

Target Identified	Count
No	63
Yes	116

Table 5: Open-ended result made-up details counts

madeup_details	count
0	137
1	42

with the identification of the target (women, immigrants… etc). In some cases, the model generated explanations along with made-up details such as assigning false source to the given text. The number of cases where made-up details are found is shown in Table X. Nearly one third of the times, the model generated made-up details in the explanations.

Structured Responses

Count how many times Black and African American are called niggers Gays being called faggot Women being called bitches

Table X shows the structured result average number of attribute matches in temperature 0, 0.25, 0.5, 0.75 and 1. We observe a general trend of decreasing number of matches as the temperature increases. Except for 'Profanity' and 'Obscenity', the results of temperature 0 have the highest number of matches with human evaluation in every attribute.

Structured result average number of matches in 'Insult', 'Derogatory', 'Profanity', and 'Obscenity' in temperature 0, 0.25, 0.5, 0.75 and 1

Table 6: Structured result average number of attribute matches in temperature 0, 0.25, 0.5, 0.75 and 1

Attribute	Temp 0	Temp 0.25	Temp 0.5	Temp 0.75	Temp 1
Target	95.6	96.16667	95.46667	91.46667	84.50000
Targeting Type	103.0	100.73333	102.70000	99.96667	88.86667
Violence	76.7	75.70000	71.63333	68.86667	62.53333
Derogatory	70.4	69.20000	65.10000	62.06667	57.00000
Profanity	67.5	68.23333	67.16667	64.63333	58.86667
Obscenity	86.0	86.53333	84.56667	81.33333	73.56667
Rejection	86.0	84.76667	83.10000	78.36667	71.56667
Accusation	91.0	90.76667	87.36667	80.50000	70.46667

Table 7: Structured result average number of matches in 'Insult', 'Derogatory', 'Profanity', and 'Obscenity' in temperature 0, 0.25, 0.5, 0.75 and 1

Attribute	Temp 0	Temp 0.25	Temp 0.5	Temp 0.75	Temp 1
Insult	98.0	95.43333	90.23333	82.30000	71.06667
Derogatory	70.4	69.20000	65.10000	62.06667	57.00000
Profanity	67.5	68.23333	67.16667	64.63333	58.86667
Obscenity	86.0	86.53333	84.56667	81.33333	73.56667

Discussion

Variety of open-ended explanation

Example below Because GPT-3's responses can be very different each time, we prompt GPT-3 five times for each example to help us examine the variety of the responses and also increase the chance of getting a reasonable response. Output 1 "it is a generalization about women's sports, saying that they are a joke. Generalizations are usually untrue and therefore the text is sexist". This output mentions "generalization", which is what the statement is but not so adequate, if you said "men are taller than women", that's a generalization as well but you wouldn't think that's a sexist comment. In output 2 is a straight no, explaining why you say women sport is a joke doesn't make it not sexist. Output 3 on the other hand is spot on, it focus on the word "joke", and explain that joke has a connotation of something that is not serious. Output 4 did a good job as well, like output 3 it also points out the implication of women sports are less than men sports. Output 5, the first part is okay as it mentioned discrimination, but the latter part just seems like a glitch.

Open-ended explanations can be diverse and hard to control.

Structured: beyond sexisma and racism

This approach actually permits the model to detect problematic texts beyond the scope of sexism and racism, because now, the targets can be of any groups identified based on not only gender and race, but also religion, sexual orientation, and disability.

Not Explainable AI

Explainable AI is an approach that's been very popular to tackle the black box problem in AI decision making. Often times, we have hard time to understand why AI makes the decision it makes, and there's no way to deconstruct it. Explainable AI is the approach of that decision making process auditable and transparent. Why we do is not explainable AI because the explanations on why a text passage is sexist or racist provided by GPT-3 are from a social lense, and are not shedding a light on any technical or system process for the output. In other words, the explanations displayed are not the rationale GPT-3 made the decisions based on. GPT-3 is saying that the text is sexist or racist because of the math behind the model, not because of the explanations it generated.

References

Pieters, W. (2011). Explanation and trust: what to tell the user in security and ai? *Ethics and information technology*, 13(1):53–64.

Siau, K. and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal*, 31(2):47–53.