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

Large language models such as GPT-3 have the capacity to generate texts that are indistinguishable from human creation. In our study, we examine GPT-3's ability to generate explanations for why a given text is sexist or racist. We prompt GPT-3 to generate explanations in two approaches: open-ended and structured.

We are interested in language models' ability in generating explanations because explanations play a crucial role in trust (Pieters, 2011). Trust is an essential factor for acceptance, which can shape the interaction between human and technology (Siau and Wang, 2018).

Stress on that this is not Explainable AI. Why it's not Explainable AI? What overlapped with Explainable AI? However, the generation of explanation to why a given passage is sexist or racist is not to be mistaken as an approach of Explainable AI.

Background

Previous paper

This paper is, to some extent, an extension of our previous study on GPT-3's ability to detect sexist and racist textual contents [**Cite previous paper**].

Sharon et al. paper

Machine generation of open-ended explanation on sexist texts found in forums [**Cite Sharon et al.**].

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

*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 31 August 2021 version of this paper are welcome at: rohan.alexander@utoronto.ca.

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.

Methods

Dataset: Ethos Hate Speech Dataset

Describe how we retrieve sexist and racist data from the data set and how many samples we obtained

- What data are included
- What data are excluded: Unclear target or targeting type is not included. Supremacy is not included.

Response Collection

- OpenAI API
- Model: Davinci
- Hyper-parameters
- Sample size

Generation Formats

What is open-ended format

- What's the prompt example: Appendix X.
- What are the metrics we used to evaluate the generations

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.

Metrics: Reasoning adequacy, target identification, and presence of made-up details.

What is structured format

- What's the prompt example: Appendix X.
- What are the metrics we used to evaluate the generations

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

Metrics: matches of attributes with human labelling.

Human Labelling

Describe how human labelling is proceed

- number of annotator
- definition of all attributes

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.

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

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

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

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.

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 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.

Table 4: Open-ended result target identification counts

Target Identified	Count
No	63
Yes	116

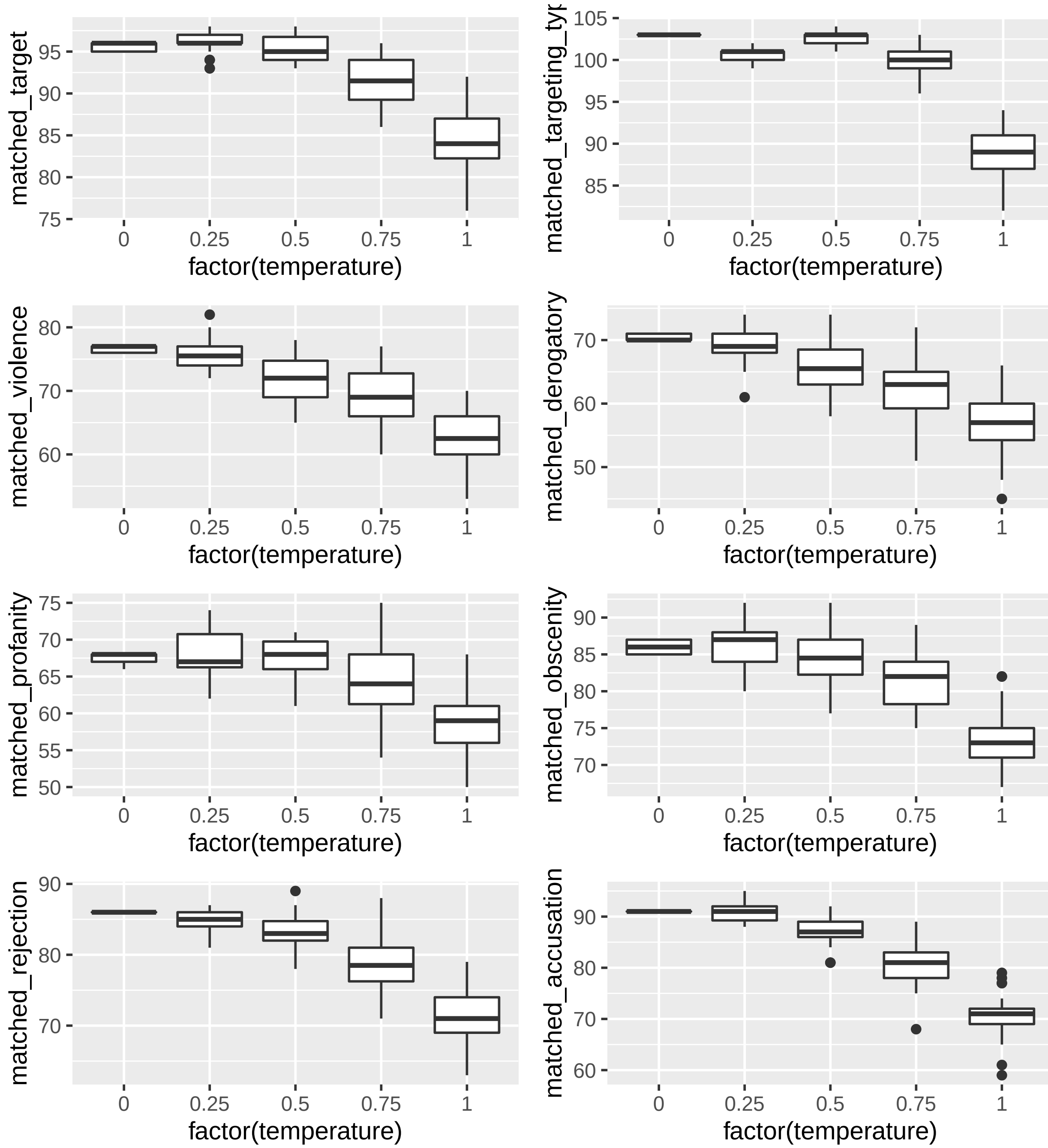
Table 5: Open-ended result made-up details counts

madeup_details	count
0	137
1	42

Structured Responses

Number of Matched Numbers within Different Temperatures

Figure 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 number of matches decreasing as the temperature increases. Except for 'Target', 'Profanity' and 'Obscenity' where the model performed the best under temperature 0.25 by a thin margin, the results of temperature 0 have the highest number of matches with human evaluation in every other attribute. Moreover, as the distribution of matched number varies within wider ranges as the temperature increases.



Comparison of All Attributes

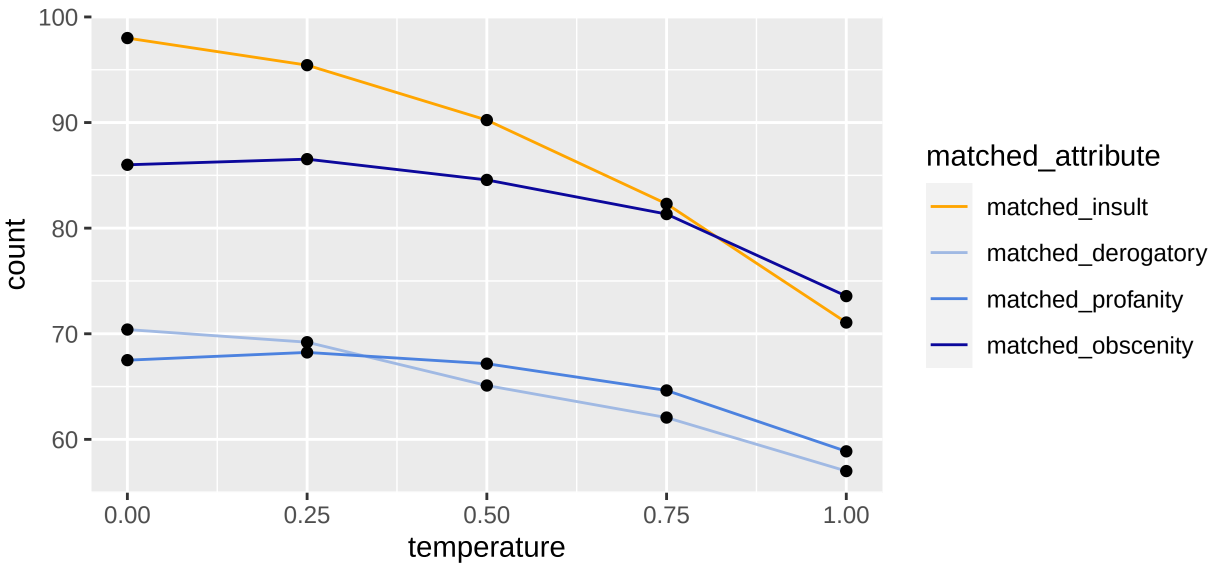
The descriptive summary is conducted and shown in Table X to assess the model' overall performances in each attribute. The results show that the model performs the best in identifying the identity categories (gender, race...etc) a sexist or racist text is targeting (sex, gender, race, origin...etc). Second to the targeting type, the model also performs well in identifying who the target is (women, African Americans...etc). The model seems to have varied capabilities in identifying the presence of violence, derogatory, profanity, obscenity, rejection, and accusation in the texts.

Table 6: Structured results attribute matches descriptive data

Target	Targeting Type	Violence	Derogatory	Profanity	Obscenity	I
Min. :84.50	Min. : 88.87	Min. :62.53	Min. :57.00	Min. :58.87	Min. :73.57	M
1st Qu.:91.47	1st Qu.: 99.97	1st Qu.:68.87	1st Qu.:62.07	1st Qu.:64.63	1st Qu.:81.33	1
Median :95.47	Median :100.73	Median :71.63	Median :65.10	Median :67.17	Median :84.57	M
Mean :92.64	Mean : 99.05	Mean :71.09	Mean :64.75	Mean :65.28	Mean :82.40	M
3rd Qu.:95.60	3rd Qu.:102.70	3rd Qu.:75.70	3rd Qu.:69.20	3rd Qu.:67.50	3rd Qu.:86.00	3
Max. :96.17	Max. :103.00	Max. :76.70	Max. :70.40	Max. :68.23	Max. :86.53	M

Number of Matched Numbers within Granular Attributes V.S. General Attribute

We define ‘Insult’ as a general attribute that encompasses ‘Derogatory’, ‘Profanity’, and ‘Obscenity’, which are what we called granular attributes. We examine how GPT-3 performs in identifying a more general attribute compared to identifying more granular attributes. Figure X shows the average numbers of matches in ‘Insult’, ‘Derogatory’, ‘Profanity’, and ‘Obscenity’ in temperature 0, 0.25, 0.5, 0.75 and 1. The results reveals that the model performs better with identifying general attributes compared to granular attributes.



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.

Appendix

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

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.