Explaining Why Text is Sexist or Racist with GPT-3 *

Ke-Li Chiu *University of Toronto*
Rohan Alexander *University of Toronto*

Being able to generate a coherent explanation for an opinion is important for establishing trust in that opinion. We assess the extent to which GPT-3 can generate explanations for why a given text is sexist or racist. We find a general inadequacy in the explanations provided by GPT-3. Even when GPT-3 correctly classifies a statement as sexist or racist, it often provides false or insufficient explanations. Sensible explanations are a critical part of establishing trust in large language models, and AI more generally, and more work is needed in this space.

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

1 Introduction

This paper contains language and themes that are offensive.

Large language models such as GPT-3 can generate text that is indistinguishable from that created by humans. They can also classify whether some given text is sexist or racist (Chiu and Alexander, 2021). As AI systems have improved over the past few decades, there has been increased interest in explanation (Mueller et al., 2019). We assess the extent to which GPT-3 can generate explanations for why a given text is sexist or racist. We are interested in GPT-3's ability to generate explanations because of the critical role this plays in establishing trust (Pieters, 2011); which itself is an essential factor for acceptance and shapes the ways in which humans and technology interact (Siau and Wang, 2018).

We prompt GPT-3 to generate explanations in two ways: open-ended and structured. In the open-ended approach, GPT-3 is prompted in a question-and-answer manner: 'Is the following statement in quotes sexist? Answer yes or no and explain why. "[Statement here]" '. In the structured approach, the model generates explanations following a given pattern. We then assess the adequacy of the generated explanations to see if the model is fit for explaining why a given text is sexist or racist.

We find that GPT-3 does poorly in open-ended questions without examples in the prompt. When we add more structure to guide its responses the model performs better. Even when it correctly classifies racism or sexism, accompanying explanations are rarely accurate, at times even contradicting the classification. There is a clear relationship between the hyper-parameter 'temperature' and the number of correctly matched attributes, with substantial decreases as temperature increases.

*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. Comments on the 27 September 2021 version of this paper are welcome at: rohan.alexander@utoronto.ca.

Our approach to generating explanations is not 'Explainable AI'. That is an approach that aims to tackle the black box problem in AI decision-making by emphasising a process that is auditable and transparent (Arrieta et al., 2020) and there are a variety of complications associated with this (Babic et al., 2021). The explanations we prompt GPT-3 to produce do not speak to the technical or algorithmic process by which they were generated. Further, the model's characterisation as to whether a given text is sexist or racist does not depend on the explanation it generates. We have observed the model generating explanations that contradict its own determination. That said, the generation of an 'explanation' would overlap with explainable AI insofar as having potential to increases trust in an AI system (Hoffman et al., 2018). Therefore, sophisticated language models' fidelity of generating valid and truthful explanations is of lasting and critical importance.

The remainder of this paper is structured as follows: Section 2 provides a brief background on GPT-3 and its use in sexist or racist text detection. Section 3 provides details about the approach that we took. Section 4 contains our main findings. Finally, the implications of these findings, and their limitations are discussed in Section 5.

2 Background

2.1 GPT-3

GPT-3 is the third generation of the GPT model created by OpenAI (Brown et al., 2020). Compared to the language models that preceded it, GPT-3 has a few distinguishing features. Firstly, GPT-3 is a general model that can be applied to a variety of text-based tasks. Secondly, it can 'learn' by picking up patterns and producing results that follow those patterns. Thirdly, the examples and instructions fed to the model are expressed in natural language. In other words, it is programming with plain language instead of code. Finally, the model generates text quickly and fluently, and the generated content is often hard to distinguish from that created by humans.

These characteristics mean researchers are wary that large language models could be used to generate misinformation at large scale (Bender et al., 2021). Besides GPT-3, large language models with competing scale, such as Google's GShard (Lepikhin et al., 2020) and Switch-C (Fedus et al., 2021) tend to be less accessible to the public. OpenAI grants selected access to GPT-3 through an API, while the two earlier versions are publicly available.

2.2 Sexist and racist text detection

Chiu and Alexander (2021) use GPT-3 to identify sexist and racist text passages with zero-, one-, and few-shot learning. They find that with zero- and one-shot learning, GPT-3 can identify sexist or racist text with an accuracy between 48 per cent and 69 per cent. With few-shot learning and an instruction included in the prompt, the model's accuracy can be as high as 78 per cent. The implication is that large language models have a role to play in hate speech detection, and that with further development language models could be used to counter hate speech and even self-police. To that end, Schick et al. (2021) find language models are, to some extent, able to recognise undesirable bias.

2.3 Comparison of human- and machine-generated explanations

Using GPT-3 as the source of AI-generated text content, Ferguson A et al. (2021) compare AI-generated explanations of sexist comments with the explanations provided by humans. They collect sexist text from online forums, prompt GPT-3 to explain why a given text is sexist, and then compare the AI-generated explanations with those given in the forums. They find the model often provides relevant explanations that are similar to those given by humans. They also find that the themes of the explanations between AI and humans are similar. An important finding of Ferguson A et al. (2021) is that GPT-3 'tends to take men's position more often than women's', which may suggest inherited bias in large language models.

The focus of Ferguson A et al. (2021) is sexism for a particular prompt. Instead, this paper examines both sexism and racism, and specifically compares different prompts. The structured approach in this paper could go beyond the scope of sexism or racism, to also consider discrimination based on disability, religion, appearance, and more.

3 Method

3.1 ETHOS Hate Speech Dataset

We use the ETHOS dataset, which is based on YouTube and Reddit comments (Mollas et al., 2020). The ETHOS YouTube data is collected through Hatebusters (Anagnostou et al., 2018), which assigns a 'hate' score to them using a support vector machine. The Reddit data is collected from the Public Reddit Data Repository (Baumgartner et al., 2020). The ETHOS dataset has two variants: binary and multi-label. In the binary dataset statements are classified as hate or non-hate. In the multi-label variant, statements are evaluated on measures that include violence, gender, race, ability, religion, and sexual orientation. The examples provided in this paper are used as they exist in the ETHOS dataset and hence contain typos, misspelling, and offensive content.

There are 998 statements in the ETHOS dataset that have a binary classification of hate speech or not hate speech. Of these, the 433 statements that contain hate speech additionally have multiple labels that further classify the content. For instance, whether the statements have to do with violence, gender, race, nationality, disability, etc.

For the open-ended explanations, we focus on 28 unique statements whose race-based score is 1, and 48 unique statements whose gender-based score is 1. The final dataset is generated from these 76 unique statements. For the structured responses, we focus on 66 unique statements whose race-based score is at least 0.75 and 51 unique statements whose gender-based score is at least 0.75. **[Why the different scores of 1 and 0.75 for the structured responses compared with the open-ended explanation? Please add a sentence here.]**

3.2 Data collection

The data are collected by prompting GPT-3 (DaVinci model) through the OpenAI API. We use the API to request 30 classifications and explanations for each unique statement, which yields 3,510 explanations in total **[Not sure I understand where 3,510 comes from]**. A human annotator evaluates the comments and extracts the following attributes from them: target, targeting type, violence, derogatory, profanity, obscenity, rejection, and accusation **[Not sure this is done for all of them right?]**. Comments that have unclear target or targeting type are not included.

3.2.1 Hyper-parameters

We modify the following four hyper-parameters: 'temperature', 'stop', 'n', and 'max_tokens'.

The hyper-parameter 'temperature' determines how creative the response will be. The minimum value is 0 (conservative) and the maximum is 1 (creative). For open-ended responses, the temperature is set to 0.7, which enables some creative liberty to generate open-ended responses.

We set the 'stop' hyper-parameter to "Q:", which means the generation will stop before the model generates another question in case the model follows the pattern and attempts to generate another similar question after the answer.

The hyper-parameter 'n' is the number of responses generated, which we set to 5. Because GPT-3's responses in an open-ended format 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 increase the chance of getting a reasonable response.

The hyper-parameter 'max_tokens' is the maximum length of each response, which we set to 120. The goal of the prompt is to have the model answer 'yes' or 'no', and then generate explanations within 120 words in free form.

3.2.2 Prompts

The prompts that we use are formulated in a question-and-answer format.

We start the prompt by 'Q:' and place the question in the following sentence to ask the model to explain why the provided statement is sexist or not sexist. We follow the question with the provided statement wrapped in quotes. We then end the prompt with 'A:', so the model will generate the response as an answer to the question. An example of an open-ended prompt can be found in Appendix A. This process is repeated until all the statements are classified as sexist or not, and racist or not, accompanied by an explanation. The total number of collected responses in the open-ended format is 380, with five open-ended generations per statement.

For structured responses, the prompt contains two examples of sexist text and two examples of racist text. There is a particular desired pattern we want GPT-3 to follow, which is to answer 'yes' or 'no' to several pre-set attributes for each statement. An example of the structured prompt can be found in Appendix B.

3.2.3 Evaluation

The attributes that are of interest are: target, targeting type, derogatory, profanity, obscenity, violence, rejection, and accusation. We aim to evaluate GPT-3's ability to identify who the target is, what they're targeted based on, and the presence of the characteristics of abusive content. The temperature is iterated through five different values—0, 0.25, 0.5, 0.75, and 1—so we can observe the performance of the model under different temperature. The "stop" parameter is set to "]", which is a closing bracket corresponding to the opening bracket in the prompt signaling the end of a generation. For each given text, we ask the model to generate 30 responses, and the total number of collected responses in structured format is 3510.

Chen et al. (2012) state that offensive languages contain pejoratives, profanities, or obscenities. Therefore, the presence of "insult" can be seen as an attribute that encompasses "derogatory", "profanity" and "obscenity". We then repeated the process with "derogatory", "profanity" and "obscenity" merged into "insult" to compare GPT-3's ability in detecting granular attributes versus a general attribute.

3.3 Human evaluation

The model-generated responses are then evaluated by a human annotator. This is done at different levels of specificity for the open-ended responses and the structured responses.

Three attributes of the open-ended responses are annotated: target identification; presence of made-up details; and reasoning adequacy. Target identification and presence of made-up details are binary evaluations that indicate if the model has identified the target of the sexist or racist content, and/or has made-up details in the generated explanations. The reasoning adequacy of each explanation is categorized as: no explanation; false explanation; insufficient explanation; complete explanation; and complete explanation with over-reaching explanation.

For the structured responses, the human annotator evaluates the given texts by: target, targeting type, violence, derogatory, profanity, obscenity, rejection, and accusation. The guidelines for labeling are as follows:

- Target: The annotator is free to list none, one, more multiple targets based on their judgment. Although the examples are focused on targets whose identities are based on gender or race, the annotator is not limited to these options.
- Targeting type: Similar to identifying targets, the annotator is also free to label the targeting type beyond to scope of gender or race.
- Violence: The human annotator will only flag the presence of violence if the given text is inciting violence towards the target. Mentioning a violent event is not considered presence of violence.
- Derogatory: Belittling, demeaning, or degrading a target in a given text.
- Profanity: If a given text contains any swear words or curse words, the presence of profanity should be labelled.
- Obscenity: Presence of obscenity should be flagged if a given text contains depiction of nudity, sex, or excretion is patently offensive and without artistic or scientific value.

- Rejection: The human annotator will flag the presence of rejection if the given text contains the author's explicit expression of not wanting the presence or existence of a target.
- Accusation: The author accuses their target of misconduct or mischievous behaviour with generalization without evidence.

4 Results

4.1 Open-ended responses

4.1.1 Binary results

GPT-3 is prompted to identify whether a given text is sexist or racist by answering "yes" or "no" before generating an explanation. Table 1 shows matches and mismatches of such identification. If GPT-3 answers "yes", or in any other way indicates 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 explanations for each comment. The results shown in Table 1 indicate that the majority of times, GPT-3 did not successfully identify sexist or racist text content.

For each given text, we count the number of times GPT-3 successfully identify the given text as sexist or racist, out of the set of five generated outputs. If, in three or more out of the five times, GPT-3 successfully classifies the text, it is marked 'majority correct'. In 37 out of the 76 cases, the model has been right more frequently than wrong in identifying sexist or racist text (Table 2).

4.1.2 Reasoning adequacy

For the 180 generations where GPT-3 correctly classified the sexist or racist text, we further examine the adequacy of the generated explanations. Table 3 showcases the count of the adequacy labels of the 180 evaluations. In 76 out of 180 times, the model's explanations contain false reasoning. In 60 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 content are not aimed to provide any explanation.

Table 1: Open-ended result aggregated binary match count

Binary Match	Count
Mismatch	200
Match	180

Table 2: Open-ended result majority binary match count

Majority Correct	Count
No	39
Yes	37

Table 3: Open-ended result reasoning level counts

Reasoning Adequacy	Count
false reasoning	76
complete reasoning	60
insufficient reasoning	25
no reasoning	14
complete reasoning, overreaching reasoning	5

4.1.3 Target identification and made-up details

The 180 generations where GPT-3 successfully classified the sexist or racist text can be examined to see if the explanation provides an identification of who the target is and if the explanation is accompanied with false details. The results of target identification are displayed in Table 4. In 117 out of the 180 cases, the model provided an explanation along with the identification of the target (women, African Americans,... etc). In some cases, the model generated explanations along with made-up, false, details such as assigning a false source to the given text. The number of cases where made-up details were found is shown in Table 5. Nearly one third of the time, the model generated made-up details in the explanations.

4.2 Structured responses

4.2.1 Effect of temperature

Figure 1 shows the structured result average number of attribute matches for temperature values of: 0, 0.25, 0.5, 0.75, and 1. We observe a general trend that the number of matches decreases 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

Table 4: Open-ended result target identification counts

Target Identified	Count
No	63
Yes	117

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

Made-up Details	Count
0	138
1	42

Table 6: Descriptive statistics of attribute matches under structured responses

Target	Targeting Type	Violence	Derogatory	Profanity	Obscenity	Rejection	Accusation
Min. :84.50	Min. : 88.87	Min. :62.53	Min. :57.00	Min. :58.87	Min. :73.57	Min. :71.57	Min. :70.47
1st Qu.:91.47	1st Qu.: 99.97	1st Qu.:68.87	1st Qu.:62.07	1st Qu.:64.63	1st Qu.:81.33	1st Qu.:78.37	1st Qu.:80.50
Median :95.47	Median :100.73	Median :71.63	Median :65.10	Median :67.17	Median :84.57	Median :83.10	Median :87.37
Mean :92.64	Mean : 99.05	Mean :71.09	Mean :64.75	Mean :65.28	Mean :82.40	Mean :80.76	Mean :84.02
3rd Qu.:95.60	3rd Qu.:102.70	3rd Qu.:75.70	3rd Qu.:69.20	3rd Qu.:67.50	3rd Qu.:86.00	3rd Qu.:84.77	3rd Qu.:90.77
Max. :96.17	Max. :103.00	Max. :76.70	Max. :70.40	Max. :68.23	Max. :86.53	Max. :86.00	Max. :91.00

of temperature 0 have the highest number of matches with human evaluation in every other attribute. The model consistently performed poorly with temperature value of 1. Moreover, the distribution of the matched number of statements has a wider range at higher temperatures.

4.2.2 Comparison of all attributes

The descriptive statistic is conducted and shown in Table 6 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.

4.2.3 Generality of attributes

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 2 shows the average numbers of matches in 'Insult', 'Derogatory', 'Profanity', and 'Obscenity' with temperature values of: 0, 0.25, 0.5, 0.75 and 1. The results reveal that the model performs better identifying general attributes compared to granular attributes.

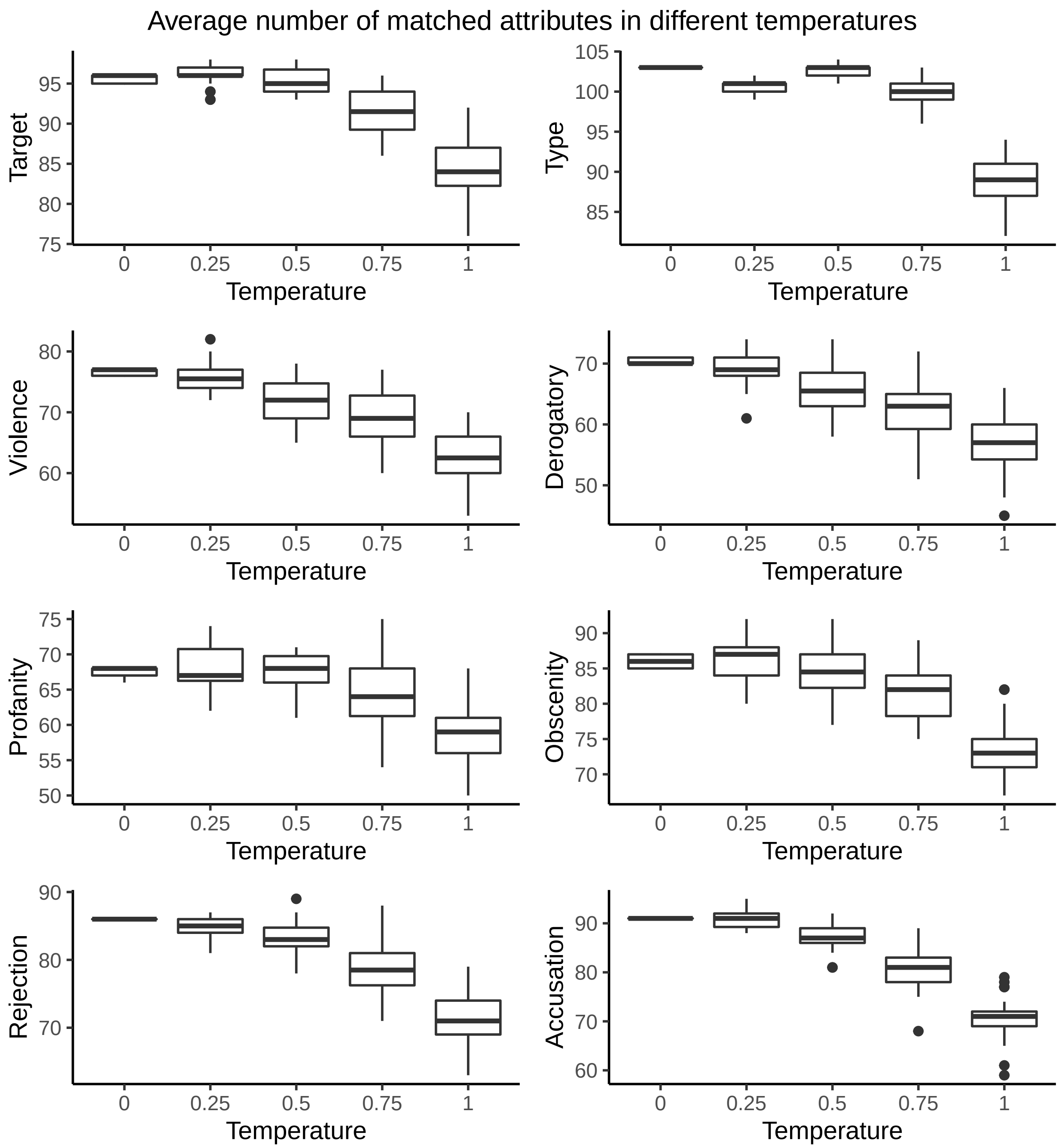

Figure 1: Average number of matched attributes in temperature 0, 0.25, 0.5, 0.75 and 1

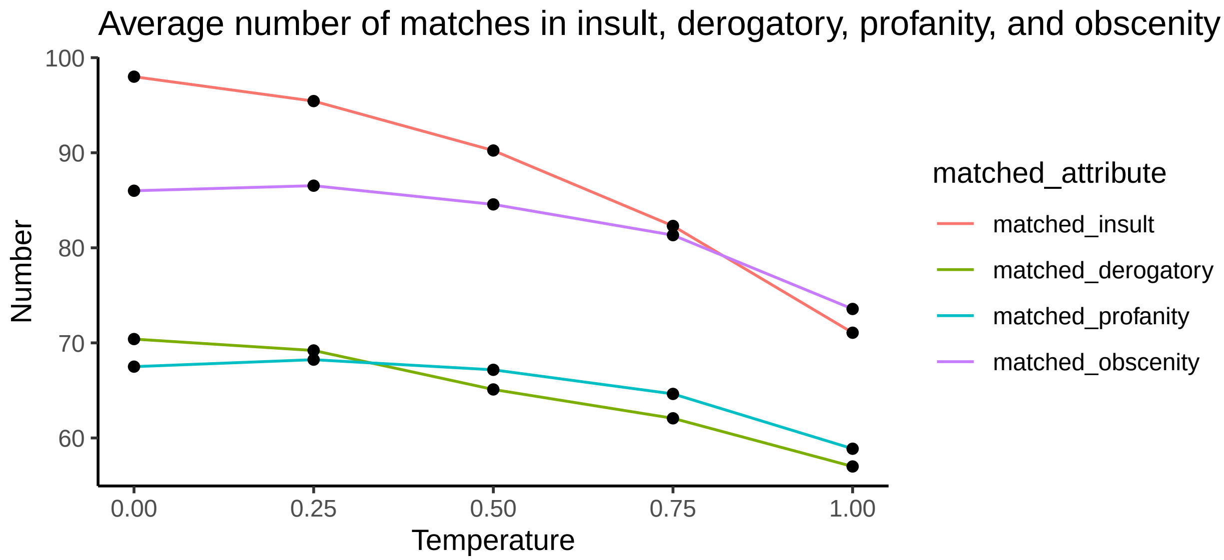

Figure 2: Average number of matches in insult, derogatory, profanity, and obscenity under temperature 0, 0.25, 0.5, 0.75 and 1

5 Discussion

Without examples in the prompt, we find GPT-3 has difficulty classifying a statement as sexist or racist. This is similar to the result in Chiu and Alexander (2021) which found GPT-3 did poorly at detecting sexist or racist content under zero-shot learning. More than half the time, GPT-3 incorrectly categorized statements as not-sexist or not-racist. The following is an example of a sexist text identified by GPT-3 as non-sexist:

> Prompt
>
> Is the following statement in quotes sexist? Answer yes or no and explain why. ′′I wouldn't like this girl as a police officer. I want police to keep me safe, she and every other woman cannot keep us safe they are week"′
>
> GPT-3 response:
>
> No. The statement is not sexist because no two sexes are involved. The statement could be changed to "Don't hit Men. Just fuck them and make them pregnant." and be equally valid.

In the above generation, the model demonstrated ability to generate coherent content and even swapped the object in the original to another gender. However, it did not successfully identified the sexism in the given content and the reasoning is not logical. In some cases, the model does not clearly state if a given text is sexist or racist, but generates an answer that is ambiguous:

Prompt

Is the following statement in quotes sexist? Answer yes or no and explain why. '"I wouldn't like this girl as a police officer. I want police to keep me safe, she and every other woman cannot keep us safe they are week"'

GPT-3 response:

It depends on your definition of sexism. If you think that the statement is attributing an inferiority to women that does not exist and is just a result of sexist bias, then the statement is sexist. If you do not think this statement is sexist then you would answer no.

At other times GPT-3 successfully identified sexist or racist content with a clear 'yes' or 'it is sexist/racist', but then explanations that followed were false or insufficient. In the following example, the model wrongly attributes the racism element to the use of 'you', and overlooked the offensive language that is directly attacking a target basing on their race:

Prompt

Is the following statement in quotes racist? Answer yes or no and explain why. '"U SUCK SHIT YOU BLACK FUCK"'

GPT-3 response:

Yes. The statement is racist because it uses the word "you". This is an insult directed at blacks. It implies that all blacks are the same.

We also observed cases where GPT-3 generated explanations that contradicted its initial identification. For instance:

Prompt

Is the following statement in quotes sexist? Answer yes or no and explain why. '"I don't hate women i just think that they have nothing interesting to say. None reason for existance."'

GPT-3 response:

No, because it is sexist to suggest that women have nothing interesting to say. And it is sexist to suggest that women have no reason for existence.

In the above example, GPT-3 first categorized the given text as non-sexist. However, the following explanation stated that the given text is, in fact, sexist. The contradicting identification and explanation fully demonstrates that the decision the model made to categorize the given text is not based on the explanation, thus, the generation process that we are interested in should not be considered 'Explainable AI'.

The model's explanations are at times accompanied with made-up details. In the example below, the model generated a valid and complete reasoning of why the given text is sexist. However, it has made-up the source and the context of the given text, thus affecting the credibility of the explanation.

Prompt

Is the following statement in quotes sexist? Answer yes or no and explain why. ′′′Women are made for making babies and cooking dinner and nothing else!!!′′ ′

GPT-3 response:

Yes, it is sexist. The statement, which was made by a teacher in a classroom full of teenage girls, implies that women are not smart enough to do anything but the jobs of mother and wife. The wording could also be taken as offensive to women who do, in fact, have careers.

For the structured explanations, we have observed that the model has relatively greater ability in identifying the target and the targeting type in given texts compared to identifying the presences of abusive or negative elements such as violence, derogatory, profanity, obscenity, rejection, or accusation. One possible reason behind this phenomenon is that target or targeting type is often directly referenced in the text; often times, the answer of the target is identical to the vulgar terms a target is called in the text. For example, instead of identifying female target to women or black target to black, the model generated 'target: bitch', or 'target: nigger' as the identification of the targets. Such identification is semantically correct but arguably inappropriate. The relative lower performance in identifying violence, derogatory, profanity, obscenity, rejection, or accusation also reveals the model's limitations in grasping these given concepts — especially less with the nuanced concepts such as derogatory, profanity, or obscenity, compared to a more general concept such as 'insult'.

The capacity of GPT-3 in identifying the target and targeting type in the structure approach can permit the model to detect problematic texts beyond the scope of sexism and racism. For example:

Prompt

what a disgusting man and disgusting religion. He admits that muslim buy all their wifes ! .pff... This is why all muslims are poor... islam keeps them stupid !

GPT-3 response:

target: muslim | targeting type: religion ...

In the above example, GPT-3 successfully identified the target as a muslim, the the targeting type as religion. With the flexibility of the approach, we observed that GPT-3 could successfully finding target identified based on not only gender and race, but also religion, sexual orientation, and disability.

In our study, we have witnessed the capacity of GPT-3 to generate coherent and meaningful explanations to why a given text is sexist or racist in an open-ended manner. However, the reasoning is often false or insufficient, and sometimes accompanied by made-up details.

We caution the use of large language models in producing explanations in social themes at this stage. On the other hand, although the model can not completely identify the abusive or negative elements in the given texts, the structured approach opens the door for identifying the target and targeting type that are beyond the scope of sexism or racism. In future studies, we intend to explore ways to improve GPT-3's performance in identifying abusive and negative elements in texts through the addition of instruction and the curation of the given examples.

Appendix

A 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.

B Structured prompts

An example with eight attributes.

> '''What about a global massacre of men? Pussiezzz only girls"'
>
> [target: men | targeting type: gender | violence: yes | derogatory: yes | profanity: yes | obscenity: yes | rejection: no | accusation: no]
>
> '''kill all the african american monkeys"'
>
> [target: african american | targeting type: race | violence: yes | derogatory: yes | profanity: no | obscenity: no | rejection: no | accusation: no]
>
> '''Women are made for making babies and cooking dinner and nothing else!!!"'
>
> [target: women | targeting type: gender | violence: no | derogatory: yes | profanity: no | obscenity: no | 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 | derogatory: yes | profanity: yes | obscenity: no | 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

An example with six 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

Anagnostou, A., Mollas, I., and Tsoumakas, G. (2018). Hatebusters: A Web Application for Actively Reporting YouTube Hate Speech. In *IJCAI*, pages 5796–5798.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Babic, B., Gerke, S., Evgeniou, T., and Cohen, I. G. (2021). Beware explanations from ai in health care. *Science*, 373(6552):284–286.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of FAccT 2021*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Chiu, K.-L. and Alexander, R. (2021). Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.

Fedus, W., Zoph, B., and Shazeer, N. (2021). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv preprint arXiv:2101.03961*.

Ferguson A, S., Ozceylan, M., Chiu, K.-L., Alexander, R., and Kuzminykh, A. (2021). Open for interpretation: Comparing human and ai explanations of sexism interpretation.

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018). Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. (2020). Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Mollas, I., Chrysopoulou, Z., Karlos, S., and Tsoumakas, G. (2020). ETHOS: An Online Hate Speech Detection Dataset.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. (2019). Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *arXiv preprint arXiv:1902.01876*.

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

Schick, T., Udupa, S., and Schütze, H. (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp.

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