

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 OpenAI’s 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). Trust 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, we prompt GPT-3 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, we prompt GPT-3 to follow a given pattern. We then assess the adequacy of the explanations generated by GPT-3. We are interested in firstly, the extent to which it correctly classifies whether the statements are sexist/racist or not;

***Acknowledgements:** We thank Monica Alexander and Thomas William Rosenthal for helpful comments. Comments on the 06 October 2021 version of this paper are welcome at: rohan.alexander@utoronto.ca. **Funding statement:** We have no direct financial disclosures. We gratefully acknowledge the indirect support of Gillian Hadfield, the Schwartz Reisman Institute for Technology and Society, and OpenAI for providing free access to GPT-3. **Data accessibility statement:** Code and data are available at: <https://github.com/kelichiu/GPT3-sexist-racist-content-explanation>. We are unable to share direct access to GPT-3, however have provided the code that generated the GPT-3 output and all GPT-3 output. Please note that we have added asterisks to racial slurs and other offensive content in this paper, however the inputs and outputs did not have these. **Author contributions:** Chiu had the original idea, obtained and created the datasets, and wrote the first draft of the paper. Chiu and Alexander analysed and interpreted the data, contributed to writing the paper, and approved the final version.

and secondly, the reasonableness of the explanation that accompanies that classification.

We find that GPT-3 does poorly in the open-ended approach. When we add more structure to guide its responses the model performs better. But even when it correctly classifies racism or sexism, the accompanying explanations are rarely accurate. At times they even contradict the classification. On a technical level, we find a clear relationship between the hyper-parameter *temperature* and the number of correctly matched attributes, with substantial decreases as *temperature* increases.

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 classification of a statement as sexist or racist does not depend on the explanation it generates and can contradict it. That said, the generation of an explanation would overlap with explainable AI insofar as having potential to increase 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 some limitations, are discussed in Section 5.

2 Background

2.1 GPT-3

GPT-3 is the third generation of the Generative Pre-trained Transformer (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). Large language models, 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, namely GPT and GPT-2, are publicly available.

2.2 Sexist and racist text detection

Chiu and Alexander (2021) use GPT-3 to identify sexist and racist hate speech 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 of that paper 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. Similarly, [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 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 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 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

The source of our racist and sexist statements is 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 each comment. 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 statements that we use are 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 labels that further classify the content. For instance, whether the statements have to do with violence, gender, race, nationality, disability, etc. The label content is the proportion of annotators that positively classified the statement with that label and ranges between 0 and 1.

For the open-ended approach we qualitatively evaluate the results, so we only consider 76 unique statements: the 28 unique statements whose race-based score is 1, and the 48 unique statements whose gender-based score is 1. For the structured approach we consider 117 unique statements: the 66 unique statements whose race-based score is at least 0.75 and the 51 unique statements whose gender-based score is at least 0.75. For the structured approach, we reduce the threshold to 0.75 to include more statements in the dataset. This increased sample size is important in the structured approach because the evaluation of the results is more quantitative compared with the open-ended results whose evaluation is more qualitative.

3.2 Data collection

The data that we are interested in are the outputs of GPT-3. We prompt GPT-3 (DaVinci model) through the OpenAI API, and obtain three datasets: open-ended results, structured results with eight attributes, and structured results with six attributes. When generating the open-ended results, we prompt GPT-3 to produce five classifications and explanations per statement, which yields 380 generations in total.

For the structured explanations, we use the API to generate two datasets, each consisting of 30 classifications and explanations for each of the 117 statements, for five different temperatures. The two datasets differ insofar as the number of attributes that we prompt GPT-3 to consider. One dataset has eight attributes: *target*, *targeting type*, *violence*, *derogatory*, *profanity*, *obscenity*, *rejection*, and *accusation*. The other dataset has only six: *target*, *targeting type*, *violence*, *insult*, *rejection*, and *accusation*. The latter dataset, where we combine *derogatory*, *profanity*, and *obscenity* into *insult*, is a complementary dataset used to assess the model’s capacity to recognize granular attributes (derogatory, profanity, and obscenity) compared with a more general attribute (insult).

To evaluate the structured explanations, we primarily consider the dataset with eight attributes. There are 17,550 generations (117 unique statements each considered 30 times for each of five temperatures) in each of these two datasets.

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 is the default temperature in the OpenAI GPT-3 Playground. We choose to maintain the default setting, which enables some creative liberty to generate open-ended responses. On the other hand, for the structured responses, the temperature is iterated through 0, 0.25, 0.5, 0.75, and 1. We do not conduct grid tuning because we are not looking for the ‘best’ performance, but to compare the variation of performance under different temperature values.

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 five. 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 free-form explanations comprising less than 120 tokens (a token is roughly equivalent to a word).

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 a sentence to ask the model to explain why the provided statement is sexist/racist or not sexist/racist. 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, and 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 a 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 are targeted based on, and the presence of the characteristics of abusive content. The *temperature* parameter is iterated through five different values—0, 0.25, 0.5, 0.75, and 1—as this changes the nature of the responses that we would expect. 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.

Chen et al. (2012) state that insulting language contains 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 a single attribute *insult* to compare GPT-3’s ability to detect granular attributes and a general attribute.

3.3 Human evaluation

The model-generated responses are then evaluated by an 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 annotator evaluates the given text by: target, targeting type, violence, derogatory, profanity, obscenity, rejection, and accusation. The guidelines for labeling are:

- 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 annotator will only flag the presence of violence if the given text incites violence towards the target. Mentioning a violent event is not considered presence of violence.
- Derogatory: The annotator flags content that is belittling, demeaning, or degrading toward a target in a given text ([Merriam-Webster, a](#)).
- Profanity: If a given text contains any swear words or curse words ([Merriam-Webster, b](#)), 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 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 ([Federal Communications Commission, 2021](#)).
- 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 of 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 GPT-3 did not successfully identify sexist or racist text content a majority of the time.

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 |

For each given text, we count the number of times GPT-3 successfully identifies it 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’. For 37 out of the 76 statements, the model was majority-correct in terms of identifying sexist or racist statements (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 shows 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 generates explanations with complete reasoning; in five cases, the model generates explanation with completed reasoning accompanied with overreaching reasoning. In 25 cases, the model produces explanations with insufficient reasoning; in 14 cases the model does not produce any explanation, or produces content that is not aimed at providing an explanation.

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 |

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 |
|-----------------|-------|
| No | 138 |
| Yes | 42 |

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 provides an explanation along with the identification of the target (women, African Americans, ...). In some cases, the model generates 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 are found is shown in Table 5. Nearly one-third of the time, the model generates 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 performs the best under temperature 0.25 by a small margin, the results of temperature 0 have the highest number of matches with the human evaluation in every other attribute. The model consistently performs 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 statistics shown in Table 6 assess overall performance in each attribute. The model performs the best in identifying the identity categories (sex, gender, race, origin, ...) that a sexist/racist statement is targeting. The model also performs well in identifying who the target is (women, African Americans, ...). The model seems to have

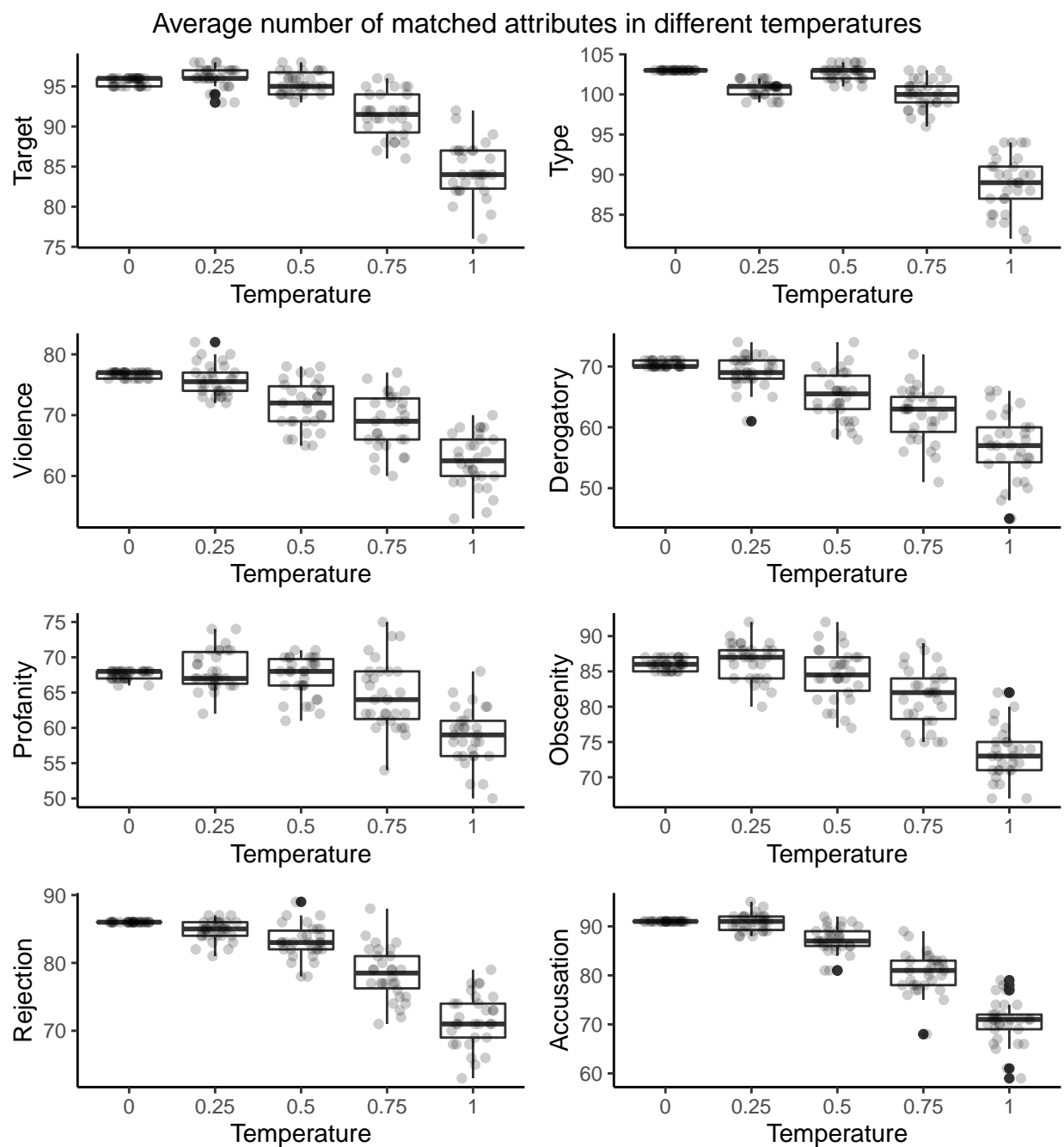


Figure 1: Average number of matched attributes in temperature 0, 0.25, 0.5, 0.75 and 1 (note the differing y-axes)

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 |

varied capabilities in identifying the presence of violence, derogatory, profanity, obscenity, rejection, and accusation in the statements.

4.2.3 Generality of attributes

The general attribute insult encompasses derogatory, profanity, and obscenity, which are 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 show that the model tends to perform better at identifying general attributes compared to granular attributes.

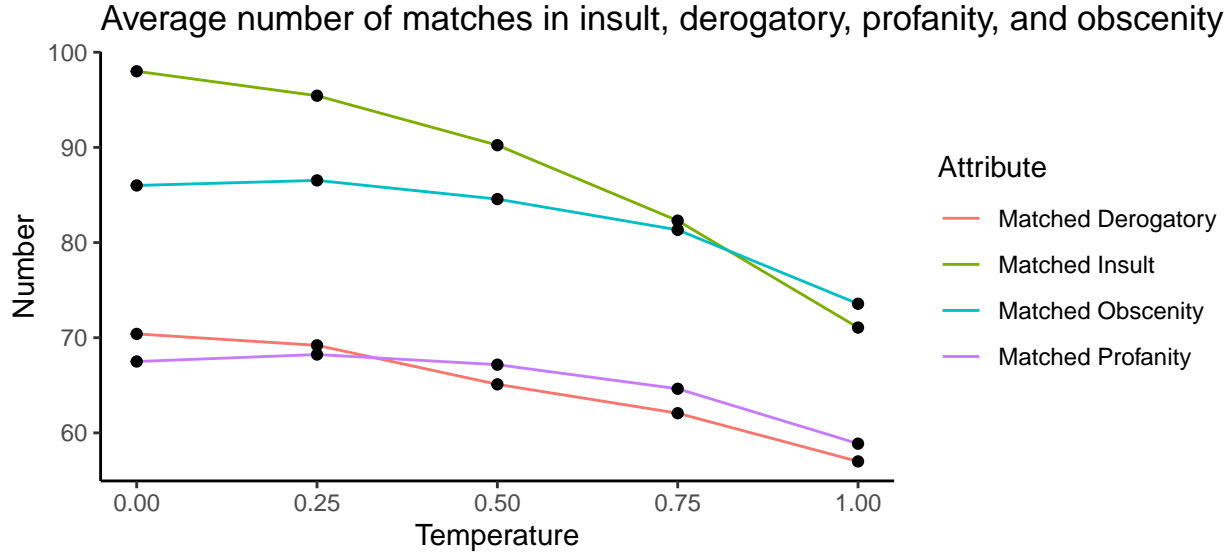


Figure 2: Average number of matches in insult, derogatory, profanity, and obscenity under temperature 0, 0.25, 0.5, 0.75 and 1

5 Discussion

5.1 GPT-3 generally struggles to provide explanations

Without examples in the prompt, we find GPT-3 has difficulty classifying a statement as sexist or racist. This is similar to [Chiu and Alexander \(2021\)](#) who 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 statement identified by GPT-3 as not-sexist (asterisks added *ex post*):

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 f*** them and make them pregnant." and be equally valid.

In the above generation, the model demonstrates its ability to generate coherent content and even swaps the object in the original to another gender. However, it does not successfully identify the sexism in the given content and the reasoning is not logical.

In some cases, the model does not clearly state if a given statement is sexist or racist, but generates an answer that is ambiguous. The following is another generation provided by GPT-3 to that same prompt:

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 the explanations that followed are false or insufficient. In the following example, the model wrongly attributes the racist element to the use of 'you', and overlooks the offensive language that is directly attacking a target based on their race:

Prompt:

Is the following statement in quotes racist? Answer yes or no and explain why. ''U SUCK S*** YOU BLACK F***''

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 observe cases where GPT-3 generates explanations that contradict 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 categorizes the given statement as non-sexist. However, the following explanation states that the given statement is, in fact, sexist. The contradicting identification and explanation demonstrates that the classification 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 statement is sexist. However, it has made-up the source and the context of the given statement, 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 observe that the model has relatively greater ability to identify the target and the targeting type in given statements compared to identifying the presence of abusive or negative elements such as violence, derogatory, profanity, obscenity, rejection, or accusation. One possible reason behind this is that target or targeting type is often directly referenced in the statement; often the answer of the target is identical to the vulgar terms a target is called in the statement. For example, instead of identifying female targets as women, or black targets as black, the model often generates ‘target: b**ch’, or ‘target: ni**er’ as the identification. Such identification is semantically correct but 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 with nuanced concepts such as derogatory, profanity, or obscenity, compared to a more general concept such as ‘insult’.

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

Prompt:

what a disgusting man and disgusting religion. He admits that muslim buy all their wives ! .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, and the targeting type as religion. With the flexibility of the approach, we have observe that GPT-3 could successfully identify the target identified based on not only gender and race, but also religion, sexual orientation, and disability.

5.2 Conclusion, weaknesses and next steps

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

One particular issue with the human evaluation stage of the GPT-3 output is that there was just one evaluator. It is possible that this evaluator is biased in certain ways, and adding multiple evaluators would go part-way to addressing these concerns.

We caution the use of large language models in producing explanations of social themes at this stage, especially in open-ended settings. The model did not completely identify the abusive or negative elements in the given statements. But we see the structured approach as having potential to identify targets and targeting types beyond sexism or racism. In future studies, we intend to explore ways to improve GPT-3’s performance in identifying abusive and negative elements in statements 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 strongner... 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 Conference 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*.
- Federal Communications Commission (2021). Obscene, Indecent and Profane Broadcasts.
- 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, 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*.
- Merriam-Webster. Derogatory.
- Merriam-Webster. Profanity.

- 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.