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 previous study, we observed that GPT-3 the model's accuracy in detection sexist or racist text contents can be as high as 78 per cent [**cite our own paper**]. The importance of explanation in AI has been emphasized in numerous popular press outlets over the past decades (Mueller et al., 2019). In this study, we aim to assess GPT-3 ability to generate explanations on why given text contents are sexist or racist. 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).

Our approach to generate explanations is not to be mistaken from Explainable AI. Explainable AI is an approach that aims to tackle the black box problem in AI decision making by making the decision making process auditable and transparent [**Citation**]. The explanations we prompt GPT-3 to produce are to be interpreted solely from a social lense, and are not shedding a light on the technical or algorithmic process of the generation. In other words, the model's decision on if a given text is sexist or racist does not depend on the explanation in generates. In fact, we have observed the model generating contradicting explanations to its own determination, discussed in the later section.

However, such generations of explanations has an overlapping effect with Explainable AI — increasing users' trust in an AI system. Hoffman et al. (2018) state that various forms on intelligent systems are trusted more if their decisions are explained. Therefore, sophisticated language models' fidelity of generating valid and truthful explanations should be evaluated.

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. In the open-ended method, GPT-3 is prompted to produce explanations on why a given text is sexist or racist in a question-answering manner. On the other hand, the model generates explanations following a given pattern under the

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

structured method. 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.

Background

GPT-3

GPT-3 is the third generation of the GPT model created by OpenAI (Brown et al., 2020). Compared to the language models before GPT-3, GPT-3 is considered groundbreaking for a few reasons. First, GPT-3 is not specialized in one or just a few tasks, but a generalist that can be trained in doing all kinds of tasks. Second, it learns like human do. Given only a few examples, GPT-3 can pick up the patterns in the examples, and produce the results following the given patterns. Third, 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. Lastly, the model generates the text quickly and fluently, and the generated contents are hard to be distinguished from human creation. Because of these powerful characteristics of this technology, researchers are wary that large language models will be used to generate misinformation at large scale (Bender et al., 2021). Besides GPT-3, large language models with competing scales such as Google G-Shard [**citation**] and G-Switch [**citation**] are as well not accessible to the public. As to date this paper is written, GPT-3 is not open to the public, and only selected users are granted the access through OpenAI API.

Sexist and Racist Text Detection

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**]. In our previous paper, we use GPT-3 to identify sexist and racist text passages with zero-, one-, and few-shot learning. We find that with zero- and one-shot learning, GPT-3 is able to 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. We conclude 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. We expand the research to explore the explanations of why a given text passage is sexist and racist in this paper.

Comparison of Human- and Machine-Generated Explanations

Using also GPT-3 as the source of AI-generated text contents, **Sharon et al citation** presented an unprecedented study that compares AI-generated explanations of subtle sexism assessment with human-creation. The authors collected a series of sexist text content from online forum, prompted GPT-3 to explain why a given text is sexist, then compared the AI-generated explanations with human creation. The authors have found that the model could often provide relevant explanations that are easily understood for human. They have also found that the themes of the explanations between AI and human are similar to each other. An important finding of **Sharon et al citation** is that GPT-3 "tends to take men's position more often than women's", which is an evidence of inherited bias in large language models. While the focus of **Sharon et al. citation** is sexism assessment, our study examines both sexism and racism assessment with both open-ended and structured approaches. In fact, the structured approach we proposed in this study could go beyond the scope of sexism or racism, but discrimination against disability, religions, appearances and more.

Methods

Dataset: Ethos Hate Speech Dataset

We use the ETHOS dataset created by Mollas et al. (2020). ETHOS is based on comments found in YouTube and Reddit. The ETHOS YouTube data is collected through Hatebusters (Anagnostou et al., 2018). Hatebusters is a platform that collects comments from YouTube and 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 dataset has two variants: binary and multi-label. In the binary dataset comments are classified as hate or non-hate. In the multi-label variant, the comments are evaluated on measures that include violence, gender, race, ability, religion, and sexual orientation. The examples provided in this paper are from 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 classify the content, for instance, does the comment have to do with violence, gender, race, nationality, disability, etc. We considered the **X** statements that contain race-based hate speech, and we focus on the **X** whose race-based score is at least 0.75. Similarly, we considered the **X** statements that contain gender-based hate speech, and again focused on the **X** whose gender-based score is at least 0.75.

A human annotator evaluates the comments and extracts the following attributes from them — target, targeting type, violence, derogatory, profanity, obscenity, rejection, and accusation. Comments that have unclear target or targeting type are not included. The final dataset contains **X** comments.

Data Collection

The data is collected from prompting GPT-3 (Davinci model) in giving explanations through OpenAI API. We modified the following hyper-parameters for desired outputs — `temperature`, `stop`, `n`, and `max_tokens`. `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 in our study for GPT-3 to obtain some creative liberty to generate open-ended responses. We set the `stop` 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. "n" is the number of responses generated, which is 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 also increase the chance of getting a reasonable responses. "max_tokens" is the maximum length of each response, which is set to 120. The goal of the prompt is to have the model answers "yes" or "no", then generates explanations within 120 words in free form. The prompts are formulated in a question-answering format. We start the prompt by "Q:" and place the question in the following sentence to ask the model to explain why the provided example is sexist or not sexist. Following the question is the provided example wrapped in quotes. We then end the prompt with "A:", so the model will generate the response as an answer to the question. The example of an open-ended prompt can be found in **Appendix X**. This process is repeated until all the examples are given an explanation of why it is sexist or not, and the total number of collected responses in open-ended format is **X**.

For structured responses, the prompt contains two examples of sexists texts and two examples of racist texts. The examples demonstrate the desired patterns we want GPT-3 to follow, which is answering "yes" or "no" to a number of preset attributes under a given text. The attributes 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 example of the structured prompt can be found in **Appendix X**. 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 **X**. Chen et al. (2012) state that offensive languages contain pejoratives, profanities, or obscenities. Therefore, presence of "insult" can be seen as an attributes 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. The example of the structured prompt with reduced attributes can be found in **Appendix X**.

Human Evaluation

The model-generated responses are then evaluated by a human annotator. The open-ended responses with three attributes — target identification, presence of made-up details, and reasoning adequacy. Target identification and presence of made-up details are binary evaluation that simply indicate if the model has identified the target of the sexist or racist content or has made up details in the generated explanations. The reasoning adequacy of each explanations are categorized as follows: 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 the same set of attributes — 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 focus 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 of a violent event is not considered presence of violence.
- Derogatory: **Belittling, diminishing**
- Profanity: If the given text contains any swear words or curse words. For example
- Obscenity: **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's accuses its target of misconducts of mischievous behaviours with generalization without evidence. For example, ""

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 37 out of the 76 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, n = 76x5 = 380

Binary Match	Count
Mistmatch	200
Match	180

Table 2: Open-ended result majority binary match count

Majority Correct	Count
No	39
Yes	37

Reasoning Adequacy

For the 180 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 180 evaluations. In 76 out of 180 times, the model's explanations are constructed with 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 contents are not aimed to provide any explanation.

Target Identification and Made-Up Details

We delved into the 180 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. 117 out of the 180 cases, the model provided an explanation along

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

madeup_details	count
0	138
1	42

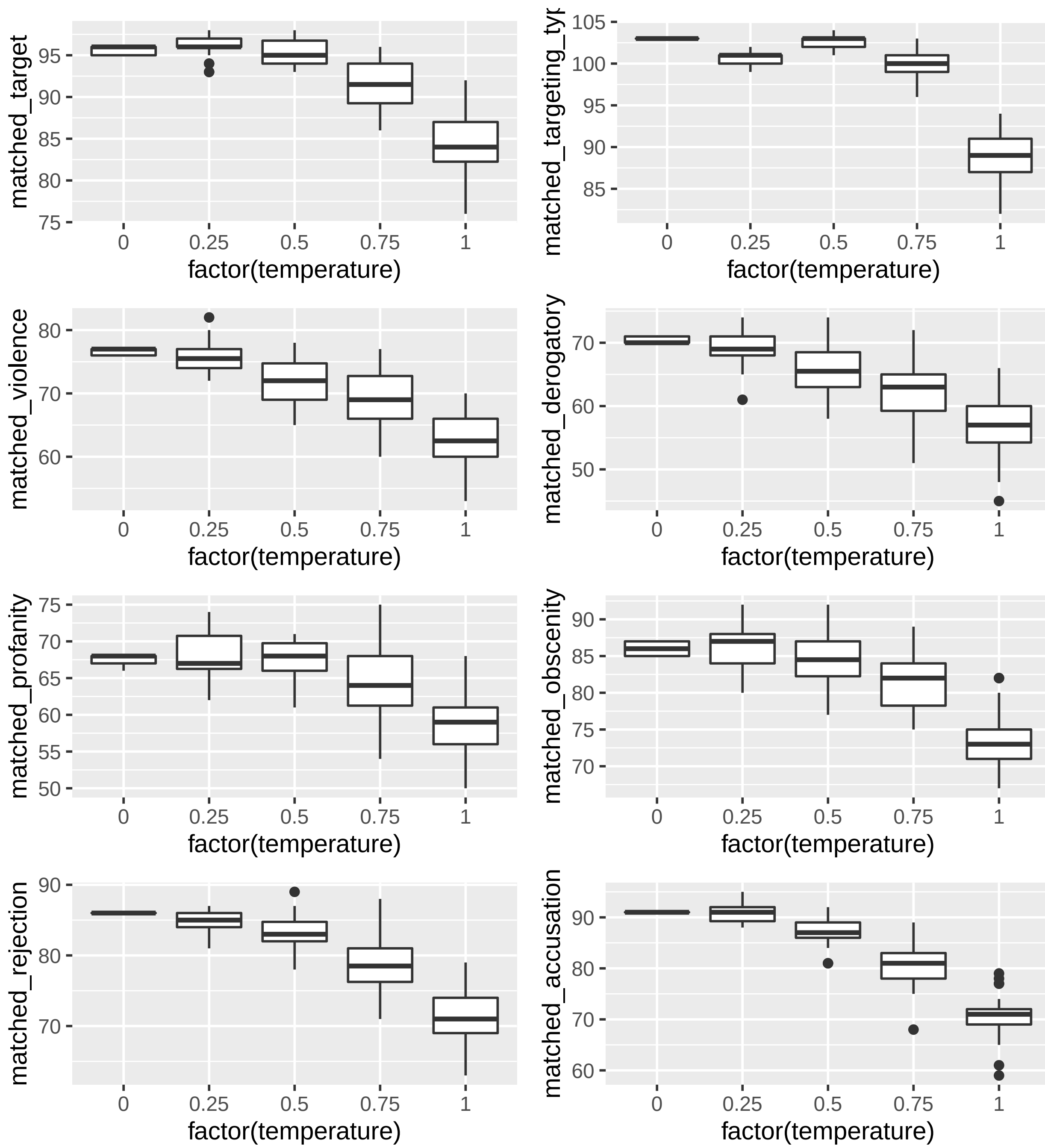
with the identification of the target (women, African Americans… etc). In some cases, the model generated explanations along with made-up details such as assigning false source to the given text. The number of cases where made-up details are found is shown in Table X. Nearly one third of the times, the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

Structured Responses

30 generations per comment.

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. The model performed consistently the poorest under temperature 1. Moreover, as the distribution of matched number varies within wider ranges as the temperature increases.



Comparison of All Attributes

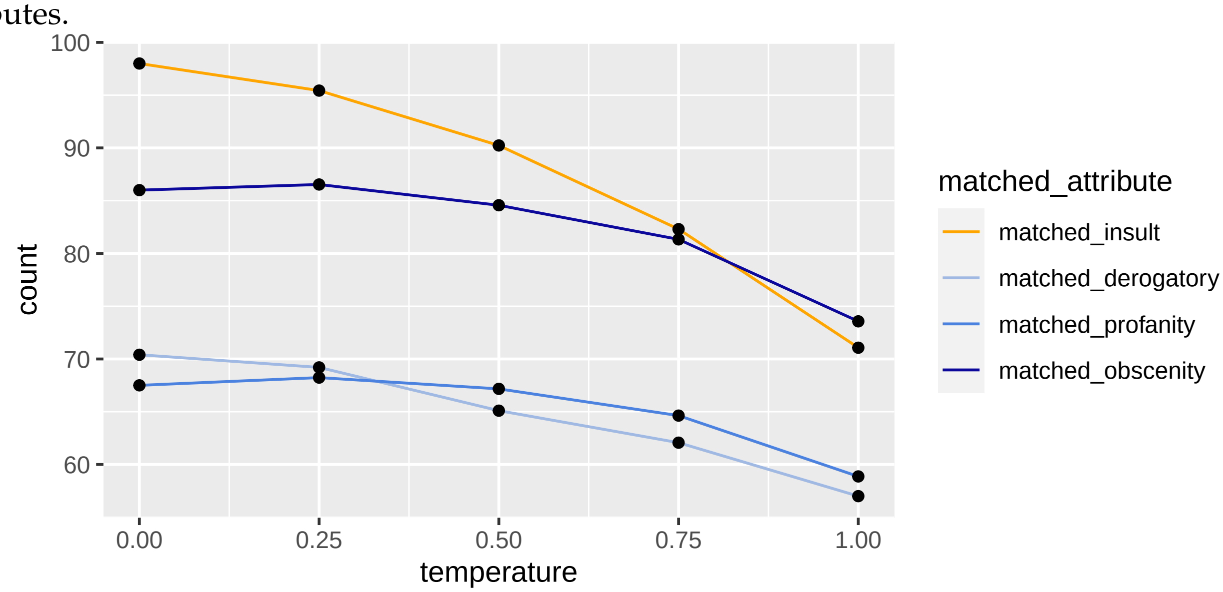
The descriptive statistic is conducted and shown in Table X to assess the model' overall performances in each attribute. The results show that the model performs the best in identifying the identity categories (gender, race…etc) a sexist or racist text is targeting (sex, gender, race, origin…etc). Second 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
Min. :84.50	Min. : 88.87	Min. :62.53	Min. :57.00	Min. :58.87	Min. :73.57
1st Qu.:91.47	1st Qu.: 99.97	1st Qu.:68.87	1st Qu.:62.07	1st Qu.:64.63	1st Qu.:81.33
Median :95.47	Median :100.73	Median :71.63	Median :65.10	Median :67.17	Median :84.57
Mean :92.64	Mean : 99.05	Mean :71.09	Mean :64.75	Mean :65.28	Mean :82.40
3rd Qu.:95.60	3rd Qu.:102.70	3rd Qu.:75.70	3rd Qu.:69.20	3rd Qu.:67.50	3rd Qu.:86.00
Max. :96.17	Max. :103.00	Max. :76.70	Max. :70.40	Max. :68.23	Max. :86.53

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

Ability to generate open-Ended explanations is sub-optimal

Open-ended explanations are often accompanied with made-up details

Structured explanations identify higher-level attributes better compared to granular attributes

Structured explanations can go beyond the scope of sexism 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.

Why this is 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.

```
**The example of contradicting explanation**
```

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 | 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

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.

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.

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

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.

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