

RESEARCH

Detecting Hate Speech with GPT-3

Ke-Li Chiu^{1*†}, Annie Collins² and Rohan Alexander^{1,2}

*Correspondence:

rohan.alexander@utoronto.ca

¹Faculty of Information, University of Toronto, Toronto, Canada

Full list of author information is available at the end of the article

[†]Equal contributor

Abstract

Sophisticated language models such as OpenAI's GPT-3 can generate hateful text that targets marginalized groups. Given this capacity, we are interested in whether large language models can be used to identify hate speech and classify text as sexist or racist. 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 can identify sexist or racist text with an accuracy between 57 per cent and 68 per cent depending on the category of text and type of learning. With few-shot learning, the model's accuracy can be as high as 88 per cent. Large language models have a role to play in hate speech detection, and with further development could eventually be used to counter hate speech.

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

1 Introduction

This paper contains language and themes that are offensive.

Natural language processing (NLP) models use normal words, often written text, as their data. For instance, one might have content from many books and want to group them into themes. Sophisticated NLP models are being increasingly embedded in society. For instance, Google Search uses an NLP model called BERT to better understand what is meant by a word given its context. Some sophisticated NLP models, such as OpenAI's Generative Pre-trained Transformer 3 (GPT-3), can additionally produce text as an output.

The text produced by sophisticated NLP models can be hateful. In particular, there have been many examples of text being generated that targets marginalized groups based on their sex, race, sexual orientation, and other characteristics. For instance, 'Tay' was a chatbot that was released by Microsoft in 2016 on Twitter. Within hours of being released, some of its tweets were sexist. Large language models are trained on enormous datasets from various sources. This means they usually contain untruthful statements, human biases, and abusive language. Further, even when the models do not possess intent, they risk producing text that is offensive or discriminatory, and thus cause unpleasant, or even triggering, interactions [1].

Often the datasets that underpin these models consist of, essentially, the whole public internet. This source raises concerns around three issues: exclusion, over-generalization, and exposure [2]. Exclusion happens due to the demographic bias in the dataset. In the case of language models that are trained on US and UK English scraped from the Internet, datasets may be disproportionately white, male, and young. Therefore, it is not surprising to see white supremacist, misogynistic, and ageist content being over-represented in training datasets [1]. Over-generalization

stems from the assumption that what we see in the dataset represents what occurs in the real world. Words such as ‘always’, ‘never’, ‘everybody’, or ‘nobody’ are frequently used for rhetorical purpose instead of their literal meanings. However, language models do not always recognize this and make inference based on generalized statements using these words. For instance, hate speech commonly uses generalized language for targeting a group such as ‘all’ and ‘every’, and a model trained on these statements may generate similarly generalized and harmful statements. Finally, exposure refers to the relative attention, and hence consideration of importance, given to something. In the context of NLP this may be reflected in the emphasis on English-language created under particular circumstances, rather than another language or circumstances that may be more prevalent.

While these issues, among others, give us pause, the dual-use problem, which is that the same technology can be applied for both good and bad uses, provides motivation. For instance, while stylometric analysis can reveal the identity of political dissenters, it can also solve the unknown authorship of historic text [2]. In this paper we are interested in whether large language models, given that they can produce harmful language, can also identify, or learn to identify, harmful language.

Even though large NLP models do not have a real understanding of language, the vocabularies and the construction patterns of harmful languages can be thought of as known to them. We show that this knowledge can be used to identify abusive language and even hate speech. We consider 120 different extracts that have been categorized as ‘racist’, ‘sexist’, or ‘neither’ in single-category settings (zero-shot, one-shot, and few-shot) and 243 different extracts in mixed-category few-shot settings. We ask GPT-3 to classify these based on zero-, one-, and few-shot learning, with and without instruction. We find that the model performs best with mixed-category few-shot learning. In that setting the model can accurately classify around 86 per cent of the racist extracts and 88 per cent of sexist extracts. If language models can be used to identify abusive language, then not only is there potential for them to counter the production of abusive language by humans, but they could also potentially self-police.

2 Background

2.1 Language models, Transformers and GPT-3

In its simplest form, a language model involves assigning a probability to a certain sequence of words. For instance, the sequence ‘the cat in the hat’ is probably more likely than ‘the cat in the computer’. We typically talk of tokens, or collections of characters, rather than words, and a sequence of tokens constitutes different linguistic units — words, sentences, and even documents [3]. Language models predict the next token based on inputs. If we consider each token in a vocabulary as a dimension, then the dimensionality of language quickly becomes large [4]. Over time a variety of statistical language models have been created to nonetheless enable prediction. The n-gram is one of the earliest language models. It works by considering the co-occurrence of tokens in a sequence. For instance, given the four-word sequence, ‘the cat in the’, it is more likely that the fifth word is ‘hat’ rather than ‘computer’. In the early 2000s, language models based on neural networks were developed, for instance [3]. These were then built on by word embeddings language models in the

2010s in which the distance between tokens represents how related those tokens are, for instance [5]. In 2017, [6] introduced the Transformer, which marked a new era for language models. The Transformer is a network architecture for neural networks that can be trained more quickly than many other approaches [6]. Now most representative pre-trained language models, such as Google's: Bidirectional Encoder Representations from Transformers (BERT) [7], as well as OpenAI's: Generative Pre-trained Transformer (GPT)-2 [8], and GPT-3 [9], are built on this architecture. These models are widely used, for instance BERT is used by Google search.

GPT-3 is the third generation of the Generative Pre-trained Transformer models created by OpenAI, a private company in California that develops artificial intelligence models. GPT-3 is an autoregressive NLP model that can perform a variety of tasks, including responding to questions, summarizing, and parsing text, translation, and classification. Interactions with the model involve inputting some text as a prompt and GPT-3 returning a text completion according to that prompt.

GPT-3 is one of the largest, publicly available, Transformer language models. GPT-3 is distinctive to its predecessors because of few-shot learning. This means that GPT-3 can 'learn' to perform a new task based on only a few examples, expressed in natural language, instead of a fine-tuning process that can require a large amount of data. GPT-3 has led to unexpected NLP applications, such as computational code generation given natural language prompts.

Like other language models, GPT-3 has also generated inappropriate or even hateful content. For instance, [10] demonstrated the use of GPT-3 in mass-producing radicalized text targeting the Islamic population. And [11] show that GPT-3 and similar language models can propagate misconceptions that could deceive human readers. For instance, when asked 'Who really caused 9/11?', they found that GPT-3 provided the false statement 'The US government caused 9/11.'

2.2 Hate speech detection

There is no commonly held definition of hate speech. Different legal jurisdictions have different definitions, as do different companies and other groups. One definition is 'the intentional verbalization of prejudice against a social group' [12]. Detecting hate speech is difficult because the definition of hate speech varies depending on a complex intersection of the topic of the assertion, the context, the timing of the post, synchronized world events, and the identity of speaker and recipient [13]. Moreover, it is difficult to distinguish hate speech from offensive language [14]. Hate speech detection is of interest to academic researchers in a variety of domains including computer science [15] and sociology [14]. It is also of interest to industry, for instance to maintain standards on social networks, and in the judiciary to help identify and prosecute crimes. Since hate speech is prohibited in several countries, misclassification of hate speech can become a legal problem. For instance, in Canada, speech that contains 'public incitement of hatred' or 'wilful promotion of hatred' is specified by the Criminal Code [16]. Policies toward hate speech are more detailed in some social media platforms. For instance, the Twitter Hateful Conduct Policy states:

You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation,

gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

[17]

There has been a large amount of research focused on detecting hate speech. As part of this process, various hate speech datasets have been created and examined. For instance, [18] create a dataset that captures hate speech in the form of racist and sexist language that includes domain expert annotation. They use Twitter data, and annotate 16,914 tweets: 3,383 as sexist, 1,972 as racist, and 11,559 as neither. They had a high degree of annotator agreement. Most of the disagreements were to do with sexism, and often explained by an annotator lacking apparent context. [14] trains a classifier to distinguish between hate speech and offensive language. To define hate speech, they use an online ‘hate speech lexicon containing words and phrases identified by internet users as hate speech’. Even these datasets have bias. For instance, [19] found racial bias in five different sets of Twitter data annotated for hate speech and abusive language. They found that tweets written in African American English are more likely to be labeled as abusive.

3 Methods

We examine the ability of GPT-3 to identify hate speech in zero-shot, one-shot, and few-shot settings. There are a variety of parameters, such as temperature, that control the degree of text variation. Temperature is a hyper-parameter between zero and one. Lower temperatures mean that the model places more weight on higher-probability tokens. To enhance consistency, the temperature is set to zero in our experiments and not varied. There are two categories of hate speech that are of interest in this paper. The first targets the race of the recipient, and the other targets the gender of the recipient. With zero-, one-, and few-shot single-category learning, the model identifies hate speech one category at a time. With few-shot mixed-category learning, the categories are mixed, and the model is asked to classify an input as sexist, racist, or neither. Zero-shot learning means an example is not provided in the prompt. One-shot learning means that one example is provided, and few-shot means that two or more examples are provided.

3.1 Dataset

We use the ETHOS dataset created by [20]. ETHOS is based on comments found in YouTube and Reddit. The ETHOS YouTube data is collected through Hatebusters [21]. Hatebusters is a platform that collects comments from YouTube and assigns a ‘hate’ score to them using a support vector machine. That hate-score is only used to decide whether to consider the comment further or not. The Reddit data is collected from the Public Reddit Data Repository [22]. The classification is done by contributors to a crowd-sourcing platform. They are first asked whether an example contains hate speech, and then, if it does, whether it incites violence and other additional details. The dataset has two variants: binary and multi-label. In the binary dataset comments are classified as hate or non-hate based. In the multi-label variant comments are evaluated on measures that include violence, gender,

race, ability, religion, and sexual orientation. The dataset that we use is from the ETHOS dataset and hence contain typos, misspelling, and offensive content.

We use all of the 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 classify the content. For instance, does the comment have to do with violence, gender, race, nationality, disability, etc. We considered all of the 136 statements that contain race-based hate speech, and we focus on the 76 whose race-based score is at least 0.5, meaning that at least 50 per cent of annotators agreed. Similarly, we considered all of the 174 statements that contain gender-based hate speech, and again focused on the 84 whose gender-based score is at least 0.5. To create a balanced dataset, we select 120 of these statements to result in 60 sexist statements and 60 racist statements. For comparison purposes, we randomly draw 120 statements that are classified as not hate speech. The balanced sample enables us to more easily understand our results.

For the purposes of one-shot and single-category few-shot learning, the dataset was split in half such that 30 racist comments, 30 sexist comments, and 60 neutral comments were used as inputs and the remainder were used as examples. This division was kept consistent for zero-shot learning to allow for the same set of 120 comments to be classified by all three approaches. With the introduction of the 10 static example sets for multi-category few-shot learning, 240 comments were classified for each example set. The comments in the example sets were drawn from the existing pool of 240 comments. For each example set, the comment from each category being used in the example set was dropped from the input dataset and an additional comment from each original dataset (one ‘racist’, one ‘sexist’, and one ‘neither’) was added in as a replacement. This replacement was kept consistent each time, for a total of 243 unique comments classified in the few-shot mixed-category setting.

3.2 Zero-shot learning

With zero-shot learning, the model is not given an example and is expected to generate an answer without additional information. In this setting, GPT-3 is asked to identify one passage of text, in one category, each time. A simple instruction is given asking the model to answer only ‘Yes’ or ‘No’. We ask the model to perform the identification of sexist and racist comments 120 times. The model considers 30 sexist comments, 30 racist comments, and 60 comments that are neither sexist nor racist. An example of zero-shot learning for recognizing a comment that is labeled as sexist is:

Prompt:

Is the following text sexist? Answer yes or no.

‘The thing is women are not equal to us men and their place is the home and kitchen’

Answer:

GPT-3 response:

Yes.

Given the prompt, the model provides this answer based on the dataset that it was trained on by OpenAI. It is not possible for us to train a similar model ourselves due to the financially resources required to create the dataset and to run the model.

3.3 One-shot learning

With one-shot learning, like in the zero-shot learning case, the model is asked to identify one passage of text in one category at a time. In the one-shot learning case, the model is given one example from a category and is asked to identify whether the next passage of text belongs to the selected category. Again, a simple instruction is also given to ask the model to answer only ‘Yes’ or ‘No’, and the same breakdown of 30 sexist comments, 30 racist comments, and 60 comments that are neither sexist nor racist is considered. That is, the model performs this task 120 times with different comments each time. An example of one-shot learning for recognizing a comment that is labelled as sexist is:

Prompt:

The following text in quotes is sexist:

‘Feminism is a very terrible disease’

Is the following text sexist? Answer yes or no.

‘She is heavily relying on him to turn the other cheek... tough talking demon infested woman.’

Answer:

GPT-3 response:

Yes.

3.4 Few-shot learning

Under the few-shot setting, three examples are provided to the model. There are two settings considered in the few-shot experiments: single-category, and mixed-category. In the single-category setting, the model is given two examples that belong to the selected category, and one example that is not hate speech. The answer is expected to be binary. For instance, the model is expected to categorize a passage of text as ‘sexist’ or ‘not sexist’. An example of this is:

Prompt:

‘Too bad women don’t know how to kill themselves’: sexist.

‘You should use your time to arrest murderers not little kids’: not-sexist.

‘Now they know better than this s*** lol they dudes. The stronger sex. The man supremacy’: sexist.

‘The thing is women are not equal to us men and their place is the home and kitchen:’

GPT-3 response:

sexist.

The assignment of sexist or not in the prompt, and our evaluation of whether GPT-3 was right, is based on the ETHOS dataset classification.

On the other hand, in the mixed model, the model is given an example of sexist text, an example of racist text, and an example of text that is not hate speech. In the mixed categories setting, the model is expected to categorize the given passage of text as ‘sexist’, ‘racist’, or ‘neither’. To simplify the classification procedure and allow for comparisons between different example inputs, we created ten sets of examples against which 240 comments were each classified. We elected to use ten example sets to allow for some variation in results while keeping the number of groupings, and the size of the resulting dataset tractable. The example sets each consist of one sexist, one racist, and one ‘neither’ comment and were created using the comments at indexes one through ten from each dataset. Each set of examples is used to assist the model in classifying 60 racist comments, 60 sexist comments, and 120 comments that are in neither category. The inputted comments were the first $n + 1$ comments from their respective datasets with the i^{th} comment dropped corresponding to the use of example set i . This results in 2,400 responses in total, with 243 unique comments classified overall. The ten sets of examples are shown in Appendix ???. An example of this is:

Prompt:

‘That girl still need a good ass whooping, stupid ass b****!!’: sexist.
 ‘U SUCK S*** YOU BLACK F***’: racist.
 ‘You deserve way more subscribers. The s*** you say is f***ing gold.’: neither.
 ‘Every girl on this planet curses her mother for their birth. They know that they are the powerless sex and they pray that they had been born as men! With d***s and stuff.’:

GPT-3 response:

sexist.

Finally, in the mixed model, we added a brief instruction to restrict the generated answers to ‘racist’, ‘sexist’ or ‘neither’. This is to prevent the model from generating answers that are out of scope. An example of this is:

Prompt:

Classify the following texts into ‘racist’, ‘sexist’ or ‘neither’.
 ‘That girl still need a good ass whooping, stupid ass b****!!’: sexist.
 ‘U SUCK S*** YOU BLACK F***’: racist.
 ‘You deserve way more subscribers. The s*** you say is f***ing gold.’: neither.
 ‘Every girl on this planet curses her mother for their birth. They know that they are the powerless sex and they pray that they had been born as men! With d***s and stuff.’:

GPT-3 response:

sexist.

4 Results

We assess GPT-3’s performance in all settings using accuracy, precision, recall, and F1 score. Accuracy is the proportion of correctly classified comments (hate speech and non-hate speech) out of all comments classified. Precision is the proportion

of hate speech comments correctly classified out of all comments classified as hate speech (both correctly and incorrectly). Recall is the proportion of hate speech comments correctly classified out of all hate speech comments in the dataset (both correctly and incorrectly classified). The F1 score is the harmonic mean of precision and recall. In the case of hate speech classification, it is better to have a model with high recall, meaning a model that can identify a relatively high proportion of the hate speech text within a dataset. But the F1 score can provide a more well-rounded metric for model performance and comparison.

4.1 Zero-shot learning

The overall results of the zero-shot experiments are presented in Table 1. Tables ??, ??, and ??, in Appendix ?? contain additional detail. The model has 34 matches (true positives and negatives) and 26 mismatches (false positives and negatives) in the sexist category, and 41 matches and 19 mismatches in the racist category. The model performs more accurately when identifying racist comments, with an accuracy of 68 per cent, compared with identifying sexist comments, with an accuracy of 57 per cent, however the F1 score for each type of hate speech is approximately the same. The overall ratio of matches and mismatches is 75:45. In other words, the accuracy in identifying hate speech in the zero-shot setting is 63 per cent. The model has an F1 score of 81 per cent in this setting, with a notably high recall of 87 per cent as well.

4.2 One-shot learning

The results of the one-shot learning experiments are presented in Table 2. Tables ??, ??, and ??, in Appendix ?? provide additional detail. The model has 37 matches and 23 mismatches in the racist category, and 36 matches and 24 mismatches in the sexist category. In contrast with the result generated from zero-shot learning, the model performs approximately the same in identifying racist comments and identifying sexist comments. The overall ratio of matches and mismatches is 73:47. In other words, the accuracy of identifying hate speech in the one-shot setting is 61 per cent. The general performance in the one-shot setting is worse than in the zero-shot setting, with lower accuracy, recall, precision, and F1 score overall.

4.3 Few-shot learning – single category

The results of the single-category, few-shot learning, experiments are presented in Table 3. Tables ??, ??, ??, in Appendix ?? provide additional detail. The model has 41 matches and 19 mismatches in the racist category, and 42 matches and 18 mismatches in the sexist category. Like the one-shot setting, the model performs almost equally well at identifying racist comments compared with identifying sexist comments. The general performance in the single-category few-shot learning setting is slightly more accurate than performance in other settings, with an accuracy of 69 per cent compared with 61 per cent in the one-shot setting and 63 per cent in the zero-shot setting. However, the model in this context has the lowest F1 score of the settings considered thus far at 61 per cent.

4.4 Few-shot learning – mixed category

The results of the mixed-category few-shot experiments are presented in Table 4. Tables ?? and Table ?? in Appendix ?? provide additional detail. Among the ten sets of examples, Example Set 1 yields the best performance in terms of accuracy (91 per cent) and F1 score (87 per cent) for racist comments and Example Set 10 yields the best performance by the same metrics for sexist comments (accuracy of 89 per cent, F1 score of 85 per cent). The example set that yields the worst results in identifying racist text in terms of F1 score is Example Set 8, which has an F1 score of 70 per cent (and the lowest accuracy at 73 per cent) for this dataset. The example set that yields the worst results in identifying sexist text in terms of F1 score is Example Set 9, which has an F1 score of 69 per cent (and the lowest accuracy at 74 per cent) for this dataset. The differences between Example Sets 1, 8, 9, and 10 suggest that, although the models are provided with same number of examples, the content of the examples also affects how the model makes inferences. Overall, the mixed-category few-shot setting performs approximately the same at identifying sexist text and racist text. It is also has higher accuracy and F1 score than the zero-shot, one-shot, and single-category few-shot settings for both racist and sexist text.

The unique generated answers are listed in Table 8. These are the response of GPT-3 that we obtain when we ask the model to classify statements, but do not provide examples that would serve to limit the responses. Under the mixed-category setting, the model is observed to generate answers that are out of scope. For instance, other than ‘sexist’, ‘racist’, and ‘neither’, we also see answers such as ‘transphobic’, ‘hypocritical’, ‘Islamophobic’, and ‘ableist’. In some cases, the model even classifies a text passage into more than one category, such as ‘sexist, racist’ and ‘sexist, misogynistic’. The full list contains eighty different answers instead of three.

The results presented for each category of text include the classifications of comments that were labelled as ‘neither’ and the category in question. For the purposes of our analysis, a classification was considered a true positive if the answer outputted by GPT-3 contained a category that matched the comment’s label. For example, if a comment was labelled ‘sexist’ and the comment was classified by the model as ‘sexist, racist’, this was considered a true positive in the classification of sexist comments. If a comment was labelled ‘sexist’ and the comment was classified by the model as ‘racist’, ‘transphobic’, ‘neither’, etc. this was considered a false negative.

Since each comment is only labelled with one hate speech category, a classification was considered a true negative if the label of the comment was ‘neither’ and the comment received a classification that did not include the category being considered. For example, if a comment was labelled ‘neither’ and the model answered ‘racist’, this is considered a true negative in the classification of sexist comments (the comment is not sexist, and the model did not classify it as sexist), but a false positive in the classification of racist comments (the comment is not racist, but the model classified it as racist).

4.5 Few-shot learning – mixed category with instruction

To reduce the chance of the model generating answers that are out of scope, a brief instruction is added to the prompt, specifying that the answers be: ‘sexist’, ‘racist’,

or ‘neither’. The addition of an instruction successfully restricts the generated answers within the specified terms. The unique generated answers are: ‘racist’, ‘sexist’, and ‘neither’.

The results of the mixed-category few-shot learning, with instruction, experiments are presented in Tables 9 and 10. And Tables ??, and ??, in Appendix ?? provide additional detail. With the addition of an instruction in the prompt, Example Set 10 remains the best performing example set in terms of accuracy (89 per cent) and F1 score (84 per cent) for sexist text, however Example Set 4 now performs best for racist text in terms of both accuracy (89 per cent) and F1 score (85 per cent). When considering the classification of racist and sexist speech overall, the models perform similarly with and without instruction when classifying racist text, but the model appears to perform slightly better at identifying sexist text when the instruction is omitted.

However, examining label-classification matches across all categories (‘sexist’, ‘racist’, and ‘neither’), mixed-category few-shot learning almost always performs better with instruction than without instruction (Figure 1). Across all example sets, the mean proportion of matching classifications (out of 240 comments) for mixed-category few-shot learning without instruction is 67 per cent. The average proportion of matching classifications rises to 75 per cent for learning with instruction.

5 Discussion

In the zero-shot learning setting where the model is given no examples, its accuracy rate for identifying sexist and racist text is 63 per cent with an F1 score of 81 per cent. In the one-shot learning setting the accuracy decreases to 61 per cent with an F1 score of 65 per cent. Accuracy increases to 69 per cent in the single-category few-shot learning setting, with an F1 score of 61 per cent. Of the three single-category approaches, zero-shot learning has a notably high recall value of 87 per cent.

It is likely that the model is not ideal for use in hate speech detection in the zero-shot learning, one-shot learning, or single-category few-shot learning settings, as the accuracy rates are between 60 per cent and 70 per cent. [14], using a different model and approach, similarly find ‘that almost 40 per cent of hate speech is misclassified’. And when [23] use GPT-2 they find a similar ability to recognize sexually explicit content, however using an alternative model - Google’s T5 [24] - they find better results.

In the mixed-category few-shot setting, different example sets yield different accuracy rates for racist and sexist comments, with noticeable improvement over the single-category approaches. Mixed-category few-shot learning without instruction had noticeably better F1 scores for both racist and sexist comments than either zero-shot or one-shot learning. With instruction added, mixed-category few-shot learning performed similarly well for racist text identification, however the model performed relatively poorly in identifying sexist speech, with an F1 score of 63 per cent overall and a recall of 50 per cent meaning nearly half of the sexist comments were wrongly classified. Overall, it appears as though GPT-3 is most effective at identifying both racist and sexist comments in the mixed-category few-shot learning setting, without instruction.

Examining the proportion of classification matches for each example set (calculated as the number of correct ‘racist’, ‘sexist’, and ‘neither’ classifications out of all comments classified), the overall proportion of matches increases in almost every example set after adding an instruction in the prompt. The differences in accuracy, precision, and recall between the example sets suggests that the content of the examples may matter as well. It may be that deliberately curated examples would enable better accuracy. Adding an instruction to the prompt is useful for restricting the output of the model to more clear-cut classifications. Depending on the purposes of classification, it appears as though adding an instruction can improve model performance by some metrics, however this may vary by the category of text considered and still produce the low recall observed in many of the example sets classified with instruction.

Looking at the misclassified examples helps us to understand what may be contributing to erroneous results. The following comment has a true label of ‘sexist’, but the model classifies it as ‘racist’ in the mixed-category few-shot setting with instruction. This comment uses the word ‘Islam’ which may often be associated with discrimination based on race or ethnicity, but its core argument is about gender-based violence. When instructed to pick a single classification, the model did not have the necessary sensitivity to categorize the comment as sexist instead of racist. The option to classify a comment under multiple hate speech categories is one reason for the model performing better without instruction in the mixed-category setting.

‘Islam! Lets beat their feminist ass’

Mixed-category few-shot classification: racist, sexist (True label: sexist)

Mixed-category few-shot classification, with instruction: racist (True label: sexist)

As another similar example, the following comment has a true label of ‘neither’, but the model classified it as ‘transphobic’ in the mixed-category few-shot setting (without instruction).

‘Transgenders have to protect each other, not fight each other’: transphobic (True label: neither.)

The comment has a seemingly positive connotation towards the transgender community; the inclusion of ‘transgenders’ may be the reason for the false classification. If this were the case, then combined with the previous example, the model may tend to generate false positive results when it encounters words that are identity terms, erroneously ignoring the context in which they are used. Understanding the circumstances under which this occurs is an important area for future research.

The following comment is an example of false negativity. This comment has a true label of ‘racist’ and was classified as ‘neither’ by the model. This is possibly because of the misspelling of the profane word ‘phuck’, where the implied word, which carries aggression, is obvious to a human reader.

‘phuck white rasist ass people’: neither (True label: racist.)

If this were the case, then it also points to a weakness of the use of this type of model for detecting hate speech in human-generated content. In particular, the

way that humans use words changes quickly, especially sexist and racist language online. Large language models take a long time to train and deploy. It may be that, unless considerable context is provided in the prompt, the model will not identify words that have become sexist and racist through their usage in the time since the model was trained.

In order to test the impact of misspellings on hate speech classification, we examined a subset of the ETHOS dataset containing the profane words or sub-strings indicated in Table ?? in Appendix ?. These words were selected due to their prevalence in the dataset and in some cases their specific racist or sexist connotation. The comments were then edited to include misspellings or censorship (including numbers, asterisks, or dashes to remove certain vowels) on a given word or sub-string and run through the zero-shot learning process. Details of the misspellings added are also included in Table ?. Of the 34 sexist comments and 27 racist comments considered, the misspellings and censorship impacted the classification of seven comments, all of which belonged to the racist category (results in Table ? in Appendix ?). Interestingly, three comments with added misspellings were classified as ‘racist’ where they had previously been classified as ‘not racist’. This speaks to potential inconsistencies in the behavior of GPT-3 in understanding profanity and censorship and presents another area for further investigation.

In conclusion, with proper settings such as the inclusion of instruction and curated examples, large natural language models such as GPT-3 can identify sexist and racist text at a similar level of specificity to other methods. However, it is possible that if a user intentionally misspells profane words, the models may be less likely to identify such content as problematic. This possibility deserves further investigation due to the tendency for language to change quickly. Furthermore, the models might classify texts that contain identity terms as problematic, as they are often associated with harmful text when being targeted. Various settings should be further explored to tackle these obstacles. To shed light on why a text is misclassified, it might be appropriate to prompt GPT-3 to deliver an accompanying explanation for the decision, or to consider a larger dataset of sexist and racist content. The extent to which the identification and classification of hate speech can be explained is especially of interest for future work and is something that could be implemented on the easily available GPT-2 from OpenAI or T5 from Google.

Availability of data and materials

Code and data are available at: <https://github.com/kelichiu/GPT3-hate-speech-detection>.

Competing interests

The authors declare that they have no competing interests.

Funding

There was no explicit funding provided for this research, but we gratefully acknowledge the in-kind support of Gillian Hadfield, the Schwartz Reisman Institute for Technology and Society, and OpenAI for providing access to GPT-3 under the academic access program.

Author's contributions

KC had the original idea, obtained and created the datasets, and wrote the first draft of the paper. KC, AC, and RA analysed and interpreted the data, contributed to writing the paper, substantially revised it, and approved the final version.

Acknowledgements


We thank two anonymous reviews and the editor, as well as Amy Farrow, Haoluan Chen, John Giorgi, Mauricio Vargas Sepúlveda, Monica Alexander, Noam Kolt, and Tom Davidson for helpful discussions and suggestions. 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.

Figure 1 Comparing classification with and without an instruction

Author details

¹Faculty of Information, University of Toronto, Toronto, Canada. ²Department of Statistical Sciences, University of Toronto, Toronto, Canada.

References

- Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big?  In: Proceedings of FAccT 2021 (2021)
- Hovy, D., Spruit, S.L.: The social impact of natural language processing. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 591–598 (2016)
- Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. *Journal of Machine Learning Research* 3(Feb), 1137–1155 (2003)
- Rosenfeld, R.: Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE* 88(8), 1270–1278 (2000)
- Turian, J., Ratinov, L., Bengio, Y.: Word representations: A simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 384–394 (2010)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: *Advances in Neural Information Processing Systems*, pp. 5998–6008 (2017)
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018)
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. *OpenAI Blog* 1(8), 9 (2019)
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020)
- McGuffie, K., Newhouse, A.: The radicalization risks of GPT-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807* (2020)
- Lin, S., Hilton, J., Evans, O.: TruthfulQA: Measuring How Models Mimic Human Falsehoods (2021). 2109.07958
- Kennedy, B., Atari, M., Davani, A.M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaladar, S., Portillo-Wightman, G., Gonzalez, E., et al.: The gab hate corpus: A collection of 27k posts annotated for hate speech (2018)
- Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pp. 1–10 (2017)
- Davidson, T., Warmesley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11 (2017)
- Srba, I., Lenzini, G., Pikuliak, M., Pecar, S.: Addressing hate speech with data science: An overview from computer science perspective. *Hate Speech - Multidisziplinäre Analysen und Handlungsoptionen*, 317–336 (2021). doi:10.1007/978-3-658-31793-5_14
- Criminal Code: Government of Canada. As viewed 19 March 2021, available at: <https://laws-lois.justice.gc.ca/eng/acts/c-46/section-319.html> (1985)
- Twitter: Hateful conduct policy. As viewed 19 March 2021, available at: <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy> (2021)
- Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: *Proceedings of the NAACL Student Research Workshop*, pp. 88–93 (2016)
- Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. In: *Proceedings of the Third Workshop on Abusive Language Online*, pp. 25–35 (2019)
- Mollas, I., Chrysopoulou, Z., Karlos, S., Tsoumakas, G.: ETHOS: An Online Hate Speech Detection Dataset (2020). 2006.08328
- Anagnostou, A., Mollas, I., Tsoumakas, G.: Hatebusters: A Web Application for Actively Reporting YouTube Hate Speech. In: *IJCAI*, pp. 5796–5798 (2018)
- Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., Blackburn, J.: The pushshift reddit dataset. In: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 830–839 (2020)
- Schick, T., Udapa, S., Schütze, H.: Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP (2021). 2103.00453
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (2020). 1910.10683

Figures
Tables

Table 1 Performance of zero-shot learning in hate speech classification

Category	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Racism	68	67	73	70
Sexism	57	54	100	70
Overall	62	58	87	81

Table 2 Performance of one-shot learning in hate speech classification

Category	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Racism	62	59	80	68
Sexism	60	57	77	66
Overall	61	58	78	65

Table 3 Performance of single-category few-shot learning in hate speech classification

Category	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Racism	68	66	77	71
Sexism	70	75	60	67
Overall	69	69	68	61

Table 4 Performance of mixed-category few-shot learning in text classification

Example set	Category	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	Racism	91	82	92	87
	Sexism	87	86	72	78
2	Racism	88	79	88	83
	Sexism	87	82	77	79
3	Racism	87	74	93	82
	Sexism	84	79	73	76
4	Racism	84	69	98	81
	Sexism	83	72	80	76
5	Racism	80	64	93	76
	Sexism	86	82	75	78
6	Racism	86	71	98	83
	Sexism	86	77	80	79
7	Racism	77	59	97	73
	Sexism	87	85	73	79
8	Racism	73	55	95	70
	Sexism	83	70	85	77
9	Racism	80	63	97	76
	Sexism	74	57	85	68
10	Racism	88	77	90	83
	Sexism	89	83	87	85
All	Racism	83	68	94	79
	Sexism	84	76	79	77

Table 5 Classifications generated by GPT-3 under mixed-category few-shot learning without instructions

racist — neither — homophobic — sexist — sexist, racist, — sexual assault — religious — sexual harassment — racist, sexist, — transphobic

Table 6 Classifications of all comments using mixed-category few-shot learning, with instruction

Actual classification	GPT-3 classification		
	Neither	Racist	Sexist
Neither	984	163	53
Racist	98	500	2
Sexist	264	35	301

Table 7 Performance of mixed-category few-shot learning in text classification, with instruction

Example set	Category	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	Racism	88	79	87	83
	Sexism	83	81	63	71
2	Racism	87	86	72	78
	Sexism	82	89	53	67
3	Racism	86	76	83	79
	Sexism	81	84	52	64
4	Racism	89	83	87	85
	Sexism	80	83	50	62
5	Racism	86	87	68	77
	Sexism	76	84	35	49
6	Racism	88	93	70	80
	Sexism	73	82	23	36
7	Racism	88	80	87	83
	Sexism	79	85	47	60
8	Racism	79	62	93	75
	Sexism	84	90	60	72
9	Racism	86	75	87	81
	Sexism	77	88	35	50
10	Racism	77	59	100	74
	Sexism	89	85	83	84
All	Racism	85	75	83	79
	Sexism	80	85	50	63

Table 8 Classifications generated by GPT-3 under mixed-category few-shot learning without instructions

racist — neither — homophobic — sexist — sexist, racist, — sexual assault — religious — sexual harassment — racist, sexist, — transphobic

Table 9 Classifications of all comments using mixed-category few-shot learning, with instruction

Actual classification	GPT-3 classification		
	Neither	Racist	Sexist
Neither	984	163	53
Racist	98	500	2
Sexist	264	35	301

Table 10 Performance of mixed-category few-shot learning in text classification, with instruction

Example set	Category	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	Racism	87.78	78.79	86.67	82.54
	Sexism	82.78	80.85	63.33	71.03
2	Racism	86.67	86.00	71.67	78.18
	Sexism	82.22	88.89	53.33	66.66
3	Racism	85.56	75.76	83.33	79.36
	Sexism	80.56	83.78	51.67	63.92
4	Racism	89.44	82.54	86.67	84.55
	Sexism	80.00	83.33	50.00	62.50
5	Racism	86.11	87.23	68.33	76.63
	Sexism	76.11	84.00	35.00	49.41
6	Racism	88.33	93.33	70.00	80.00
	Sexism	72.78	82.35	23.33	36.36
7	Racism	88.33	80.00	86.67	83.20
	Sexism	79.44	84.85	46.67	60.22
8	Racism	78.89	62.22	93.33	74.66
	Sexism	84.44	90.00	60.00	72.00
9	Racism	86.11	75.36	86.67	80.62
	Sexism	76.67	87.50	35.00	50.00
10	Racism	76.67	58.82	100.00	74.07
	Sexism	89.44	84.75	83.33	84.03
All	Racism	85.39	75.41	83.33	79.17
	Sexism	80.44	85.03	50.17	63.11