Detecting Hate Speech with GPT-3 *

Ke-Li Chiu *University of Toronto*
Rohan Alexander *University of Toronto and Schwartz Reisman Institute*

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 56.7 per cent and 68.3 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 88.3 percent. 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.

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

Introduction

This paper contains language and themes that are offensive.

Natural language processing (NLP) models focus on using normal words, often written text, as their data. For instance, one might have content from a large number of 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 natural language processing (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 that untruthful statements, human biases, and abusive language are inevitably included. Further, even when the models do not possess intent, they risk producing synthetic texts that are offensive or discriminatory and thus cause unpleasant, or even triggering, interaction experiences (Bender et al., 2021).

*Code and data are available at: https://github.com/kelichiu/GPT3-hate-speech-detection. We gratefully acknowledge the 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. We thank 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. Comments on the 25 February 2022 version of this paper are welcome at: rohan.alexander@utoronto.ca.

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 (Hovy and Spruit, 2016). 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 (Bender et al., 2021). 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 considerations 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 languages 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 to 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 (Hovy and Spruit, 2016). 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. In particular we consider 120 different extracts that have been categorized as 'racist', 'sexist', or 'neither' in the single-category settings (zero-shot, one-shot, and few-shot) and 243 different extracts in the 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 85.6 percent of the racist extracts and 88.3 percent 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.

Background

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 (Bengio et al., 2003). 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 (Rosenfeld, 2000). 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 Bengio et al. (2003). 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 Turian et al. (2010). In 2017, Vaswani et al. (2017) 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 (Vaswani et al., 2017). Now most representative pre-trained language models, such as Google's model: Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), as well as OpenAI's: Generative Pre-trained Transformer (GPT)-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020), 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.

Until January 11, when Fedus et al. (2021) announced a Transformer model with a trillion parameters, GPT-3 was the largest, publicly-known, Transformer language model. GPT-3 is distinctive from 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, McGuffie and Newhouse (2020) demonstrated the use of GPT-3 in mass-producing radicalized text targeting the Islamic population. And Lin et al. (2021) 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.'

Hate speech detection

There is no commonly held definition of hate speech and 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' (Kennedy et al., 2018). Detecting hate speech is difficult because the definition of hate speech varies depending on the complex intersection of the topic of the assertion, the context, the timing of the post, syn-

chronized world events, and the identity of speaker and recipient (Schmidt and Wiegand, 2017). Moreover, it is difficult to discern hate speech from merely offensive language (Davidson et al., 2017). Hate speech detection is of interest to academic researchers in a variety of domains including computer science (Srba et al., 2021) and sociology (Davidson et al., 2017). 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 (Criminal Code, 1985). 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.
>
> Twitter (2021)

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, Waseem and Hovy (2016) create a dataset that captures hate speech in the form of racist and sexist language that includes domain expert annotation. The 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, and most of the disagreements were to do with sexism, and often explained by an annotator lacking apparent context. And, Davidson et al. (2017) trains a classifier to distinguish between hate speech and offensive language. To define hate speech, they use online 'hate speech lexicon containing words and phrases identified by internet users as hate speech'. It is important to note that even these datasets have bias. For instance, Davidson et al. (2019) 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.

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.

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. 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 (Baumgartner et al., 2020). The classification is done by contributors to a crowd-sourcing platform. They are first asked with 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, 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.

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 multiple 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. Similarly, we considered the all of 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 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.

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.

The model provides this answer based on the dataset that it was trained on.

One-shot learning

With one-shot learning, similarly to 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 break-down 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.

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 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 classified each. We elected to use ten example sets to allow for some variation in results while keeping the number of groupings and size of the resulting data set reasonably small. 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 data sets 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 available in Table 21 in Appendix A. 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.

Result

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 identified out of all comments identified as hate speech (both correctly and incorrectly). Recall is the number of hate speech comments correctly identified out of all hate speech comments in the data set (both correctly and incorrectly identified). F1 score is the harmonic mean of precision and recall. In the case of hate speech identification it is desirable to have a model with high recall, meaning a model that can identify a relatively high proportion of the hate speech text within a data set, however the F1 score can provide a more well-rounded metric for model performance and comparison.

Zero-shot learning

The results of the zero-shot experiments are presented in Tables 1, 2, 3, and 4. 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 with higher accuracy and precision when identifying racist comments compared with identifying sexist comments. 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 62.5 per cent.

Table 1: Classification of racist statements with zero-shot learning

	GPT-3 classification	
Actual classification	Not racist	Racist
Not racist	19	11
Racist	8	22

Table 2: Classification of sexist statements with zero-shot learning

	GPT-3 classification	
Actual classification	Not sexist	Sexist
Not sexist	4	26
Sexist	0	30

Table 3: Classification of hate speech with zero-shot learning

	GPT-3 classification	
Actual classification	Not hate speech	Hate speech
Not hate speech	23	37
Hate speech	8	52

Table 4: Performance of zero-shot learning in hate speech classification (percent)

Category	Accuracy	Precision	Recall	F1
Racism	68.33	66.67	73.33	69.84
Sexism	56.67	53.57	100.00	69.77
Overall	62.50	58.43	86.67	80.54

Table 5: Classification of racist statements with one-shot learning

	GPT-3 classification	
Actual classification	Not racist	Racist
Not racist	13	17
Racist	6	24

Table 6: Classification of sexist statements with one-shot learning

	GPT-3 classification	
Actual classification	Not sexist	Sexist
Not sexist	13	17
Sexist	7	23

One-shot learning

The results of the one-shot learning experiments are presented in Table 5, Table 6, Table 7, and Table 8. 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 approximately the same in identifying racist comments compared with identifying sexist comments. The general performance in the one-shot setting is worse than in the zero-shot setting, with lower accuracy, recall, and precision overall. 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 60.8 per cent.

Few-shot learning – single category

The results of the single-category, few-shot learning, experiments are presented in Table 9, Table 10, Table 11, and Table 12. The model has 41 matches and 19 mismatches in the racist category, and 42 matches and 18 mismatches in the sexist category. Similar to the one-shot setting, the model performs almost equally well at identifying racist comments compared with identifying sexist comments. The general performance in the few-shot

Table 7: Classification of hate speech with one-shot learning

	GPT-3 classification	
Actual classification	Not hate speech	Hate speech
Not hate speech	26	34
Hate speech	13	47

Table 8: Performance of one-shot learning in hate speech classification

Category	Accuracy	Precision	Recall	F1
Racism	61.67	58.54	80.00	67.61
Sexism	60.00	57.50	76.67	65.72
Overall	60.83	58.02	78.33	65.25

Table 9: Classification of racist statements with single-category few-shot learning

	GPT-3 classification	
Actual classification	Not racist	Racist
Not racist	18	12
Racist	7	23

learning setting is similar to the performance in the one-shot learning setting as well. The overall ratio of matches and mismatches is 83:37, or around 69.2 per cent, with higher precision and lower recall compared with the one-shot learning setting.

Few-shot learning – mixed category

The results of the mixed-category few-shot experiments are presented in Table 13, Table 14, and Table 15.

Among the ten sets of examples, Example Set 10 yields the best performance in terms of F1 score for both racist and sexist comments. It also has the highest accuracy for identifying sexist comments at 91.25 percent and is tied with Example Set 2 for the highest accuracy in identifying racist comments at 90.42 percent. 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 68.26 percent (and the lowest accuracy at 77.92 percent) 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 68.45 percent (and the lowest accuracy at 80.42 percent) for this dataset. The differences between Example Sets 8, 9, and 10 suggest that, although the models are provided with same number of examples, the content of the

Table 10: Classification of sexist statements with single-category few-shot learning

	GPT-3 classification	
Actual classification	Not sexist	Sexist
Not sexist	24	6
Sexist	12	18

Table 11: Classification of hate speech with single-category few-shot learning

	GPT-3 classification	
Actual classification	Not hate speech	Hate speech
Not hate speech	42	18
Hate speech	19	41

Table 12: Performance of single-category few-shot learning in hate speech classification

Category	Accuracy	Precision	Recall	F1
Racism	68.33	65.71	76.67	70.77
Sexism	70.00	75.00	60.00	66.67
Overall	69.17	69.49	68.33	60.51

examples also affects how the model makes inferences. Overall, the mixed-category few-shot setting performs slightly better at identifying sexist text versus racist text. It is also performs with higher accuracy and a higher 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 16. 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.

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 (the comment is sexist, and the model classified it as sexist) and a false positive in the classification of racist comments (the comment is not racist, and the model classified it as racist). 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 either 'neither' or the category not being analyzed, and the comment received a classification that did not include the category being considered. For example, if a comment was labelled 'racist' and the model answered 'neither', 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 negative in the classification of racist comments (the comment is racist, but the model did

Table 13: Classification of racist statements with mixed-category few-shot learning

		GPT-3 classification	
Example set	Actual classification	Not racist	Racist
1	Not racist	160	20
	Racist	5	55
2	Not racist	164	16
	Racist	7	53
3	Not racist	155	25
	Racist	4	56
4	Not racist	149	31
	Racist	1	59
5	Not racist	142	38
	Racist	4	56
6	Not racist	152	28
	Racist	1	59
7	Not racist	132	48
	Racist	2	58
8	Not racist	130	50
	Racist	3	57
9	Not racist	142	38
	Racist	2	58
10	Not racist	163	17
	Racist	6	54
All	Not racist	1489	311
	Racist	35	565

Table 14: Classification of sexist statements with mixed-category few-shot learning

		GPT-3 classification	
Example set	Actual classification	Not sexist	Sexist
1	Not sexist	172	8
	Sexist	17	43
2	Not sexist	170	10
	Sexist	14	46
3	Not sexist	168	12
	Sexist	16	44
4	Not sexist	161	19
	Sexist	12	48
5	Not sexist	170	10
	Sexist	15	45
6	Not sexist	166	14
	Sexist	12	48
7	Not sexist	172	8
	Sexist	16	44
8	Not sexist	158	22
	Sexist	9	51
9	Not sexist	142	38
	Sexist	9	51
10	Not sexist	167	13
	Sexist	8	52
All	Not sexist	1646	154
	Sexist	128	472

Table 15: Performance of mixed-category few-shot learning in text classification

Example set	Category	Accuracy	Precision	Recall	F1
1	Racism	89.58	73.33	91.67	81.48
	Sexism	89.58	84.31	71.67	77.48
2	Racism	90.42	76.81	88.33	82.17
	Sexism	90.00	82.14	76.67	79.31
3	Racism	87.92	69.14	93.33	79.43
	Sexism	88.33	78.57	73.33	75.86
4	Racism	86.67	65.56	98.33	78.67
	Sexism	87.08	71.64	80.00	75.59
5	Racism	82.50	59.57	93.33	72.72
	Sexism	89.58	81.82	75.00	78.26
6	Racism	87.92	67.82	98.33	80.27
	Sexism	89.17	77.42	80.00	78.69
7	Racism	79.17	54.72	96.67	69.88
	Sexism	90.00	84.62	73.33	78.57
8	Racism	77.92	53.27	95.00	68.26
	Sexism	87.08	69.86	85.00	76.69
9	Racism	83.33	60.42	96.67	74.36
	Sexism	80.42	57.30	85.00	68.45
10	Racism	90.42	76.06	90.00	82.44
	Sexism	91.25	80.00	86.67	83.20
All	Racism	85.58	64.50	94.17	76.56
	Sexism	88.25	75.40	78.67	77.00

Table 16: 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 | hypocritical | I'm not a | I'm not brave | none of your business | I didn't | I'm not offended | you're not | sexual | I don't play | I don't care | I don't know | victim blaming | it's a question | irrelevant | I'm not crazy | creepy | romantic | I was taught to | I didn't say | ignorant | I'm not sure | I agree | not true | not even wrong | YouTube doesn't remove | socialist | conspiracy theorist | overpopulation | ableist | Islamophobic | conspiracy theory | hippie | sexist, rape ap | cliche | not even close to | nostalgic | I don't think | not a question | hippy | terrorist | hate speech | he was a terrorist | preachy | not a valid argument | not sexist | brave | Islamophobe | because you're a | troll | mental | never | pedophilia | mental illness is not | I'm sure you | I'm not going | not even close | grammar | rape apologist | cliché | vegan | sexist, misogynistic | I do | wrong | funny | question | when you were taught | opinion | emotional | conspiracy | nostalgia

classify it as racist).

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 Table 17, Table 18, Table 19, and Table 20. With the addition of an instruction in the prompt, Example Set 10 remains the best performing example set in terms of accuracy (92.08 percent) and F1 score (84.03 percent) for sexist text, however Example Set 4 now performs best for racist text in terms of both accuracy (91.25 percent) and F1 score (83.20 percent). When considering the classification of racist and sexist speech independently, the models perform similarly with and without instruction in the mixed-category few-shot setting.

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 66.9 percent. The average proportion of matching classifications rises to 74.5 percent for learning with instruction.

Table 17: Classifications of all comments using mixed-category few-short learning, with instruction

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

Table 18: Classification of racist statements with mixed-category few-shot learning, with instruction

		GPT-3 classification	
Example set	Actual classification	Not racist	Racist
1	Not racist	163	17
	Racist	8	52
2	Not racist	171	9
	Racist	17	43
3	Not racist	159	21
	Racist	10	50
4	Not racist	167	13
	Racist	8	52
5	Not racist	173	7
	Racist	19	41
6	Not racist	176	4
	Racist	18	42
7	Not racist	163	17
	Racist	8	52
8	Not racist	139	41
	Racist	4	56
9	Not racist	159	21
	Racist	8	52
10	Not racist	132	48
	Racist	0	60
All	Not racist	1602	198
	Racist	100	500

Table 19: Classification of sexist statements with mixed-category few-shot learning, with instruction

		GPT-3 classification	
Example set	Actual classification	Not sexist	Sexist
1	Not sexist	170	10
	Sexist	22	38
2	Not sexist	175	5
	Sexist	28	32
3	Not sexist	174	6
	Sexist	29	31
4	Not sexist	174	6
	Sexist	30	30
5	Not sexist	176	4
	Sexist	39	21
6	Not sexist	177	3
	Sexist	46	14
7	Not sexist	175	5
	Sexist	32	28
8	Not sexist	176	4
	Sexist	24	36
9	Not sexist	177	3
	Sexist	39	21
10	Not sexist	171	9
	Sexist	10	50
All	Not sexist	1745	55
	Sexist	299	301

Table 20: Performance of mixed-category few-shot learning in text classification, with instruction

Example set	Category	Accuracy	Precision	Recall	F1
1	Racism	89.58	75.36	86.67	80.62
	Sexism	86.67	79.17	63.33	70.37
2	Racism	89.17	82.69	71.67	76.79
	Sexism	86.25	86.49	53.33	65.98
3	Racism	87.08	70.42	83.33	76.33
	Sexism	85.42	83.78	51.67	63.92
4	Racism	91.25	80.00	86.67	83.20
	Sexism	85.00	83.33	50.00	62.50
5	Racism	89.17	85.42	68.33	75.93
	Sexism	82.08	84.00	35.00	49.41
6	Racism	90.83	91.30	70.00	79.24
	Sexism	79.58	82.35	23.33	36.36
7	Racism	89.58	75.36	86.67	80.62
	Sexism	84.58	84.85	46.67	60.22
8	Racism	81.25	57.73	93.33	71.34
	Sexism	88.33	90.00	60.00	72.00
9	Racism	87.92	71.23	86.67	78.20
	Sexism	82.50	87.50	35.00	50.00
10	Racism	80.00	55.56	100.00	71.43
	Sexism	92.08	84.75	83.33	84.03
All	Racism	87.58	71.63	83.33	77.04
	Sexism	85.25	84.55	50.17	62.97

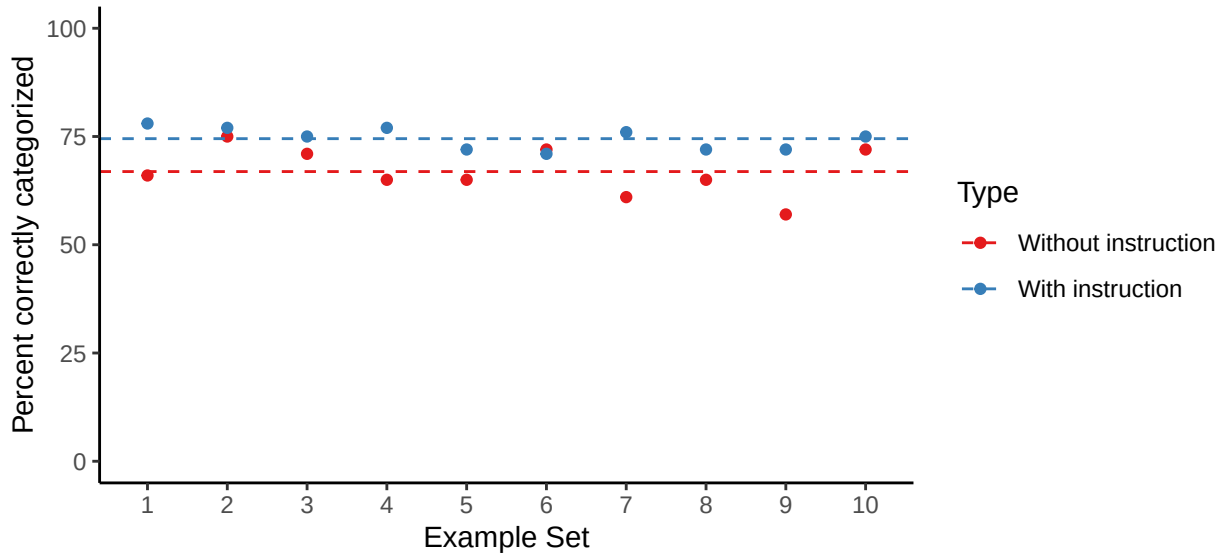

Figure 1: Comparing classification with and without an instruction

Discussion

In the zero-shot learning setting where the model is given no examples, its accuracy rate for identifying sexist and racist text is 62.5 per cent with an F1 score of 80.54 percent. In the one-shot learning setting the accuracy decreases to 60.83 percent with an F1 score of 65.25 percent. Accuracy increases to 69.2 percent in the single-category few-shot learning setting, with an F1 score of 60.5 percent. Of the three single-category approaches, zero-shot learning has a notably high recall value of 86.67 percent.

It is likely that the model is not suitable for use in hate speech detection in either 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. Davidson et al. (2017), using a different model and approach, similarly find 'that almost 40 per cent of hate speech is misclassified'. And when Schick et al. (2021) use GPT-2 they find a similar ability to recognise sexually explicit content, however using an alternative model - Google's T5 (Raffel et al., 2020) - 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. Our definition of true negative in the mixed-category few-shot setting (see Section **??**) resulted in a large number of true negatives in our analysis, thus recall and F1 score may be better metrics for measuring model performance in this context. 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 62.97 percent overall and a recall of 50.17 percent. 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 number of classification matches for each example set, 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 'racist', but the model classifies it as 'sexist'. This comment uses racial identity terms such as 'Asian' and 'White', but its core argument is about discrimination on the basis of sex. The model did not have the necessary sensitivity to categorize the comment as sexist instead of racist.

> 'The male gene is stronger, as it should be of course. Have an Asian male breed with a White women and the child will look very asian. Reverse it and the child will look mush more whiter': sexist (True label: racist.)

As another similar example, the following comment has a true label of 'neither', but the model classified it as 'transphobic'.

> 'Civil rights movement for transgendered people.': transphobic (True label: neither.)

The comment is a plain statement without judgment; the inclusion of 'transgendered people' may be the reason for the false classification (specifically a false positive in the context of detecting hate speech in general). If this were the case, then combined with the previous example, the model may have a tendency 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 data set containing the profane words or sub-strings indicated in Table 22 in Appendix B. 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 22. 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 23 in Appendix B). 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 behaviour 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 are capable of identifying 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 publicly available GPT-2 from OpenAI or T5 from Google.

Appendix A

Table 21: The ten example sets for the mixed-category, few-shot learning, experiments

Set	Example 1 (sexist)	Example 2 (racist)	Example 3 (not hate speech)
1	You should know women's sports are a joke	why does that african american monkey hold gun like that?	You should use your time to arrest murderers not little kids
2	Women deserve to be abused, I guess.	Why do negroes tear down their own neighborhoods when a black thug gets killed by law enforcement? Must be because of low IQ.	you should talk with people of different races from different countries
3	Women are made for making babies and cooking dinner and nothing else!!!	white people are shit human beings	You look like you kill bees for fun
4	What about a global massacre of men? Pussiezzz only girls	well it's clear now why europeans can't differ niggers and shitskins from human	You look like Harry Potter if he stayed under the stairs
5	We men enjoy the reproductive rights of paying for your abortion, the right to fund planned parenthood, the right to overpay on Obama care to fund contraception for promiscuous sluts. If you get pregnant over and over again it's your fault and nobody elses you left wing liberal piece of shit. Go be a whore. Keep spreading disease. Half of all black women have herpes. And we have the right to pay for it. You are outrageous.	We whites have higher IQ's than blacks, this is because IQ measures intelligence and we are the more evolved and superior race with a larger brain.	You look kinda dead inside, you parents must not haved loved you.
6	The men are turtles...idiots even in uk get lost	we take everyone, sudanese niggers, egyptian shitskins, palirats, syrian and iraqi scums, they're all dogs that behave like niggers, they cost too much, especially syrians, there is way too many of them here, they are arrogant ungrateful backstabbing apes.	You have the grace and composure of a rabid hyena
7	That girl still need a good ass whooping, stupid ass bitch!!	U SUCK SHIT YOU BLACK FUCK	You deserve way more subscribers . The shit you say is fucking gold.
8	stone throwing must be legal when against women	This sick, disgusting, white terrorist/fascist killed more than 50 innocent people.	You can see the country is fertile and green.ă just need the right governance
9	She is heavily relying on him to turn the other cheek...tough talking demon infested woman.	This is considered WHITE in Sweden. No wonder they accepted all those black and muslim immigrants. They are literally color blind.	you are a great brother to abbie, that family is changing the world!
10	Really, I hope she get raped in the future. Raping women should be legal I think	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 adressňüňÜňóØňňèňüňíňØňüňôňÇ	You are a brave man.....for letting them keep the comments enabled

Table 22: Profane strings used to test extract comments to test impacts of misspelling and censorship on hate speech classificiation

Original	Edits
nigg	n1gg, nig, n*gg, n-gg
fuck	fck, phuck, f*ck, f-ck
whore	wh0re, whor3, wh0r3, wh*re, wh-re
kill	k1ll, kil, k-ll, k*ll
rape	r@pe, rap3, r@p3, rap, r*pe, r-pe
die	d1e, di3, d13, dye, d*e, d-e
bitch	b1tch, bich, bithc, bicht, b*tch, b-tch

Appendix B

Table 23: Zero-shot learning classifications impacted by misspelling and censoring racist comments

Label	Status	Comment	GPT-3 classification
Racist	Unedited	Fuck these generate vodka niggers. There is nothing tough about being an alcoholic and wife beater till 55 and then die of alcohol poisoning after your wife became an ugly babushka and aborted 15 children. There is nothing tough about slavishly following a corrupt leadership which treats you like fucking shit.	Yes
Racist	Edited	fck these generate vodka niggers. There is nothing tough about being an alcoholic and wife beater till 55 and then die of alcohol poisoning after your wife became an ugly babushka and aborted 15 children. There is nothing tough about slavishly following a corrupt leadership which treats you like fcking shit.	No
Racist	Edited	phuck these generate vodka niggers. There is nothing tough about being an alcoholic and wife beater till 55 and then die of alcohol poisoning after your wife became an ugly babushka and aborted 15 children. There is nothing tough about slavishly following a corrupt leadership which treats you like phucking shit.	No
Racist	Unedited	Fuck off niggers WV is full	Yes
Racist	Edited	phuck off niggers WV is full	No
Racist	Unedited	Why are niggers superior athletically to all other races?	Yes
Racist	Edited	Why are n*ggers superior athletically to all other races?	No
Racist	Unedited	That fucking black cunt gets fucking cross hairs on the fucking head! Look I owe not shit to anyone who dosent owe me!!! LOOK WHORE!! FUCK YOU TO DEATH!!! 11 PERCENTILE AND FALLING!! WHORE BLACK RACIST WHORE! TAKE THE RED PILL!	No
Racist	Edited	That phucking black cunt gets phucking cross hairs on the phucking head! Look I owe not shit to anyone who dosent owe me!!! LOOK WHORE!! phuck YOU TO DEATH!!! 11 PERCENTILE AND FALLING!! WHORE BLACK RACIST WHORE! TAKE THE RED PILL!	Yes
Racist	Unedited	Stand the fuck up and take them out all you fucking lazy white males! fuck being called a racist protect your woman and children. If you find it in the street kill it and bury it right there these animals will wipe you out if your not careful	No
Racist	Edited	Stand the phuck up and take them out all you phucking lazy white males! phuck being called a racist protect your woman and children. If you find it in the street kill it and bury it right there these animals will wipe you out if your not careful	Yes
Racist	Unedited	Ima fucking stab u up bruv. Fucking white prick bruv.. fucking racist. Gammon	Yes
Racist	Edited	Ima fcking stab u up bruv. fcking white prick bruv.. fcking racist. Gammon	No
Racist	Edited	Ima f*cking stab u up bruv. f*cking white prick bruv.. f*cking racist. Gammon	No
Racist	Unedited	fucking monkeys kill them all	No
Racist	Edited	fucking monkeys k-ll them all	Yes

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*.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

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*.

Criminal Code (1985). Government of Canada. As viewed 19 March 2021, available at: https://laws-lois.justice.gc.ca/eng/acts/c-46/section-319.html.

Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*.

Hovy, D. and Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., et al. (2018). The gab hate corpus: A collection of 27k posts annotated for hate speech.

Lin, S., Hilton, J., and Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods.

McGuffie, K. and Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer.

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, 88(8):1270–1278.

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

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Srba, I., Lenzini, G., Pikuliak, M., and Pecar, S. (2021). Addressing hate speech with data science: An overview from computer science perspective. *Hate Speech - Multidisziplinäre Analysen und Handlungsoptionen*, page 317336.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394.

Twitter (2021). Hateful conduct policy. As viewed 19 March 2021, available at: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.