

NLP Project: Hate Speech And Target Detection Using Prompting & ChatGPT

Shahar Shcheranski

Ben-Gurion University of the Negev
shcheran@post.bgu.ac.il

Sarit Hollander

Ben-Gurion University of the Negev
sarithol@post.bgu.ac.il

Aviv Rovshitz

Ben-Gurion University of the Negev
avivrov@post.bgu.ac.il

Leor Ariel Rose

Ben-Gurion University of the Negev
leorro@post.bgu.ac.il

1 Introduction

Generally speaking, hate speech refers to any form of communication expressing prejudice or discrimination against another group in some way. It is often used as a tool to spread fear, hatred, and intolerance among different social groups. Hate speech is not limited to any particular language, race, or religion, and can take the form of verbal, written, or symbolic expressions. It can also be spread through the use of social media, websites, and other digital channels. The most commonly targeted groups for hate speech include women, people of color, the LGBT community, Muslims, immigrants, and Jews.

There is a growing prevalence of hate speech in society today, and it can have wide-ranging and damaging effects on members of the community as well as individuals. As part of our efforts to combat hate speech, we must identify hateful language and those who are targeted by it. In this way, we can take proactive measures to combat it. For us to be able to create an environment that is safe and inclusive for all, we must be cognizant of the language that is used as well as the individuals or groups who are being targeted.

ChatGPT¹ as answered by it is a variant of OpenAI's GPT-3 model, fine-tuned for the task of conversational response generation. It has been trained to respond to user input in a conversational manner, providing answers to questions, generating text based on prompts, and engaging in casual conversation. It is designed to be used as a chatbot in various applications such as customer service, online counselling, and entertainment.

The purpose of this project is to use ChatGPT and prompting in order to analyse hate speech and target detection in a quantitative and qualitative way.

2 Related Work

In this project, we will build on previous work on automating hate speech detection using transformer-based architectures (Mozafari et al., 2020; Aluru et al., 2020; Kennedy et al., 2020; Chiu et al., 2021; Arviv et al., 2021) to contribute to a long line of research. The most similar work was conducted by (Chiu et al., 2021) which used GPT-3 and prompting for racism and sexism detection without any target detection. There is no doubt that all of these large-scale language model works are based on earlier works that used simpler machine and deep learning techniques, such as (Davidson et al., 2017; Waseem and Hovy, 2016; Nobata et al., 2016; Magu et al., 2017) which used SVM and logistic regression or (Ziqi et al., 2019; Gambäck and Sikdar, 2017; Del Vigna12 et al., 2017; Park and Fung, 2017) which used RNNs and CNNs.

Many of these works attempt to identify hate speech in textual information rather than the hate target. On the other hand (Markov and Daelemans, 2022) focused on detecting the target of hate speech in Dutch social media by BERTje and (Pérez et al., 2022) focused on identifying the target of hate speech on Twitter by BETO.

There are probably a lot more related works that we missed. However, from our search, we did not come across any that used ChatGPT for hate speech and target detection.

3 Data

We use a subset of "Parler Hate Speech" annotated data from Israeli and Tsur (2022) containing 55 not-hate posts and 49 hate posts (Fig. 1). Posts were sampled from the full annotated data and were annotated to include hate targets. We randomly selected posts that were not hateful, and hand-picked hateful posts from randomly selected posts in order to ensure that there existed a hate target as defined

¹<https://chat.openai.com/>

by [Stevenson \(2010\)](#) oxford dictionary of english.

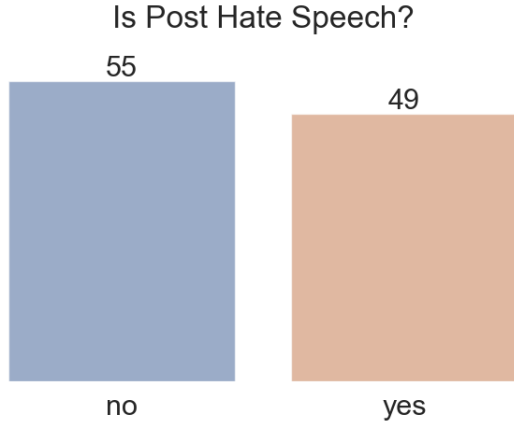


Figure 1: Hate and not-hate posts in our subset data.

Annotations of hate targets were done by each project member and were included only if everyone agreed. The posts were annotated into five categories: "jews", "muslims", "blacks", "females", "lgbt" and were also noted by implicit or explicit hate. Although many of the posts are explicit and against Jews, Muslims, and Blacks, some do contain implicit hatred (Fig. 2). We make this annotated subset available in the project's repository ².

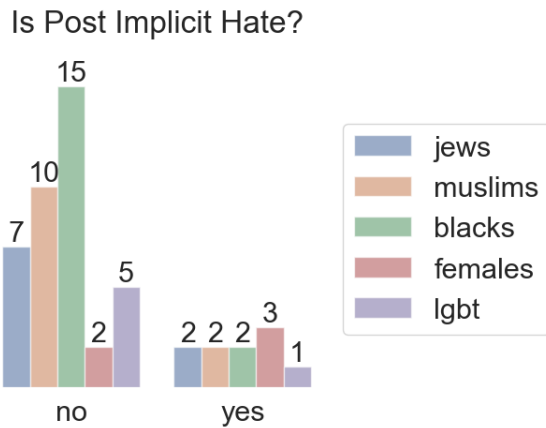


Figure 2: Hate target by implicit and explicit hate posts in the subset data.

4 Computational Setting

The purpose of this project is to determine whether ChatGPT can identify hate speech and hate targets. There are no controllable parameters in ChatGPT, but prompting can be used to control it. Our

²<https://github.com/leorroze/ChatGPT-Hate-Speech>

prompting techniques include zero-shot, one-shot, and few-shot.

4.1 Hate Speech Detection

As part of hate speech detection using prompting, ChatGPT is asked to identify a single post, every time. ChatGPT is given a context sentence describing what is hate speech as defined by [Stevenson \(2010\)](#) oxford dictionary of english. Additionally, number of example are provided depending on the prompting type and a simple instruction is given to identify hate speech only with the answers "Yes" or "No".

4.2 Hate Speech Target Detection

As part of hate speech target detection using prompting, ChatGPT is asked to identify a single post, every time. ChatGPT is given a context sentence describing what is hate speech as defined by [Stevenson \(2010\)](#) oxford dictionary of english. Additionally, number of example are provided depending on the prompting type and a simple instruction is given to identify hate speech target with an open answer.

4.3 Zero-Shot

In zero-shot learning, no examples are provided and the model is expected to generate an answer without any additional information. Example of zero-shot prompt can be seen in Appendix A.

4.4 One-Shot

In one-shot learning, a single example from the same class is provided and the model is expected to generate an answer. Example of one-shot prompt can be seen in Appendix A.

4.5 Few-Shot

In few-shot learning, N examples are provided and the model is expected to generate an answer. we provided ChatGPT $0.5N$ examples from the same class as well as $0.5N$ examples from different classes. We choose $N=4$ to insure enough and balanced number of examples. Example of few-shot prompt can be seen in Appendix A.

5 Results and analysis

5.1 Quantitative Results

A quantitative evaluation was conducted using all the results from each prompt type generated by ChatGPT.

5.1.1 Hate Speech Detection

There are 98 matches and 6 mismatches in the zero-shot model (Fig. 3a). There are 89 matches and 15 mismatches in the one-shot model (Fig. 3b). There are 99 matches and 5 mismatches in the few-shot model (Fig. 3c). All models identify hate posts better than not-hate posts. In quantitative terms (Table 1), the zero-shot model achieves 0.891/1/0.942 precision/recall/F1-score respectively, the one-shot model achieves 0.774/0.98/0.865 precision/recall/F1-score respectively, and the few-shot model achieves 0.923/0.98/0.95 precision/recall/F1-score respectively. There is no surprise that few-shot performs better when it comes to hate speech detection. We were surprised to discover that one-shot performs poorly than zero-shot.

Prompt	Precision	Recall	F1 Score
Zero	0.891	1.0	0.942
One	0.774	0.98	0.865
Few	0.923	0.98	0.95

Table 1: Hate speech detection quantitative comparison using precision, recall and F1-score.

5.1.2 Hate Speech Target Detection

There are 40 matches and 9 mismatches in the zero-shot model (Fig. 4a). There are 44 matches and 5 mismatches in the one-shot model (Fig. 4b). There are 46 matches and 3 mismatches in the Few shot model (Fig. 4c). In quantitative terms (Table 2), the zero-shot model achieves 0.816/0.816/0.816 precision/recall/F1-score respectively, the one-shot model achieves 0.898/0.898/0.898 precision/recall/F1-score respectively, and the few-shot model achieves 0.939/0.939/0.939 precision/recall/F1-score respectively. There is no surprise that few-shot performs better when it comes to hate speech target detection. We were surprised to discover that one-shot performs better than zero-shot after we saw poor performance on hate speech detection.

Prompt	Precision	Recall	F1 Score
Zero	0.816	0.816	0.816
One	0.898	0.898	0.898
Few	0.939	0.939	0.939

Table 2: Hate speech target detection quantitative comparison using precision, recall and F1-score.

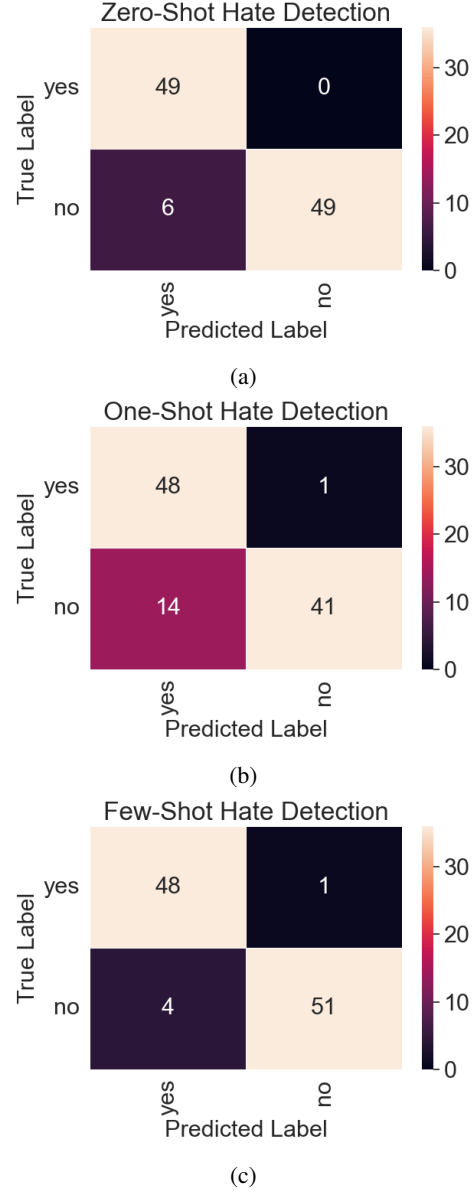


Figure 3: Hate speech detection quantitative comparison using confusion matrices.

5.2 Qualitative Results

We analyzed some errors qualitatively to gain a better understanding of the results from each prompt type generated by ChatGPT.

5.2.1 Mistakes Can Be Fixed Using A Different Example

In order to understand why one-shot performed poorly than zero-shot in hate speech detection, we applied the same one-shot process with different examples (Fig. 5). By changing the example, we observed a different ChatGPT response. This suggests that the response is dependent on the example, which explains why one-shot performed poorly

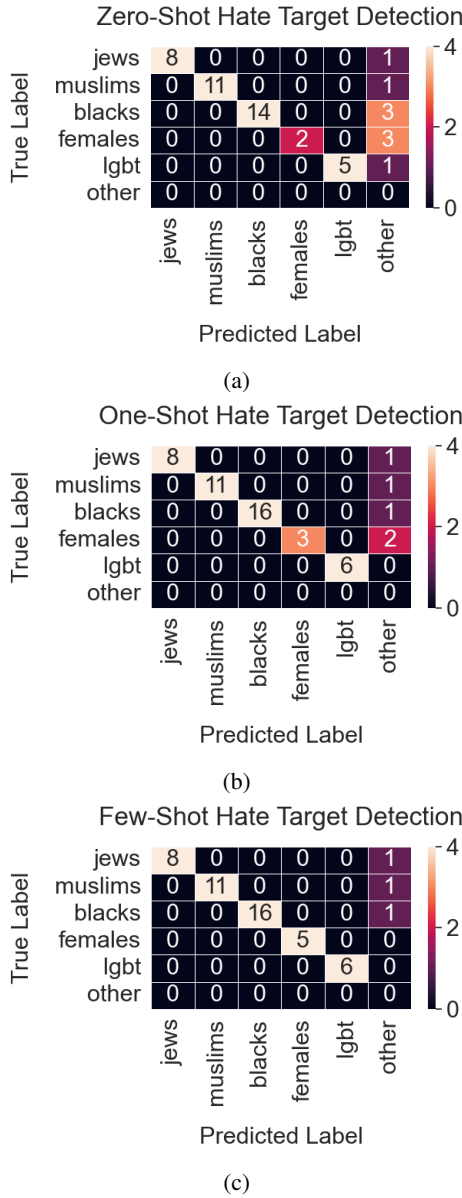


Figure 4: Hate speech target detection quantitative comparison using confusion matrices.

compared to zero-shot - over sensitivity.

5.2.2 Mistakes Can Be Fixed Using Dialog

A dialog was used to assist ChatGPT with the right identification (Fig. 6) so that we could understand if it could be optimised. In the event of a mismatch, we attempted to get an explanation from ChatGPT on how he identifies hate or targets. As a result, we corrected his explanation, which led to a match. It indicates that ChatGPT sometimes has difficulty understanding the prompt and may require additional dialog assistance.

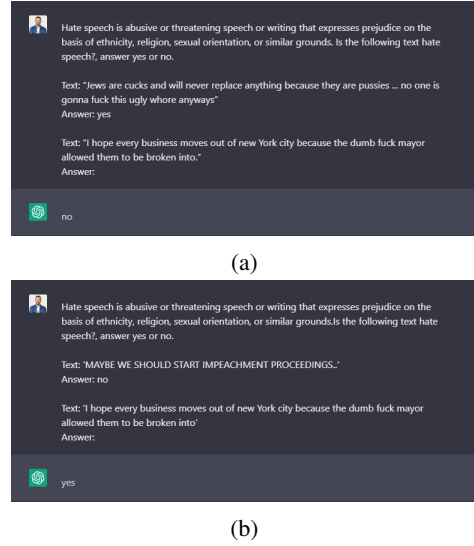


Figure 5: One-shot hate speech detection using different examples.

5.2.3 Target Can Be Explained After Specifying It

A dialog was used to assess ChatGPT's understanding of the prompt (Fig. 7). In cases where target identification was mismatched, we stated the correct target to ChatGPT to get an explanation why the given prompt refers to a certain hate target. Despite not being able to identify the hate target by itself, ChatGPT was able to explain it. This indicates that ChatGPT understands the task well, but is unable to accomplish it.

6 Conclusion

We performed an extensive quantitative and qualitative analysis of ChatGPT for hate speech and target identification using three different prompting techniques - zero-shot, one-shot and few-shot.

ChatGPT performed surprisingly well in both hate speech and target detection in the zero-shot learning setting. The one-shot learning setting performed poorly in identifying hate speech due to its sensitivity to specific examples. One-shot can probably be improved by using some example ensemble to reduce sensitivity. As expected, few-shot performed the most impressively in both hate speech and target detection. ChatGPT can also be used semi-supervised (human in the loop) in order to achieve more accurate and quality results.

To conclude, the large natural language model ChatGPT, although black-box and only prompt-based, is really powerful for detecting hate speech and targets with really high performance, and it can



Figure 6: ChatGPT additional dialog for better identification in mismatch event.

be improved even further with further research.

References

- Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.
- Eyal Arviv, Simo Hanouna, and Oren Tsur. 2021. It's a thin line between love and hate: Using the echo in modeling dynamics of racist online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 61–70.
- Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2021. Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.
- Thomas Davidson, Dana Warmesley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In



Figure 7: ChatGPT additional dialog for better assessment in mismatch target identification.

Proceedings of the international AAAI conference on web and social media, volume 11, pages 512–515.

Fabio Del Vigna¹², Andrea Cimino²³, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Abraham Israeli and Oren Tsur. 2022. Free speech or free hate speech? analyzing the proliferation of hate speech in parler. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 109–121.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439*.

Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the hate code on social media. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 608–611.

- Iliia Markov and Walter Daelemans. 2022. The role of context in detecting the target of hate speech. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 37–42.
- Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019* 8, pages 928–940. Springer.
- Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.
- Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.
- Juan Manuel Pérez, Franco Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, et al. 2022. Assessing the impact of contextual information in hate speech detection. *arXiv preprint arXiv:2210.00465*.
- Angus Stevenson. 2010. *Oxford dictionary of English*. Oxford University Press, USA.
- Zeera Waseem and Dirk Hovy. 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.
- Z Ziqi, D Robinson, and T Jonathan. 2019. Hate speech detection using a convolution-lstm based deep neural network. *IJCCS*, 11816:2546–2553.

A Appendix

A.1 Prompting Examples

A.1.1 Zero-Shot Hate Detection

Prompt:

Hate speech is abusive or threatening speech or writing that expresses prejudice on the basis of ethnicity, religion, sexual orientation, or similar grounds. Is the following text hate speech?, answer yes or no.

Text: 'Now we know where Muslims come from. They are spawned in sewers and crawl out like rats to consume everyone else's stuff.'

Answer:

ChatGPT response:

yes

A.1.2 Zero-Shot Hate Target Detection

Prompt:

Hate speech is abusive or threatening speech or writing that expresses prejudice on the basis of ethnicity, religion, sexual orientation, or similar grounds. The following text contains hate, towards who is the hate directed?

Text: 'Now we know where Muslims come from. They are spawned in sewers and crawl out like rats to consume everyone else's stuff.'

Answer:

ChatGPT response:

The hate in this text is directed towards Muslims.

A.1.3 One-Shot Hate Detection

Prompt:

Hate speech is abusive or threatening speech or writing that expresses prejudice on the basis of ethnicity, religion, sexual orientation, or similar grounds. Is the following text hate speech?, answer yes or no.

Text: 'Man accused of shoving elderly NYC woman has been arrested over 100 times. Another Fricking Black Thug

Hero for the Left. This is the Scum walking our streets that are now dictating our police policy.'

Answer: yes

Text: 'Now we know where Muslims come from. They are spawned in sewers and crawl out like rats to consume everyone else's stuff.'

Answer:

ChatGPT response:

yes

A.1.4 One-Shot Hate Target Detection

Prompt:

Hate speech is abusive or threatening speech or writing that expresses prejudice on the basis of ethnicity, religion, sexual orientation, or similar grounds. The following text contains hate, towards who is the hate directed?

Text: 'Man accused of shoving elderly NYC woman has been arrested over 100 times. Another Fricking Black Thug Hero for the Left. This is the Scum walking our streets that are now dictating our police policy.'

Answer: blacks

Text: 'Now we know where Muslims come from. They are spawned in sewers and crawl out like rats to consume everyone else's stuff.'

Answer:

ChatGPT response:

muslims

A.1.5 Few-Shot Hate Detection

Prompt:

Hate speech is abusive or threatening speech or writing that expresses prejudice on the basis of ethnicity, religion, sexual orientation, or similar grounds. Is the following text hate speech?, answer yes or no.

Text: 'LOOOOOOOOOOOOL... The US is an illegitimate state? What a cuck-old lesbian!'

Answer: yes

Text: 'Cuomo threatens to shut down the Synagogues in NYC. So where the fuck are you Schumer and that fat fucking umpa loompa Nadler. You're both fake fucking Jews.'

Answer: yes

Text: 'If liberals don't like it, don't read it, but they might enrich their lives by reading. #JordanPeterson'

Answer: no

Text: 'If Trump loses the election he won't be loser the people of this country will be the losers'

Answer: no

Text: 'Now we know where Muslims come from. They are spawned in sewers and crawl out like rats to consume everyone else's stuff.'

Answer:

ChatGPT response:

yes

Text: 'Gender Specifics I was born a male and will always be a male. My preferred pronouns are Bite, Me. There is only one other gender: FEMALE. EAT SHIT TRANNIES!!'

Answer: lgbt

Text: 'Now we know where Muslims come from. They are spawned in sewers and crawl out like rats to consume everyone else's stuff.'

Answer:

ChatGPT response:

muslims

A.1.6 Few-Shot Hate Target Detection

Prompt:

Hate speech is abusive or threatening speech or writing that expresses prejudice on the basis of ethnicity, religion, sexual orientation, or similar grounds. The following text contains hate, towards who is the hate directed?

Text: 'Tell me again how muslims are hard working and good people. Sure, they work hard at killing good people and living off our benefits!'

Answer: muslims

Text: 'Twit finally got me AGAIN M banned I called Omar a Muslim piece of shit But it's TRUE! they HATE TRUTH THX PARLER for keeping me on'

Answer: muslims

Text: 'That fat nigger Oprah in the racism commercial says the niggers have demands. I demand they all go back towards Africa in duck boats and sink halfway there.'

Answer: blacks