Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech

Fan Huang
huangfan@acm.org
Indiana University Bloomington
Bloomington, IN, United States

Haewoon Kwak
haewoon@acm.org
Indiana University Bloomington
Bloomington, IN, United States

Jisun An
jisun.an@acm.org
Indiana University Bloomington
Bloomington, IN, United States

ABSTRACT

Recent studies have alarmed that many online hate speeches are *implicit*. With its subtle nature, the explainability of the detection of such hateful speech has been a challenging problem. In this work, we examine whether ChatGPT can be used for providing natural language explanations (NLEs) for implicit hateful speech detection. We design our prompt to elicit concise ChatGPT-generated NLEs and conduct user studies to evaluate their qualities by comparison with human-written NLEs. We discuss the potential and limitations of ChatGPT in the context of implicit hateful speech research.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**.

KEYWORDS

Hate Speech, Toxicity Detection, Natural Language Explanation, ChatGPT, Large Language Models, Human Annotation

1 INTRODUCTION

Warning: *This paper contains offensive content and may be upsetting.*

In November 2022, OpenAI launched a new chatbot model, Chat-GPT [15]. Just two months after its debut, ChatGPT acquired 100 million monthly active users in January 2023, reportedly making it the fastest-growing AI tool in history [8]. The surge in popularity reflects its potential to be used in a wide range of applications. As the most powerful text generation model, not surprisingly, researchers and practitioners have been evaluating its capability for various tasks, including question-and-answer in financial, medical, legal, and psychological areas [10], medical report simplification [13], bug fixing in computer programs [18], and stance detection from texts [20]. While ChatGPT has shown impressive performances for those 'objective' tasks, it is underexplored how ChatGPT would

respond to more 'subjective' tasks, which are crucial elements in understanding social phenomena and require social judgment and decision-making. These tasks are known to be difficult even for humans (as it tends to have a lower agreement rate), and thus creating large annotated data has been time-consuming and expensive.

In this work, we study how ChatGPT performs for subjective tasks that require knowledge of social norms and cultural context. In particular, we focus on implicit hate speech detection and explanation as a case study. Online hate speech is known to be one of the most significant societal issues with many real-world negative consequences [1, 11]. Recent studies have discovered that many online hate speeches are *implicit*, and even advanced machine learning models cannot achieve high accuracy in their detection [7]. Moreover, it has been a challenging problem to provide the explainability of those models by using natural language explanations (NLEs) [4, 7].

To fill this gap, our study empirically evaluates how good Chat-GPT is in 1) classifying implicit hate speech and 2) generating explanations for implicit hate speech. In particular, our two research questions are as follows:

- RQ1: Can ChatGPT detect implicit hateful tweets well?
- RQ2: Does ChatGPT generate quality NLEs?

Using the LatentHatred dataset [7], which is one of the most widely used datasets in the domain, we compare the responses of ChatGPT with the human-written data via human evaluations, highlighting the potentials and limitations of ChatGPT responses.

2 RELATED WORK

Hate Speech Explanation Generation To hinder the spread of toxic speech among online social platforms, researchers developed machine learning models in recent years [2, 5, 16]. Explicit hate speech is easy to understand and detect, while implicit hate speech can be hard to capture due to its nuanced nature [7]. Human-written NLEs to explain why a given text is hateful can be useful in both AI-assisted systems and ML model fine-tuning pipelines [12]. Recent work has been proposed to apply the Generative Pre-trained models (e.g., GPT-2, BART, OPT, and T5 [7, 12, 19]), to create NLEs describing why a given text is considered to be biased [17] or hateful [7, 17].

ChatGPT and its Evaluation ChatGPT is built based on the Reinforcement Learning from Human Feedback (RLHF) approach, which takes advantage of the extensive human annotations [15]. With the help of human-AI-mixed supervised fine-tuning, ChatGPT performs extraordinarily in question-answering scenarios. Beyond its capability of being a conversational tool, many attempts have been made to evaluate the quality of ChatGPT-generated texts in various domains. Guo et al. [10] construct a Human ChatGPT Comparison

Corpus (HC3), which compiles a handful of question-and-answer datasets ranging from financial, medical, to psychological areas, and study the characteristics of ChatGPT's responses compared with that of humans. Jeblick et al. [13] use ChatGPT to generate a simplified version of a radiology report and assess its quality by radiologists. Sobania et al. [18] evaluate ChatGPT on the standard bug fixing benchmark set, QuixBugs. They find that ChatGPT's bug-fixing performance is competitive with the other existing approaches. Gilson et al. [9] examine the performance of ChatGPT on questions within the scope of the United States Medical Licensing Examination (USMLE) Step 1 and Step 2 exams and show that the ChatGPT is comparable to a third-year medical student. Zhang et al. [20] examine how ChatGPT performs on stance detection task (i.e., inferring the standpoint (Favor, Against, or Neither) towards a target in a given text). The above works similarly find out that, in most cases, ChatGPT's answers are comparable to existing human annotations.

In this work, we apply ChatGPT to a new problem that provides a concise NLE for implicit hateful speech, which is a nuanced and context-dependent task. The inherent nature of unclear boundaries of implicit hateful speech and online toxicity makes the problem more challenging.

3 CHATGPT-BASED EXPLANATIONS

3.1 Case Study: LatentHatred Dataset

As a case study, we conduct our analysis by using the *LatentHatred* dataset [7], including 6,358 implicit hateful tweets with their corresponding human annotated target group (i.e., the specific group as the target of hatefulness) and implied statement (i.e., the implications of the expressed hatefulness). For our experiment, we randomly select 12.5% of the whole data (i.e., 795 instances), which is the equal number of the test instances for the implicit hate explanation generation task in [7]). We call these implied statements the *human-written NLEs* as explained in [7].

3.2 ChatGPT Response Collection

Based on the 795 implicit hateful tweets, we use ChatGPT (Jan 9 version (2023)) to generate classification answers (whether a given tweet is hateful or not) and explanations for the reasoning. After testing several variations, we finalize our prompt as follows:

> Given Post: 'AN IMPLICIT HATEFUL TWEET'. Answer yes or no for whether the post is implicitly hateful. Explain why in one sentence.

Specifically, "Given Post: 'AN IMPLICIT HATEFUL TWEET'" is to point out what the given tweet is, "Answer yes or no for whether the post is implicitly hateful" is to ensure that ChatGPT provides a clear answer for implicit hate classification, and "Explain why in one sentence" is to generate a concise explanation for its decision.

Since ChatGPT does not provide official APIs as of January 2023, we use one of the unofficial APIs[1] to collect the ChatGPT responses. For each implicit hate tweet post, we collect three responses from ChatGPT. As the responses by ChatGPT can be influenced by history, we create a new chat session for every implicit hateful tweet to ensure no context is provided to ChatGPT. Table 1 shows an

[1]https://github.com/acheong08/ChatGPT

Prompt	Given Post: 'do you consider identitarianism more of a viable future option than the alt right label identity evropa seem to be going that way'. Answer yes or no for whether the post is implicitly hateful. Explain why in one sentence.
Human-written NLEs [7]	non-white people are enemies
ChatGPT 1	**It's unclear.** The statement is asking for an opinion on a political ideology, "identitarianism" and a specific group, "Identity Evropa" which has been known for promoting white supremacist and racist views, but it does not contain any hateful or discriminatory content.
ChatGPT 2	**Yes.** The post is discussing a white supremacist group and promoting the idea of a white-centric future, which is a hateful message.
ChatGPT 3	**No.** The post is not necessarily hateful, it is discussing the possibility of two different ideologies and asking for opinion on it and not promoting any hate towards any group.

Table 1: Example of our prompt and ChatGPT's responses

example of an input and responses by ChatGPT. We regard the Yes/No response as the binary classification result and the following explanations as the ChatGPT-generated NLE. We notice that, even though we try to limit the answer to Yes or No (i.e., "Answer yes or no for whether the post is implicitly hateful"), ChatGPT sometimes expresses its confusion (e.g., 'It's unclear') when a given tweet falls in a gray area.

For each tweet post, we aggregate the three responses from ChatGPT in the following way. We assign a numeric score of +1 to Yes (i.e., ChatGPT classifies a given post as implicitly hateful), -1 to No (not hateful), and 0 to other responses. We then get the average of the three scores and call it *ChatGPT score* of a given tweet. We label a tweet post 'Hateful' if the ChatGPT score is greater than 0, 'Non-Hateful' if it is less than 0, and 'Uncertain' if it is 0.

3.3 Human Evaluation

We conduct human evaluations to measure the quality of ChatGPT on the classification task and the generated NLEs.

Quality of ChatGPT-based Classification. While all our test instances are from the implicit hate tweet dataset, we find that ChatGPT disagrees (i.e., responding non-hateful or uncertain) for a sizable number of tweet posts. Our manual inspection shows that those posts are potentially debatable or extremely implicit/subtle to be hateful. Thus, we turn to re-evaluating those disagreeing cases to examine to what extent human agrees with ChatGPT.

We annotate the data using Amazon Mechanical Turk (Mturk). For each disagreeing instance, we ask Mturk workers to classify whether a given tweet is hateful, not hateful, or uncertain. We further design three types of evaluations by providing different

contexts. We give a respondent 1) a post only, 2) a post with its human-written NLE, and 3) a post with its ChatGPT-generated NLE. For 3), since we collect three responses from ChatGPT, we choose one NLE from the corresponding label. For example, for a post labeled as Non-Hateful by ChatGPT, if it has two NLEs with the answer No and one NLE with the answer Yes, we randomly select one NLE with the answer No. When a post is labeled Uncertain, we provide the NLE with a score of 0. We collect the three responses from three different Mturk workers for each instance in each evaluation experiment.

To ensure the quality of the collected human annotations, we recruit Mturk Masters who: i) have an approval rate greater than 98%; ii) have more than 5000 HITs approved; and iii) are located in the United States. We also note that no demographic information is collected during this process. All the respondents are monetarily compensated at a rate above the minimum hourly wage in the United States. Since our experiments contain hateful content, we provided extra mental wellness consultant information to prevent the participants from being exposed to unexpected pressures and getting hurt.

Quality of ChatGPT-generated Explanation. To evaluate and compare the quality of human-written NLEs and ChatGPT-generated ones, we use *Informativeness* [6] and *Clarity* [3], which are commonly used metrics that reflect the human perception of generated texts. In our context, Informativeness captures the relevance of the NLE in explaining why a tweet would be considered hateful (7-point Likert scale ranging from 1 (Completely Not Informative) to 7 (Very Informative)). Clarity measures how clear the NLE is (7-point Likert scale ranging from 1 (Completely Unclear) to 7 (Very Clear)).

In this evaluation, we randomly sample 100 implicit hateful tweets. For each of those 100 cases, we provide a Mturk worker with an original tweet and its corresponding NLE and ask to rate the Informativeness and Clarity of the given NLE. We conduct two experiments, one with human-written NLEs and another with ChatGPT-generated NLEs. Since we aim to inspect the general quality of the ChatGPT-generated NLEs, we randomly select one NLE from the three responses of the ChatGPT.

We follow the same procedure as the above Mturk experiment. We also collect three annotations for each tweet-NLE pair. To ensure the quality of responses, we hire experienced research assistants to review them and resolve any disputing cases.

4 RESULT

4.1 RQ1: Does ChatGPT detect implicit hateful tweets well?

Our first research question is whether ChatGPT can detect implicit hateful tweets well. It is a prerequisite for generating correct and quality NLEs. Among the 795 instances from the LatentHatred dataset [7], ChatGPT recognizes 636 instances as 'Implicitly Hateful (80%).' In other words, ChatGPT recognizes 146 instances as 'Not Implicitly Hateful (18.4%)' and 13 instances as 'Uncertain (1.6%),' which conflicts with the original labels in [7]. To examine which is correct, we conduct a user study explained in §3.3. We ask MTurk workers whether a given post is implicitly hateful but with 1) a

post only, 2) a post and human-written NLE, and 3) a post and ChatGPT-generated NLE. Considering the possibility that it is not straightforward to decide whether a given post is hateful or not, we provide the option of 'Not Sure' in addition to 'Yes' and 'No.' We collect three responses for each instance.

Similar to the computation of the ChatGPT score, we assign +1 when a Mturk worker answers Yes (hateful), -1 when the worker answers No (non-hateful), and 0 otherwise. We compute the mean of the three scores for each instance and call the *average hatefulness score*.
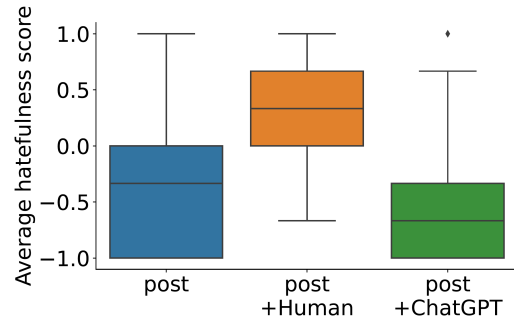

Figure 1: Human evaluation for disagreement between [7] and ChatGPT. We give MTurk workers 1) a post only, 2) a post and human-written NLE ("post+Human"), and 3) a post and ChatGPT-generated NLE ("post+ChatGPT").

There are several interesting observations from the result in Figure 1. First, when only a post is given, we find the average hatefulness score is negative (mean: -0.41), which means lay people (i.e., Mturk workers) are less likely to consider a tweet hateful when ChatGPT could not find hate in it. Second, the average hatefulness score is negative (mean: -0.52) when a post is given with the ChatGPT-generated NLE. Interestingly, lay people are more likely to find a given tweet non-hateful when ChatGPT explains why the given tweet is non-hateful (mean: -0.52) compared to when only a post is given (mean: -0.41). The difference is statistically significant, confirmed by the t-test (t(316) = -13.75, p<0.001). Third, the average hatefulness score is positive (mean: 0.29) when a post is given with the human-written NLE. This shows that lay people's decisions can be drifted by the additional explanation; however, the lower absolute score indicates that providing human-written NLE is not as convincing as the other two cases.

Considering the diversity and nuanced nature of the implicit hate speech [7] and overall toxic behavior online [14], it is not surprising that lay people's decisions can be affected by additional context. Nevertheless, what we observe from the above experiment clearly demonstrates the potential of ChatGPT for nuanced, subjective tasks. Lay people tend to agree with ChatGPT that those tweets are not likely to be hateful. However, on the other hand, ChatGPT-generated NLEs *reinforce* the human perception. In a case when ChatGPT's decision is wrong, such capability might be a danger to misleading lay people and potentially constructing wrong labels.

In summary, our answer to RQ1 is *yes*. ChatGPT shows 80% agreement with [7]. Furthermore, for the 20% disagreement cases,

lay people are highly likely to lean toward ChatGPT's classification results.

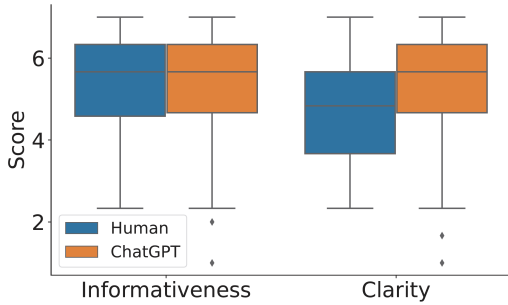## 4.2 RQ2: Does ChatGPT generate quality NLEs?

Figure 2: Quality evaluation of human-written NLEs and ChatGPT-generated NLEs

For the randomly selected 100 instances from our dataset, we collect the Informativeness and Clarity scores as we explained in §3.3. In Figure 2, ChatGPT-generated NLEs show significantly higher Clarity scores (mean: 5.39) than human-written NLEs (mean: 4.68). As confirmed by the t-test, the difference is statistically significant (t(197) = -3.73, p < 0.001). However, there is no statistically significant difference in Informativeness (t(198) = -0.69, p = 0.49).

Our findings demonstrate that, for end users, both human-written NLEs and ChatGPT-generated NLEs provide information to help users identify implicit hatefulness. However, ChatGPT provides more comprehensive illustrations so users can easily confirm implicit hatefulness from a given tweet.

In summary, our answer to RQ2 is also *yes*. ChatGPT generates quality NLEs comparable to human annotators for implicit hate speech.

5 DISCUSSIONS AND CONCLUSION

In this paper, we examine ChatGPT-generated NLEs for the implicit hateful speech from the perspectives of their accuracy and quality.

ChatGPT correctly identifies 80% of the implicit hateful tweets in our experiment setting. The results demonstrate the great potential of ChatGPT as a data annotation tool using a simple prompt design. Further studies can be done to investigate the effect of different prompt designs. Furthermore, for those 20% disagreement cases, our experiments show that ChatGPT's results are more likely to align with lay people's perceptions. Moreover, we also find that the ChatGPT-generated NLEs can reinforce human perception, and they tend to be perceived as clearer than human-written NLEs. These results pose an important question. Since ChatGPT can be convincing, its capability would become a risk of misleading lay people if its decision is wrong. This highlights that it would require extra caution when using ChatGPT as a tool for assisting data annotation.

6 ETHICAL CONSIDERATIONS

Our data collection design via Amazon Mechanical Turk (Mturk) has been approved by Singapore Management University (Approval No.: IRB-22-076-A043(622)). To prevent human annotators from feeling offended or having mental discomfort, we highlight the possible hatefulness in the headline of the Mturk tasks. We also suggest annotators stop labeling anytime they felt uncomfortable and provide the consultant hotlines in our task instructions.

REFERENCES

[1] Jisun An, Haewoon Kwak, Claire Seungeun Lee, Bogang Jun, and Yong-Yeol Ahn. 2021. Predicting Anti-Asian Hateful Users on Twitter during COVID-19. In *EMNLP Findings*.

[2] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *13th International Workshop on Semantic Evaluation*.

[3] Anja Belz and Eric Kow. 2009. System Building Cost vs. Output Quality in Data-to-Text Generation. In *Proceedings of European Workshop on Natural Language Generation*.

[4] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *NeurIPS*.

[5] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online*.

[6] Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Computer Speech & Language* (2020).

[7] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *EMNLP*.

[8] Engadget. 2023. ChatGPT reportedly reached 100 million users in January. https://www.engadget.com/chatgpt-100-million-users-january-130619073.html.

[9] Aidan Gilson, Conrad Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. 2022. How Well Does ChatGPT Do When Taking the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *medRxiv* (2022), 2022–12.

[10] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv:2301.07597* (2023).

[11] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *ICWSM*.

[12] Fan Huang, Haewoon Kwak, and Jisun An. 2023. Chain of Explanation: New Prompting Method to Generate Higher Quality Natural Language Explanation for Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference*.

[13] Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. 2022. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. *arXiv:2212.14882* (2022).

[14] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*.

[15] OpenAI. 2022. ChatGPT: Optimizing Language Model for Dialogue. *OpenAI blog* (2022).

[16] Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *ICWSM*.

[17] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *ACL*.

[18] Dominik Sobania, Martin Briesch, Carol Hanna, and Justyna Petke. 2023. An Analysis of the Automatic Bug Fixing Performance of ChatGPT. *arXiv:2301.08653* (2023).

[19] Rohit Sridhar and Diyi Yang. 2022. Explaining Toxic Text via Knowledge Enhanced Text Generation. In *NAACL*.

[20] Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would Stance Detection Techniques Evolve after the Launch of ChatGPT? *arXiv:2212.14548* (2022).