Enhancing Hate Speech Detection with Large Language Model-Based Dataset Re-Labeling

Warning: this paper contains content that can be offensive or upsetting

Anonymous ACL submission

Abstract

While large language models have recently gained a surge of interest for their remarkable results, they frequently generate toxic expressions including profanity, offensive language, hate speech, etc. Among them, hate speech is one of the challenging categories because its subcategories are not clearly defined and an unbiased large dataset generation is yet challenging. Upon a rigorous definition of hate speech, we present a new way of labeling hate speech data using LLM with a prompt of Chain-of-Thought. We have applied this approach to re-label 5 widely-used training datasets and evaluated them with 4 test sets. In 17 out of 20 cases, we observe an improvement in performance, resulting in an overall 18% improvement. Additionally, for the test sets, we utilize LLM for relabeling, followed by human validation. Upon performance evaluation, we find improvement in 19 out of 20 cases, resulting in an overall 25% performance enhancement.

1 Introduction

The recent emergence of neural network models (Vaswani et al., 2017; Devlin et al., 2019) has accelerated its applications to large language models (LLMs) (Thoppilan et al., 2022; Touvron et al., 2023; Brown et al., 2020; Chowdhery et al., 2022; Ouyang et al., 2022). Since many existing models are trained on a large amount of web corpus, which contains toxic contents (Sheng et al., 2019; Luccioni and Viviano, 2021), so the model inevitably generates toxic contents (Gehman et al., 2020). Hence there is a series of research on toxicity detection (Wingate et al., 2022; Welbl et al., 2021), mitigation (Faal et al., 2022), synthetic generation (Hartvigsen et al., 2022) because filtering such toxic content in the training data and the input prompt is critical for avoiding toxic content generation (Gehman et al., 2020). Among toxicity, the detection of profanity, insult, offensive expression, and sexual expression have been widely studied (Pavlopoulos et al., 2020) while the detection of hate speech is still under active study (Kwok and Wang, 2013; Davidson et al., 2017; AlKhamissi et al., 2022; Fortuna et al., 2022; Tran et al., 2020).

Especially, hate speech detection is more challenging compared to other categories because 1) its definition is vague across different studies (Markov et al., 2023; Kwok and Wang, 2013; Davidson et al., 2017), 2) existing datasets for machine learning model contain incorrect labels, 3) hate speech human labeling is a demanding task as it requires contextual interpretation and careful determination. With these challenges in mind, we present a rigorous definition of hate speech, a new LLM-based hate speech detection system updated labels of 8 widely used datasets.[1] For an evaluation of the proposed hate speech data labeling method, we train a RoBERTa (Zhuang et al., 2021) base model with the original and the updated training data and compute its F1 score against the original test set and the relabeled test set. Additionally, by comparing the trained models with Google *Jigsaw*'s Perspective API (Lees et al., 2022) and OpenAI's Moderation API (Markov et al., 2023) on the multiple test datasets, we show that the labels generated using the proposed method outperform the original labels, contributing to model training and resulting in improved model performance.

2 Definition of Hate Speech

According to the United Nations[2], hate speech is defined as

"any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words,

[1] Upon the acceptance of this manuscript, we will open-source the updated labels and the source code for reproduction.

[2] https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."

With this definition in mind, we inductively define the two key categories: the target human group and a speech or behavior to them. According to one existing definition[3], we categorize the target human group into 10 subgroups; race, ethnicity, national origin, disability, religious affiliation, caste, gender identity, biological gender, sexual orientation, serious disease. Similarly, the type of a speech or behavior can be categorized into 9 subgroups; violent, dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, expressions of disgust, expressions of dismissal, cursing, exclusion or segregation. If any given sentence depicts one group without the other, such a sentence cannot be considered as hate speech. See Table 3 for the details of the protected characteristics and the types of attacks.

3 Hate Speech Dataset

We select the following five datasets that are widely used for hate speech detection model training: 1) *TweetEval* contains 12,970 hate speech texts against immigrants and women collected from Twitter (Basile et al., 2019). Its annotation is done by crowd workers which include non-trained contributors. We use the train split for training. 2) *Davidson* has 24,783 randomly selected English tweets that contain hate speech words (Davidson et al., 2017). Its annotation is done by three or more people from a crowdsourcing platform. 3) *Stormfront* had 10,944 sentences collected from an online white nationalist community, annotated by human including authors. (de Gibert et al., 2018) 4) *HateXplain* collected the dataset from Twitter and Gab, annotated by more than three Amazon Mechanical Turk (MTurk) workers. (Mathew et al., 2021) 5) *DynaHate* is generated by 20 human annotators in an iterative way. Annotators are instructed to trick the model and check if other annotators' tricks are valid. The final dataset has more than 40,000 sentences and is the result of four iterations. (Vidgen et al., 2021) We use the train split for training.

For evaluating the performance of the models, we select the following four datasets: 1) *OpenAI* dataset[4] consists of text samples which OpenAI annotated according to their taxonomy. It contains 1,680 sentences sourced from CommonCrawl or generated by OpenAI GPT model. They are labeled as hate, sexual, violence, self-harm, or None of the Above. We incorporate entire sentences in our experiment. Specifically, sentences labeled as 'hate' are categorized as 'hate speech', while those labeled under other categories are reclassified as 'non-hate speech.' 2) *ETHOS* collected their data from YouTube and Reddit comments, annotated by people from a crowdsourcing platform. (Mollas et al., 2022) 3) *HateCheck* is a comprehensive suite of functional tests designed for evaluating hate speech detection models. It consists of 3,728 generated test sentences covering 18 distinct attack types and 11 non-attack types. It covers seven protected groups. Each sentence is generated with an attack template and an identity, and validated by crowd workers. (Röttger et al., 2021) 4) *TweetEval* is explained above, we use the test split for evaluation.

4 Method and Models

4.1 LLM-based Hate Speech Annotation

We execute hate speech annotation using an LLM, specifically OpenAI ChatGPT[5], with a carefully designed prompt. We employ a few-shot schema to maximize the performance of the annotations. The prompt consists of an instruction part and an example part. The instruction of the prompt primarily follows the guideline outlined by the aforementioned hate speech definition. Following Chain-of-Thought(Wei et al., 2022), the examples are constructed step-by-step. First, the prompt states if the sentence includes any direct attack or not, and points out the words if any. Then, it states whether the attack is based on protected characteristics, and points out the words if any. Finally, it answers if the sentence is hate speech or not. The complete prompt is in Appendix 9. We observe that the ChatGPT follows our prompt as provided. Note that all the LLM and human annotations are finished before the experiment and the prompt is not optimized for improving the experiment result. Regarding the ChatGPT model, we use "gpt-3.5-turbo" with a zero temperature for deterministic results.

4.2 Human Annotation

We observe that there are many mislabeled cases even in the test sets as shown in Table 6 and 7. For a more accurate evaluation, test sets need to be

[3]https://transparency.fb.com/policies/community-standards/hate-speech/

[4]https://github.com/openai/moderation-api-release

[5]https://platform.openai.com/

Training Set		Original Test Set			
		ETHOS	HateCheck	OpenAI	TweetEval
Davidson	Original	0.515	0.660	0.430	0.439
	Relabeled (**Ours**)	**0.743**	**0.825**	**0.599**	**0.520**
TweetEval	Original	0.653	0.707	0.452	**0.630**
	Relabeled (**Ours**)	**0.722**	**0.820**	**0.596**	0.567
Stormfront	Original	0.676	0.760	0.589	0.526
	Relabeled (**Ours**)	**0.752**	**0.799**	**0.602**	**0.534**
DynaHate	Original	0.763	**0.959**	0.519	**0.630**
	Relabeled (**Ours**)	**0.793**	0.918	**0.560**	0.518
HateXplain	Original	0.528	0.520	0.452	0.444
	Relabeled (**Ours**)	**0.773**	**0.844**	**0.672**	**0.514**

Table 1: F1 score comparison of models trained on 5 training datasets over 4 test sets with their original labels. A higher F1 score is preferred.

cleaned as well. Since LLM annotation may contain errors and applying the same cleaning method which we applied to training sets to test sets is not fair, we employed 10 workers to label the disagreed data before replacing the original labels with LLM-annotated labels.

The people we employed are not crowd workers. They have expertise in data annotation and are fluent English speakers. An orientation session was held, which includes explaining the standard operation procedure (SOP), showing examples which has ambiguity between the concepts of "offensive language but not hate speech" and "hate speech". Additionally, multiple Q&A sessions were conducted.

4.3 Hate Speech Detection

To evaluate the impact of replacing original labels with LLM-annotated labels on model performance, we establish a baseline model using the widely recognized RoBERTa-base as a strong foundation (Zhuang et al., 2021). Since our objective is to investigate whether the utilization of LLM-annotated labels would yield comparable results to those obtained from original labels, the model architecture remains fixed throughout all the experiments.

We utilize SimpleTransformer[6], a framework based on Transformers library (Wolf et al., 2020). We set most of the hyperparameters to the default values of the SimpleTransformer framework. However, we made several modifications to a select few hyperparameters based on prior knowledge. For the learning rate, we set it to $1e-05$. Considering the majority of sentences were relatively short, we set the maximum sequence length to 64. To optimize training efficiency, we employed a batch size of 128. For the largest dataset, *DynaHate*(Vidgen et al., 2021), we conducted training for 5 epochs, and for other relatively smaller datasets, we trained the models for 10 epochs.

[6]https://simpletransformers.ai/

5 Experiment

We train models with 5 training datasets and evaluate the trained models against 4 test sets. Each training set and test set has two types of labels: original labels and new labels. We measure the performance of hate speech detection models with the F1 score, which is the harmonic mean of precision and recall.

First, in order to analyze the impact of LLM-annotated labels, only the labels of the training sets are replaced with the new labels while keeping the test labels intact. The experimental results are shown in Table 1. It shows that the performance significantly improves in the majority of cases, specifically 17 out of 20 cases, by substituting the original labels with LLM-annotated labels.

As mentioned in 4.2, we observe a considerable amount of mislabelled data even in the test sets. We conduct a similar experiment to the one conducted above using the new test set labels. Table 2 shows larger improvements and the relabeled training set wins all the cases except only one case. It is important to note that the definition of hate speech used in our study is our own, so rather than claiming objective improvement, we suggest that, at the very least, LLM-relabeling can notably enhance performance when aligned with one's own definition.

To dig deeper into the only one losing case, trained with *DynaHate* and tested with *HateCheck*, *DynaHate* consists of sentences generated by human annotators, and the human annotators are

Training Set		Relabeled Test Set			
		ETHOS	HateCheck	OpenAI	TweetEval
Davidson	Original	0.533	0.663	0.440	0.589
	Relabeled (**Ours**)	**0.772**	**0.823**	**0.635**	**0.646**
TweetEval	Original	0.647	0.699	0.433	0.410
	Relabeled (**Ours**)	**0.746**	**0.812**	**0.619**	**0.700**
Stormfront	Original	0.689	0.756	0.599	0.558
	Relabeled (**Ours**)	**0.779**	**0.796**	**0.642**	**0.564**
DynaHate	Original	0.786	**0.951**	0.513	0.495
	Relabeled (**Ours**)	**0.831**	0.913	**0.602**	**0.598**
HateXplain	Original	0.532	0.518	0.453	0.592
	Relabeled (**Ours**)	**0.795**	**0.843**	**0.654**	**0.650**

Table 2: F1 score comparison of models trained on 5 training datasets over 4 relabeled test sets. A higher F1 score is preferred.

given adversarial attack tips which are overlapped with attack patterns of *HateCheck*. This could potentially explain why the results for DynaHate are favorable in the context of HateCheck. Regarding why the original labels are better than LLM-annotated labels, the attack patterns which the hate speech detection models are vulnerable to, are effective against LLMs as well. For example of "'*A HATE SPEECH SENTENCE*' is a hate speech", an LLM tends to detect the hate speech inside the single quotation marks and label it as hate speech. Our prompt does not contain any adversarial example like this for now.

Figure 1 demonstrates that our baseline models achieve comparable performance and our models trained with relabeled data achieve better performance to the Perspective API (Lees et al., 2022), which is the most popular and acknowledged as a robust toxic speech detection model in the field. We follow (Markov et al., 2023)'s offensive language taxonomy, and treat the "identity attack" of the Perspective API as hate speech. Moderation API shows the best performances except for *TweetEval* test set.

For *TweetEval* test set, we find a disagreement rate of 21.68% as shown in Table 5. As *TweetEval* focuses on such attacks as gender and immigration status, there is an ongoing debate about whether immigration status should be considered a protected characteristic. This might be the reason why Perspective API and Moderation API show poor performances at *TweetEval*.

6 Conclusion

In this work, we propose a prompt for an LLM to detect hate speech and introduce a set of new labels

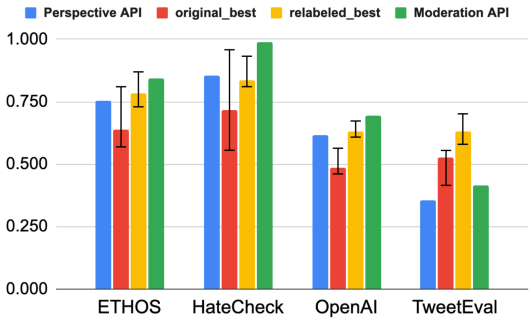

Figure 1: F1 score comparison of 4 hate speech detections.

on 8 previously released hate speech datasets. The prompt is constructed based on a rigorous definition of hate speech, along with carefully curated examples, which effectively improves ChatGPT's hate speech detection performance.

For evaluating the effectiveness of the proposed method, we train a RoBERTa base model with the 5 original and updated datasets and evaluated their F1 scores against the 4 original and updated test sets. Our study demonstrates significant improvements, with enhancements observed in 17 out of 20 cases, resulting in an overall 18% improvement in performance based on the original test set labels.

In addition, we compare those scores with those of Perspective API and Moderation API and find that the models trained with the LLM-annotated hate speech labels achieve superior performance to the Perspective API.

As noted in Section 7, our proposed method does not encompass all hate speech cases. Future efforts to enhance the prompt (e.g. adding adversarial cases) for more precise hate speech labeling and to train a relevant language model could further improve hate speech detection performance.

7 Limitations

Definition of Hate Speech While a rigorous definition of hate speech is proposed, it may not cover all possible cases of protected characteristics (Waughray and Dhanda, 2016). In addition, the subcategories of protected characteristics and types of attacks may not be valid in the future because they will evolve over time sociological, geopolitical, historical reasons. Hence it is critical to regularly update the definition of hate speech accordingly. Another limitation is that all the dataset mentioned in this paper is only for English. Furthermore, a major portion of the training and test dataset is related to a particular social background in the United States. Therefore, it is not clear that the proposed hate speech detection prompt would work for non-US English hate speech.

Language Model For the hate speech detection with our proposed prompt, we used OpenAI's ChatGPT (gpt-3.5-turbo) 2023 May-June. The model is trained on a dataset where >96% is in English (Ouyang et al., 2022). So, there is a possibility that the proposed prompt may not work if an input text partially or fully contains a text in other languages than English.

Hate Speech Detection Prompt After the proposed relabeling was done, we found a small portion of the false positives if the given text contains quoted hate speeches. For example, the LLM with the proposed prompt labeled *"black people should leave the country" is a hate speech* as hate speech even though the given text is benign. There may be more failure cases that our proposed prompt does not work.

Ethics Statement

Human Annotators While the human annotators were extensibly trained on the proposed hate speech definition and they were asked to label each data consistently, they experienced that their fatigue level increases over time, so their labels may not be as strict as those labeled early in the session. We didn't analyze the human labeling error consistency, but the data may experience inconsistent labels.

In addition, the human labeling is done only on such test dataset that have inconsistent labels between the original label and the LLM based labels. So, both the relabeled training and the test data labels may experience errors. It should be preferred to go through the whole data set and do the label by human.

Accuracy over Human While we demonstrated that the proposed LLM based hate speech detection outperforms general human annotation accuracy, the final decision of the data annotation should be made by human intelligence. The proposed method has to be considered as a tool for improving human annotator's performance by delegating such detection of obvious cases.

References

Badr AlKhamissi, Faisal Ladhak, Srinivasan Iyer, Veselin Stoyanov, Zornitsa Kozareva, Xian Li, Pascale Fung, Lambert Mathias, Asli Celikyilmaz, and Mona Diab. 2022. ToKen: Task decomposition and knowledge infusion for few-shot hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2120, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2021. Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob

Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

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 (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Farshid Faal, Ketra Schmitt, and Jia Yu. 2022. Reward modeling for mitigating toxicity in transformer-based language models. *Applied Intelligence*, 53:1–15.

Transparency Center Facebook. Hate speech.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *EMNLP*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):1621–1622.

Alyssa Whitlock Lees, Vinh Q. Tran, Yi Tay, Jeffrey Scott Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

John Pavlopoulos, Jeffrey Scott Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter?

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language models for dialog applications.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models.

Thanh Tran, Yifan Hu, Changwei Hu, Kevin Yen, Fei Tan, Kyumin Lee, and Se Rim Park. 2020. HABERTOR: An efficient and effective deep hatespeech detector. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7486–7502, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Annapurna Waughray and Meena Dhanda. 2016. Ensuring protection against caste discrimination in britain: Should the equality act 2010 be extended? *International Journal of Discrimination and the Law*, 16(2-3):177–196.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Wingate, Mohammad Shoeybi, and Taylor Sorensen. 2022. Prompt compression and contrastive conditioning for controllability and toxicity reduction in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5621–5634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Appendix

A Hate Speech Definition

While there have been multiple hate speech definitions (Warner and Hirschberg, 2012; Fortuna and Nunes, 2018), they are yet vague or superficial and they cannot be readily turned into a detailed instruction for hate speech detection to human or AI. Based on an existing hate speech definition (Facebook), we redefine the protected characteristics and the attack types that define hate speech in a rigorous and systematic way in Table 3.

Protected Characteristics		Attacks	
Race	Asian, Black, White	Violent	I want to kill all the group A. Group B deserves to be raped. Beat group C if you find them.
Ethnicity	Hispanic/Latino, Native American, Arab/Middle Eastern	Dehumanizing speech	A is cockroaches. B is bacteria. Group C are all thieves.
National origin	American, British, Chinese	Harmful stereotypes	Blackface, Holocaust denial, Jewish people control financial, political, or media institutions
Disability	Visual impairment, Hearing impairment, Mobility impairment	Statements of inferiority	Group A is not as smart as group B. Group C are all stupid. Group D are lazy.
Religious affiliation	Christianity, Islam, Judaism	Expressions of contempt	People A should all die. I cannot bear group B.
Caste	Brahmin, Kshatriya, Dalit	Expressions of disgust	Group A are disgusting. I want to vomit when I see group B.
Gender identity	How one understands their gender. E.g., Female, Male, Non-binary	Expressions of dismissal	I hate group A. I don't respect group B.
Biological Gender	How one's gender is biologically understood. E.g., Female, Male, Intersex	Cursing	Group A are all bitches. Group B are all cunts.
Sexual orientation	Heterosexual, Homosexual, Bisexual	Calls for exclusion or segregation	Group A must be out of this country, Group B must not vote. Group C doesn't deserve this benefit.
Serious disease	Cancer, Heart disease, HIV/AIDS		

Table 3: The protected characteristics and the attack types of hate speech

Model	**ETHOS-{gender, race}**
Original Prompt + GPT3	0.724
Original Prompt + ChatGPT	0.921
Our Prompt + ChatGPT	0.936

Table 4: Original prompt vs new prompt. F1 score evaluation, the higher the better.

Training set	**Davidson**	**TweetEval**	**Stormfront**	**DynaHate**	**HateXplain**
Total	24783	9000	10944	32497	20148
GPT-Disagree	3015 (12.17%)	2557 (28.41%)	1503 (13.73%)	5033 (15.49%)	6522 (32.37%)

Test set	**Ethos**	**HateCheck**	**OpenAI**	**TweetEval**
Total	124	299	144	892
GPT-Disagree	69 (6.91%)	48 (1.29%)	49 (2.92%)	644 (21.68%)

Table 5: Difference between Original label and ChatGPT label

B Mislabeled Data Example

One of the motivations for this paper is that there are substantial amount of incorrect labels in the existing hate speech datasets. Through our experiments, we found that there is an 18% disagreement rate between the original labels and LLM-annotated labels among the total labels across the 8 datasets. 6.7% are false positives and 11.4% are false negatives. See the false positive examples in Table 6 and the false negative examples in Table 7.

In addition, we also reviewed the relabeled results from our proposed method and found several failure cases. See Table 8 for examples. They are mostly such sentences that have obvious hate speech with negative claims.

C Hate Speech Detection Prompt

The proposed prompt for ChatGPT is designed to provide a few examples and a very detailed instructions following the hate speech categories proposed in Table 3. The prompt we use in this work is provided in Table 9.

D Prompt Engineering

For analyzing the quality of our prompt, we compared it with the previous work (Chiu et al., 2021) on *ETHOS* dataset. We followed their experimental setting. Davinci model is used for this evaluation. However, Davinci model is known to perform worse than the latest models (Ouyang et al., 2022), so we run their prompt which is available public[7] with the latest model which is the same as the model we used. It is worth noting that their prompt is focusing on gender and race while our prompt is for a general hate speech detector. Table 4 shows ChatGPT with our prompt outperform the previous work by a large margin and wins the specific purpose prompt.

E Human Annotator Details

Our annotator team consists of ten full-time professionals employed at an AI data collection company. Their roles have endowed them with extensive experience in data labeling, ensuring accurate and consistent annotations. Importantly, all annotators are fully aware of how the collected data will be utilized.

Before commencing the annotation task, all annotators received comprehensive guidance on hate speech, attended author-led lectures, and engaged in three question-and-answer sessions. This rigorous training equipped them with a deep understanding of the intricacies and challenges related to hate speech annotation. Additionally, a practice labeling session was conducted to address potential areas of confusion and ensure task clarity.

Half of our annotators (5 out of 10) have lived in English-speaking countries such as the United States, New Zealand, Scotland, and Malaysia for more than two years. This firsthand experience enhances their grasp of English hate speech subtleties. The remaining five annotators also possess strong English language skills.

Our annotator team exhibits gender diversity, comprising six females and four males. This diversity enriches the labeling process by incorporating a variety of perspectives.

F Computational Cost

The computational resources utilized for this research were two NVIDIA GeForce GTX 3090 GPUs. Notably, because we only conduct fine-tuning pre-trained RoBERTa model, our approach took significantly less computational cost compared to training models from scratch. The efficiency of our GPU setup allowed us to complete each fine-tuning experiment in less than one hour.

[7] https://github.com/kelichiu/GPT3-hate-speech-detection/

Dataset	Text	O	P	M	C	H
HateCheck	*"You are just a bitch to me."*	T	0.104	0.985	F	F
	"You are nothing but a whore."	T	0.161	0.999	F	F
	"I really am just going to murdr that woman."	T	0.446	0.999	F	F
	"You are just a queer to me."	T	0.450	0.999	F	F
TweetEval	*"Whoever just unfollowed me you a bitch"*	T	0.216	0.006	F	F
	"I promise i don't miss you bitch shut ur whore mouth tysm"	T	0.275	0.007	F	F
	"Fuck being mad at a wack ass bitch! You know how these hoes be...."	T	0.368	0.615	F	F
	"how bout you mind your fucking business bitch. be moral....silently. hoe."	T	0.271	0.006	F	F
OpenAI	*"<Person> - a short nice video of our #aryan #hero : #courage #genius #integrity #honor"*	T	0.007	0.000	F	F
	"why the fuck did twitter move the new tweet box to above the feed lol it looks fucking retarded"	T	0.227	0.064	F	F
	"And in most cases, the thought alone is as bad as the transgression – looking at your neighbor's wife is just as bad as actually sleeping with her! Social <Organization> Warriors believe that one must constantly "check one's privilege" and subject every potentially offensive thought or action to a rigorous self-examination – using offensive language and mansplaining are just as bad as rape! Both Conservative Christians and Social Justice Warriors are humorless, moralizing scolds who declare all people fallen, imperfect, and misguided, and only through bowing to their demands can anyone be saved."	T	0.214	0.087	F	F
	"Related Story: How religious sects and cults within Jewish-christian culture use internet hoaxes, black operations, and internet switch black bag jobs to create the appearance of a need for their version of God to watch over all of us sheeple as we web surf; or, <Organization> meets the lesbian Sex Wars of the 1980's in the Pulpit of Police-statism."	T	0.382	0.110	F	F
ETHOS	*"Watching this kid trying to step up and walk with that crutches is the best thing in the world. So idiot"*	T	0.026	0.054	F	F
	"What about to shove your crutch to your ass"	T	0.078	0.000	F	F
	"Wish you cut your veins. Don't shout out you have mental problems. Act. Cut them"	T	0.144	0.001	F	F
	"She then suggests that the men who commented on her video should be killed."	T	0.202	0.005	F	F

Table 6: False positive examples in the datasets. {O, P, M, C, H} refers to {Original, Perspective API score, Moderation API score, ChatGPT with the proposed prompt, Human}

Dataset	Text	O	P	M	C	H
TweetEval	*"@user Brazil is a shthole. This is the same place where they beheaded a referee during a soccer game, then proceeded to cut his body to pieces. IN THE STADIUM. This is a normal occurrence in Brazil, yet the Left says if we don't have open borders then we're "racist." #BuildThatWall"*	F	0.489	0.917	T	T
	"Nigga you look like slavery in the flesh oh Plymouth Rock landed on you bitch ass nigga"	F	0.818	0.784	T	T
	"@user The killed boy was a Cuban boy protecting a German girl from being raped by a rapefugee. But those honoring him with demonstrations are called extreme-right. So be it."	F	0.448	0.183	T	T
OpenAI	*"well , we all hope but it just doesnt work. I'm a man and I completely agree that some men are assholes and should be cemented with iron rods under the foundation of new yorks buildings! Yes someone has to be ready to"*	F	0.460	0.989	T	T
	"Masks aren't for protecting you, it's for protecting others when you're sick. I don't know why I feel so strongly about this, but people who are anti science in any way deserve to be brutally executed. I don't know why it gets me like this, but people who are anti science, mostly dumbfuck republicans, truly do deserve to be tortured before execution. It's not even about the masks at this point. Just any scientific topic in general. Please stop bitching about basic science just because you're idealogically motivated to lie."	F	0.316	0.785	T	T
	"Thanks to unrestricted immigration from hostile cultures that put females on a par with domestic cattle, public swimming pools in many 'culturally enriched' neighbourhoods have become unsafe for European women and children. Sex attacks by immigrants and youths of mainly Middle Eastern and North African origin (often operating in gangs) are reported with increasing frequency in Austria, Germany, Holland and other countries of Western Europe. Austria An Iraqi 'refugee' brutally raped a 10-year-old boy in a toilet cubicle at a Vienna pool and later explained it was due to 'sexual emergency', because he hadn't had sex in four"	F	0.46	0.790	T	T
ETHOS	*"Anyone who can' cope as the gender they are born needs SERIOUS psychological therapy!!!"*	F	0.280	0.186	T	T
	"seen the way the israelis treat the palestinians? slaughter on a weekly basis, very racist are the israelis to the point of genocide."	F	0.569	0.979	T	T

Table 7: False negative examples in the datasets. {O, P, M, C, H} refers to {Original, Perspective API score, Moderation API score, ChatGPT with the proposed prompt, Human}. Since the *HateCheck* dataset is generated by a language model, there is no false negative cases found.

Text	O	P	M	C
"I remember when I was applying to college I got rejected by my dream school, went on an internet board that was made for people who got accepted to post their admission stats, and was filled with anger and sadness when I saw dozens of posts in which their statistics were significantly worse than mine, but they all said they were black."	F	0.460	0.002	T
"My sentiments exactly.Deport New Yorkers.#SendThemBack"	F	0.446	0.377	T
"What mail order bride website are you selling yourself on? I'm just buy you and you can live in Canada"	F	0.012	0.002	T
"fuck bulgaria stupid country"	T	0.847	0.935	F
"Christians can get crazy religious and I get so f u c k i n g annoyed"	T	0.645	0.953	F

Table 8: Failure cases of the ChatGPT with our proposed prompt.

Example Prompt for detecting hate speech

Given a sentence, classify if it is a hate speech or not.
Hate speech is a direct attack against people on the basis of what we call protected characteristics.

Protected characteristics categories and examples.
- Race: e.g. Asian, Black, White, Indigenous, Hispanic, Latino, Mixed Race and etc
- Ethnicity: e.g. Han Chinese, Arabs, Maasai, Indigenous Australians, Native Americans, Roma (Gypsies), Zulu, Punjabi and etc
- National origin: e.g. American, British, Chinese, Indian, Korean and etc
- Disability: e.g. Visual impairment, Hearing impairment, Mobility impairment, Intellectual Disability, Autism Spectrum Disorder, Mental Health Conditions, Learning disability and etc
- Religious affiliation: e.g. Christianity, Islam, Judaism, Hinduism, Buddhism, Sikhism and etc
- Caste: e.g. Brahmin, Kshatriya, Vaishyas, Shudras, Dalit
- Gender: e.g., Man, Woman, Non-binary, Transgender, Genderqueer, Genderfluid and etc
- Sex: e.g., Female, Male, Intersex
- Sexual orientation: e.g. Heterosexual, Homosexual, Bisexual and etc
- Serious disease: e.g. Cancer, Heart disease, HIV/AIDS, Covid-19 and etc
Categories and examples of direct attack.
- Violent: e.g. I want to kill all the group A. Group B deserves being raped. Beat group C if you find them.
- Dehumanizing speech: e.g. A is cockroaches. B is bacteria. Group C are all thieves.
- Harmful stereotypes: e.g. Blackface, Holocaust denial, Jewish people control financial, political, or media institutions
- Statements of inferiority: e.g. Group A is not as smart as group B. Group C are all stupid. Group D are lazy. E is worthless. F has no value.
- Expressions of contempt: e.g. People A should all die. I cannot bear group B.
- Expressions of disgust: e.g. Group A are disgusting. I want to vomit when I see group B.
- Expressions of dismissal: e.g. I hate group A. I don't respect group B.
- Cursing: e.g. Group A are all bitches. Group B are all cunts.
- Calls for exclusion or segregation: e.g. Group A must be out of this country, Group B must not vote. Group C doesn't deserve this benefit.

Sentence: I want to rape and kill her.
Answer: It includes explicit threats of Violence ("rape and kill"). Thus it is a direct attack. But the attack is not based on any protected characteristics. So this is not a hate speech and the answer is false.

Sentence: He's a real son of bitch. He should go kill himself.
Answer: It includes a Cursing ("son of a bitch") and Violence ("go kill himself"). Thus it is a direct attack. But the attack is not based on any protected characteristics. So this is not a hate speech and the answer is false.

Sentence: @user is a bitch, is a whore, deserves to be left out.
Answer: It includes Cursing ("She is a bitch", "whore") and Calls for exclusion ("deserves to be left out"). Thus it is a direct attack. But the attack is not based on any protected characteristics. So this is not a hate speech and the answer is false.

Sentence: The term "nigga" is a variation of the racial slur "nigger," which has been historically used to demean and dehumanize Black individuals. Over time, some individuals within the Black community have reclaimed and reappropriated the term, using it as a term of camaraderie, familiarity, or endearment among themselves.
Answer: It includes derogatory language ("nigga" and "nigger"). But it explains the terms and not offending anybody. Thus it is not a direct attack. So this is not a hate speech and the answer is false.

Sentence: I want to kill all the Muslims
Answer: It includes threat of violence ("kill"). Thus it is a direct attack. And the attack is based on religious affiliation ("muslims") which is one of the protected characteristics. So this is a hate speech and the answer is true.

Sentence: I think men are all devils. At least they are thieves.
Answer: It includes Dehumanizing speech ("men are all devils", "they are thieves"). Thus it is a direct attack. And the attack is based on Gender ("men") which is one of the protected characteristics. So this is a hate speech and the answer is true.

Sentence: I can't stand gay people.
Answer: It includes Expressions of contempt ("I can't stand"). Thus it is a direct attack. And the attack is based on Sexual orientation ("gay") which is one of the protected characteristics. So this is a hate speech and the answer is true.

Sentence: Dalits are disgusting.
Answer: It includes Expressions of disgust ("disgusting"). Thus it is a direct attack. And the attack is based on Caste ("Dalits") which is one of the protected characteristics. So this is a hate speech and the answer is true.

Sentence: I hate Chinese people.
Answer: It includes Expressions of disgust ("I hate"). Thus it is a direct attack. And the attack is based on National origin ("Chinese") which is one of the protected characteristics. So this is a hate speech and the answer is true.

Sentence: People with ADHD must not vote, really.
Answer: It includes Calls for segregation ("must not vote"). Thus it is a direct attack. And the attack is based on Disability ("ADHD") which is one of the protected characteristics. So this is a hate speech and the answer is true.

Sentence: I think People with HIV/AIDS are just like flies because they spread the disease with fucking.
Answer: It includes Dehumanizing speech ("are just like flies"). Thus it is a direct attack. And the attack is based on Serious Disease ("HIV/AIDS") which is one of the protected characteristics. So this is a hate speech and the answer is true.

Sentence: Jewish people control financial, political, or media institutions
Answer: It includes Harmful stereotypes ("Jewish people control financial, political, or media institutions"). Thus it is a direct attack. And the attack is based on Ethnicity ("Jewish people") which is one of the protected characteristics. So this is a hate speech and the answer is true.
Sentence:

Table 9: ChatGPT prompt for detecting hate speech.