Exploring the Capability of ChatGPT to Reproduce Human Labels for Social Computing Tasks

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and
Gareth Tyson

The Hong Kong University of Science and Technology (Guangzhou)

Abstract. Harnessing the potential of large language models (LLMs) like ChatGPT can help address social challenges through inclusive, ethical, and sustainable means. In this paper, we investigate the extent to which ChatGPT can annotate data for social computing tasks, aiming to reduce the complexity and cost of undertaking web research. To evaluate ChatGPT's potential, we re-annotate seven datasets using ChatGPT, covering topics related to pressing social issues like COVID-19 misinformation, social bot deception, cyberbully, clickbait news, and the Russo-Ukrainian War. Our findings demonstrate that ChatGPT exhibits promise in handling these data annotation tasks, albeit with some challenges. Across the seven datasets, ChatGPT achieves an average annotation F1-score of 72.00%. Its performance excels in clickbait news annotation, correctly labeling 89.66% of the data. However, we also observe significant variations in performance across individual labels. We believe that this research opens new avenues for analysis and can reduce barriers to engaging in social computing research.

Keywords: ChatGPT · Crowdsourcing · Social Data Annotations

1 Introduction

Crowd-sourced human intelligence is commonly used for text data annotation [9]. Such annotations are then used for training various models, including stance detection [8], hate speech detection [11], and bot detection [7]. While unsupervised methods are being introduced for classification tasks, such methods usually require large data samples. Thus, social computing research still relies heavily on human annotations. This, however, creates significant barriers for less well-funded research labs. For example, the annotation of a 10,000-post social media dataset by three human annotators would take approximately five hours. With a rate of $25 per individual worker, this would cost hundreds of dollars [5]. When performing comparative analyses across multiple datasets, these costs can easily escalate to thousands of dollars.

Recently, the release of ChatGPT has uncovered a range of cases where large language models (LLMs) can help substitute human intelligence [24]. Several works compare the use of ChatGPT to human methods [22]. For instance, researchers have investigated the use of ChatGPT for automatic misinformation detection [20], and even generating academic writing [2]. In this paper, we explore the potential of using ChatGPT for five text-based social computing annotation tasks. We seek to understand whether ChatGPT has the potential to reproduce human-generated annotations. ChatGPT's annotations can highlight its usefulness against crowd-sourced annotations. To achieve this, we first use ChatGPT to label seven annotation datasets on five distinct social problems – COVID-19 controversies (3x), social bot deception, cyberbully, clickbait news, and Russo-Ukrainian War stance detection. We compare ChatGPT's labels with the human assigned labels on those datasets. Our results show that ChatGPT *does* have the potential to perform data annotation tasks. Performance is highest for the clickbait headlines dataset, with ChatGPT correctly annotating 89.66% of headlines. In contrast, performance is worst for the COVID-19 hate speech task, which only correctly annotates 52.24% of posts. Closer inspection reveals that performance varies substantially across individual labels. We observe significant gaps (over 25%) exist between labels' accuracy on five out of seven datasets. For instance, in social bot detection, while ChatGPT can identify 81.1% of human tweets, it only identifies 45.5% of bot tweets.

We hope that this work can open up new lines of analysis and can act as a basis for future research into the use of ChatGPT for human annotation tasks. Our contributions are as follows:

1. We evaluate the efficacy of ChatGPT at replacing human annotators for five important social computing data tasks, spanning seven datasets.
2. We show that ChatGPT *can* replace human annotators, yet this varies across tasks and datasets. Performance is highest for clickbait headlines detection (89.66% accuracy), and lowest for hate speech detection (52.24% accuracy).

2 Related Work

Recent research has looked at using ChatGPT for annotating social computing data. Huang et al. [12] report that ChatGPT is able to correctly annotate 80% of the implicit hateful tweets from the `LatentHatred` dataset [6]. In addition, the authors show that ChatGPT's explanations can reinforce human annotators' perception of the target text in explaining why tweets would be annotated as hateful or not. Our study also examines how well ChatGPT performs in annotating hate speech. However, we do not limit the annotation to a binary decision for whether tweets are hateful or not, and further include a neutral label. Note, existing literature has highlighted the importance of this, stating that tweets with neutral expressions can defense the spread of hateful content [15].

Similar to us, others have used ChatGPT for performing stance detection [1]. Zhang et al. [22] evaluate ChatGPT's performance on detecting political stance on two prevalent datasets, `SemEval-2016` [16] and `P-Stance` [14]. The authors report that ChatGPT can outperform most state-of-art stance detection models in zero-shot settings, suggesting ChatGPT's potential to handle stance annotation tests. Major challenges still remain though. Aiyappa et al. [1] show that ChatGPT's performance varies across different model versions. The authors also point out that such variance is due to the possibility of data leakage, where past prompts are used for training the next ChatGPT generation. Nonetheless, it remains unclear how well ChatGPT performs in annotating individuals' stances in a broader context, such as those who are in favor or against a particular issue. Our analysis also finds further challenges, e.g., that ChatGPT has a tendency to overestimate neutral stances.

In addition to the aforementioned themes, others have experimented with ChatGPT to perform tasks such as sentiment analysis and fake news detection [3]. However, most of this literature only focuses on a single annotation task. In contrast, we seek to perform a comparative analysis across different data annotation tasks aiming to solve social problems.

3 Methodology

3.1 Overview of annotation tasks and datasets

We follow a comparative approach to analyze the differences in human annotations and ChatGPT annotations by utilizing seven different datasets. We select seven annotation datasets covering five social problems that are commonly used in academic research: ***(i)*** COVID-19 controversies (vaccine arguments, anti-Asian hate speech, and COVID-19 fake news) [19, 11, 17], ***(ii)*** social bot deception [7], ***(iii)*** cyberbully [13], ***(iv)*** clickbait news [4], and ***(v)*** Russo-Ukrainian War [23, 10, 18]. For these seven datasets, we then attempt to recreate the human annotations using ChatGPT.

Our dataset selection strategy is based on the following requirements: ***(i)*** The datasets must be in English to avoid differences in language provision, and ***(ii)*** The datasets must be annotated by human annotators, as we wish to compare the human annotations with ChatGPT. We list our targeted annotation tasks and datasets statistics in Table 1.

3.2 ChatGPT Annotation

For each annotated dataset, we try to recreate the same annotations using ChatGPT. We utilize OpenAI API, configured with module `gpt-3.5-turbo-0631`, to annotate each target dataset. We rely on an official prompt example for classification tasks from OpenAI.[2]

In the official document, most prompts are imperative sentences starting with a verb. As such, we choose "Classify", which is frequently used in annotation work. We use this verb to design our prompt. According to the official template, we find that starting a new row with a word describing the subject and object in the prompt is effective. Thus, we follow this pattern, injecting the subject and objective of the annotation task here. Another benefit of this template is that it is flexible to inject text input to annotate and specify a desired format for ChatGPT to respond with its annotation.

Based on this template, we modify this example into a generalized prompt template applicable to our seven distinctive datasets. The template can be adjusted to be applied for annotating text in different context as shown as follows:

[2] `https://platform.openai.com/docs/guides/completion/prompt-design`

Dataset	**Size**	**ChatGPT annotation**	**Labels (Human/ChatGPT)**
Vaccine Stance [19]	5,926	5,832 (98.41%)	Anti-vaccine (21.56%/10.62%) Pro-vaccine (39.72%/26.86%) Neutral (38.72%/62.52%)
COVID-19 Hate Speech [11]	2,290	2,280 (99.56%)	Hate (18.73%/34.30%) Counterhate (22.58%/23.73%) Neutral (58.69%/41.97%)
COVID-19 Fake News [17]	10,700	10,540 (98.50%)	Real news (52.34%/64.00%) Fake news (47.66%/36.00%)
Social Bot [7]	16,824	16,235 (96.50%)	Human (54.08%/69.01%) Bot (45.92%/30.99%)
Anti-LGBT Cyberbully [13][1]	4,299	4,193 (97.53%)	Cyberbully (29.22%/49.10%) Not cyberbully (70.78%/50.90%)
Clickbait Headlines [4]	32,000	31,084 (97.14%)	Clickbait (50.00%/40.54%) Not clickbait (50.00%/59.46%)
Russo-Ukrainian Stance [18]	1,460	1,321 (90.48%)	Pro-Ukraine (63.90%/48.52%) Pro-Russia (36.10%/51.48%)

Table 1: A summary of selected datasets and ChatGPT's annotation. The column "ChatGPT annotation" shows the volume of ChatGPT's responses matching any candidate label of the dataset, with the percentage representing its proportion to dataset size. The column "Label (Human/ChatGPT)" details the proportion of each label's size to dataset size, annotated by human or ChatGPT respectively.

"*Classify the text about {Topic} with a label from [Label 1, Label 2, ...]. Text: "{text to classify}". Desired format: <label_for_ classification>*", where *{Topic}* refers to the topic or background of the text; The *[label1, label2, ...]* refers to the set of candidate labels for ChatGPT to annotate the text; and *{text to classify}* refers to the text input for ChatGPT to produce label. In addition, the plain-language index *Desired format* indicates that ChatGPT should only respond using the label without any other text. We then apply this template to generate a ChatGPT prompt according to the dataset's original annotation strategy. Following this, we pass all data to ChatGPT for annotation. When ChatGPT responds to such a prompt, it is necessary to parse the response and extract its annotation. We consider a response parsable only if it follows the desired format – only providing a label without any other text. Thus, we extract ChatGPT's annotation by matching its response to any candidate labels for the dataset.

In all, we only encounter an average of 3.13% ($SD = 2.99\%$) responses per-dataset that fail to be parsed. Note, a small number (0.1%) of failed cases are because ChatGPT states there is not enough information for it to make a decision. For example, a failed response from Russo-Ukrainian Stance dataset states "*Cannot classify the given text with the label [Pro-Russia, Pro-Ukraine] as it does not provide any relevant information about the topic.*"

We emphasize that there are many ways in which our methodology could be expanded and refined. Our future work will involve exploring alternative forms of prompt formulation and response mining.

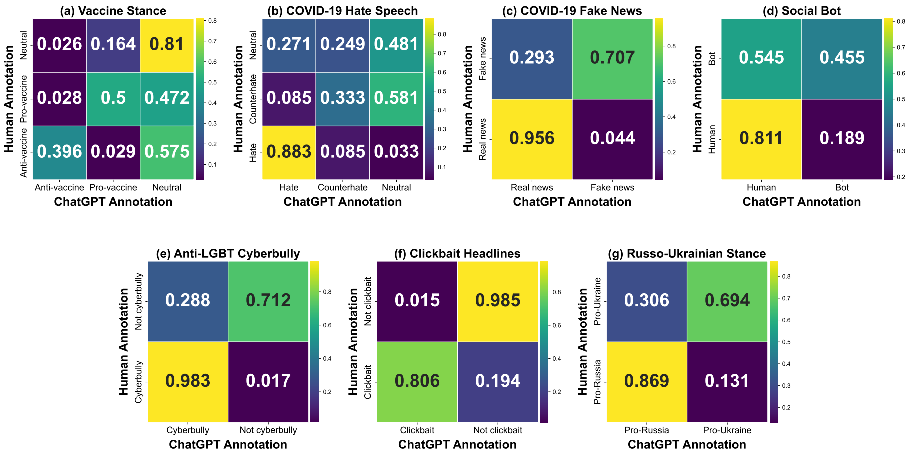

Fig. 1: The confusion matrices of ChatGPT's annotations for the seven annotation datasets. The y-axis label for a given row shows human label of text, and values in each cell show the percentage of those text annotated in the corresponding x-axis label by ChatGPT.

4 Results and Analysis

To evaluate the performance of ChatGPT, we compare ChatGPT's annotation against the original human annotations contained within each dataset. We treat the original human annotations as the gold standard that ChatGPT must predict. As such, we treat ChatGPT as a prediction engine, which we can then evaluate using traditional classifier metrics. We use weighted F1-score to evaluate ChatGPT's performance on data annotation. Given a dataset, a higher F1-score indicates that ChatGPT provides annotations *more* similar to humans.

4.1 Results Summary

Table 1 presents statistics for the ChatGPT annotation results. The annotated datasets contain 73,493 text items. ChatGPT annotates 71,484 (97.27%) items. For the remainder, 2% of responses do not match any candidate label in the dataset, and 0.73% of responses fail due to API errors. This confirms that ChatGPT can generate easily extractable annotation labels in desired format.

Table 2 presents the weighted recall, precision, and F1-score for ChatGPT's predictions for each dataset. For the seven datasets, ChatGPT achieves an average weighted F1-score of 72.00% (SD = 13.90%). This suggests that, *as a data annotator, ChatGPT has the potential to generate annotations similar to humans.* However, its performance varies across different domains. ChatGPT performs well on COVID-19 Fake News, Anti-LGBT Cyberbully, Clickbait Headlines, and Russo-Ukrainian Stance datasets (weighted F1-score > 75%). In contrast, ChatGPT performs poorly on Vaccine Stance, COVID-19 Hate Speech, and Social Bot datasets (weighted F1-score < 65%).

Dataset	**w-Recall**	**w-Precision**	**w-F1**	**Label**	**Recall**	**Precision**	**F1**
Vaccine Stance	59.81%	66.11%	59.17%	Anti-vaccine	39.65%	80.13%	53.05%
				Pro-vaccine	50.02%	74.03%	59.70%
				Neutral	80.99%	50.25%	62.02%
COVID-19 Hate Speech	52.24%	55.61%	51.88%	Hate speech	88.27%	48.08%	62.25%
				Counterhate speech	33.33%	31.79%	32.54%
				Neutral speech	67.19%	31.79%	56.03%
COVID-19 Fake News	83.75%	85.55%	83.43%	Real news	95.63%	78.18%	86.03%
				Fake news	70.70%	93.65%	80.57%
Social Bot	64.96%	65.33%	63.70%	Human	81.12%	64.16%	71.65%
				Bot	45.54%	66.75%	54.14%
Anti-LGBT Cyberbully	79.08%	87.17%	80.03%	Cyberbully	98.28%	58.43%	73.29%
				Not cyberbully	71.17%	99.02%	82.81%
Clickbait Headlines	89.66%	90.92%	89.56%	Clickbait	80.57%	98.17%	88.50%
				Not clickbait	98.53%	83.86%	90.60%
Russo-Ukrainian Stance	75.85%	79.83%	76.26%	Pro-Ukraine	69.35%	90.02%	78.34%
				Pro-Russia	86.91%	62.50%	72.71%

Table 2: A summary of ChatGPT's overall and label-wise annotation performance on selected datasets. The prefix "w-" notes that the measurements' calculations are weighted average by the number of human-annotation for the label.

We next explore how ChatGPT performs on different labels in each annotation dataset. Figure 1 presents the confusion matrices for the seven datasets. For each matrix, the y-axis refers to the ground truth human labels, and the x-axis refers to ChatGPT's labels. The value in a cell, row $\boldsymbol{i}$ and column $\boldsymbol{j}$, presents the proportion of text with label $\boldsymbol{i}$ that are annotated with label $\boldsymbol{j}$ by ChatGPT. For all annotation tasks, the proportions of correctly annotated tweets by ChatGPT vary per distinct label. In all, *five out of seven* datasets present such an imbalance, where a gap of more than 25% exists between labels' accuracy. These results suggest that *for a given annotation task, ChatGPT's accuracy varies heavily across different labels.*

4.2 Task Analysis and Implications

Next, we dive into each annotation task to present implications according to ChatGPT's performance. To assist the following analyses on ChatGPT's label-wise performance, we also present ChatGPT's recall, precision, and F1-score for each annotation label in Table 2.

Rank 1st: Clickbait Headlines. For Clickbait Headlines, ChatGPT's overall performance ranks *first* out of seven annotation tasks by a weighted F1-score of 89.56%. ChatGPT correctly identifies 89.66% clickbait or non-clickbait news headlines. This is the highest accuracy among the annotation datasets. In addition, the weighted precision rate of 90.92% suggests that ChatGPT can effectively annotate clickbait news headlines, while only introducing a small number of false positives. In addition, ChatGPT attains recall, precision, and F1-score exceeding 80% across all labels. Specifically, the corresponding confusion matrix (Figure 1(f)) shows that ChatGPT can correctly identify 80.6% of clickbait and 98.5% of non-clickbait headlines. In conclusion, *ChatGPT shows better overall performance on Clickbait Headlines than the other six annotation datasets.*

Rank 2nd: COVID-19 Fake News. For COVID-19 Fake News, ChatGPT's overall performance ranks *second* out of seven annotation tasks, with a weighted F1-score of 83.43%. ChatGPT correctly distinguish 83.75% of news as real or fake. However, ChatGPT's performance across different news varies when measured by recall and precision. For real news, ChatGPT achieves a high recall of 95.63%, but a lower precision of 78.18%. For fake news, ChatGPT only attains a recall of 70.7%, but a higher precision of 93.65%. Such a pattern is caused by the existence of many false positives of real news annotated by ChatGPT. Moreover, corresponding confusion matrix (Figure 1(c)) shows that, while ChatGPT can identify 95.6% real news, 29.3% fake news are misidentified as real. *These results suggest that, for COVID-19 Fake News, ChatGPT is able to annotate real news with high accuracy, but also has a tendency to misidentify fake news as real.* As fake news detection mainly relies on news content when training models, we conjecture *such a limitation is caused by a lack of comprehensive COVID-19 fake news data for training ChatGPT.*

Rank 3rd: Anti-LGBT Cyberbully. For Anti-LGBT Cyberbully annotation, ChatGPT's overall performance ranks *third* out of seven tasks, with a weighted F1-score of 80.03%. Overall, ChatGPT correctly annotates 79.08% posts. However, when annotating cyberbullying posts, ChatGPT achieves 98.28% recall, but only attain 58.43% precision. Such a pattern indicates the existence of false positives among cyberbullying posts annotated by ChatGPT. The confusion matrix supports this by showing that ChatGPT only correctly annotates 71.2% non-cyberbully posts and 28.8% non-cyberbullying posts are misidentified as cyberbully. To summarize, *ChatGPT is better at annotating cyberbully posts (using our prompt) when compared against other posts containing non-cyberbully expressions. However, ChatGPT often misclassifies non-cyberbullying posts as cyberbully.*

Rank 4th: Russo-Ukrainian Stance. For the Russo-Ukrainian Stance detection dataset, ChatGPT's overall performance ranks *fourth* out of seven tasks, with a weighted F1-score of 76.26%. Note, the invasion of Ukraine took place after the training date of our ChatGPT version. Overall, ChatGPT correctly annotates 75.85% of tweets' stance. We find that, for the pro-Ukraine tweets, ChatGPT reports a high precision of 90.02%, with a low recall of 69.35%. In contrast, for the pro-Russia tweets, ChatGPT reports a high recall of 86.91%, but with a low precision of 62.50%. This means that, when ChatGPT annotates pro-Ukraine tweets, it is usually correct (high precision), but the same is not true for annotating a tweet pro-Russia (low precision). The corresponding confusion matrix (Figure 1(e)) shows that 30.6% pro-Ukraine tweets are mis-annotated as pro-Russia by ChatGPT. In summary, *when ChatGPT labels a tweet as expressing a pro-Ukraine stance, it is usually correct.* Yet, this comes at the cost of a low recall rate. In contrast, *ChatGPT has more false positives when labeling tweets as pro-Russian, but does have a higher recall.*

Rank 5th: Social Bots. For the Social Bot dataset, ChatGPT's overall annotation performance ranks *fifth* out of seven annotation datasets, with a weighted F1-score of 63.70%. Overall, ChatGPT correctly annotates 64.96% of tweets as posted by humans or bots. For precision, ChatGPT only attains 64.16% for human tweets and 66.75% for bot tweets. This suggests ChatGPT's ability to reproduce precise annotations is still limited on social bot detection, i.e., ChatGPT often reports false positives for both human and bot tweets. For recall, ChatGPT attains 81.12% for human tweets but only 45.54% for bot tweets. Such a significant difference implies that, while ChatGPT manages to identify most human tweets, it often fails to distinguish bot tweets from human ones. The corresponding confusion matrix supports this by showing that 54.5% of bot tweets are mis-annotated, compared to only 18.9% human tweets. In summary, *ChatGPT has the potential to act as an annotator to distinguish most human content from bot content. Yet, ChatGPT's capability is limited when identifying content generated by social bots. In our case, ChatGPT often accidentally annotates social bot tweets as human-generated. As a result, its annotation for human tweets can involves lots of false positives.*

Rank 6th: Vaccine Stance. For the Vaccine Stance dataset, ChatGPT's overall performance ranks *sixth* out of seven datasets, with a weighted F1-score of 59.17%. Overall, ChatGPT correctly annotates 59.81% of tweets' stance towards COVID-19 vaccine. ChatGPT is more precise when detecting tweets expressing anti-vaccine (precision = 80.13%) and pro-vaccine stance (precision = 74.03%), compared to tweets with neutral expression (precision = 50.25%). In the meantime, ChatGPT's recall varies across these three labels. While ChatGPT achieves a high recall of 80.99% for neutral tweets, it attains only 39.65% for anti-vaccine tweets and 50.02% for pro-vaccine tweets. This indicates that ChatGPT is conservative when annotating tweets' stance towards COVID-19 vaccine and often mis-labels tweets' stance as neutral. According to the corresponding confusion matrix (Figure 1(a)), ChatGPT mis-annotates 57.5% of anti-vaccine tweets and 47.2% pro-vaccine tweets as neutral tweets. Therefore, *ChatGPT performs poorly in labeling tweets' stance on the COVID-19 vaccine. This is because many anti-vaccine or pro-vaccine tweets are mislabeled as neutral by ChatGPT.*

Rank 7th: COVID-19 Hate Speech. For the COVID-19 Hate Speech dataset, ChatGPT's performance ranks *seventh* out of seven, with a weighted F1-score of 51.88%. Overall, ChatGPT correctly annotates 52.24% of tweets as expressing or countering anti-Asia hate. Specifically, for hate speech, ChatGPT attains a F1-score of 62.25% with a high recall of 88.27%. For counterhate speech, ChatGPT attains a very low F1-score of 32.54%, with a recall of 33.33%. This suggests that ChatGPT's performance varies when annotating hate and counterhate content on COVID-19. Importantly, ChatGPT's precision for all three labels is lower than 50%, confirming its annotation for each label involves many false positives. As highlighted by the corresponding confusion matrix (Figure 1(b)), ChatGPT mis-labels 27.1% of neutral tweets as hate tweets, and 24.9% neutral tweets as counterhate tweets. Meanwhile, it mis-annotates 58.1% of counterhate tweets as neutral. In summary, *ChatGPT's annotations are inaccurate for the COVID-19 Hate Speech task. ChatGPT often mis-annotates neutral content as hate speech, and fails to distinguish counterhate speech from neutral content.*

5 Conclusion and Discussion

This study has investigated ChatGPT's potential to act as a data annotator for social computing data. We have compared ChatGPT's performance on seven datasets against the original human annotations. ChatGPT is often able to reproduce human labels, achieving an average F1-score of 72.00%. That said, we discover significant variations in its performance across different domains.

Our work has a number of limitations, which form the basis of our future work. *First*, this study only examines ChatGPT's annotation performance with a small number of datasets covering five social issues. We would like to further inspect ChatGPT's potential as a annotator on other domains (e.g., named-entity recognition), or on other social problems (e.g., political polarization). *Second*, we acknowledge that we only use a single prompt for annotation, and there are many variants that could be experimented with. Prompt design is a major theme of future work, which we believe we yield better results. We are keen to explore how state-of-art prompt-tuning methods, like few-shot prompting and Chain-of-Thought technique [21], can assist with prompt formulation.

References

1. Aiyappa, R., An, J., Kwak, H., Ahn, Y.Y.: Can we trust the evaluation on chatgpt? arXiv preprint arXiv:2303.12767 (2023)
2. Aydın, Ö., Karaarslan, E.: Openai chatgpt generated literature review: Digital twin in healthcare. Available at SSRN 4308687 (2022)
3. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al.: A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023 (2023)
4. Chakraborty, A., Paranjape, B., Kakarla, S., Ganguly, N.: Stop clickbait: Detecting and preventing clickbaits in online news media. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). pp. 9–16. IEEE (2016)
5. Díaz, M., Kivlichan, I., Rosen, R., Baker, D., Amironesei, R., Prabhakaran, V., Denton, E.: Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 2342–2351 (2022)
6. ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., Yang, D.: Latent hatred: A benchmark for understanding implicit hate speech. arXiv preprint arXiv:2109.05322 (2021)
7. Fagni, T., Falchi, F., Gambini, M., Martella, A., Tesconi, M.: Tweepfake: About detecting deepfake tweets. Plos one **16**(5), e0251415 (2021)
8. Glandt, K., Khanal, S., Li, Y., Caragea, D., Caragea, C.: Stance detection in covid-19 tweets. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Long Papers). vol. 1 (2021)
9. Haq, E.U., Lu, Y.K., Hui, P.: It's all relative! a method to counter human bias in crowdsourced stance detection of news articles **6**(CSCW2) (nov 2022). https://doi.org/10.1145/3555636, `https://doi.org/10.1145/3555636`

10. Haq, E.U., Tyson, G., Lee, L.H., Braud, T., Hui, P.: Twitter dataset for 2022 russo-ukrainian crisis. arXiv preprint arXiv:2203.02955 (2022)
11. He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., Kumar, S.: Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 90–94 (2021)
12. Huang, F., Kwak, H., An, J.: Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. arXiv preprint arXiv:2302.07736 (2023)
13. Kennedy, C.J., Bacon, G., Sahn, A., von Vacano, C.: Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. arXiv preprint arXiv:2009.10277 (2020)
14. Li, Y., Sosea, T., Sawant, A., Nair, A.J., Inkpen, D., Caragea, C.: P-stance: A large dataset for stance detection in political domain. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 2355–2365 (2021)
15. Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S.K., Goyal, P., Mukherjee, A.: Thou shalt not hate: Countering online hate speech. In: Proceedings of the international AAAI conference on web and social media. vol. 13, pp. 369–380 (2019)
16. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: Detecting stance in tweets. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). pp. 31–41 (2016)
17. Patwa, P., Sharma, S., PYKL, S., Guptha, V., Kumari, G., Akhtar, M.S., Ekbal, A., Das, A., Chakraborty, T.: Fighting an infodemic: Covid-19 fake news dataset (2020)
18. Peixian, Z., Ehsan-Ul, H., Yiming, Z., Pan, H., Gareth, T.: Echo chambers within the russo-ukrainian war: The role of bipartisan users. In: 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (2023). https://doi.org/10.1145/3625007.3627475
19. Poddar, S., Mondal, M., Misra, J., Ganguly, N., Ghosh, S.: Winds of change: Impact of covid-19 on vaccine-related opinions of twitter users. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 16, pp. 782–793 (2022)
20. Sallam, M., Salim, N.A., Ala'a, B., Barakat, M., Fayyad, D., Hallit, S., Harapan, H., Hallit, R., Mahafzah, A., Ala'a, B.: Chatgpt output regarding compulsory vaccination and covid-19 vaccine conspiracy: A descriptive study at the outset of a paradigm shift in online search for information. Cureus **15**(2) (2023)
21. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
22. Zhang, B., Ding, D., Jing, L.: How would stance detection techniques evolve after the launch of chatgpt? arXiv preprint arXiv:2212.14548 (2022)
23. Zhu, Y., Haq, E.u., Lee, L.H., Tyson, G., Hui, P.: A reddit dataset for the russo-ukrainian conflict in 2022. arXiv preprint arXiv:2206.05107 (2022)
24. Zhu, Y., Yin, Z., Tyson, G., Haq, E.U., Lee, L.H., Hui, P.: Apt-pipe: A prompt-tuning tool for social data annotation using chatgpt. In: Proceedings of the ACM on Web Conference 2024. pp. 245–255 (2024)