Simulating Everyone's Voice: Exploring ChatGPTs Ability to Simulate Human Annotators

ELMosts

Abdirizak Yussuf yussu037@umn.edu
Claire Chen chen6242@umn.edu
Dinesh Challa chall125@umn.edu
Venkata Sai Krishna abbar005@umn.edu

Abstract

We explores ChatGPT's ability to simulate human annotators through zero-shot prompts with demographic information for controversial topics. Human annotation is critical for NLP model building but has limitations, particularly in the context of controversial topics, which include potential biases from human annotators, and budget constraints for large-scale annotations. ChatGPT could be a potential alternative, but it is essential that it captures human biases accurately. To evaluate its annotation performance, we curated a data set of 50 controversial news article titles dealing with topics like abortion, immigration, politics, race, etc. We used metrics such as Inter-annotator reliability and continuous disagreement labels to evaluate ChatGPT's performance against human annotators and found that ChatGPT responses have a higher degree of agreement or lower disagreement across topics.

1 Introduction

The development of large language models, such as GPT-4, has revolutionized the field of natural language processing, making it possible to automate repetitive human tasks like proofreading, essay writing etc. GPT-4 has undergone various benchmarks and evaluations by its developers, but as NLP enthusiasts, we are curious to benchmark its performance against that of human annotators. If ChatGPT can replicate human annotators with high accuracy, it has the potential to capture a lot of diverse opinions and viewpoints in a relatively short amount of time. This would be particularly useful in cases where large amounts of text need to be annotated, such as in social media analysis, opinion mining, or content moderation. By understanding how these models compare to human annotators, we can identify ways to improve their performance and enhance their usefulness in a variety of applications.

A typical annotation pipeline starts with collecting raw data which is then annotated by multiple human annotators. Agreements between annotators are used to come up with a consensus annotation. This manual task requires a lot of money, human resources and time to come up with initial labels for the texts. Later, the large language models are trained on this data to automate the process. Lots of studies were conducted to see if humans can be replaced with language models in different fields a few are as follows.

A study by (Jungwirth and Haluza, 2023), experimented on the "text-davinci-003" model of GPT-3 by prompting questions on "how chatbots can be applied to health research?" and concluded that it can potentially contribute as a team member in public health research but needed further study on exploring it to full capabilities and examining the ethical aspects of it. In the paper (Spitale et al., 2023), GPT-3 was asked to generate tweets on topics like Covid-19, vaccination, theory of evolution etc. The results show that GPT-3 can produce accurate information that is easy to understand as well as producing compelling disinformation. The study also shows that humans cannot distinguish tweets generated by GPT-3 from tweets written by human users. In the paper (Lund and Wang, 2023), authors interview chatGPT on the impact it could have on academia and libraries. Their findings reveal that chatGPT has the power to advance in academia or librarianship. However, it also states that it should be used in a professional and ethical way rather than harming future professionals. These studies encourage us to understand the capabilities of the latest chatGPT model.

More recently, a study by (Gilardi et al., 2023), on chatGPT showed that it can outperform crowdworkers for twitter data text annotation tasks. The results from the study revealed an interesting insight where inter-annotator agreement was better in chatGPT, when prompted in zero-shot fashion,

than the humans. In the paper "Everyone's voice matters" by et.al. Wan (Wan et al., 2023)., tries to predict the level of disagreement on an input text after being trained on 5 datasets like SBIC, SocialChem etc. The main objective here is that the training set is annotated manually by 140 annotator opinions and their demographics information is also fed to the model to learn the person. The experiment proves that demographics like age, ethnicity and gender do affect an annotator's opinion and cause disagreement between groups on a topic.

The above studies focused on exploring chatGPT through zero-shot prompts. The limitation here is they ignored who annotated the ground truth labels i.e., the demographics of the human annotators were not considered. In our paper we plan to include human demographics like age, gender, political ideology, race, and religious beliefs and see if chatGPT can reflect them. By this approach it could be possible to conclude that chatGPT might not be able to reflect a specific group of humans on a certain topic rather than broadening the scope by saying chatGPT cannot mimic all the humans on a topic. This whole process can help us know for what type of texts chatGPT can be used to annotate the raw texts instead of humans. This study can be useful to all those who work on text data for which human annotation is essential be it in academia or in the industry.

2 Methods

The data flow for this project starts with 83,000 news article titles scraped from the St. Cloud Times Newspaper. We then performed zero-shot classification using bart-large-mnli with class labels {controversial, uncontroversial, other}. From the classified titles, we manually sampled 50 controversial titles (highly disagreeable instances). Human annotators were asked to assign a label to each title, based on their perspective. We also collected their demographic information. We prompted ChatGPT to simulate the perspectives of individuals given only their demographic information. To do this, we injected individual demographic information into ChatGPT prompts. We then prompted the model to perform the same annotation task as the human annotators. We use the disagreement metric from [everyone's voice matters] paper to compare the annotations produced by human annotators and ChatGPT personas.

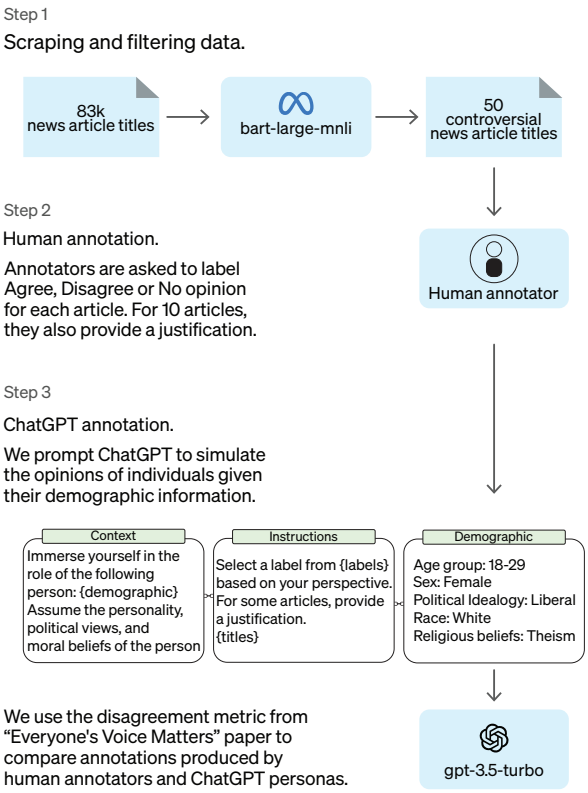

Figure 1: Pipeline for data flow

2.1 Inter-annotator Agreement

Inter-rater reliability is an important measure of agreement between different individuals. High inter-rater reliability indicates that multiple raters' answers for the same question are consistent, whereas low reliability indicates they are inconsistent. Various evaluation metrics, such as Cohen's Kappa statistics, Fleiss's Kappa, Scott's π, and Krippendorff's α, can be used to quantify inter-rater reliability. Krippendorff's alpha is particularly applicable, as it can be applied to all measurement levels, sample sizes, and incomplete data. For our project, we chose to use Krippendorff's α to measure the agreement among human annotators and ChatGPT annotators. Krippendorff's α (Krippendorff, 2011) is calculated by taking the ratio of observed weighted percent agreement p_a to weighted percent agreement expected by chance p_e (Eq.1). Specifically, p_a indicates how often the reviewers actually agreed, while p_e indicates the percent agreement the reviewers would achieve guessing randomly. As a result, Krippendorff's α ranges from -1 to 1, with a value of 1 indicating perfect agreement among the raters, 0 indicating agreement that is no better than chance, and -1 indicating systematic disagreement among the raters.

Since subjective annotations on controversial titles have no clear or objective answers, we are comparing the inter-rater reliability between humans and ChatGPT for each issue topic. By doing so, we can assess the consistency of the annotations provided by ChatGPT in comparison to those provided by human annotators, and identify any patterns or trends in the data that may be of interest.

$$\alpha = \frac{p_a - p_e}{1 - p_e} \tag{1}$$

2.2 Disagreement Analysis

The continuous disagreement label introduced by "Everyone's Voice Matters" (Wan et al., 2023) provides a measure of the degree of disagreement among annotators for a given text. A high disagreement label suggests that the title is very controversial, whereas a low label indicates less controversy. During the annotation process, we collected discrete annotations of "agree", "no opinion", and "disagree". To compute the continuous disagreement labels, we first calculate the majority voting result r_{ymaj}, which is the highest agreement rate r_k among the three annotations. The agreement rate is determined by Eq.2, where N is the number of annotators, x is the controversial title, and k is the annotation. Next, we define the continuous disagreement label as 1 minus the majority result (Eq.3). Thus, the continuous disagreement label ranges from 0 to 1, with 0 indicating perfect agreement among the raters and 1 indicating significant disagreement.

$$r_{ymaj} = argmax_k \, r_k(x) \tag{2}$$

$$r_k = \frac{1}{N} \sum_{i=1}^{N} 1[y_i(x) = k] \tag{3}$$

3 Results

Observations on the inter-annotator agreement for responses from human annotators and ChatGPT personas were tabulated in Table 1, using Krippendorff's alpha as the measure, across different controversial topics such as Abortion, Immigration, Social Issues, Political Issues, Racial Justice, and Religion. The results show that human annotators had a low agreement, with alpha values ranging from 0.017 to 0.22. In contrast, ChatGPT personas had higher alpha values, ranging from 0.36 to 0.50, indicating moderate agreement. The

Topic	Human Annotators	ChatGPT Personas
Abortion	0.22	0.32
Immigration	0.15	0.40
Social Issues	0.11	0.40
Political Issues	0.017	0.50
Racial Justice	0.19	0.40
Religion	0.18	0.36
All Topics Combined	**0.15**	**0.42**

Table 1: Krippendorff's alpha values for each annotator group

Krippendorff's alpha values for the entire data set were 0.15 for human annotators and 0.42 for ChatGPT personas. This indicates minimal agreement among human annotators, validating the controversial nature of the data set. In contrast, ChatGPT personas showed a higher level of consistency, with Krippendorff's alpha value of 0.42, indicating moderate agreement among them.

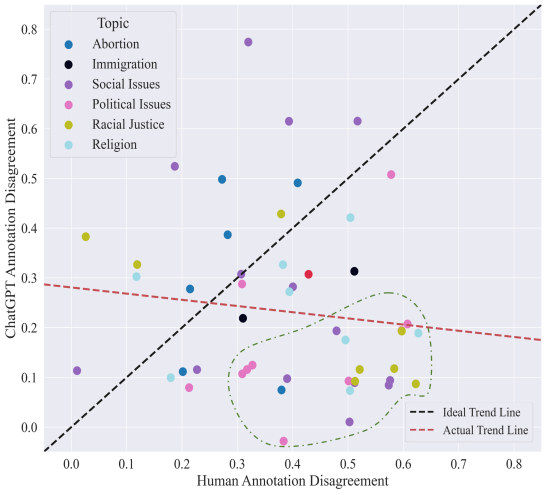

Figure 2: Disagreement label comparison

In addition to computing inter-annotator agreement, we also calculated a continuous disagreement label for both human and ChatGPT annotations for each item/title in the dataset. We illustrated our findings using a scatter plot in Figure 2, where data points on or around the diagonal line (in black) represent ideal scenarios where ChatGPT annotations replicate human annotations. However, we observed that the actual regression line (in red) is relatively flat, owing to certain data points (in green) corresponding to Social Issues, Political Issues, and Racial Justice. These data points have

Topic	F1 (%)	MSE
Abortion	68.25	0.028
Immigration	54.16	0.025
Social Issues	43.73	0.108
Political Issues	57.14	0.078
Racial Justice	69.44	0.147
Religion	59.52	0.061
All Topics Combined	56.77	0.084

Table 2: Evaluation Metrics

lower levels of disagreement compared to human annotations, causing a deviation in the actual regression line. The performance of ChatGPT annotations relative to human annotations with regard to disagreement is quantified using Mean Square Error (MSE), tabulated in Table 2, and it is observed that MSE corresponding to certain topics like Social Issues and Racial Justice are relatively higher compared to other topics.

Finally, the labeling process of the data set was carried out through a majority voting approach with the help of both human annotators and ChatGPT personas. The F1-score metric was used to evaluate the obtained labels, and the results showed that ChatGPT's performance was nominal with an F1-score of 56.77% for the entire data set. However, it is important to note that the F1-score varied across different topics, ranging from 43%-69%. Despite the nominal performance, the use of ChatGPT as a labeling tool can still be considered a useful and efficient approach for certain types of data sets.

4 Discussion

4.1 Limitations

This study has several limitations that should be considered when interpreting the results. First, the dataset used in this study was limited to only 50 titles. This was due to the nature of the text corpus, which consisted of news articles that did not cover a lot of controversial topics. As a result, the generalizability of the findings to other domains may be limited. Secondly, the human annotations for this study were obtained from only 10 individuals. This was due to time and budget constraints, which prevented us from recruiting a larger number of annotators. As a result, we may not have captured the full range of diverse opinions on the topic under investigation. Future studies could benefit from obtaining a larger and more diverse sample of annotators to enhance the validity and reliability of the results.

4.2 Datasets

The selection and annotation of 50 controversial news titles from St. Cloud Times, along with agreement annotations and short justifications provided by both humans and ChatGPT, can expand the choices available to researchers in the field of natural language processing (NLP). The unique domain of our dataset, as compared to other commonly used datasets such as Social Chemistry 101 or Social Bias Inference Corpus, offers researchers an opportunity to train and test NLP models on news titles that are specific to St. Cloud Times. In addition, with the record of annotators' demographic information, our dataset can be used to gain a better understanding of the specific issues or topics generating controversy among different communities, as well as to analyze how news titles can be made less offensive for some groups of people. While the number of annotations is limited due to time constraints, the creation of a high-quality annotated dataset can have a great impact on the direction and focus of future research and development projects in NLP, social, or psychology fields.

4.3 Ethics

Our project aims to explore whether ChatGPT can replace human annotators in the annotation process. While ChatGPT has the potential to reduce the cost and time required for collecting annotated data, which is beneficial for large-scale projects, there are some concerns when using it for complex issues that require subjective human judgment. One major limitation of using ChatGPT is the lack of diversity of perspectives and experiences represented in the annotated data, as the model is trained on a specific set of data and may not be able to accurately capture all the perspectives of underrepresented groups. Additionally, ChatGPT may produce unreliable annotations due to its inherent biases or lack of understanding of certain contexts or cultural issues. It is difficult to determine whether ChatGPT has enough and updated knowledge to understand new and emerging laws or incidents. Thus, it is crucial to carefully consider the trade-offs and potential limitations before deciding to use ChatGPT as a replacement for human annotators in the annotation process.

5 Conclusions

In this study, we curated a dataset of 50 controversial news article titles covering various topics such as abortion, immigration, politics, and race, among others. We presented a study that explores the ability of ChatGPT to simulate human annotators through zero-shot prompts with demographic information for controversial topics.

Our findings suggest that ChatGPT annotations perform well in capturing the aggregated labels, as evidenced by the moderate F1 scores. However, they perform poorly when response mixture is taken into consideration due to high IAA or lower disagreement compared to human annotations. It is important to note that this observation is based on a single experiment and should be interpreted with caution.

Furthermore, the experiment setup was constrained by time and resource limitations, which only allowed for 10 human annotations per title and a total of 50 titles. Future work could expand the scale and diversity of human annotators to obtain more authoritative results. Additionally, analyzing the demographics side of things from an analysis standpoint would be an interesting area to explore. Capturing human and ChatGPT justifications using a similarity metric would also be a valuable avenue for future research.In conclusion, this study contributes to our understanding of the capabilities and limitations of ChatGPT in simulating human annotators for controversial topics. Future research could build on these findings to enhance our understanding of how natural language processing models can be used in social science research.

Acknowledgments

We thank Dr. Dongyeop Kang and Risako Owan for their invaluable input, guidance, and support throughout the development of this project. We also thank the annotators who responded to our survey.

References

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks.

David Jungwirth and Daniela Haluza. 2023. Artificial intelligence and public health: An exploratory study. *International Journal of Environmental Research and Public Health*, 20(5).

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Brady D Lund and Ting Wang. 2023. Chatting about chatgpt: how may ai and gpt impact academia and libraries? *Library Hi Tech News*.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis) informs us better than humans. *arXiv preprint arXiv:2301.11924*.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information.