Evaluation of LLM-Based Justification on Downvoted Reddit Comments

Akhilan Gurumoorthy
UC San Diego

Chih-Lin Wang
UC San Diego

Jaewoong Yun
UC San Diego

1 Introduction

Reddit is one of the world's largest social media platforms continues to grow among millions of adolescents and young adults. It serves as a platform for a large variety of information, and its "subreddit" system ensures that all users can get the chance to either dive deeply into topics that they are quite familiar with, while also having the opportunity to converse with others about topics that they might not be quite familiar with. From current events to proven facts, Reddit serves as a source of information, and because of this, it is important that a system to maintain and indicate credibility is in place.

To determine the popularity of comments, Reddit implements one of its most notable features: its voting system. Through this system, users are able to "upvote" or "downvote" a post or comment. An upvote raises the score of a comment by 1, while a downvote lowers the score of a comment by 1. To many users, the amount of upvotes or downvotes evaluates the credibility of the post or comment, which can be effective against spam, trolls, or bot posts. In theory, comments that have a large number of upvotes are assumed to be certified by the community to be a reliable source of information. In contrast, comments with heavy downvotes can be disregarded, as, theoretically, the community has decided that its information is not valid or adds harm to the platform.

In practice, however, this system can be abused, especially in controversial subreddits; any user that posts a comment, regardless of whether that comment is true or false, will likely be downvoted in a subreddit that disagrees with that comment. In this situation, the comment is not being downvoted based on its credibility; instead, it is being downvoted due to other external factors. This can have multiple negative consequences, both for the platform and for its users. Commentors that have their comments unfairly may move away from the platform as a whole, and users who read the unfairly downvoted comments may dismiss them as incorrect. In this case, if the comments contain correct information and help the platform, the users may become unconsciously misinformed or may be drawn to write out comments that negatively impact the platform.

In our paper, we examine this issue by analyzing highly downvoted comments and determining whether or not they have been downvoted fairly. We examine comments in 13 different subreddits that span across a wide range of topics with varying levels of controversy. We attempt to implement automated content moderators through the use of several different LLM models. After determining the reliability of our automated content moderators, we will then be able to determine whether or not the most downvoted comments are downvoted fairly. In the end, if we found a significant portion of the most downvoted comments on Reddit to be legitimate yet downvoted unfairly, we can then conclude that mass downvoting of reasonable comments is indeed a phenomenon on Reddit that needs to be addressed. The results and scripts used in our project are publicly available at: https://github.com/jaewoongy/Sociotechnical-Cybersecurity-Code.

2 Background and Related Work

2.1 Content Moderation

One of our primary objectives is to implement an LLM as an automated content moderator. We also plan to feed the LLM information about the guidelines on the chosen subreddits to moderate. A similar paper has been published that implements these techniques. [1] Other work includes a study on Reddit user behavior that evaluates the effectiveness of subreddits that remove the "downvoting" feature. It states that one of the most common misconceptions is that downvoting posts can be a good content moderator. However, subreddits that do not allow downvoting show higher levels of quality and open-minded discussions, and have experienced a significant decrease in hostile, closed-minded arguments. [2]

[1] https://kumarde.com/papers/llm-contentmod.pdf
[2] https://arxiv.org/pdf/1705.02673

2.2 Hate Speech Detection

One of our biggest limitations in our implementation are the different semantic meanings that are carried behind harmful words. It could not be a toxic comment even though it may contain profanity or vulgarism, which could then be classified as a "fairly downvoted" comment. A related paper has completed a study on the importance of separating offensive language with actual hate speech. [3] There have also been evaluations using LLMs in detecting sexist or racist text using few-shot learning. [4]

3 Methodology

Subreddit	True	False
WorldNews	97	3
LegalAdvice	83	17
AskScience	60	40
NBA	98	2
IAmA	72	28
ScienceUncensored	86	14
Technology	69	31
Sports	87	13
Politics	84	16
UCSD	89	11
MarkMyWords	69	31
AskHistorians	82	18
AskTrumpSupporters	81	19

Table 1: Human-annotated Labels for each Subreddit

3.1 Research Plan

We have decided to divide our research plan into three steps. The first step involves data collection, the second step involves configuring the automated detector, and the final step involves data analysis. We chose this method of separation because we felt that it highlighted the three main aspects of our project. Our overall goal is to analyze Reddit comments through an automated detector. With this split, we are able to take the time to first focus on the comments, then the detector, and finally the analysis. The first task in our research plan is to get the most downvoted comments from multiple subreddits. We tried to find a mix of controversial subreddits and general topic subreddits, as this variety could help verify the robustness of our detector. We then needed to find the most downvoted comments of these subreddits. To do this, we needed to access some version of a Reddit API that would allow us to retrieve the comments of the respective subreddits. Using the API query, we can create a dataframe consisting of the comment and labels that we will annotate. Using this dataframe, we can combine it with the our models' predicted labels and rules as columns for final evaluation.

The second task is to configure several LLMs to serve as an automated content moderator. Using GLHF.chat, we can leverage several high-performance models through api queries. We chose the three largest models from the site: Meta-Llama-3.1-405B-Instruct, Qwen2.5-72B-Instruct, and Llama-3.3-70B-Instruct. Using these fine-tuned models, we prompted each model to serve as a content moderator analyzing comments for rule violations. In our prompt, we also specified each LLM to analyze 20 comments in batches and judge them based on each specified subreddit rules. Then, the LLM should return a JSON object containing a key of "comment number" and corresponding items "unfairly downvoted" and "rule violated". "Unfairly downvoted" represents a boolean value where "True" represents the comment was unfairly downvoted, and "False" represents that the comment broke one of the subreddit rules. "Rule violated" represents the rule which was broken if the comment broke a rule. Finally, we create a dataframe combining the two items as columns, this allows us to create our metrics and our visualizations.

Our goal is to decide whether the downvoted comments are done so in a "fair" or "unfair" manner. We were advised to come up with a fairly detailed description of what constitutes an unfairly downvoted comment. To address the issue of different subreddits having different topics, we decided to utilize the subreddit's guidelines to dictate the definition of fairness. Our definition of fairness will be based on whether or not the comments meet the guidelines of the subreddit. Fairly downvoted comments are those that have content that breaks or nearly breaks the guidelines of the subreddit, and unfairly downvoted comments are those that have contents that very clearly stays within the guidelines of the subreddit. This definition allows us to implicitly use expert opinions in our definitions of fairness, which should provide a more accurate representation of the subreddits we are examining. We feed this definition of fairness into our LLMs to each serve as an automated content moderator to find whether or not comments were downvoted fairly.

The final task is to perform statistical analysis on the automated content moderators. We go through the most downvoted comments and manually label them as fairly or unfairly downvoted. We utilize the same guideline-based criteria that was used to configure the LLMs. Using these manually labeled comments, we compare the performance among the different GPT-based automated detection systems. This allows us to determine the effectiveness between each automated detection system and within each subreddit. In terms of statistics, we plan to report the accuracy rate of the system for each subreddit. We want to place more importance on ensuring that the model correctly detects unfairly downvoted comments, so we want to maximize our recall (true positive rate). Using both statistics, we can determine both the effectiveness of the automated content moderator and the prevalence of unfairly downvoted comments on Reddit.

[3] https://arxiv.org/abs/1703.04009
[4] https://arxiv.org/abs/2103.12407

Subreddit	Llama-3.3-70B		Qwen2.5-72B		Llama-3.1-405B		Subreddit Average	
	Acc.	Recall	Acc.	Recall	Acc.	Recall	Acc.	Recall
AskHistorians	0.732	0.747	0.257	0.145	0.832	**0.976**	0.607	**0.623**
Technology	0.897	**0.985**	0.564	0.657	0.802	0.929	0.754	**0.857**
MarkMyWords	0.866	0.909	0.356	0.300	0.713	0.914	0.645	0.708
AskTrumpSupporters	0.705	0.718	0.545	0.511	0.743	0.878	0.664	0.702
IAmA	0.749	0.794	0.713	0.699	0.822	0.877	0.761	0.790
ScienceUncensored	0.786	0.831	0.347	0.345	0.822	0.874	0.652	0.683
WorldNews	0.762	0.765	0.347	0.327	0.822	0.827	0.644	0.640
Politics	0.762	0.831	0.594	0.600	0.743	0.812	0.700	0.748
UCSD	0.806	0.838	0.614	0.578	0.792	0.811	0.737	0.742
Sports	0.705	0.691	0.614	0.625	0.733	0.795	0.684	0.704
NBA	0.921	0.919	0.792	0.788	0.743	0.737	0.819	**0.815**
LegalAdvice	0.703	0.726	0.515	0.500	0.693	0.726	0.637	0.651
AskScience	0.505	0.623	0.554	0.656	0.505	0.557	0.521	**0.612**
Model Average	0.761	0.798	0.524	0.518	0.751	**0.824**	0.679	0.713

Table 2: Comparison of Accuracy and Recall across Models for Different Subreddits

The subreddits that we have tested are: r/ScienceUncensored, r/AskTrumpSupporters, r/politics, r/UCSD, r/legaladvice, r/NBA, r/worldnews, r/askscience, r/AskHistorians, r/sports, r/IAmA, r/MarkMyWords, r/technology. These subreddits allow us to access a variety of topics, such as politics, sports, technology, and science. Furthermore, we have several subreddits that can be viewed as controversial, which allows us to examine both general topic subreddits and controversial subreddits.

4 Results

Using each of the three models as predictors, accuracy and recall were calculated for each subreddit. The accuracy may provide some information about the performance of each LLM. However, because our primary research objective is to identify how effective our model is in identifying unfairly downvoted comments, we must also consider the recall. In this case, the recall represents the proportion of unfairly downvoted comments that the LLM correctly predicted as unfairly downvoted.

As shown in Table 2 we see a significant performance difference among the models. The Qwen2.5-72B-Instruct model performed very poorly in subreddits like AskHistorians, WorldNews, and ScienceUncensored. The case for its poor performance could be a result of its lower capable context window compared to the Llama models. Feeding the LLM in chunks of 20 comments per query for some subreddits might have caused the LLM to struggle comprehending both the sophisticated prompt and the inputs, resulting in labeling most subreddit comments as false. However, it performed fairly well in subreddits like IAmA and NBA with recalls of 0.7 and 0.788, respectively. In these two subreddits, we noticed it is fairly easy to determine whether a comment is unfairly downvoted or against subreddit rules. For example, in the IAmA subreddit, comments are usually directed to the original poster aiming to answer questions. Comments that are not directly "attacking" the poster yet downvoted tend to be "unfairly downvoted" - these comments tend to be a redundant or off-topic comment or question. Nevertheless, the model achieved an average of an overall accuracy of 0.505 and recall of 0.518, which is only slightly better performing than a randomly predicting algorithm.

The second-best performing model in terms of recall was the Llama3.3-70B-Instruct model. In some subreddits like Technology, NBA, and AskScience, the model outperformed the Llama3.1-405B model when correctly predicting the positive true labels. The model also achieved the highest overall accuracy. While being significantly smaller, the Llama 3.3 70B model is a newer version model with more optimized for instruction following and comprehension. Therefore, it is expected to perform only slightly worse than the 400B model. Surprisingly, while it only predicted 68% of total comments in the technology subreddit as unfairly downvoted, it achieved a near 100% recall. The technology subreddit does not specify rules for comments that are wrong about science or misinformative, signifying that the model does well in labeling non-toxic comments. On the other hand, a similar subreddit, AskScience enforces rules against misinformation, which the model failed to perform well on.

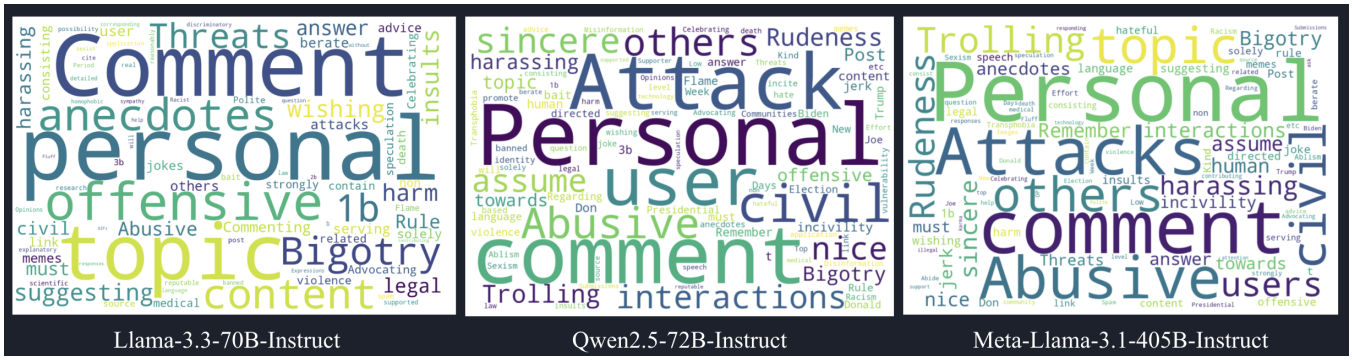

Figure 1: Word Cloud of LLM-predicted Rules Violated

Subreddit	Llama-3.3-70B	Qwen2.5-72B	Llama-3.1-405B	Subreddit Average
AskHistorians	81.4%	35.0%	88.8%	68.4%
Technology	68.0%	49.2%	83.2%	66.8%
MarkMyWords	69.0%	39.2%	72.4%	60.2%
AskTrumpSupporters	67.2%	44.7%	79.9%	63.9%
IAmA	54.4%	51.4%	62.0%	55.9%
ScienceUncensored	78.8%	44.0%	78.0%	66.9%
WorldNews	76.2%	35.9%	74.4%	62.2%
Politics	71.4%	52.7%	73.0%	65.7%
UCSD	69.9%	51.0%	82.0%	67.6%
Sports	59.2%	54.1%	71.4%	61.6%
NBA	82.7%	59.3%	74.7%	72.2%
LegalAdvice	60.2%	39.4%	53.3%	51.0%
AskScience	65.8%	46.9%	65.6%	59.4%
Model Average	69.6%	46.4%	73.7%	Total Average: 63.2%

Table 3: Percentage of Unfairly Downvoted Comments by Model and Subreddit

The best performing model for correctly classifying unfairly downvoted comments is the Llama3.1-405B model. The model achieved a recall of at least 0.8 for 9 out of the 13 subreddits, which is a significant improvement over the 70B model, which only achieved a recall of at least 0.8 for 6 out of the 13. It performed extremely well in the AskHistorians subreddit with a recall score of 0.976. The model performed poorly on only one subreddit, the AskScience subreddit, which states rules against misinformation. Nevertheless, for sentiment and contextual comments, the model performs exceptionally well.

Ultimately, the results have largely varied among the 13 subreddits, with the best performing model correctly classifying subreddits with clear guidelines on comments. Either Llama models are appropriate for this task as they achieved a reasonable accuracy and recall score for each subreddit, despite taking in hundreds of tokens of context for each prompt.

One of the biggest challenges of evaluating these results is when determining misinformative comments and classifying them as against some subreddit rules. It seems none of the models are capable of identifying misinformation especially in the AskScience subreddit, which highlights the complexity of evaluating these subreddits. Human annotators are also a significant bias in the evaluation process, as we need to fact-check the comment using online sources or from other replies to the comment.

As shown in Figure 1, each model produced a word cloud for all rules violated for all subreddits. Notably, most downvoted comments that are against subreddit policy tend to be personal/offensive attacks to other commentors in the subreddit. However, there are some words that represent "memes", "jokes" or "[off] topic". These represent the subreddits that have more serious rules in their subreddit, limited not just to un-related comments, but also misinformative comments that the LLM is not able to accurately predict. The subreddits with these rules include LegalAdvice, AskScience, and WorldNews, all of which scored the lowest average recall across all subreddits.

Lastly, Table 3 allows us to determine how many comments are classified as unfairly downvoted for each subreddit. A significant portion can be accredited to more lenient subreddits that have a lot of downvoted comments that are not against the rules, such as the AskHistorians, UCSD, and Technology subreddits. Meanwhile, in the LegalAdvice and AskScience subreddits, more stricter rules are posted, yet not much is done in enforcing these rules.

5 Discussion

5.1 Conclusions

Through our results, we can draw some conclusions about both of our research questions. We conclude that the predictor based on the Llama-3.1-405B model is our most effective predictor. It has a 75% accuracy rate, and more importantly, it has a recall score of 0.824. This shows that our predictor is working fairly well on a dataset that it has not seen before, and as a result, we feel that it is reliable enough to allow us to draw at least general conclusions about our main research question.

However, the evaluation of LLMs ultimately depend on how specific the rules of each subreddits are. Lesser defined rules would cause a difference in viewpoints for both the annotator and the LLM, which most likely resulted in an ineffective prediction for some subreddits. In addition, the evaluation also depends on how strict the moderation team is in removing comments for each subreddit. One probable outcome in subreddits with lower percentages of downvoted comments could be accredited to the lower amount of comments that are removed for breaking the rules.

Furthermore, we are able to conclude that a very large percentage of the downvoted comments are unfairly downvoted. When examining all 13 subreddits while using all three predictive models, 63.2% of the examined comments were judged to be unfairly downvoted. When looking at the Llama-3.1-405B based model, this percentage jumps even higher to 73.7%, with some subreddits showing as high as 88% of downvoted comments being done so in an unfair manner. All subreddits using this predictive model showed that over half of their heavily downvoted comments were done so unfairly.

Finally, we looked at the specifics of the LLM results to attempt to determine the reason that these fairly downvoted comments were downvoted. By examining the word cloud generated by the three LLM predictors, we see that the words "personal" and "attack" are generated the most. Therefore, it seems that the most common reason for comments being downvoted fairly is to limit personal attacks, which is a rule that seems to be present in most, if not all, of the subreddits that were examined in this project.

5.2 Future Considerations

5.2.1 Larger Dataset

In the future, we feel that it would be better to have a larger dataset. The dataset that we generated was quite large, with over 16,000 comments, but this is an incredibly small portion of the full set of comments on Reddit. Furthermore, even though we attempted to choose subreddits that we felt represented a wide variety of topics and varying levels of controversy, the 13 subreddits that we examined in this project may not serve as the most accurate representation of the platform as a whole, and a larger dataset across more subreddits could do this better.

5.2.2 Better LLM Predictor

While our LLM predictor had a very good accuracy and recall rate, there is still quite a bit of room for improvement. Ideally, our predictor is able to accurately analyze the fairness of the comment's downvote over 90% of the time, and we were only able to do 75% of the time. Semantic analysis is a growing field, and using newer models could allow us to generate a better predictor, which could in turn lend more credibility to the conclusions that we derive from this project. In addition, further worth with prompting could significantly increase the results of the paper, as demonstrated in several papers mentioned in the related works section.

5.2.3 Increased Labeling Consistency

In the process of annotating the comments to test the LLM predictor, we split up the annotation among the three authors. Because of this, there is inherent bias in how each author determines the fairness of the comment's downvote. In the future, we would like to find a way to eliminate this bias from the annotation process. Unfortunately, because of the sheer size of the dataset, we are not sure how feasible this is, but if this bias is able to be eliminated, we could generate more credible conclusions about our predictor's reliability.

5.2.4 Adding More Context to Comments

One of the biggest challenges in our project was identifying what exactly was meant by the comment. To do this, we needed to look through lesser-implied comments in the reddit post and identify who the comment was referring to and what comment it was replying to. However, because of a higher complexity in the prompting, this method was not implemented. By implementing a chaining-technique to incorporate all the comments in one branch, the LLM could have a more accurate prediction of whether the comment is unfairly downvoted or not. It could also prove quite effective, especially for misinformative comments, because replies to the comment may question the credibility of the evaluated comment.

5.2.5 Identifying Misinformative Comments

Our models performed the worst in subreddits that had explicit rules against misinformation and disinformation. However, it is not only incredibly difficult for the LLM to predict whether a comment is misinformative, but it's also very tedious for the human annotator to fact-check these comments. Ultimately, we are unsure of whether our annotated comments were truly misinformative. By using a fact-checking tool as a helper, we can more easily classify comments in stricter subreddits.

References

Chiu, K.-L., Collins, A., Alexander, R. (2022, March 24). Detecting hate speech with GPT-3. arXiv.org. https://arxiv.org/abs/2103.12407

Davidson, T., Warmsley, D., Macy, M., Weber, I. (2017, March 11). Automated hate speech detection and the problem of offensive language. arXiv.org. https://arxiv.org/abs/1703.04009

Horne, B. D., Adali, S., Sikdar, S. (2017, May 7). Identifying the social signals that drive online discussions: A case study of reddit communities. arXiv.org. https://arxiv.org/abs/1705.02673

Kumar, D., AbuHashem, Y., Durumeric, Z. (2024, January 17). Watch your language: Investigating content moderation with large language models. https://kumarde.com/papers/llm-contentmod.pdf