Large-Scale Quantitative Evaluation of Dialogue Agents' Response Strategies against Offensive Users

Haojun Li, Dilara Soylu, Christopher D. Manning
Department of Computer Science
Stanford University
{haojun,soylu,manning}@stanford.edu

Abstract

As voice assistants and dialogue agents grow in popularity, so does the abuse they receive. We conducted a large-scale quantitative evaluation of the effectiveness of 4 response types (`avoidance`, `why`, `empathetic`, and `counter`), and 2 additional factors (using a `redirect` or a voluntarily provided `name`) that have not been tested by prior work. We measured their direct effectiveness on real users in-the-wild by the re-offense ratio, length of conversation after the initial response, and number of turns until the next re-offense. Our experiments confirm prior lab studies in showing that `empathetic` responses perform better than generic `avoidance` responses as well as `counter` responses. We show that dialogue agents should almost always guide offensive users to a new topic through the use of redirects and use the user's name if provided. As compared to a baseline avoidance strategy employed by commercial agents, our best strategy is able to reduce the re-offense ratio from 92% to 43%.

1 Introduction

Conversational bots are increasingly popular among the general population which is correlated with an increase in bot abuse (Cercas Curry and Rieser, 2018). Analysis of the chat logs of an Alexa Prize[1] competition social bot shows that more than 10% of the conversations contain some level of offensiveness. Recently, researchers begin to measure the appropriateness of virtual agent responses to abuse. However, prior work either use self-reported scales of emotions by non-anonymous volunteers (Chin et al., 2020) or perceived quality of the conversation from crowd workers (Cercas Curry and Rieser, 2019). However, these qualitative metrics only measure the appropriateness of the response rather than the actual effect of the responses in a real conversation. Unlike the participants recruited for controlled lab studies or crowdsourced studies, real users abuse agents voluntarily, anonymously, and repeatedly.

To address these limitations, we conducted a large scale study similar to Cohn et al. (2019) to quantitatively measure the effectiveness of response strategies. As opposed to Cohn et al. (2019) which uses user ratings as the evaluation metric, we measured 1) the re-offense ratio; 2) the number of turns until the next offense; 3) the number of turns until the end of the conversation after the initial response. These metrics measure offensive behavior directly as opposed to user ratings which measures the quality of conversations as a whole. We show that using a redirection is significantly better than not using one, and using empathetic responses and user names is also effective at mitigating abuse, but only in combination with a redirection.

2 Related Work

There's a large body of research on physical agent abuse (Bartneck et al., 2005, 2007), particularly by children (Brščić et al., 2015; Nomura et al., 2016; Tan et al., 2018; Gallego Pérez et al., 2019; Yamada et al., 2020). There has also been much work on understanding the reason behind bot abuse (Angeli and Carpenter, 2005; Angeli, 2006; Brahnam, 2006). More recently, Cercas Curry and Rieser (2019) found that "polite refusal" responses are the most appropriate compared to many other responses by commercial bots. Similarly, Chin and Yi (2019); Chin et al. (2020) further evaluated the effectiveness of empathetic and counter-attacking response strategies by measuring their impact on cultivating emotions that are known to reduce ag-

[1]The Alexa Prize is a competition organized by Amazon Science to advance Conversational Artificial Intelligence, allowing university teams to develop conversational bots and get feedback from real users.

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 556–561
July 29–31, 2021.

Strategy	Description	Example Script
AVOIDANCE	The bot politely avoids talking about the offensive topic.	*I'd rather not talk about that.*
AVOIDANCE + REDIRECT	Same as AVOIDANCE, but the bot also gives a REDIRECT to change the topic.	*I'd rather not talk about that. So, who's your favorite musician?*
AVOIDANCE + NAME	Same as AVOIDANCE, but the bot also appends the user's name at the end of its utterance.	*I'd rather not talk about that, Peter.*
AVOIDANCE + NAME + REDIRECT	Same as AVOIDANCE + NAME, but the bot also gives a REDIRECT to change the topic.	*I'd rather not talk about that, Peter. So, who's your favorite musician?*
WHY	The bot asks the user why they made an offensive utterance.	*Why did you say that?*
WHY + NAME	Same as WHY, but the bot also appends the user's name at the end of its WHY utterance.	*Why did you say that, Peter?*
COUNTER + REDIRECT	The bot points out the inappropriate nature of the user utterance to the user, similar to Gallego Pérez et al. (2019).	*That is a very suggestive thing to say. I don't think we should be talking about that. Let's move on. So, who's your favorite musician?*
EMPATHETIC + REDIRECT	The bot empathizes with the user's desire to talk about inappropriate topics, and attempts to move on to a different topic.	*If I could talk about it I would, but I really can't. Sorry to disappoint. So, who's your favorite musician?*

Table 1: Response strategies we tested along with their descriptions and example scripts.

gression. Contrary to these end-of-conversation responses, strategies employed by human call center agents reviewed by Brahnam (2005) found that actively redirecting the conversation is more effective at mitigating on-going offenses than passively ignoring the offensive behavior, a factor not yet examined by prior work. Inspired further by Chen and Williams (2020), who showed that user engagement is improved when robots refer to users with their names, and Suler (2004), who showed that anonymity may expose bad user behaviors, we investigate whether using users' voluntarily provided names would also mitigate offensive behavior. Finally, informed by prior research showing the use of contemplation in improving children's learning (Shapiro et al., 2014), we test the hypothesis that a response strategy inviting the offensive users to reflect on why they made an offensive remark can reduce offensiveness.

3 Hypotheses

We test 4 hypotheses in our work:

1. **REDIRECT** Informed by Brahnam (2005), we hypothesize that using an explicit redirection when responding to an offensive user utterance is more effective than not using one as doing so actively redirects the user to a different discussion topic.

2. **NAME** Informed by Suler (2004) and Chen and Williams (2020), we hypothesize that including the user's name in the bot's response is more effective than not including it as doing so increases engagement with the user and provides a sense of identification.

3. **WHY** Informed by Shapiro et al. (2014), we hypothesize that asking the user the reason why they made an offensive remark would invite them to reflect on their behavior, and help reduce future offenses.

4. **EMPATHETIC & COUNTER** Informed by Chin et al. (2020), we hypothesize that empathetic responses are more effective in mitigating agent abuse than plain avoidance, while counter responses make no difference.

In order to test these hypotheses as well as interactions between different factors influencing the effectiveness of the response strategies, we cross multiple conditions with each other. Full description can be found at table 1.

Response Strategy	Sample Size	Re-offense	CI	Next	CI	End	CI
AVOIDANCE	1724	0.918	±0.0066	1.01	±0.0056	1.08	±0.2
AVOIDANCE+NAME	867	0.938	±0.0082	1.02	±0.017	1.11	±0.26
AVOIDANCE+NAME+REDIRECT	860	0.406	±0.017	8.6	±0.81	16.3	±0.98
AVOIDANCE+REDIRECT	1759	0.466	±0.012	7.32	±0.43	13.5	±0.58
COUNTER+REDIRECT	1859	0.471	±0.012	6.83	±0.41	12.3	±0.62
EMPATHETIC+REDIRECT	1814	0.432	±0.012	6.72	±0.37	13.1	±0.56
WHY	1755	0.952	±0.0051	1.05	±0.031	1.09	±0.33
WHY+NAME	836	0.947	±0.0077	1.33	±0.32	2.41	±1.53

Table 2: Response strategies and their measurements and confidence intervals (CI). Notice that sample size for strategies using user's name is significantly smaller than other strategies. This is because we can only select those strategies when the user volunteered a name.

	Base	Alternative	ΔRe-offense	ΔEnd	ΔNext
1	AVOIDANCE	AVOIDANCE+REDIRECT	**-0.452**†	**12.421***	**6.311**‡
2	AVOIDANCE+NAME	AVOIDANCE+NAME+REDIRECT	**-0.532**†	**15.202***	**7.584**‡
3	AVOIDANCE+REDIRECT	AVOIDANCE+NAME+REDIRECT	**-0.060**	**2.814**	1.281
4	AVOIDANCE	AVOIDANCE+NAME	0.020	0.033	0.007
5	WHY	WHY+NAME	-0.004	1.315	0.288
6	AVOIDANCE+NAME	WHY+NAME	0.010	1.298	0.316
7	AVOIDANCE	WHY	**0.033**	0.016	0.035
8	AVOIDANCE+REDIRECT	COUNTER+REDIRECT	0.005	-1.162	-0.486
9	AVOIDANCE+REDIRECT	EMPATHETIC+REDIRECT	*-0.035*	-0.373	-0.603

Table 3: Differences of metrics between pairs of strategies. Very Significant results ($p < 0.005$, stricter than p-value adjusted for Bonferroni correction 0.0125) are noted in bold. Significant results ($p < 0.05$) are italicized. † Odds Ratio p-value < 0.005. ‡ Cohen's d value > 0.8. *Cohen's d value > 0.7

4 Data Collection

We built our experiments into a custom open-domain conversational chatbot developed as part of the Alexa Prize competition. During the competition, Alexa users can invoke a competition bot by saying *"alexa lets chat"* or just *"lets chat"* to an Alexa-enabled device, after which Alexa hands off the conversation to a randomly assigned competition bot.

4.1 Stage 1: Offensiveness Detection

Before we test response strategies, we need to describe what counts as "Offensive User Behavior". Defining clear boundaries for offensive speech is a challenging task (Chen et al., 2012; Xiang et al., 2012; Khatri et al., 2018). As a practical way forward, we first classified user utterances by whether they contain any of the offensive phrases listed in the *"Offensive/Profane Word List"* shared by Dr. Luis von Ahn's research group at Carnegie Mellon University.[2] After around a month of collection (about 6000 conversations), we hand-selected the 500 most common overtly offensive user utterances. To increase recall, we built regexes that catch utterances that end in these 500 offensive phrases (such as "i want to talk about ***") and only trigger our experiments (described later) when these utterances or regexes are detected. To verify the efficacy of this regex classifier, we separately sampled 500 utterances from the first round of collection and manually labeled them for overt offensiveness. We found that this simple classifier achieves 64.4% recall and 91.7% precision, which is intended since we would like to trigger our experiments with very high precision. However, during our evaluation in section 6, we used a different offensive classifier that looks for utterances containing any offensive phrases which achieved 100% recall and 82.6% precision. This is also intended since it is better to over-classify offensive behavior during our evalua-

[2]Data can be found at https://www.cs.cmu.edu/~biglou/resources/.

tion to be conservative.

4.2 Stage 2: Response Experiments

We conducted our experiments from May 23, 2020 to August 23, 2020, during which we collected a total of 13276 offensive conversations with a total of 49511 categorized offensive utterances.[3] After detecting an offensive utterance and depending on whether the user offered a name in the beginning, the bot selects a strategy from table 1 for the entire conversation, and then randomly selects a response from a set of scripted responses for that strategy. We will also make a dataset containing attributes (i.e. the offensiveness) of each utterance of each conversation, a notebook to reproduce our results, as well as a csv of all of the bot's actual responses available on GitHub: https://github.com/LithiumH/offensive.

5 Proposed Metrics

We propose 3 metrics that directly measure strategy effectiveness from conversation logs. The first metric is the re-offense ratio (a.k.a. Re-offense), measured as the number of conversations that contained another offensive utterance after the initial bot response over the total number of conversations that used the same strategy. Intuitively, the responses leading to a smaller number of re-offenses more are effective at reducing user abuse. We also measure the length of the conversation after the response assuming there are no more re-offenses (a.k.a. End) to understand *how* a strategy stopped abuse. When the strategy is unable to stop re-offense, we are interested to know how many turns passed until the user offended again (a.k.a. Next). We believe that strategies that are able to delay offense longer are more effective at mitigating user abuse.

6 Hypothesis Testing and Discussion

All the metrics measured are shown in table 2. To test the hypotheses laid out in section 3, we run several pair-wise one-way T-tests on different strategies and different metrics in table 3.

6.1 H1: REDIRECT

Rows 1 and 2 in Table 3 show that, controlling the base strategy and whether the bot includes the user's name in its response, using a redirection gives a large, statistically significant improvement over not using one, halving the re-offense rate. Statistically significant differences in the End metric in table 2 and 3 show that when the user stopped their abusive behavior, `REDIRECT` is able to prolong a non-offensive conversation effectively on average while `no REDIRECT` stopped the conversation immediately. Similar differences can also be seen in the Next metric, which shows that offensive users almost always immediately re-offend without a `REDIRECT`, but delay their re-offense when given a redirection.

This suggests that *active avoidance is better than passive avoidance* and that social bots should always make an attempt to actively redirect the course of the conversation when facing an offensive remark.

6.2 H2: NAME

Though the effect sizes are small, rows 3, 4, and 5 of Table 3 show that including a user's name in the response is only effective when used together with a REDIRECT. This suggests that *including a user's name does not discourage re-offense by itself, but rather encourages the user to follow the new direction that the bot proposes*. It can be further corroborated by the statistically significant increase in the End metric, which shows an increase in the average number of non-offensive turns until the end of the conversation.

6.3 H3: WHY

Rows 6 and 7 of table 3 suggest that using the `WHY` strategy yielded a significant 3% increase in the re-offense ratio. Contrary to our belief that users will give an honest answer and reflect on their actions, asking why invites the users to repeat their abuse. Qualitative analysis of users' responses to the `why` question yields similar conclusions. This further supports section 6.1 that it is much better to quickly move on to a new topic than dwell on the current abuse. However, the effect sizes are small, which suggests that the main contributor for re-offense behavior is still the absence of a redirection.

6.4 H4: EMPATHETIC & COUNTER

Table 3 rows 8 and 9 suggest a 3.5% statistically significant reduction[4] in the re-offense ratio when

[3] More than half of the offensive user utterances are sexual in nature, potentially due to the fact that Alexa has a female voice by default. Similar observations were made previously (Cercas Curry and Rieser, 2018)

[4] Not adjusted for Bonferroni correction; more data is needed to fully justify this significance. We will leave this to followup work.

using the EMPATHETIC strategy together with a REDIRECT. There do not seem to be any significant differences between AVOIDANCE strategies and COUNTER strategies. We thus validated the conclusion drawn in prior research (Chin et al., 2020) in the wild.

7 Future Directions

The main limitation of our work was keeping customer satisfaction in mind when designing our experiments under Alexa Prize competition rules. This prevented us from replicating strategies such as joke strategies mentioned in Cercas Curry and Rieser (2019) and parenting strategies such as love-withdrawal as mentioned in Gallego Pérez et al. (2019). We were similarly unable to test the effectiveness of de-anonymization and peer-listening strategies similar to Tan et al. (2018) that would test how would the users respond if they were told that their conversations were not anonymous/private. It would also be useful to gather metadata about our participants such as age and gender (while maintaining anonymity). However, this is not allowed under Alexa Prize competition rules.

8 Ethical Concerns

Despite the empirical effectiveness of the AVOIDANCE + REDIRECT strategy as detailed in this work, we would like to remind researchers of the societal dangers of adopting similar strategies. Alexa has a default female voice and the majority of offensive responses we receive are sexual in nature as stated before. As pointed out by prior work (Cercas Curry and Rieser, 2019; West et al., 2019; Cercas Curry et al., 2020), inappropriate responses further gender stereotypes and set unreasonable expectations of how women would react to verbal abuse. Without pointing out the inappropriateness of user offenses, these response strategies could cause users to believe their offenses will go unnoticed in the real world as well. Thus, we urge researchers to consider the greater impact of deploying such strategies in voice-based dialogue agents beyond the proposed effectiveness metrics.

9 Conclusion

We present the first study on automatically measuring conversational agent offense mitigation strategies in-the-wild using 3 intuitive and novel metrics: re-offense ratio, length of the conversation after bot response, and number of turns until the next offensive utterance. We believe the automatic metrics we proposed make it easier to quickly evaluate response strategies, and thus allow researchers to experiment with more factors for constructing a successful response.

We evaluated 4 response strategies (AVOIDANCE, WHY, EMPATHETIC, and COUNTER) with 2 additional factors (REDIRECT and NAME). We showed that to mitigate offensiveness, the bot should almost always empathetically and actively move on to a different topic, and while doing so use the offending user's name whenever possible. We found that the bot should never ask a user why they made offensive utterances, as doing so causes the user to almost always repeat their offense immediately.

We hope our systematic evaluation of response strategies raises awareness of bot abuse as social bots become more popular and accessible.

Acknowledgements

We thank the anonymous reviewers of both CHI 2021 and SIGDial 2021 for their thoughtful comments that improved the paper. We would also like to thank professor Dan Jurafsky for his help in reviewing the paper as well as support and feedback from the Stanford NLP group, especially Peter Henderson, Abi See, and Ashwin Paranjape.

References

Antonella De Angeli. 2006. On verbal abuse towards chatterbots. In *Proceedings of CHI06 Workshop On the Misuse and Abuse of Interactive Technologies, Montréal, Québec, Canada*, pages 21–24.

Antonella De Angeli and Rollo Carpenter. 2005. Stupid computer! Abuse and social identities. In *Proc. Interact 2005 workshop Abuse: The darker side of Human-Computer Interaction*, pages 19–25. http://www.agentabuse.org.

Christoph Bartneck, Chioke Rosalia, Rutger Menges, and Inèz Deckers. 2005. Robot abuse-a limitation of the media equation. In *Proc. Interact 2005 workshop Abuse: The darker side of Human-Computer Interaction*, pages 54–57. http://www.agentabuse.org.

Christoph Bartneck, Marcel Verbunt, Omar Mubin, and Abdullah Al Mahmud. 2007. To kill a mockingbird robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 81–87.

Sheryl Brahnam. 2005. Strategies for handling customer abuse of ECAs. In *Proc. Interact 2005 workshop Abuse: The darker side of Human-Computer Interaction*, pages 62–67.

Sheryl Brahnam. 2006. Gendered bods and bot abuse. In *Proceedings of CHI06 Workshop On the Misuse and Abuse of Interactive Technologies, Montréal, Québec, Canada*, pages 13–17.

Dražen Brščić, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Takayuki Kanda. 2015. Escaping from children's abuse of social robots. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*, pages 59–66.

Amanda Cercas Curry and Verena Rieser. 2018. #MeToo Alexa: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Amanda Cercas Curry and Verena Rieser. 2019. A crowd-based evaluation of abuse response strategies in conversational agents. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 361–366, Stockholm, Sweden. Association for Computational Linguistics.

Amanda Cercas Curry, Judy Robertson, and Verena Rieser. 2020. Conversational assistants and gender stereotypes: Public perceptions and desiderata for voice personas. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 72–78, Barcelona, Spain (Online). Association for Computational Linguistics.

Xiangyu Chen and Andrew Williams. 2020. Improving Engagement by Letting Social Robots Learn and Call Your Name. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 160–162, New York, NY, USA. Association for Computing Machinery.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy Is All You Need: How a Conversational Agent Should Sespond to Verbal Abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Hyojin Chin and Mun Yong Yi. 2019. Should an Agent Be Ignoring It? A Study of Verbal Abuse Types and Conversational Agents' Response Styles. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–6, New York, NY, USA. Association for Computing Machinery.

Michelle Cohn, Chun-Yen Chen, and Zhou Yu. 2019. A large-scale user study of an Alexa Prize chatbot: Effect of TTS dynamism on perceived quality of social dialog. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 293–306, Stockholm, Sweden. Association for Computational Linguistics.

Jorge Gallego Pérez, Kazuo Hiraki, Yasuhiro Kanakogi, and Takayuki Kanda. 2019. Parent Disciplining Styles to Prevent Children's Misbehaviors toward a Social Robot. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 162–170.

Chandra Khatri, Behnam Hedayatnia, Rahul Goel, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal. 2018. Detecting offensive content in open-domain conversations using two stage semi-supervision. *arXiv preprint arXiv:1811.12900*.

Tatsuya Nomura, Takayuki Kanda, Hiroyoshi Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. 2016. Why do children abuse robots? *Interaction Studies*, 17(3):347–369.

Shauna Shapiro, Kristen Lyons, Richard Miller, Britta Butler, Cassandra Vieten, and Philip Zelazo. 2014. Contemplation in the Classroom: a New Direction for Improving Childhood Education. *Educational Psychology Review*, 27.

John Suler. 2004. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326.

Xiang Zhi Tan, Marynel Vázquez, Elizabeth J Carter, Cecilia G Morales, and Aaron Steinfeld. 2018. Inducing bystander interventions during robot abuse with social mechanisms. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, pages 169–177.

Mark West, Rebecca Kraut, and Han Ei Chew. 2019. I'd blush if i could: closing gender divides in digital skills through education.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984.

Sachie Yamada, Takayuki Kanda, and Kanako Tomita. 2020. An escalating model of children's robot abuse. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 191–199.