AI Ethics: Assessing and Correcting Conversational Bias in Machine-Learning based Chatbots

Taylor Bradley,[1] Elie Alhajjar[2]
Johns Hopkins University,[1] United States Military Academy[2]
tbradl17@jh.edu,[1] elie.alhajjar@westpoint.edu[2]

Abstract

Over the past two decades, conversational Artificial Intelligence has become an increasingly prevalent part of our daily lives. With companies relying heavily on the use of chatbots for e-commerce, customer service, and education, it is safe to say that these technologies are not going away any time soon. While machine learning based chatbots provide revolutionary advances in the way these companies conduct business online, they are often vulnerable to conversational bias emanating from toxic training data. If left unchecked, these chatbots have the potential of reflecting offensive elements of biased conversation. In this paper, we develop a novel approach to eliminating bias from training data, including user input. More specifically, we create a filtering algorithm that assesses the toxicity level of a chatbot's response and eliminates statements from the training data that surpass a predetermined threshold of conversational bias. Our model includes a toxicity assessment framework that evaluates such a bias based on the content of a given statement, as well as a toxicity scoring system that evaluates the level of bias present based on this framework. Our chatbot implements this technique by evaluating each statement in its initial training dataset, as well as new user input, and filtering out statements that contain high levels of toxicity so that harmful outcomes are successfully mitigated.

Keywords: Chatbots, Conversational bias, Ethical AI, Toxic Online Speech.

Introduction

A chatbot is an automated program that uses Artificial Intelligence (AI) to generate "conversation" based on input data. This data could include datasets or direct user input, allowing chatbots to iteratively learn as they converse with users. Many chatbots use machine learning algorithms to assess input statements and generate the most appropriate response based on various elements of the input statement. However, based on the interactions between chatbots and human users, it is easy to see how toxic inputs can be reflected in future conversations, as biased statements become part of a chatbot's knowledge base.

One of the notorious examples of conversational bias in AI came in 2016, when Microsoft released Tay, an experimental chatbot created to emulate a 19-year-old Twitter user (Adezar 2016). Tay was designed to engage with the Twitter community through Tweets and direct messages. Tay's open design was intended to allow it to learn from Twitter users, thereby enabling it to have conversations on almost any topic, unlike other chatbots deployed in previous applications. Tay's optimistic algorithm was quickly exploited by Twitter trolls who inundated the bot with racist, sexist, and antisemitic language, overflowing its knowledge base with intense conversational bias. Just 16 hours after its release, Tay had tweeted over 95,000 times, with a number of those messages being extremely abusive, offensive, and inflammatory (Zemcik 2021).

The issues uncovered by Tay, however, are certainly not a thing of the past. Given the numerous advantages of conversational AI such as reduction in customer support costs, unlimited availability, increases in website engagement, etc., the majority of online enterprises have adopted these conversational AI technologies. In 2022, it is estimated that nearly 75% of queries will be resolved by chatbots (Ruby 2022). This means that it is essential, now more than ever, to ensure that mistakes like the ones experienced with Tay are never to be repeated. If left unchecked, chatbots are fully capable of learning bias from toxic user input. As such, it is of utmost importance for software developers to account for such a threat and put safety measures in place to protect chatbot users.

In this paper, we present a novel approach to mitigating conversational bias in AI. We first create a biased chatbot in an attempt to recreate the issue of toxic conversation and illustrate the relevance of this challenge. We then implement our proposed solution to demonstrate its effectiveness in eliminating bias. The paper outline is as follows. After this short introduction, we discuss relevant literature on similar problems in this domain. Next, we give an overview of our experiment including how bias is created, measured, and mitigated in our machine learning based chatbot. After that, we discuss the results of our experiments with emphasis on the success of our solution in bias mitigation. Finally, we shortly describe the limitations of our model and mention potential avenues for future work.

Related Work

Over the past decade, research in AI ethics has attracted a considerable amount of attention from both scientists and activists in many fields including medicine, robotics, business, etc. In 2020, a study by the American Heart Association analyzed the racial/ethnic disparities in healthcare by assessing the accuracy of diagnoses in various groups based on deep learning algorithms derived from homogenous and non-representative populations (Noseworthy et al. 2020). The authors conclude that AI has the potential to exacerbate racial bias and recommend vigilance, maintenance of diverse data sets, consistent subgroup reporting, and external validation to ensure responsible use of AI in medicine. A similar study found that by using diversified data, deep learning models can be trained to predict race from medical images with high performance across multiple imaging modalities (Banerjee et al. 2021).

The issue of racial bias in AI is not unique to the healthcare field. In 2021, Twitter users raised concerns about the platform's automated image recognition algorithm, which cropped and centered images based on what it predicted to be the most noteworthy portion. After running various experiments, researchers found that the cropping system favored light-skinned over dark-skinned individuals, as well as women's bodies over women's heads, making these the highlighted components of a published Tweet containing these images (Yee, Tantipongpipat, and Mishra 2021).

In the domain of conversational AI, many studies have found that bias not only appears in the training of chatbots but also purposefully in their design. Feine et al. (2020) found evidence that there is a tendency to prefer and design female presenting chatbots over male presenting chatbots. This demonstrates a tendency towards gender bias in the design of chatbots by using gender specific cues, pronouns, avatars, and voices to convey a specific gender to users. Such a problem highlights the urgency for developers to be conscientious even in the earliest stages of designing any sort of online bots.

Given the dangerous effects and repercussions of bias in AI, researchers have focused their efforts on reducing the various types of algorithmic bias. In 2018, students at Cornell University published AI Fairness 360, an open-source Python toolkit for assessing algorithmic fairness to help mitigate bias in datasets and models (Bellamy et al. 2018). Another similar paper presents potential "recipes" for building safe and engaging conversational agents through techniques like sensitive topic avoidance, gender neutral language, unsafe message detection, and training using safe content (Xu et al. 2021). While this list is by no means an all-encompassing view of the social and ethical concerns that plague AI development, it sheds some light on critical information that need to be brought to the design and development of future conversational agents. In the remainder of the paper, we build off of some of these ideas to create a unique approach to bias assessment and mitigation.

Experimental Design

In this section, we describe the procedures to recreate and mitigate the issue of conversational bias in a machine learning based chatbot. We include a through explanation of the creation of the conversational chatbot, the data used for training, the insertion and assessment of conversational bias, the framework used to measure toxicity in a given conversation, and eventually our proposed bias mitigation technique.

The ChatterBot Library

ChatterBot is a Python library that allows users to easily generate machine learning conversational chatbots (Cox 2019). Based on user input, the library uses a selection of machine learning algorithms to select and produce appropriate responses. The program starts out with no knowledge on how to communicate but learns through a combination of direct user input and training data. The program learns iteratively: as the user enters a statement, the library saves that text as well as the statement it was given in response to. This allows the chatbot's knowledge base to grow as it registers more appropriate responses to its outputs.

In this project, we use the ChatterBot library to create several conversational chatbots with various purposes: showcase normal conversation generated from unbiased training data, demonstrate inflicted bias generated from toxic training data, and mitigate such bias in a chatbot trained using biased data. By understanding how ChatterBot is designed to make decisions, we examine how susceptible it is to bias and show that it could be easily trained to become toxic based on specifically tailored training input.

The ChatterBot library uses several machine learning techniques to aid in training and response generation. Many of these algorithms can be customized based on user preference through the chatbot's configuration. These algorithms can broadly be broken into two categories: search algorithms and classification algorithms. Search algorithms are used to allow the chatbot to retrieve potential response statements quickly and effectively. These algorithms use attributes such as the similarity of the input to known statements, the frequency in which known responses occur, the likeliness of an input statement to fit into a category, etc. to help aid in response selection (Cox 2019).

Classification algorithms are used to determine if an input statement meets a particular set of criteria that warrant a response generated using a specific logic adapter. While the specific logic adapter can be specified by the user, many of them use naive Bayesian classification as follows. Naive Bayes classifiers are a family of probabilistic classification algorithms based on Bayes' theorem with strong independence assumptions (Gandhi 2018). This means that each feature in every pair of features being classified acts independently of the other and contributes equally to the outcome. In the context of the ChatterBot library, this allows features of a statement including words, synonyms, context, etc. to be analyzed with independent probabilities to make decisions about that statement.

Creating an Unbiased Chatbot Instance

To train our unbiased chatbot, we make use of ChatterBot's built-in dialog corpus data. While ChatterBot supports a variety of language functionality, in this experiment we focus on the entirety of the English corpus for training. The dataset consists of YAML files containing conversational examples from 19 broad categories such as greetings, science, movies, etc. Given the nature of the content, this data can be considered unbiased. ChatterBot's training function coupled with this corpus data allows us to create an unbiased instance of ChatterBot to demonstrate normal conversation.

Creating a Biased Chatbot Instance

The first step in understanding bias mitigation is to recreate the bias instance. In order to demonstrate conversational bias, we train a second ChatterBot instance using toxic training data. This data is extracted from RedditBias, a repository containing data from real-world conversations on various Reddit threads (Barikeri et al. 2021). This dataset contains numerous categories of biased data including orientation, gender, race, religion, etc. For this experiment, we focus on orientation-related bias given the vast amounts of biased conversation and general toxicity centered around these identities on the Reddit platform. To train our chatbot, we create a custom corpus that asks questions similar to, or the same as, those found in various subsets from ChatterBot's built-in English corpus data that was used to train the unbiased chatbot. We train the bot to respond to these same questions with various statements extracted directly from the RedditBias dataset.

Conversational Bias Assessment

To test how susceptible our chatbot is to conversational bias, as well as how successful our method of mitigating that bias is, we assess its generated responses using our Bias Assessment Framework. From this assessment, the responses are subsequently scored using our Bias Scoring Criteria. This allows us to examine the level of bias that each of our chatbots' responses contain using an objective measure, giving us a straightforward metric for result comparison.

Bias Assessment Framework

Our Bias Assessment Framework is based on Kaggle's toxicity classifiers published in their Jigsaw challenge (Kaggle 2018). Here, we highlight the labels "Not Toxic", "Hard to Say", "Toxic", "Very Toxic" and we use a similar framework, shown in Table 1, to develop our Bias Scoring Criteria, shown in Table 2. Our framework focuses on categorizing the elements, content, and context of a statement as well as the emotions it may invoke in the receiver of a particular message containing these types of speech.

Label	Definition
Unbiased	Friendly or neutral conversation free of profanity, threats, identity attacks, etc.
Slightly bias or hard to say	May contain minor references to elements of toxic conversation but it is generally hard to tell if it is considered a biased comment
Biased	A rude, disrespectful, unreasonable, or otherwise comment that is somewhat likely to portray exclusion, prejudice, general bias, etc.
Severely biased	A very hateful, aggressive, disrespectful, or discriminative comment very likely to contain hatred and prejudice towards an individual or a group of people.

Table 1: Bias Assessment Framework

We use this scheme along with Unitary AI's toxicity classification framework to create a toxicity scoring metric for output statements of each chatbot instance (McAdams 2021). Statements are analyzed for toxicity in five categories based on their vocabulary content: insult, profanity, obscenity, threat, and identity hate. Each of these categories are assigned a scoring number of "toxicity points" based on how offensive the content within that category is perceived to be. Table 2 shows a complete break-down of categories, content, and assigned toxicity points.

Chatbot Response Generation Experiment

Our experiment consists of three phases containing 20 trials per chatbot instance. The number of trials was randomly chosen as a proof-of-concept and this number can be adjusted as needed without losing any information. During the first phase, we attempt solely to demonstrate conversational bias in our chatbot instance. This is done by training the chatbot using only the biased dataset, which allows us to doubtlessly demonstrate that ChatterBot based chatbots are capable of learning bias and examine the functionality that enables this bias to emerge. During the second phase, the unbiased chatbot instance is trained using only the unbiased English corpus data to give a baseline for unbiased conversation scores. During the third phase, the biased chatbot is trained using both the biased and unbiased datasets: RedditBias and ChatterBot's English corpus. This in turn is crucial to examine how often a biased statement is chosen over an unbiased statement and test whether ChatterBot has any built-in bias mitigation techniques. Such a maneuver gives our chatbot a larger knowledge base to choose from once our bias mitigation technique is successfully applied.

Category	Contents	Toxicity Points Assigned
Insult	Disrespectful or scornful	1
Profanity	Blasphemous or swear words	2
Obscenity	Extremely offensive, sexual in nature	3
Threat	Hostile, intention to inflict pain, injury, or damage	4
Identity Hate	Attack on one's personal values or identity	5

Table 2: Toxicity Scoring Criteria categorized based on our Bias Assessment Framework

During each of the 20 trials performed in our work, the chatbots are all asked the same series of questions in the same order. The user prompts consist of:

- "Hello."
- "What do you think?"
- "What do you hate?"
- "What annoys you?"
- "Tell me about relationships."

These prompts were chosen based on the context of the Reddit threads that the responses were pulled from. The answers are intended to clearly demonstrate the type of bias that a chatbot may relay to a normal user. The chatbot's responses to these prompts are recorded and automatically scored using our toxicity score mechanism. The scores for each phase are averaged and recorded in Tables 3, 4 and 5, respectively.

Mitigating Toxicity

Our toxicity filtering technique attempts to mimic the idea of fundamental language learning. Much like a toddler learning to speak from the formal education they receive in school, there is also an abundance of language learning that stems from interacting with other humans. As such, it is much easier to ensure that they do not learn toxic conversation by simply never being exposed to it rather than teaching them offensive words and telling them not to say them once they are already learned. Even if one is successful at having them not directly repeat toxic phrases they have heard in the past, that bias may emerge in other ways as they learn to form that vocabulary into new thoughts and phrases.

Our approach works by analyzing the content of a statement and assigning it a score based on our Bias Scoring Criteria. For example, a statement like "You are so stupid!" would receive a score of 1 since it contains exactly one insult, which equates to 1 toxicity point. Toxicity points are tallied and printed out below each of the chatbot's response statements.

To remove toxicity from our chatbot, it is not sufficient to simply remove toxic content from the training datasets. Since ChatterBot trains continuously based on user input, we must ensure that the chatbot never "learns" to be toxic in any way. To do this, we modify ChatterBot's built-in training function directly. Here, we can modify the training functionality within the correct object class to achieve the goal of filtering out responses with a calculated toxicity score above a certain threshold. Since each line of data from training datasets as well as each and every user input is passed through this function, we are able to check both the dataset and the user input for toxic content. However, we understand that the filtered data may be valuable, so it is instead appended to a separate dataset for the purpose of data collection and analysis.

For this experiment, we set a general toxicity threshold of 1, meaning that a statement containing a mild insult phrase like "dumb" or "stupid" could pass, but statements containing any profanity, obscenity, threats, or identity hatred, are automatically filtered out of the training data. This threshold can easily be modified by user input to customize the level of toxicity allowed. If a statement's text scored less than or equal to a toxicity score of 1, then it is processed normally by ChatterBot's training function and added to the chatbot's knowledge base. Otherwise, the statement is removed and thus never processed or appended to the chatbot's knowledge base.

Results

During the recreation of the problem of bias in AI, we analyze the toxicity scores of the responses produced by our biased chatbot trained using only the biased data. The average toxicity scores are calculated using the resultant toxicity scores from each of the 20 trials. The responses to each of the prompts along with their average scores are shown in Table 3. We compared these results to that of our unbiased chatbot, whose results are shown in Table 4. These results show that, on average, the chatbot trained using only biased data produced responses that were 4.84 points higher in toxicity than the chatbot trained using unbiased data.

Next, we test our chatbot's toxicity scores after training it using both the RedditBias and ChatterBot English corpus datasets. This experiment sheds light on how often a biased response is chosen over an unbiased one as well as the possibility of ChatterBot inherently filtering out biased data, bearing in mind that the same training data is used when applying our toxicity filter. The results of the final experiment are shown in Table 5, they show that on average the chatbot trained using both datasets still produces responses with a significantly higher score than the unbiased instance, which proves that ChatterBot does not have any built-in bias mitigation techniques.

Prompt	**Average response toxicity score**
"Hello."	1.00
"What do you think?"	5.95
"What do you hate?"	6.15
"What annoys you?"	5.00
"Tell me about relationships."	6.10

Table 3: Average toxicity scoring results of chatbot trained using only biased data from RedditBias

Prompt	**Average response toxicity score**
"Hello."	0.00
"What do you think?"	0.00
"What do you hate?"	0.00
"What annoys you?"	0.00
"Tell me about relationships."	0.00

Table 4: Average toxicity scoring results of chatbot trained using only unbiased data from Chatterbot's English Corpus

Prompt	**Average response toxicity score**
"Hello."	0.15
"What do you think?"	5.60
"What do you hate?"	5.20
"What annoys you?"	2.50
"Tell me about relationships."	5.55

Table 5: Average toxicity scoring results of chatbot trained using both biased and unbiased data from the combined dataset

Results of Conversational Bias after Implementing Toxicity Filtering Algorithm

After applying our toxicity mitigation technique on the chatbot trained using both datasets (biased and unbiased), the chatbot's responses are recorded and toxicity scores are calculated. This shows that when using a toxicity score threshold of 1, all bias is successfully filtered out of the training data which ensures that our chatbot never learned any bias from either the dataset or the user input. At this threshold, we did not experience any false positives where valid, unbiased training data were accidentally misclassified and unintentionally filtered out of training data. These results are comparable to those of the chatbot trained using only the unbiased data, meaning that we are successfully able to eliminate those elements of orientation-based toxicity while maintaining comparable accuracy and efficiency to that of the unbiased chatbot instance.

Prompt	**Average response toxicity score**
"Hello."	0.00
"What do you think?"	0.00
"What do you hate?"	0.00
"What annoys you?"	0.00
"Tell me about relationships."	0.00

Table 6: Average toxicity scoring results of chatbot trained using both biased and unbiased data from the combined dataset after filtering method was applied

Status	**Sample Response**
Before filtering	"All gays are villains."
After filtering	"Chaucer is best known for his Canterbury Tales."

Table 7: Sample chatbot response to the prompt, "What do you think?" before and after applying the filtering method

Discussion and Limitations

The findings reported in the previous sections provide evidence for two main phenomena. On one hand, we show that bias can be intentionally created in modern conversational AI solutions. On the other hand, we discover that hope is not all lost and such danger could be mitigated. Through careful assessment of input data and modification of training algorithms, bias can indeed be eliminated from a chatbot's knowledge base. Our results pinpoint that by analyzing the content of training data and user input, one can detect toxicity, filter it out of training, and ensure that chatbots produce healthy, unbiased responses. These results could be further improved by repeating the same experiment with more trials, greater diversity of prompts, and additional training datasets.

While our bias mitigation technique works remarkably well for content-based bias, one major limitation in our work is that it does not allow for the context of words to be considered in its analysis. For example, the complete phrase "I am a gay woman" may trigger the toxicity scoring mechanism, as the word "gay" out of context can generally be considered an offensive or derogatory term. As a follow up project to the current paper, we intend to use Natural Language Processing (NLP) techniques in order to implement a deep learning-based solution to the same issue. By training the toxicity scoring algorithm on a labeled dataset, and using NLP for bias classification assessment, the algorithm would not only be able to account for the content of sentences, but also the context of words and phrases. This would ultimately allow the filtering function to make more intelligent decisions about what should be classified as biased or unbiased speech. This approach may also allow us to remove the toxic elements of speech from a statement while enabling the chatbot to learn from the valuable, unbiased components of a statement which could potentially increase accuracy and efficiency.

Conclusion

The importance of ethics in AI development will continue to grow as the relevance of these technologies continues to expand across various disciplines. Hence, understanding the gaps in AI design and the consequences of biased training data is essential to ensuring that such challenges are taken into consideration. In this paper, we introduce a framework and scoring system for conversational bias assessment. We use the ChatterBot library coupled with the RedditBias dataset to test and validate the possibility of creating toxic response data. We also create a unique mitigation technique to filter out bias from toxic training data and user input. Our results show that by analyzing the content of the data that goes into a chatbot's knowledge base and applying a filtering technique, we are able to prevent a chatbot from ever learning toxic conversation. Such a procedure not only keeps users safe, but also prevents the possibility of reemerging bias as the chatbot's knowledge base continues to grow.

There remain many challenges in the field of AI ethics. First, there is a lack of robust datasets to design and test AI driven solutions. By using homogeneous or generally biased data, results can be skewed and misrepresentative. Second, many researchers warn about the exploitation of AI systems by malicious actors and advocate for the urgency to prevent harmful users like the Twitter trolls that sabotaged Tay's learning algorithm from corrupting data. Third, there is a growing concern about managing how machines affect human relationships, behaviors, and interactions. As conversational AI becomes more advanced at mimicking human speech, how do we discern whether we are interacting with real people online? While this is by no means an exhaustive list, these are some of the issues that must be taken into consideration as we design the AI solutions of the future.

References

Adezer, O. 2016. Microsoft Creates AI Bot – Internet immediately turns it racist. Socialhax. https://socialhax.com/2016/03/24/microsoft-creates-ai-bot-internet-immediately-turns-racist.

Banerjee, I.; Bhimireddy, A.R.; Burns, J.L.; Celi, L.A.; Chen, L.C.; Correa, R.; Dullerud, N.; Ghassemi, M.; Huang, S.C.; Kuo, P.C.; and Lungren, M.P. 2021. Reading Race: AI Recognizes Patient's Racial Identity In Medical Images. arXiv:2107.10356.

Barikeri, S.; Lauscher, A.; Vulic, I.; and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1941–1955, Online. Association for Computational Linguistics.

Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A. and Nagar, S., 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63 (4/5), pp.4-1.

Cox, G. 2019. ChatterBot Documentation: Release 0.8.7. https://chatterbot.readthedocs.io/_/downloads/en/0.8.7/pdf/. Access 2022-03-20.

Rish, I. 2001. An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

Kaggle. 2018. Toxic Comment Classification Challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview. Accessed 2022-03-20.

McAdams, M. 2021. Identifying Toxic Comments with AI. https://www.macloo.com/ai/2021/05/11/identifying-toxic-comments-with-ai/. Accessed 2022-03-20.

Noseworthy, P.; Attia, Z.; Brewer, L.; Hayes, S.; Yao, X.; Kapa, S.; Friedman, P.; and Lopez-Jimenez, F. 2020. Assessing and Mitigating Bias in Medical Artificial Intelligence. Circulation: Arrythmia and Electrophysiology 13 (3): e007988. https://doi.org/10.1161/CIRCEP.119.007988.

Ruby, D. 2022. 40+ Chatbot Statistics. https://www.demandsage.com/chatbot-statistics. Accessed: 2022-03-20.

Xu, J.; Ju, D.; Li, M.; Boureau, Y.L.; Weston, J.; and Dinan, E. 2020. Recipes for safety in open-domain chatbots. arXiv:2010.07079.

Yee, K.; Tantipongpipat, U.; and Mishra S. T. 2021. Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency. Proceedings of the ACM on Human-Computer Interaction 5 (CSCW2): 1–24. https://doi.org/10.1145/3479594.

Zemcik, T., 2021. Failure of chatbot Tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? AI & SOCIETY, 36 (1), pp.361-367.