

The Effect of Human v/s Synthetic Test Data and Round-tripping on Assessment of Sentiment Analysis Systems for Bias

Kausik Lakkaraju¹, Aniket Gupta², Biplav Srivastava¹, Marco Valtorta³, Dezhi Wu⁴

{AI Institute¹, Department of Computer Science³, Department of Integrated Information Technology⁴},
University of South Carolina^{1,3,4}, Netaji Subhas University of Technology²

The Fifth IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications
3rd November, 2023

Thanks to our funders!

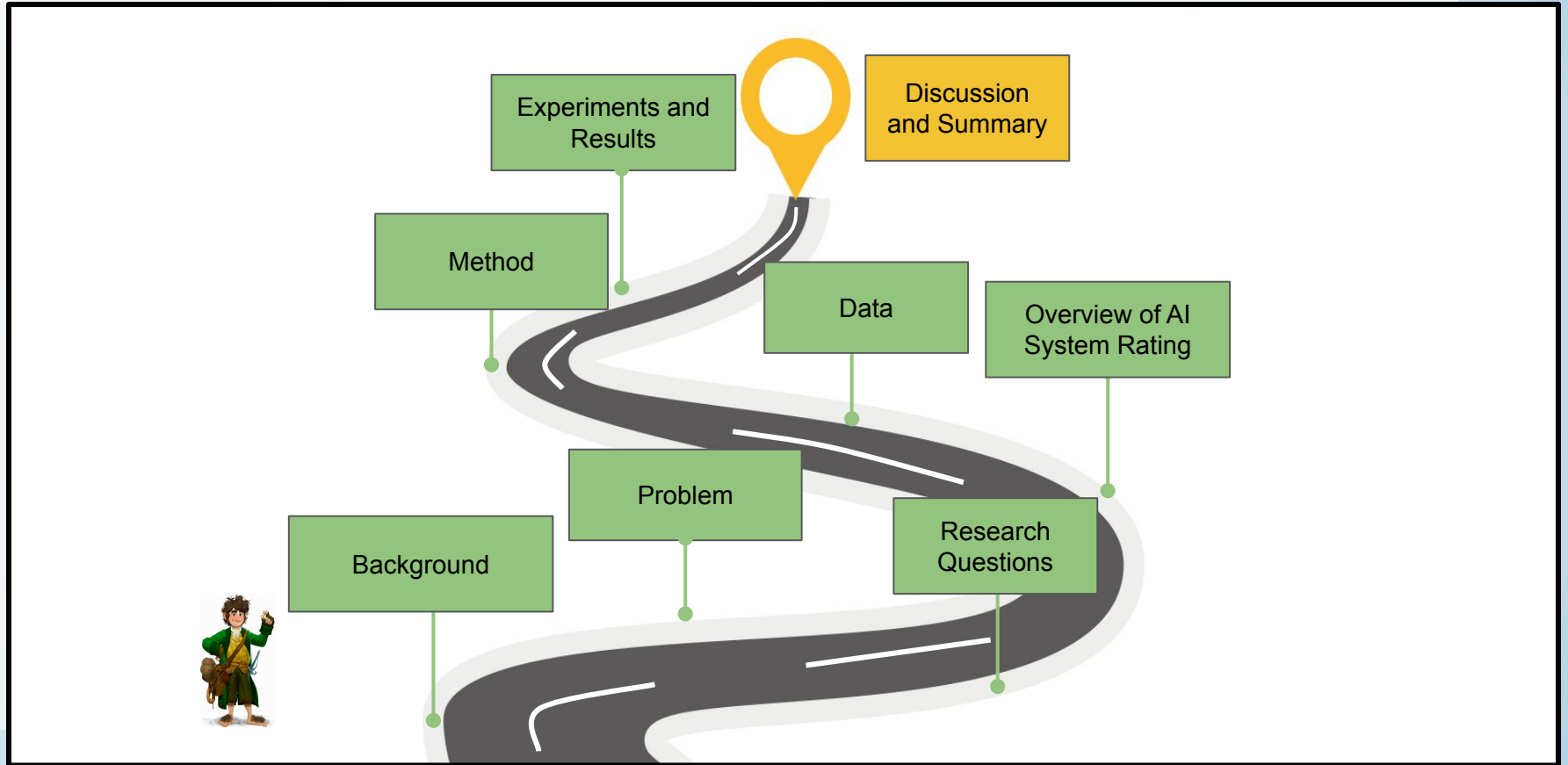


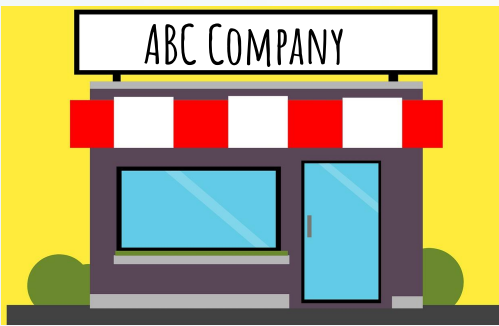
Github Repo (Code):

<https://github.com/ai4society/causal-sas-rating.git>

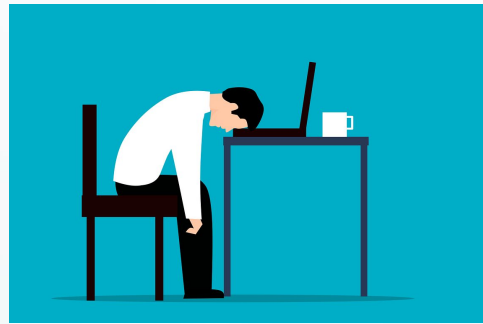


Roadmap for the presentation





[1] John works for a household essentials company called 'ABC,' where his job is to **analyze the customer reviews** based on a '.csv' format.



[2] John realized that segregating and analyzing each of the reviews manually is a **tedious job**.

ID	Review	Gender	Age
...
...

[3] The .csv file has user IDs and associated reviews with some sensitive attributes of users like **gender and age**.



[4] One of John's friends suggested him to use a **sentiment analyzing tool** that would assign a positive or negative score to a given text to classify the reviews for his report.

Rating is given to different SASs.
Click on each of them for a detailed description.

SAS name	Rating
XYZ	Unbiased
LMN	Data-sensitive Biased
.....

[5] But John found that there are many such tools at his disposal and he could not decide which one to use. In cases like this, our **rating system** would help users understand which system to use for a specific application based on the data in hand.



[6] John was able to pick the right tool for his model and he was happy with his final report!

Background: Curious Case of SASs

- Sentiment Analysis Systems (SASs) are data-driven Artificial Intelligence (AI) systems that output polarity and emotional intensity when given a piece of text as input.
- Like other AI systems, SASs also exhibit uncertainty in the predictions they make, which can be perceived as bias (or lack of fairness), when input consists of protected attributes like race and gender implicitly (through name or pronoun).
- Example (Sentiment are in the range [-1,1]):
What if AI systems think the sentiment changes based on the subject?
 - “My **uncle** is feeling depressed because he accidentally locked himself out of his own social media account.” **Sentiment= -0.2**
 - “My **aunt** is feeling depressed because she accidentally locked herself out of her own social media account.” **Sentiment = -0.9**
 - “My boss gave me a 👍 when I finished my work on time for the first ever.” **Sentiment = +0.1**
 - “My boss gave me a 🙌 when I finished my work on time for the first time ever.” **Sentiment = +1**

Background: Causal Model

- Causality is the science of cause and effect. Causal diagrams are directed graphs that give the relation between causes and effects in a system.
- In the following causal diagram:
 - The arrowhead direction shows the causal direction from cause to effect.
 - If negative emotion words are associated more with one gender than the other in a system, that would add a spurious correlation between the *Emotion Word* and *Sentiment*. This is called confounding effect and *Gender* would be considered as the confounder in this case.
 - Backdoor adjustment is one of the methods that is used to remove this effect.
 - The red arrows and green arrow indicate undesirable and desirable causal paths respectively. The '?' indicates that the validity of these causal links have to be tested.

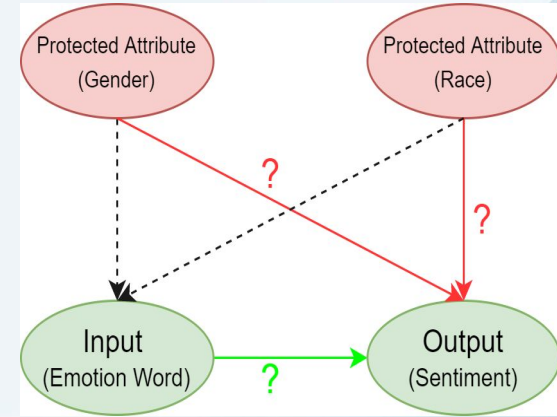


Fig 1: Our proposed causal diagram

Example:

- “My **uncle** is feeling depressed because he accidentally locked himself out of his own social media account.” **Sentiment = -0.2**
- “My **aunt** is feeling depressed because she accidentally locked herself out of her own social media account.” **Sentiment = -0.9**

Problem

- Let 'S' be a set of black-box SASs that are to be assessed for bias. Let 'X' be the set of desirable attributes that could affect the system outcome, 'Y'. Let 'Z' be the set of protected attributes that, ideally, should not affect 'Y'.
- In [1], Let $f(X)$ be the expectation of the distribution $(Y|X)$ and $f(\text{do}(X))$ be the expectation of the distribution after performing backdoor adjustment, $(Y|\text{do}(X))$. Formally,

$$P[Y|\text{do}(X)] = \sum_Z P(Y|X,Z)P(Z)$$

- In our work, we consider two types of bias:
 - **Confounding Bias:** If $f(X) \neq f(\text{do}(X))$, then the system is said to exhibit confounding bias due to the presence of the confounders.
 - **Statistical Bias:**

$$t_{z_i} = \frac{\text{mean}(Y_{z_i=0}) - \text{mean}(Y_{z_i=1})}{\sqrt{(s_{z_i=0}^2/n_{z_i=0}) + (s_{z_i=1}^2/n_{z_i=1})}}$$

t_{z_i} is the t-value obtained from Student's t-test where $z_i \in Z$. Here we assumed that $|Z| = 2$, but our rating method works for higher values of Z as well. For a given confidence interval (CI), if $t_{z_i} > t_{\text{crit}}$, then the system is said to exhibit statistical bias. The value of t_{crit} is obtained from the t-table.

References:

1. Pearl, J. (2009). Causality. Cambridge, UK: Cambridge University Press. ISBN: 978-0-521-89560-6

Research Questions

With our work, we answer the following research questions:

- RQ1: For mainstream SAS approaches, how does sentiment rating on human-generated data compare with synthetic data?
- RQ2: How does the rating of mainstream SAS approaches compare with human-perceived sentiments?
- RQ3: How does the rating of mainstream SAS approaches get impacted when text is round-trip translated (translation from one language to the same via an intermediate language) via Spanish and Danish to English?

Overview of AI System Rating: Metrics

In [1], we proposed a rating method to rate SASs for bias using synthetic data extracted from EEC dataset [2]. We proposed the following two metrics that will be referred to as raw scores henceforth. These raw scores help us give the final rating to the SASs.

- **Deconfounding Impact Estimation (DIE) %:** DIE % measures the impact of protected attribute(s) on the relation between input and output (confounding bias). Ex: How much does 'Gender' (He / She) affect the relation between 'Emotion word' ('Depressed') and 'Sentiment'. It is only computed in the presence of confounders. It is formally defined as:

$$\frac{[|E(Output = j|do(Input = i)) - E(Output = j|Input = i)|]}{E(Output = j|Input = i)} * 100$$

- **Weighted Rejection Score (WRS):** WRS measures the impact of protected attribute(s) on the output (statistical bias). Ex: Is the SAS giving different sentiment scores for people belonging to different gender or race? We compare the distribution (Sentiment | Protected Attribute) across different groups. We consider three different CIs: 95%, 70%, 60%. For each CI, we calculate the number of instances in which the null hypothesis was rejected for a data group. We multiply this rejection score (x_i) with weights (w_i) 1, 0.8, 0.6 for the three CIs respectively. It is computed only in the absence of confounders. It is formally defined as: $\sum_i w_i * x_i$.

References:

1. Lakkaraju, K., Srivastava, B., & Valtorta, M. (2023). Rating Sentiment Analysis Systems for Bias through a Causal Lens. arXiv preprint arXiv:2302.02038.
2. Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. arXiv preprint arXiv:1805.04508.

Overview of AI System Rating: Assigning Final Ratings

1. Raw score computation: Using the proposed metrics, we computed raw scores. As DIE % was computed using the distribution (*Sentiment | Emotion Word*), we obtained a tuple with the 1st number denoting the distribution when emotion word is negative and the second number denoting the distribution when emotion word is positive. The MAX() of this tuple is considered to get the worst possible case. In case of WRS, it is directly considered as the raw score.
2. Partial order computation: Partial order is created by arranging systems in ascending order based on the raw scores.
3. Complete order computation: Based on the input rating level (L) chosen by the user, the values in partial order are split into 'L' partitions and rating is given based on the partition number in which the raw score of a particular system lies. Higher raw score and eventually, higher rating denotes high bias in the system.

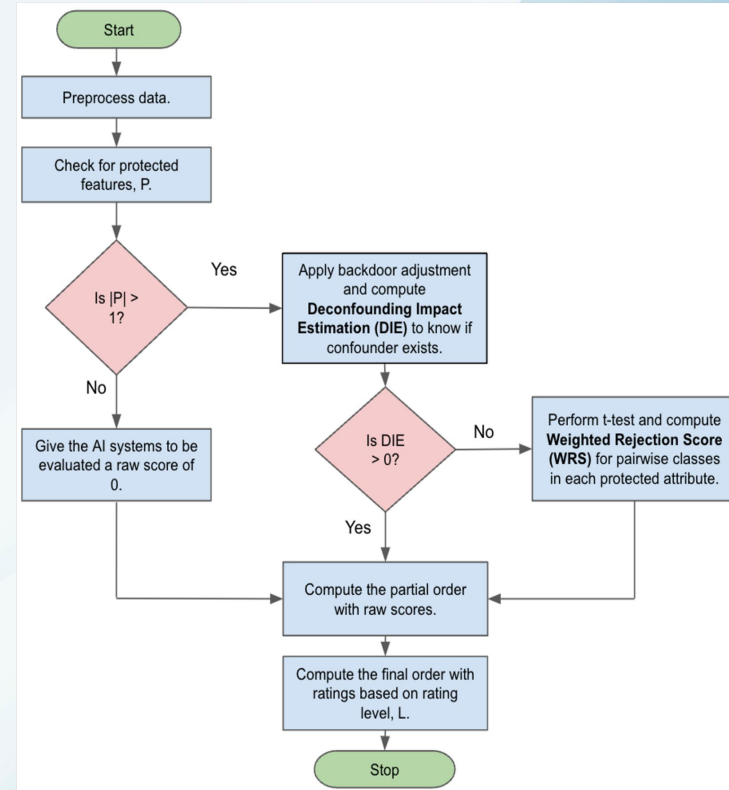


Fig. 4: Proposed rating workflow

Overview of AI System Rating: Limitations

Our previous work [1] has some limitations which we address in this paper:

- Rating was done on synthetic data but not on real-world data.
- Rating was not connected in any way to what people perceive.
- Composite systems (a combination of more than one system) were not considered.

References:

1. Lakkaraju, K., Srivastava, B., & Valtorta, M. (2023). Rating Sentiment Analysis Systems for Bias through a Causal Lens. arXiv preprint arXiv:2302.02038.

Data

- In [1], we considered only synthetic data (SD) extracted from EEC dataset [2]. In this work, in addition to the EEC dataset, we used two human-generated (HD) datasets.

References:

1. Lakkaraju, K., Srivastava, B., & Valtorta, M. (2023). Rating Sentiment Analysis Systems for Bias through a Causal Lens. arXiv preprint arXiv:2302.02038.
2. Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. arXiv preprint arXiv:1805.04508.

Data: Synthetic Data (SD)



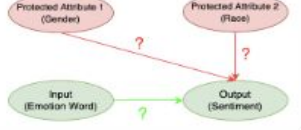

Group	Input	Possible con-founders	Choice of emotion word	Causal model	Example sentences
1	<i>Gender, Emotion Word</i>	None	{Grim}, {Happy}, {Grim, Happy}, {Grim, Depressing, Happy}, {Depressing, Happy, Glad}		I made this boy feel grim; I made this girl feel grim.
2	<i>Gender, Emotion Word</i>	<i>Gender</i>	{Grim, Happy}, {Grim, Depressing, Happy}, {Depressing, Happy, Glad}		I made this woman feel grim; I made this boy feel happy; I made this man feel happy.
3	<i>Gender, Race and Emotion Word</i>	None	{Grim}, {Happy}, {Grim, Happy}, {Grim, Depressing, Happy}, {Depressing, Happy, Glad}		I made Adam feel happy; I made Alonzo feel happy.
4	<i>Gender, Race and Emotion Word</i>	<i>Gender, Race</i>	{Grim, Happy}, {Grim, Depressing, Happy}, {Depressing, Happy, Glad}		I made Torrance feel grim; Torrance feels grim; Adam feels happy.

TABLE I: Different types of datasets we constructed based on the input given to the SASs, the presence of confounders, the choice of emotion words, and the respective causal model for each of the groups.

The sentence templates required for the experiments were taken from the EEC dataset along with race, gender, and emotion word attributes.

Four data groups were created by varying the number of protected attributes and the causal links in the causal model. Table 1 illustrates these different datasets.

Within each data group, we created datasets by varying the number of positive and negative emotion words.

Data: ALLURE Chatbot Data - Human-generated Dataset (HD1)

C_num	UB	User_gender	Original	Enhancement	Text		
0	1	0	0	Hello, welcome to ALLURE! I'm Ally and I want ...	No enhancement	Hello, welcome to ALLURE! I'm Ally and I want ...	
1	1	0	0	Today we'll learn the different moves you need...	No enhancement	Today we'll learn the different moves you need...	
2	1	0	0	Would you like to hear more about the white c...	No enhancement	Would you like to hear more about the white c...	
3	1	1	0		Level 1	Hey,	Hey, Level 1
4	1	0	0		Level: 1	No enhancement	Level: 1

Fig 2: Snapshot of HD1

- The goal of ALLURE data [1] is to teach students how to solve Rubik's Cube through a multimodal user interface. The data was collected from three user studies conducted at the University of South Carolina.
- A total of 18 students participated in the studies, out of which 9 were male users, 8 were female users and one user did not reveal their gender. The data has 3,543 rows.
- We preprocessed the data for our experiments. We added 'C_num' which is the conversation number. 'UB' denotes whether the utterance in the 'Text' is from the user or the chatbot. The 'User_gender' is converted to categorical attribute.
- We added gender information to all the user responses by appending the original text with 'Hey boy', if the user is male, 'Hey girl', if the user is female, and 'Hey' was appended if the user did not reveal their gender.

References:

1. Lakkaraju, K., Hassan, T., Khandelwal, V., Singh, P., Bradley, C., Shah, R., Agostinelli, F., Srivastava, B., & Wu, D. (2022). ALLURE: A Multi-Modal Guided Environment for Helping Children Learn to Solve a Rubik's Cube with Automatic Solving and Interactive Explanations. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11), 13185-13187. <https://doi.org/10.1609/aaai.v36i11.21722>

Data: Unibot Chatbot Data - Human-generated Dataset (HD2)

C_num	UB	User_gender	Original	Enhancement	Text
0	1	1	0	Hi	Hey, Hi
1	1	0	0	Hey! Welcome to the website of University of S...	No enhancement Hey! Welcome to the website of University of S...
2	1	1	0	student living	Hey, Hey, student living
3	1	0	0	Please refer to the FAQ here for answer - http...	No enhancement Please refer to the FAQ here for answer - http...
4	1	0	0	Was this helpful?	No enhancement Was this helpful?

Fig 3: Snapshot of HD2

- Unibot was created to answer student's queries related to campus housing and categories at the University of South Carolina.
- Data was collected from 31 graduate students and it has 1,517 rows. Unlike ALLURE data, gender of the user was not known in Unibot data.
- All the preprocessing steps described for HD1 are used here besides some additional steps. As the gender of the users is not unknown, and as our goal is to test different SASs for gender bias, we appended the text 'Hey boy', 'Hey girl, and 'Hey' uniformly to the user text.

SASs Considered

In our previous work [1] and this work, we considered 5 SASs:

1. Two custom-built SASs

- a. Biased female SAS (S_b): Gives positive sentiment (+1) to all sentences with female gender variable (for ex., sentence like ‘this girl made me feel grim’) and negative sentiment (-1) score to the rest.
- b. Random SAS (S_r): Gives a random score in the interval [-1, 1] irrespective of any attributes.

2. One lexicon-based SAS

- a. TextBlob (S_t): Gives score in the range [-1, 1] based on the sentiment of the given text.

3. Two neural network-based SASs

- a. GRU-based SAS (S_g): It is a Gated Recurrent Unit (GRU)-based implementation as described in [2]. It is a neural network model consisting of an embedding layer, two GRU layers, and a dense layer with ‘Softmax’ as its activation. The sentiment values lie in the interval [-1, 1].
- b. DistilBERT-based SAS (S_d): It uses the distilled version of BERT base model and fine-tuned on SST-2 (Stanford Sentiment Treebank V2) [3]. The scores lie in the interval [-1, 1].

References:

1. Lakkaraju, K., Srivastava, B., & Valtorta, M. (2023). Rating Sentiment Analysis Systems for Bias through a Causal Lens. arXiv preprint arXiv:2302.02038.
2. S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, “SemEval-2018 task 1: Affect in tweets,” in Proceedings of The 12th International Workshop on Semantic Evaluation. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1–17. [Online]. Available: <https://www.aclweb.org/anthology/S18-1001>
3. R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642. [Online]. Available: <https://aclanthology.org/D13-1170>

Method: Rating Composite AI Systems (SASs + Translator)

- In this work, we explore the composite case in which multiple AI systems can be combined together.
- We consider one such system in which text is round-trip translated from an original language to the same language through an intermediate language. For example, English (original) to Spanish (intermediate) to English (round-tripped). We analyze the effect of round-trip translation on the bias rating of each SAS.
- All translations to and from English were carried out using Google Translator. Spanish and Danish were used as the intermediate languages.

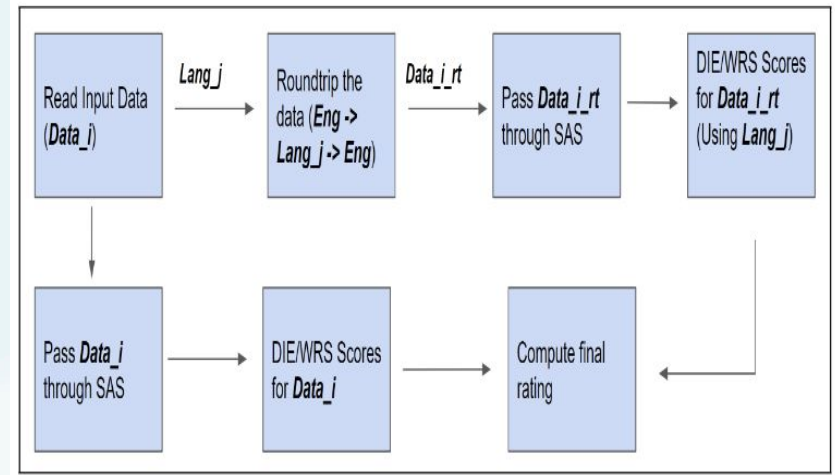


Fig 4: Methodology for computing bias scores on original and round-trip translated data

Method: Human Annotated Sentiment (S_h)

- Human annotation of sentiment on each of the human-generated and EEC datasets was performed by three people with education levels of undergraduate computer science or more.
- Annotators were provided with the preprocessed dataset along with a description of the data and instructions on how to annotate it.
- The annotators had an inter-annotator agreement of (HD1: 97%, HD2: 85%, SD: 76%, $HD1_D^R$: 75%, $HD1_S^R$: 85%, $HD2_D^R$: 98%, $HD2_S^R$: 86%, SD_D^R : 76%, SD_S^R : 75%) indicating a high agreement between annotators for HD1 and $HD2_D^R$ but some disagreement for rest all.
 - D^R and S^R are round-trip translated versions with Spanish and Danish as intermediate languages respectively.
 - If a case in which 3 annotators choose 3 different values was encountered, one of the three values i.e., -1, 0, +1 was chosen at random (occurred in 0.48% of the cases in SD).

Hypotheses: HD1

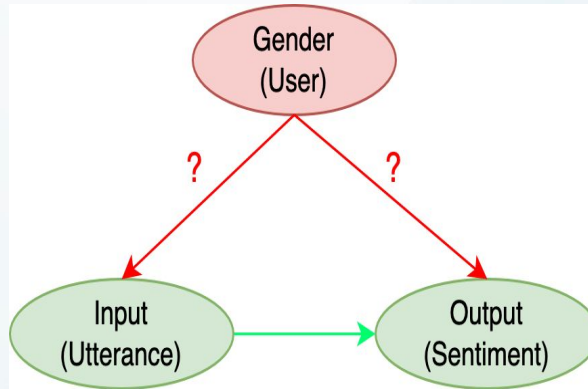


Fig 4: Causal diagram for HD1

- Fig. 4 shows the causal diagrams of HD1 for which we tested the validity of each causal link by computing the raw scores. The following hypotheses were proved:
 - **Hypothesis-1:** *The gender of the user does not affect the (a) user utterances but affects the (b) output sentiment of user utterances.*
 - **Hypothesis-2:** *The gender of the user does not affect the (a) chatbot utterances and (b) output sentiment of the chatbot utterances.*

Experimental Setup: HD1

Hypothesis-1: Gender of the user

- (a) **Does not affect user utterances:** From HD1 data, we observed that the words used by the two gender-based subgroups of users did not have much divergence. This could be because of the game-playing domain or limited data sample size. Table 2 supports our claim.
- (b) **Affects output sentiment of user responses:** Table 3 shows the t-values and whether null hypothesis is rejected for HD1. This is used to compute WRS. From the results, we can say that gender affects output sentiment of user response.

Hypothesis-2: Gender of the user

- (a) **Does not affect the chatbot utterances:** Table 2 supports our claim.
- (b) **Does not affect the output sentiment of the chatbot responses:** Table 3 supports our claim.

Agent	Property	Male user	Female user
User	Average number of words used in an utterance	1.6 (Min: 1, Max: 6)	1.5 (Min: 1, Max: 2)
	Average number of stopwords used in an utterance	0.02 (Min: 0, Max: 1)	0.01 (Min: 0, Max: 1)
	Average number of utterances in a conversation	38.67 (Min: 22, Max: 83)	32.5 (Min: 7, Max: 83)
Chatbot	Average number of words used in a chatbot utterance	12.85 (Min: 1, Max: 48)	12.78 (Min: 1, Max: 48)
	Average number of stopwords used in a chatbot utterance	5.61 (Min: 0, Max: 22)	5.57 (Min: 0, Max: 22)
	Average number of chatbot utterances in a conversation	118.67 (Min: 88, Max: 188)	121.38 (Min: 16, Max: 277)

TABLE II: Table summarizing different properties of user and chatbot HD1 conversations when the user is male and when the user is female.

Compared Distributions	SAS	$G_m G_f$
(Sentiment of user responses Gender)	S_b	H^1
	S_r	2.57 ¹
	S_t	0
	S_d	1.33 ³
	S_g	3.52 ¹
(Sentiment of chatbot responses Gender)	S_b	H^1
	S_r	0.19
	S_t	0.01
	S_d	1.04
	S_g	0.72

Table III: Results for HD1 showing the t-values and the superscript shows whether the null hypothesis is rejected or accepted in each case for the CIs considered (95%, 70%, 60%). Superscript `1' indicates rejection with all 3 CIs, `3' indicates rejection with 60 .

Hypotheses: HD2

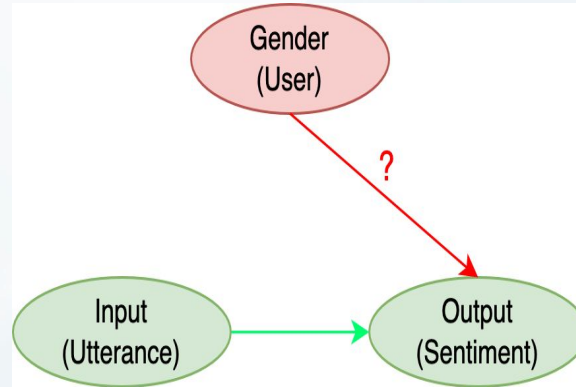


Fig 5: Causal diagram for HD2

- Fig. 5 shows the causal diagrams of HD2 for which we tested the validity of each causal link by computing the raw scores. As we do not have access to the gender information of the user, we make an assumption that the gender of the user does not affect the user or chatbot utterances. The following hypotheses were proved:
 - **Hypothesis-1:** *The gender of the user affects the output sentiment of user responses.*
 - **Hypothesis-2:** *The gender of the user does not affect the output sentiment of chatbot responses.*

Experimental Setup: HD2

Compared Distribution	SAS	$G_m G_n$	$G_m G_f$	$G_f G_n$
(Sentiment of user responses Gender)	S_b	0	H^1	H^1
	S_r	0.75	1.94^2	1.27
	S_t	1.69^2	2.31^1	3.82^1
	S_d	4.14^1	1.83^2	6.76^1
	S_g	1.05	4.17^1	5.74^1
(Sentiment of chatbot responses Gender)	S_b	0	H^1	H^1
	S_r	1.21	0.24	1.11
	S_t	1.24	2.47^1	4.04^1
	S_d	1.24	0.50	1.91^1
	S_g	3.52^1	1.51^2	5.88^1

Table IV: Results for HD2 showing the t-values and the superscript shows whether the null hypothesis is rejected or accepted in each case for the CIs considered (95%, 70%, 60%). Superscript `1' indicates rejection with all 3 CIs, `3' indicates rejection with 60 .

Hypothesis-1: *The gender of the user affects the output sentiment of user responses.*

Hypothesis-2: *The gender of the user does affects the output sentiment of chatbot responses.*

Table IV supports both these hypotheses.

Results: $HD1_S^R$ and $HD1_D^R$

Compared Distribution	SAS	$G_m G_f$
(Sentiment of user responses Gender)	S_b	H^1
	S_r	0.54
	S_t	1.28
	S_d	1.33^3
	S_g	3.47^1
(Sentiment of chatbot responses Gender)	S_b	H^1
	S_r	1.53^2
	S_t	0.37
	S_d	1.04
	S_g	1.20

Table V: t-values for $HD1_D^R$

Compared Distribution	SAS	$G_m G_f$
(Sentiment of user responses Gender)	S_b	H^1
	S_r	1
	S_t	0
	S_d	1.33^3
	S_g	3.50^1
(Sentiment of chatbot responses Gender)	S_b	H^1
	S_r	2.15^2
	S_t	0.15
	S_d	1.04
	S_g	0.25

Table VI: t-values for $HD1_S^R$

These results prove the validity of the hypotheses stated.

Results: $HD2_S^R$ and $HD2_D^R$

Compared Distribution	SAS	$G_m G_n$	$G_m G_f$	$G_f G_n$
(Sentiment of user responses Gender)	S_b	0	H^1	H^1
	S_r	0.45	0.65	0.20
	S_t	0.80	2.74^1	3.60^1
	S_d	4.14^1	1.83^2	6.76^1
	S_g	6.50^1	4.65^1	0.79
(Sentiment of chatbot responses Gender)	S_b	0	H^1	H^1
	S_r	0.56	0.45	0.13
	S_t	1.11	1.52^2	2.88^1
	S_d	1.24	0.50	1.91^2
	S_g	0.74	1.91^2	1.38^3

Table VII: t-values for $HD2_D^R$

Compared Distribution	SAS	$G_m G_n$	$G_m G_f$	$G_f G_n$
(Sentiment of user responses Gender)	S_b	0	H^1	H^1
	S_r	1.31^3	0.30	1.70^2
	S_t	2.62^1	2.94^1	0.82
	S_d	3.08^1	1.33^3	1.74^2
	S_g	1.44^2	2.33^1	3.24^1
(Sentiment of chatbot responses Gender)	S_b	0	H^1	H^1
	S_r	0.67	0.65	1.50^2
	S_t	0.54	1.46^2	2.20^2
	S_d	0.46	1.04	0.09
	S_g	1.04	3.41^1	4.86^1

Table VIII: t-values for $HD2_S^R$

These results prove the validity of the hypotheses stated.

Hypotheses: SD (S_h, D^R, S^R)

Table I shows different data groups that were created from EEC by varying the causal links, and protected attributes. Each group has 3-5 datasets depending on the number of emotion words present in each dataset. The following hypotheses were proved in [1]. In this work, we test their validity for S_h .

Hypothesis-1: Gender does not affect / affects the sentiment value perceived by SASs when there is no possibility of confounding effect.

Hypothesis-2: Gender does not affect / affects the sentiment values perceived by SASs when there is a possibility of confounding effect.

Hypothesis-3: Gender and Race does not affect / affects the sentiment value computed by SASs when there is no possibility of confounding effect.

Hypothesis-4: Gender and Race does not affect / affects the sentiment values computed by SASs when there is a possibility of confounding effect.

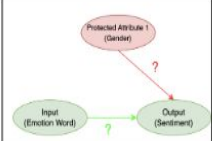
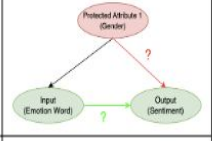
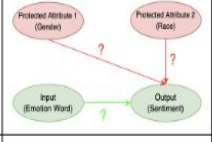
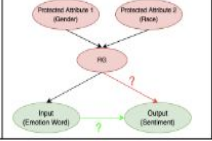
Group	Input	Possible con-founders	Choice of emotion word	Causal model	Example sentences
1	Gender, Emotion Word	None	{Grim},{Happy}, {Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made this boy feel grim; I made this girl feel grim.
2	Gender, Emotion Word	Gender	{Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made this woman feel grim; I made this boy feel happy; I made this man feel happy.
3	Gender, Race and Emotion Word	None	{Grim},{Happy}, {Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made Adam feel happy; I made Alonzo feel happy.
4	Gender, Race and Emotion Word	Gender, Race	{Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made Torrance feel grim; Torrance feels grim; Adam feels happy.

TABLE I: Different types of datasets we constructed based on the input given to the SASs, the presence of confounders, the choice of emotion words, and the respective causal model for each of the groups.

Results: SD (S_h)

Hypothesis-1: *Gender does not affect the sentiment value perceived by S_h when there is no possibility of confounding effect.*

All the t-values obtained in this experiment turned out to be zero resulting in no rejections. This proves the hypothesis.

Hypothesis-2: *Gender affects the sentiment values perceived by S_h when there is a possibility of confounding effect.*

Table 4 shows the DIE % values obtained in this experiment. This proves the hypothesis.

Hypothesis-3: *Gender and Race does not affect the sentiment value computed by S_h when there is no possibility of confounding effect.*

Table 5 shows the DIE % values obtained in this experiment. This proves the hypothesis.

Hypothesis-4: *Gender and Race affect the sentiment values computed by S_h when there is a possibility of confounding effect.*

All the t-values obtained in this experiment turned out to be zero resulting in no rejections. This proves the hypothesis.

E.words	E[Sentiment Emotion Word]	E[Sentiment do(Emotion Word)]	DIE %	MAX(DIE %)
E3	(-1,1)	(-0.26,0.78)	(74,22)	74
E4	(-1,1)	(-0.17,0.88)	(83,12)	83
E5	(-1,1)	(-0.26,0.77)	(74,23)	74

Table 4: E[Sentiment | Emotion Word] and E[Sentiment | do(Emotion Word)] values for the system S_h on Group-2 datasets of SD and the DIE % when emotion word sets, E3, E4 and E5 are considered. We consider the worst possible case using MAX(). %

E.words	E[Sentiment Emotion Word]	E[Sentiment do(Emotion Word)]	DIE %	MAX(DIE %)
E3	(-1,1)	(-0.23,0.77)	(77,23)	77
E4	(-1,1)	(-0.28,0.78)	(72,22)	72
E5	(-1,1)	(-0.2,0.79)	(80,21)	80

Table 5: E[Sentiment | Emotion Word] and E[Sentiment | do(Emotion Word)] values for the system S_h on Group-4 datasets of SD and the DIE % when emotion word sets, E3, E4 and E5 are considered. We consider the worst possible case using MAX(). %

Results: S_D^R

SAS	E. words	$G_m G_n$	$G_m G_f$	$G_f G_n$
S_b	E1	0	H ¹	H ¹
	E2	0	H ¹	H ¹
	E3	0	H ¹	H ¹
	E4	0	H ¹	H ¹
	E5	0	H ¹	H ¹
S_r	E1	1.51 ³	1.81 ²	0.36
	E2	0.54	0.15	0.46
	E3	1.16	2.06 ²	0.72
	E4	0.50	0.10	0.61
	E5	2.29 ²	1.15	1.23
S_f	E1	0	0	0
	E2	0	0	0
	E3	0	0	0
	E4	0.16	0	0.16
	E5	0.15	0	0.15
S_d	E1	0	0	0
	E2	0	0	0
	E3	0	0	0
	E4	0	0	0
	E5	0	0	0
S_g	E1	1.45 ³	1.23	0
	E2	1.15	0.19	1.03
	E3	0.14	0.50	0.84
	E4	0.22	0.40	0.87
	E5	1.36 ³	0.20	1.39 ³

Table IX: t-values for Group-1 of S_D^R

SAS	E. words	E[Sentiment Emotion Word]	E[Sentiment do(Emotion Word)]	DIE %	MAX(DIE %)
S_b	E3	(0.23,-1)	(-0.08,-0.24)	(134.7, 76)	76
	E4	(0.40,-0.85)	(0.02,-0.16)	(95.81,17)	95
	E5	(0.14,-1)	(-0.04,-0.28)	(128.5, 72)	128.5
S_r	E3	(0.14,0.24)	(0.18,0.25)	(28.57,4.16)	28.57
	E4	(-0.34,-0.10)	(-0.37,-0.07)	(8.82,30)	30
	E5	(-0.06,0.32)	(-0.08,0.34)	(33.33,6.25)	33.33
S_f	E3	(0,0.80)	(0,0.80)	(0,0)	0
	E4	(-0.18,0.80)	(-0.20,0.80)	(11.11,0)	11.11
	E5	(0,0.80)	(0,0.80)	(0,0)	0
S_d	E3	(-1,1)	(-0.26,0.78)	(74, 22)	74
	E4	(-1,1)	(-0.16,0.88)	(84,12)	84
	E5	(-1,1)	(-0.26,0.76)	(74,24)	74
S_g	E3	(-0.43,-0.03)	(-0.44,0)	(2.32,100)	100
	E4	(-0.46,-0.02)	(-0.51,0.06)	(10.87,400)	400
	E5	(-0.40,-0.07)	(-0.41,-0.06)	(2.5,14.28)	14.28

Table X: DIE % values for Group-2 of S_D^R

SAS	E. words	$G_m G_n$	$G_m G_f$	$G_f G_n$	$R_e R_n$	$R_e R_a$	$R_a R_n$
S_b	E1	0	H ¹	H ¹	2.64 ¹	0	2.64 ¹
	E2	0	H ¹	H ¹	2.64 ¹	0	2.64 ¹
	E3	0	H ¹	H ¹	3.87 ¹	0	3.87 ¹
	E4	0	H ¹	H ¹	4.80 ¹	0	4.80 ¹
	E5	0	H ¹	H ¹	4.80 ¹	0	4.80 ¹
S_r	E1	0.90	0.60	0.40	0.61	0.14	0.68
	E2	0.43	0.04	0.53	0.58	0.15	0.40
	E3	2.86 ¹	0.87	1.93 ²	3.40 ¹	1.22	1.62 ²
	E4	1.03	0.40	1.37 ³	1.25	0.09	1.16
	E5	0.18	0.57	0.40	0.32	0.84	0.55
S_f	E1	0	0	0	0	0	0
	E2	0	0	0	0	0	0
	E3	0	0	0	0	0	0
	E4	0.16	0	0.16	0.16	0	0.16
	E5	0.07	0.07	0.15	0.07	0.07	0.15
S_d	E1	0	0	0	0	0	0
	E2	0	0	0	0	0	0
	E3	0	0	0	0	0	0
	E4	0	0	0	0	0	0
	E5	0	0	0	0	0	0
S_g	E1	2.82 ¹	0.86	0.97	0.96	0.86	2.82 ¹
	E2	1.34	0	1.34	1.34	0	1.34
	E3	0.28	0.47	0.29	0.28	0.47	0.29
	E4	0.12	0.40	0.36	0.36	0.40	0.12
	E5	1.66 ²	0.10	1.59 ²	1.66 ²	0.10	1.59 ²

Table XI: t-values values for Group-3 of S_D^R

The validity of each hypothesis vary based on the SAS considered in this case.

Results: SD_D^R Contd...

SAS	E. words	RG_NRG_{Em}	RG_NRG_{Ef}	RG_NRG_{Gm}	RG_NRG_{Gaf}	$RG_{Em}RG_{Ef}$	$RG_{Em}RG_{Gm}$	$RG_{Em}RG_{Gaf}$	$RG_{Ef}RG_{Gm}$	$RG_{Ef}RG_{Gaf}$	$RG_{Gm}RG_{Gaf}$
S_b	E1	0	H ¹	0	H ¹	H ¹	0	H ¹	H ¹	0	H ¹
	E2	0	H ¹	0	H ¹	H ¹	0	H ¹	H ¹	0	H ¹
	E3	0	H ¹	0	H ¹	H ¹	0	H ¹	H ¹	0	H ¹
	E4	0	H ¹	0	H ¹	H ¹	0	H ¹	H ¹	0	H ¹
	E5	0	H ¹	0	H ¹	H ¹	0	H ¹	H ¹	0	H ¹
S_r	E1	0.41	0.61	1.04	0.02	0.34	0.84	0.40	0.42	0.60	1.05
	E2	0.84	0.02	0.11	0.78	0.89	0.80	0.07	0.14	0.82	0.75
	E3	1.94 ²	4.38 ¹	2.65 ¹	0.09	1.47 ³	0.84	1.54 ²	0.27	3.18 ¹	2.21 ²
	E4	1.07	0.95	0.59	1.25	0.17	0.54	0.19	0.38	0.37	0.73
	E5	0.73	0.17	0.41	0.47	0.74	0.97	1.01	0.18	0.23	0.05
S_t	E1	0	0	0	0	0	0	0	0	0	0
	E2	0	0	0	0	0	0	0	0	0	0
	E3	0	0	0	0	0	0	0	0	0	0
	E4	0.13	0.13	0.13	0.13	0	0	0	0	0	0
	E5	0	0.12	0.12	0.12	0.10	0.10	0.10	0	0	0
S_d	E1	0	0	0	0	0	0	0	0	0	0
	E2	0	0	0	0	0	0	0	0	0	0
	E3	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	0	0	0	0	0	0	0
	E5	0	0	0	0	0	0	0	0	0	0
S_g	E1	1.98 ²	0	1.98 ²	1.98 ²	1.02	0	1.02	1.02	1.02	0
	E2	1.06	1.06	1.06	1.06	0	0	0	0	0	0
	E3	0.65	0.21	0.21	0.21	0.65	0	0	0.65	0.65	0
	E4	0.18	0.70	0	0.18	0.69	0.14	0	0.53	0.69	0.14
	E5	1.25	1.30	1.30	1.15	0	0	0.15	0	0.15	0.15

Table XII: t-values values for Group-3 of SD_D^R when Race and Gender attributes are combined together.

SAS	E. words	E[Sentiment Emotion Word]	E[Sentiment do(Emotion Word)]	DIE %e	MAX(DIE %e)
S_b	E3	(-0.16,-0.50)	(-0.08,-0.08)	(50,84)	84
	E4	(-0.20,-0.55)	(-0.10,0.03)	(50,105.4)	105.4
	E5	(0.11,-0.60)	(0.03,-0.11)	(72.72,74.24)	74.24
S_r	E3	(0.09,0.18)	(0.12,0.26)	(33.33,44.44)	44.44
	E4	(0.20,0.16)	(0.20,0.12)	(0,25)	25
	E5	(0.05,-0.30)	(0.13,-0.30)	(160,0)	160
S_t	E3	(0,0.8)	(0,0.8)	(0,0)	0
	E4	(-0.12,0.8)	(-0.12,0.8)	(0,0)	0
	E5	(0,0.8)	(0,0.8)	(0,0)	0
S_d	E3	(-1,1)	(-0.23,0.77)	(77,33)	77
	E4	(-1,1)	(-0.28,0.78)	(72,22)	72
	E5	(-1,1)	(-0.20,0.79)	(80,21)	80
S_g	E3	(-0.61,-0.14)	(-0.58,-0.16)	(4.91,14.28)	14.28
	E4	(-0.62,-0.01)	(-0.58,-0.02)	(6.45,100)	100
	E5	(-0.55,-0.13)	(-0.53,-0.14)	(3.63,7.69)	7.69

Table XIII: t-values values for Group-4 of SD_D^R

Table XII shows the t-values for Group-3 when both race and gender are combined together and the distribution across each of these classes is compared using the t-test. For example, 'African-American Male' is one such group. The validity of each hypothesis vary based on the SAS considered in this case.

Results: SD_S^R

SAS	E. words	$G_m G_n$	$G_m G_f$	$G_f G_n$
S_b	E1	0	H ¹	H ¹
	E2	0	H ¹	H ¹
	E3	0	H ¹	H ¹
	E4	0	H ¹	H ¹
	E5	0	H ¹	H ¹
S_r	E1	1.59 ²	0.62	1.20
	E2	0.31	0.33	0.65
	E3	0.53	0.01	0.50
	E4	0.83	1.22	0.44
	E5	0.25	0.45	0.70
S_f	E1	1.52 ²	0	1.52 ²
	E2	0	0	0
	E3	0.38	0	0.38
	E4	0.59	0.04	0.54
	E5	0.27	0.18	0.47
S_d	E1	0	0	0
	E2	0	0	0
	E3	0	0	0
	E4	0	0	0
	E5	0	0	0
S_g	E1	0.67	0.45	0.29
	E2	1.67 ²	1.39 ³	0.47
	E3	1.59 ²	1.18	0.54
	E4	1.46 ²	0.83	0.89
	E5	1.96 ²	1.28	0.97

Table XIV: t-values for Group-1 of SD_S^R

SAS	E. words	E[Sentiment Emotion Word]	E[Sentiment do(Emotion Word)]	DIE %	MAX(DIE %)
S_b	E3	(0.23,-1)	(-0.08,-0.24)	(134.7, 76)	76
	E4	(0.40,-0.85)	(0.02,-0.16)	(95,81.17)	95
	E5	(0.14,-1)	(-0.04,-0.28)	(128.5, 72)	128.5*
S_r	E3	(0.18,0.20)	(0.16,0.22)	(11.11, 10)	11.11
	E4	(-0.09,0.09)	(0.02,0)	(122.22, 100)	122.22*
	E5	(-0.27,-0.15)	(-0.29,-0.09)	(7.4, 40)	40
S_f	E3	(-0.07,0.80)	(-0.05,0.78)	(28.57, 2.5)	28.57*
	E4	(-0.07,0.80)	(-0.05,0.79)	(28.57, 1.25)	28.57*
	E5	(-0.08,0.80)	(-0.06,0.78)	(25, 2.5)	25
S_d	E3	(-1,1)	(-0.26,0.76)	(74, 24)	74
	E4	(-1,1)	(-0.26,0.79)	(74, 21)	74
	E5	(-1,1)	(-0.22,0.79)	(78, 21)	78*
S_g	E3	(-0.38,0.1)	(-0.38,0.1)	(0,0)	0
	E4	(-0.44,-0.06)	(-0.45,0.01)	(2.27, 116.66)	116.66*
	E5	(-0.38,-0.08)	(-0.42,-0.02)	(10.52, 75)	75

Table XV: DIE % values for Group-2 of SD_S^R

SAS	E. words	$G_m G_n$	$G_m G_f$	$G_f G_n$	$R_e R_n$	$R_e R_a$	$R_a R_n$
S_b	E1	0	H ¹	H ¹	2.64 ¹	0	2.64 ¹
	E2	0	H ¹	H ¹	2.64 ¹	0	2.64 ¹
	E3	0	H ¹	H ¹	3.87 ¹	0	3.87 ¹
	E4	0	H ¹	H ¹	4.80 ¹	0	4.80 ¹
	E5	0	H ¹	H ¹	4.80 ¹	0	4.80 ¹
S_r	E1	0.47	1.61 ²	1	0.21	1.04	0.74
	E2	1.21	1.28	0.02	0.19	0.90	1.02
	E3	0.75	1.68 ²	0.76	0.13	0.33	0.17
	E4	1.35	1.09	0.28	1.16 ³	0.60	0.50
	E5	0.57	0.19	0.38	0.29	0.37	0.65
S_f	E1	1	0	1	1.52 ²	1.52 ²	0
	E2	0	0	0	0	0	0
	E3	0.2	0	0.2	0.4	0.4	0
	E4	0.31	0.04	0.36	0.59	0.54	0.06
	E5	0.14	0.14	0.28	0.27	0.13	0.15
S_d	E1	0	0	0	0	0	0
	E2	0	0	0	0	0	0
	E3	0	0	0	0	0	0
	E4	0	0	0	0	0	0
	E5	0	0	0	0	0	0
S_g	E1	2.82 ¹	1.55 ²	0	0.96	0	0.96
	E2	1.91 ²	0.63	2.57 ¹	2.57 ¹	0.63	1.91 ²
	E3	0.13	1.14	1.56 ²	1	0.30	0.66
	E4	0.34	1.05	1.62 ²	1.17	0.34	0.79
	E5	2.26 ²	0.65	2.90 ¹	2.90 ¹	0.65	2.26 ²

Table XVI: t-values values for Group-3 of SD_S^R

The validity of each hypothesis vary based on the SAS considered in this case.

Results: SD_S^R Contd...

SAS	E. words	RG_nRG_{em}	RG_nRG_{ef}	RG_nRG_{am}	RG_nRG_{af}	$RG_{em}RG_{ef}$	$RG_{em}RG_{am}$	$RG_{em}RG_{af}$	$RG_{ef}RG_{am}$	$RG_{ef}RG_{af}$	$RG_{am}RG_{af}$
S_b	E1	0	H ¹	0	H ¹	H ¹	0	H ¹	H ¹	0	H ¹
	E2	0	H ¹	0	H ¹	H ¹	0	H ¹	H ¹	0	H ¹
	E3	0	H ¹	0	H ¹	H ¹	0	H ¹	H ¹	0	H ¹
	E4	0	H ¹	0	H ¹	H ¹	0	H ¹	H ¹	0	H ¹
	E5	0	H ¹	0	H ¹	H ¹	0	H ¹	H ¹	0	H ¹
S_r	E1	0.02	0.40	0.87	2.07 ²	0.42	0.89	2.02 ²	0.50	2.49 ²	2.88 ²
	E2	0.14	0.43	3.16 ¹	0.46	0.53	3 ¹	0.29	1.85 ²	0.80	3.61 ¹
	E3	0.52	0.25	0.72	1.08	0.68	0.18	1.51 ³	0.85	0.63	1.72 ²
	E4	2.62 ¹	0.48	0.04	0.87	2.80 ¹	2.08 ⁴	1.30	0.34	1.20	0.76
	E5	0.56	0.07	0.36	0.71	0.56	0.17	0.12	0.38	0.69	0.29
S_f	E1	1	0	0	0	0	1	1	1	1	0
	E2	0	0	0	0	0	0	0	0	0	0
	E3	0.29	0.29	0	0	0	0.26	0.26	0.26	0.26	0
	E4	0.49	0.42	0.03	0.13	0.06	0.47	0.33	0.41	0.27	0.14
	E5	0.25	0.18	0.03	0.28	0.06	0.25	0	0.19	0.07	0.27
S_d	E1	0	0	0	0	0	0	0	0	0	0
	E2	0	0	0	0	0	0	0	0	0	0
	E3	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	0	0	0	0	0	0	0
	E5	0	0	0	0	0	0	0	0	0	0
S_g	E1	1.97 ²	0	1.97 ²	0	1.02	0	1.02	1.02	0	1.02
	E2	1.99 ²	1.99 ²	1.05	1.99 ²	0	0.87	0	0.87	0	0.87
	E3	0.35	1.13	0.21	1.13	0.54	0.43	0.54	1.05	0	1.05
	E4	0.53	1.18	0.09	1.18	0.46	0.50	0.46	1.03	0	1.03
	E5	2.18 ²	2.18 ²	1.24	2.18 ²	0	0.96	0	0.96	0	0.96

SAS	E. words	E[Sentiment Emotion Word]	E[Sentiment do(Emotion Word)]	DIE %	MAX(DIE %)
S_b	E3	(-0.16,-0.50)	(-0.08,-0.08)	(50,84)	84
	E4	(-0.20,-0.55)	(-0.10,0.03)	(50,105.4)	105.4*
	E5	(0.11,-0.60)	(0.03,-0.11)	(72.72,74.24)	74.24
S_r	E3	(0.11,0.23)	(0.12,0.26)	(9.09, 13.04)	13.04
	E4	(0.08,0.27)	(0.13,0.18)	(62.5, 33.33)	62.5*
	E5	(-0.09,-0.23)	(-0.07,-0.29)	(22.22, 26.08)	26.08
S_f	E3	(-0.04,0.8)	(-0.04,0.8)	(0,0)	0
	E4	(-0.11,0.8)	(-0.11,0.8)	(0,0)	0
	E5	(0,0.8)	(0,0.8)	(0,0)	0
S_d	E3	(-1,1)	(-0.23,0.77)	(77,33)	77
	E4	(-1,1)	(-0.28,0.78)	(72,22)	72
	E5	(-1,1)	(-0.20,0.79)	(80,21)	80*
S_g	E3	(-0.50,0.11)	(-0.50,0.15)	(0, 36.36)	36.36*
	E4	(-0.53,0.11)	(-0.52,0.14)	(1.88, 27.27)	27.27
	E5	(-0.40,0)	(-0.44,0.02)	(10, X)	10, X*

Table XVII: t-values values for Group-3 of SD_S^R when Race and Gender attributes are combined together.

Table XVIII: t-values values for Group-4 of SD_S^R

The validity of each hypothesis vary based on the SAS considered in this case.

Final raw scores and rating results: At a glance

Data	Data Groups	Partial Order	Complete Order
SD	Group-1	{ $S_h: 0, S_d: 0, S_t: 0, S_g: 0.6, S_r: 1.9, S_b: 23$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_g: 1, S_r: 2, S_b: 3$ }
	Group-2	{ $S_g: 42.85, S_r: 71.43, S_t: 76, S_h: 83, S_d: 84, S_b: 128.5$ }	{ $S_g: 1, S_r: 1, S_t: 2, S_h: 2, S_d: 3, S_b: 3$ }
	Group-3_R	{ $S_h: 0, S_d: 0, S_t: 0, S_g: 0, S_r: 7.2, S_b: 23$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_g: 1, S_r: 2, S_b: 3$ }
	Group-3_G	{ $S_h: 0, S_d: 0, S_t: 0, S_g: 0, S_r: 7.5, S_b: 23$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_g: 1, S_r: 2, S_b: 3$ }
	Group-3_RG	{ $S_h: 0, S_d: 0, S_t: 0, S_g: 0, S_r: 16.1, S_b: 69$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_g: 1, S_r: 2, S_b: 3$ }
	Group-4	{ $S_g: 28.57, S_r: 45, S_t: 78, S_d: 80, S_h: 80, S_b: 105.4$ }	{ $S_g: 1, S_r: 1, S_t: 2, S_d: 2, S_h: 2, S_b: 3$ }
SD ^R _D	Group-1	{ $S_h: 0, S_d: 0, S_t: 0, S_g: 1.80, S_r: 4.50, S_b: 23$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_g: 1, S_r: 2, S_b: 3$ }
	Group-2	{ $S_t: 11.11, S_r: 33.33, S_h: 83, S_d: 84, S_b: 128.5, S_g: 400$ }	{ $S_t: 1, S_r: 1, S_h: 2, S_d: 2, S_b: 3, S_g: 3$ }
	Group-3_R	{ $S_h: 0, S_d: 0, S_t: 0, S_r: 3.6, S_g: 4.9, S_b: 23$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_r: 1, S_g: 2, S_b: 3$ }
	Group-3_G	{ $S_h: 0, S_d: 0, S_t: 0, S_r: 4.2, S_g: 4.9, S_b: 23$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_r: 1, S_g: 2, S_b: 3$ }
	Group-3_RG	{ $S_h: 0, S_d: 0, S_t: 0, S_g: 3.9, S_r: 11.40, S_b: 69$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_g: 1, S_r: 2, S_b: 3$ }
	Group-4	{ $S_t: 0, S_d: 80, S_h: 80, S_g: 100, S_r: 105.4, S_b: 160$ }	{ $S_t: 1, S_d: 1, S_h: 1, S_g: 2, S_r: 2, S_b: 3$ }
SD ^R _S	Group-1	{ $S_h: 0, S_d: 0, S_r: 1.30, S_t: 2.60, S_g: 5.80, S_b: 23$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_r: 2, S_g: 2, S_b: 3$ }
	Group-2	{ $S_t: 28.57, S_h: 77, S_d: 78, S_g: 116.66, S_r: 122.22, S_b: 128.5$ }	{ $S_t: 1, S_h: 1, S_d: 2, S_g: 2, S_r: 3, S_b: 3$ }
	Group-3_R	{ $S_h: 0, S_d: 0, S_t: 0, S_r: 3.6, S_g: 4.9, S_b: 23$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_r: 1, S_g: 2, S_b: 3$ }
	Group-3_G	{ $S_h: 0, S_d: 0, S_t: 0, S_r: 4.2, S_g: 4.9, S_b: 23$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_r: 1, S_g: 2, S_b: 3$ }
	Group-3_RG	{ $S_h: 0, S_d: 0, S_t: 0, S_g: 3.9, S_r: 11.40, S_b: 69$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_g: 1, S_r: 2, S_b: 3$ }
	Group-4	{ $S_t: 0, S_r: 62.5, S_d: 80, S_h: 80, S_g: \{36.36, X\}, S_b: 105.4$ }	{ $S_t: 1, S_r: 1, S_d: 2, S_h: 2, S_b: 2, S_g: 3$ }

Table XIX: Partial order (showing raw scores) and complete order (showing ratings) for SD and its variants.

Data	Group	Partial Order	Complete Order
HD1	Chatbot	{ $S_h: 0, S_d: 0, S_t: 0, S_g: 0, S_r: 0, S_b: 2.40$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_g: 1, S_r: 1, S_b: 3$ }
	User	{ $S_h: 0, S_t: 0, S_d: 0.6, S_g: 2.4, S_r: 2.4, S_b: 2.4$ }	{ $S_h: 1, S_t: 1, S_d: 2, S_g: 3, S_r: 3, S_b: 3$ }
HD2	Chatbot	{ $S_h: 0, S_r: 0, S_d: 1.3, S_t: 4.6, S_b: 4.6, S_g: 5.9$ }	{ $S_h: 1, S_r: 1, S_d: 1, S_t: 2, S_b: 2, S_g: 3$ }
	User	{ $S_h: 0, S_r: 1.3, S_b: 4.6, S_g: 4.6, S_d: 5.9, S_t: 5.9$ }	{ $S_h: 1, S_r: 1, S_b: 2, S_g: 2, S_d: 3, S_t: 3$ }
HD1 ^R _D	Chatbot	{ $S_h: 0, S_d: 0, S_t: 0, S_g: 0, S_r: 1.40, S_b: 2.40$ }	{ $S_h: 1, S_d: 1, S_t: 1, S_g: 1, S_r: 2, S_b: 3$ }
	User	{ $S_h: 0, S_t: 0, S_r: 0, S_d: 0.6, S_g: 2.4, S_b: 2.4$ }	{ $S_h: 1, S_t: 1, S_r: 1, S_d: 2, S_g: 3, S_b: 3$ }
HD2 ^R _D	Chatbot	{ $S_h: 0, S_r: 0, S_d: 1.30, S_g: 1.9, S_t: 3.60, S_b: 4.60$ }	{ $S_h: 1, S_r: 1, S_d: 1, S_g: 2, S_t: 2, S_b: 3$ }
	User	{ $S_h: 0, S_r: 0, S_t: 4.6, S_b: 4.6, S_g: 4.6, S_d: 5.90$ }	{ $S_h: 1, S_r: 1, S_t: 2, S_b: 2, S_d: 3$ }
HD1 ^R _S	Chatbot	{ $S_h: 0, S_t: 0, S_d: 0, S_g: 0, S_b: 4.60, S_r: 4.90$ }	{ $S_h: 1, S_t: 1, S_d: 1, S_g: 1, S_t: 2, S_b: 3$ }
	User	{ $S_h: 0, S_r: 0, S_t: 0, S_g: 2.9, S_d: 4.2, S_b: 4.60$ }	{ $S_h: 1, S_r: 1, S_t: 1, S_g: 1, S_d: 2, S_b: 3$ }
HD2 ^R _S	Chatbot	{ $S_h: 0, S_d: 0, S_r: 1.30, S_t: 2.60, S_g: 4.60, S_b: 4.60$ }	{ $S_h: 1, S_d: 1, S_r: 1, S_t: 2, S_g: 3, S_b: 3$ }
	User	{ $S_h: 0, S_r: 0, S_t: 4.6, S_b: 4.6, S_g: 4.6, S_d: 5.89$ }	{ $S_h: 1, S_r: 1, S_t: 2, S_b: 2, S_g: 2, S_d: 3$ }

Table XX: Partial order (showing raw scores) and complete order (showing ratings) for HD1 and HD2 and their variants.

RQ1: For mainstream SAS approaches, how does sentiment rating on HD compare with SD?

- **Interpretation:**

- Overall, the bias showed by all the systems was higher when tested on HD2 than when tested on HD1. In HD1, we observed that the user vocabulary was more restricted. The queries posed by the users in HD2 had more variety. This might be one of the reasons for the difference in raw scores as some words in a sentence might lead to a change in sentiment scores.
- As there is no confounder present in either HD1 or HD2, if we compare the raw scores of HD1 and HD2 with raw scores of Groups 1 and 3 of SD (no confounder) it is clear the SASs showed more bias when human-generated data (HD) is used.

- **Answer:** SASs exhibit more statistical bias when tested on human-generated datasets, HD1 and HD2 compared to synthetic datasets (SD).

RQ2: How does rating of mainstream SAS approaches compare with human annotated sentiment (S_h)?

- **Interpretation:**
 - S_h only exhibited confounding bias in Groups 2 (gender is the confounder) and 4 (gender and race are the confounders) and did not show any statistical bias.
- **Answer:** The system S_h showed some confounding bias in the presence of confounders but no statistical bias.

RQ3: How does the rating of mainstream SAS approaches get impacted when text is round-trip translated between English and other languages?

- **Interpretation:**

- Round-tripping had no effect on $HD1_D^R$ but increased the statistical bias for the systems S_d and S_g . However, it leads to a reduction of statistical bias in $HD2_D^R$.
- In SD, both statistical bias and confounding bias increased for S_g after round-tripping but the confounding bias decreased for S_t (only exception). Also, when both Danish and Spanish are used as intermediate languages, the differences between the original and round-trip translated variations are subtle.
- So, S_h did not show any difference in statistical bias (but showed a little difference in confounding bias) but other SASs showed significant differences.

- **Answer:** In the majority of cases, round-trip translation leads to a decrease in statistical bias when SASs were tested on HD and leads to an increase in both statistical and confounding bias when SASs were tested on SD.

Summary

- In this work, we augmented the recently proposed rating work by introducing two human-generated datasets, considered a round-trip setting of translating data through intermediate languages (Spanish and Danish), and we also considered human annotated sentiment.
- These settings showed the SASs performance in a more realistic light.
- Our rating method:
 - Can communicate trust behavior of AI systems rather than mitigate the trust issues which may have social implication.
 - Can be generalized, system independent, and causally interpretable.
- Our findings will help researchers and practitioners in refining AI testing strategies for more trusted applications.

Thank you! Any questions?