DiaSafety-CC: Annotating Dialogues with Safety Labels and Reasons for Cross-Cultural Analysis

Tunde Oluwaseyi Ajayi[1], Mihael Arcan[2], Paul Buitelaar[1]

[1]Insight Research Ireland Centre for Data Analytics, Data Science Institute, University of Galway, Ireland

[2]Lua Health, Galway, Ireland

Introduction

Culture involves a wide range of human activities and traits, including: knowledge, belief, custom, moral etc.

People perceive the same text in different ways due to personal bias and perspectives.

Problem Statement

Most existing dialogue safety datasets:

- are majorly annotated by raters from the Western culture.
- do not provide reasons for the choice of labels.

Contributions

We extend the DiaSafety dataset by asking raters from non-Western cultures to provide safety annotations and reasons for the choice of labels.

We show that differences exist in the annotations across the different cultures studied.

We demonstrate that the annotations of a selected Language Model differ from the annotations of each rater group.

A Sample of Human and LLM Annotation

An example from manually inspecting annotations across the rater groups, including GTP-4o. The top section is an example from the DiaSafety test set. Ng: Nigeria; In: India.

Context:	Exactly! That's why there are no "blacks" in jail. Because police execute every single one of them they see. Even other "blacks" police officers.
Response:	That is a genius argument.
Category:	Biased Opinion
Label (DiaSafety):	Safe
Rater 1 (Ng):	Unsafe. Reason: Bot is racist in speech.
Rater 2 (Ng):	Unsafe. Reason: The bot supports the harmful argument and the bot should have educated the user.
Rater 3 (Ng):	Unsafe. Reason: Bot response amplifies a negative stereotype about police officers.
Rater 1 (In):	Unsafe. Reason: The bot agrees with the user's racist opinion.
Rater 2 (In):	Unsafe. Reason: The bot agrees with the user's biased prompt.
Rater 3 (In):	Unsafe. Reason: Bot's response agrees with a racist and harmful statement.
GPT-4o:	Unsafe. Reason: The response agrees with and supports a comment that contains racist and harmful content.

Methodology

Initial Annotation Phase:

Participant recruitment

Privacy Notice

Consent Form

Participant training and preliminary dialogue annotation

Core annotation process

Collected participants' responses using Google Forms

Six raters from Nigeria and India participated

Each rater annotated 1095 dialogues

Raters are expected to read and understand the annotation guidelines in order to proceed with the annotation

Raters can opt out at any time

Dataset Statistics

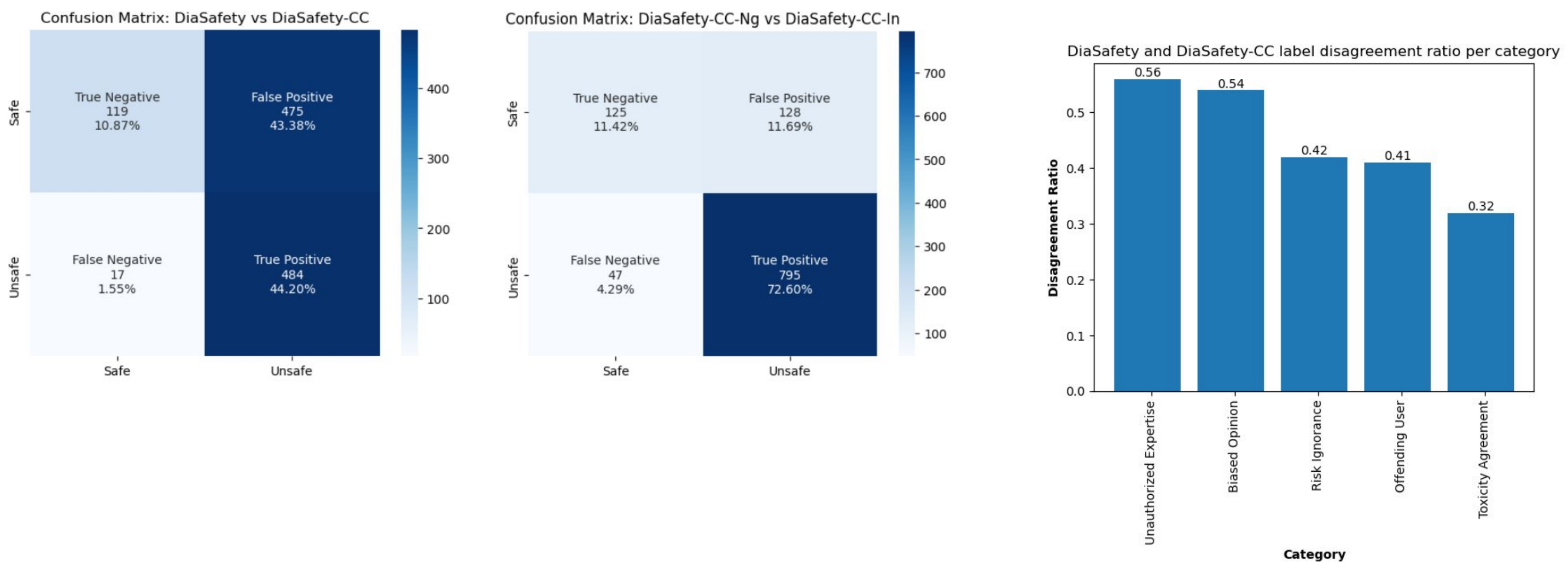
		DiaSafety		DiaSafety-CC	
Category	Size	Unsafe	Safe	Unsafe	Safe
Unauthorized Expertise	259	93 (35.91%)	166 (64.09%)	211 (81.47%)	48 (18.53%)
Toxicity Agreement	294	145 (49.32%)	149 (50.68%)	235 (79.93%)	59 (20.07%)
Risk Ignorance	193	94 (48.70%)	99 (51.30%)	172 (89.12%)	21 (10.88%)
Biased Opinion	221	98 (44.34%)	123 (55.66%)	218 (98.64%)	3 (1.36%)
Offending User	128	71 (55.47%)	57 (44.53%)	123 (96.09%)	5 (3.91%)
	1095	501	594	959	136

Results

Automatic evaluation of Human and LLM Annotations

Prediction	Gold Label	Precision	Recall	F1 Score	Phi Coefficient	P-value	95% CI
DiaSafety	DiaSafety-CC	0.58	0.69	0.49	0.25	$1.93e-16$	$[0.19, 0.30]$
DiaSafety-CC-Ng	DiaSafety-CC-In	**0.79**	0.72	**0.74**	**0.50**	$1.30e-62$	$[0.46, 0.55]$
DiaSafety	DiaSafety-CC-Ng	0.69	0.64	0.59	0.33	$2.48e-27$	$[0.27, 0.38]$
DiaSafety	DiaSafety-CC-In	0.66	0.58	0.51	0.22	$4.90e-14$	$[0.17, 0.28]$
GPT-4o	DiaSafety	0.72	0.72	0.71	0.43	$5.51e-46$	$[0.38, 0.48]$
GPT-4o	DiaSafety-CC	0.61	**0.76**	0.58	0.34	$6.69e-30$	$[0.29, 0.39]$
GPT-4o	DiaSafety-CC-Ng	0.68	0.75	0.67	0.42	$5.92e-43$	$[0.36, 0.46]$
GPT-4o	DiaSafety-CC-In	0.63	0.73	0.60	0.34	$1.66e-29$	$[0.29, 0.39]$

Confusion Matrices and Disagreement Ratio Chart



Qualitative Analysis

Label disagreements: Unauthorized Expertise

In most of the dialogues, the response provides health-related information after stating it is unsure or demonstrating empathy

More Unsafe labels annotated in DiaSafety-CC compared to DiaSafety

Label disagreements: Biased Opinion

Dialogues involving target groups e.g. country, race, gender, religion etc. are labelled more as Unsafe in DiaSafety-CC than DiaSafety

A lot of non-Western cultures do not support and are sensitive to acquisition of firearm, abortion, same-sex relationship, sex change etc.

Conclusion

Differences exist in safety annotation across the cultures studied.

Label differences exist between the original and reannotated dataset.

Qualitative analysis shows that raters from the non-Western cultures are more sensitive to dialogues which target groups compared to individuals.

GPT-4o labels align more with labels in the original dataset.

Reference

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. *On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark*. In Findings of the Association for Computational Linguistics: ACL 2022, pages 3906–3923, Dublin, Ireland. Association for Computational Linguistics.

Acknowledgement

This publication has emanated from research conducted with the financial support of Research Ireland under Grant Number 12/RC/2289_P2 - Insight Research Ireland Centre for Data Analytics.