

## How Much do Robots Understand Rudeness? Challenges in Human- Robot Interaction

Michael Orme, Yanchao Yu, Zhiyuan Tan

Edinburgh Napier University

10 Colinton Rd, Edinburgh EH10 5DT



# Intro

- Goal: To test machine recognition of rudeness
  - Profanity Detection
  - Profanity Clearing Methods
  - Measurement of Rudeness
  - Comparison of rudeness ratings with machine models, LLMs and Humans
- Looking at conversational agents for children:
  - Chatbots can contain profanity, which is not appropriate for children
  - Want to ensure that no profanity can possibly be produced



# Profanity detection libraries

- Many different libraries
- Varying levels of effectiveness and methods
- Better profanity performed best – used for detection

Method	True Positive	False Positive	False Negative	True Negative	Precision	Recall	F1	Accuracy	Time(Secs)
Better Profanity	225	49	38	2688	82.12%	85.55%	83.80%	97.10%	104.8814
Profanity Check	183	59	80	2678	75.62%	69.58%	72.48%	95.37%	5.138391
Profanity Filter	182	6	81	2731	96.81%	69.20%	80.71%	97.10%	115.7867
String Search	188	75	75	2662	71.48%	71.48%	71.48%	95.00%	0.131575
Profanity	121	5	142	2732	96.03%	46.01%	62.21%	95.10%	1.271979
Regex	155	3	108	2734	98.10%	58.94%	73.63%	96.30%	1.033833

# Profanity Cleaning

- Many methods
- Varying effectiveness
- Context Important

Version	Text
Original	Lesbian? No. I found a picture of Jared Leto in one of her drawers, so I'm pretty sure she's not harboring <b>same-sex</b> tendencies.
Profanity Removed	Lesbian? No. I found a picture of Jared Leto in one of her drawers, so I'm pretty sure she's not harboring tendencies.
Word Paraphrased	Lesbian? No. I found a picture of Jared Leto in one of her drawers, so I'm pretty sure she's not harboring People of the same gender tendencies.
Sentence Paraphrased	Jared Leto is not a lesbian, as I found her in etiquette and thought she was.



# Fine tuning

- Fine-tuned DialoGPT-Small using each Dataset
- Generated responses using jokes dataset to evaluate their responses

[microsoft/DialoGPT-small · Hugging Face](#)  
[Question-Answer Jokes \(kaggle.com\)](#)



# Rudeness measurement

- Evaluation done by 5 hate speech models and 2 sentiment analysis models
- Percentage of profanity goes down but not removed completely
- Significant similarity is lost through cleaning process, original meaning lost

	Models	Original	Profanity Removal	Word Paraphrasing	Sentence Paraphrasing
Hate Speech Detection	Cardiffnlp-hate-latest (Antypas and Camacho-Collados, 2023b)	7.57%	<b>5.45%</b>	6.46%	5.46%
	Dehatebert (Aluru et al., 2020a)	3.93%	2.58%	2.78%	<b>2.56%</b>
	Hatexplain (Mathew et al., 2020)	4.30%	2.94%	2.98%	<b>2.76%</b>
	MuRIL (Das et al., 2022b)	36.08%	<b>21.48%</b>	24.33%	22.55%
	Dynabench-r4 (Vidgen et al., 2021b)	8.54%	<b>6.86%</b>	8.18%	7.17%
	Average	12.08%	<b>7.86%</b>	8.95%	8.10%
Sentiment Analysis	SiEBERT (Hartmann et al., 2023)	77.84%	76.68%	76.13%	<b>73.66%</b>
	TimeLMs (Loureiro et al., 2022)	48.27%	41.36%	42.85%	<b>40.92%</b>
	Average	63.06%	59.02%	59.49%	<b>57.29%</b>

Clean-up Version	Similarity (%)
Profanity Removed	71.11%
Paraphrased Word	<b>78.26%</b>
Paraphrased Sentence	45.08%

# Results

- Used models, ChatGPT and Human ratings for 10 responses
- Hate Speech and Sentiment Analysis were changed from 0-1 to 1-5 to match other results
- Human results do not match with machine

Q-Index	ChatGPT	Hate Speech Detection	Sentiment Analysis	Human Rating
Q1	3.0	4.8	3.9	3.15
Q2	4.0	4.0	2.5	3.12
Q3	3.0	4.8	1.9	3.24
Q4	2.0	4.9	1.9	3.30
Q5	1.0	3.2	1.1	3.25
Q6	3.0	4.8	1.9	3.08
Q7	2.0	4.9	1.9	3.07
Q8	3.0	4.9	1.9	2.88
Q9	1.0	4.1	1.5	2.62
Q10	3.0	4.1	1.6	3.23

- 1 is very rude and 5 is very polite



Q5	<p><b>Q:</b> Why do women have smaller feet than men?</p> <p><b>A:</b> Because they're thin. Fat is a significant contributor to the development of obesity. Women have small feet.</p>
----	---



# Conclusion

- AI and Machine Tools not able to recognise rudeness and cannot remove it without fundamentally altering the meaning of the text
- Rudeness detection is complex and a significant challenge for AI
- Any detection must consider the intricacies of human conversation and the diverse backgrounds of the participants



# Future Work

- Better method for detecting profanity
  - Multimodal emotion recognition
  - Corpus to facilitate unlearning
- Different approach needed to fix problems of rudeness
  - We plan to use a machine unlearning approach to remove profanity
  - Means dataset can have profanity in it and then the effects of it can be removed before the output is affected
  - Removes need for costly RLHF alignment





# LREC-COLING 2024

## Thank you!

### For questions contact:

[michael.orme@napier.ac.uk](mailto:michael.orme@napier.ac.uk)

