Evaluating ChatGPT Against Functionality Tests for Hate Speech Detection

- Mithun Das, Saurabh Kumar Pandey, Animesh Mukherjee

Warning!

⚠️ This presentation contains material that many will find offensive or hateful; however this cannot be avoided owing to the nature of the work.

Working Definition of Hate Speech

"We define hate speech as an expression of direct hatred that targets a protected group or its members for being part of that group. Protected groups include those based on age, disability, gender identity, race, national or ethnic origin, religion, sex, or sexual orientation, which reflects the international legal consensus."[1]

[1] Röttger, Paul, et al. "HateCheck: Functional Tests for Hate Speech Detection Models." 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). 2021.

Mitigating Hate Speech

- Several datasets have been proposed in various languages.
- Several models have been developed to detect hate speech automatically.
- Models were evaluated by measuring their performance on the held-out test data using different metrics

Recent Advancement - ChatGPT

- Recently, pre-trained language models, such as **ChatGPT** have shown **great potential** in performing several tasks, including **hate speech detection**.
- ChatGPT can achieve an **accuracy** of approximately **80%** when compared to MTurker annotations.

While ChatGPT has shown promising results, is this model reliable and robust? Can we deploy it directly in the wild?

Research Questions

- **RQ1:** How effective is ChatGPT based on a diverse set of *functionality tests* in detecting hate speech across languages?
- **RQ2:** What are the weaknesses of ChatGPT in detecting emoji-based hate speech?

Functional testing

- Functional testing refers to the ability of a hate speech detection model to classify hateful and non-hateful posts of different types.
- Each functionality is defined by a set of test cases sharing a common gold-standard label.

Testcase	**Label**	**Expected Label**
I hate Women	**Hate**	**Hate**
I hate apples	**Nonhate**	**Nonhate**

Datasets for Testing

- HateCheck (Röttger et al., 2021)
  - 29 model functionalities motivated by a review of previous research and a series of interviews with civil society stakeholders.
- Multilingual HateCheck (MHC) (Röttger et al., 2022)
  - Extending HateCheck, MHC covers 34 functionalities across ten languages.
- HatemojiCheck (Kirk et al., 2022)
  - Provide seven functionalities covering six identities.

	Hateful	Non-hateful	Total
HateCheck	2,563	1,165	3,728
MHC	25,511	11,071	36,582
HatemojiCheck	2,654	1,276	3,930

Model Setup

- We employ the **gpt-3.5-turbo** model, a chatbot based on the GPT-3.5 language model.
- OpenAI provides several parameters, such as **max_tokens**, **temperature**, and **top_p**, which users can modify in the API request according to the task that needs to be performed.’
- As we intend to perform a binary classification task, we set the **temperature value to 0** to reduce the randomness.
- We use the default values for other controllable parameters.

Prompt Design

- A prompt is a set of instructions provided to an LLM that customizes and enhances its capabilities for a specific task.
- In our case, the purpose is to classify statements as hate speech or not. We therefore craft the following prompt:
    - *Can you determine if the given statement would be classified as hate speech according to the provided definition?* ***[hate speech definition]****. Answer only 'Yes' or 'No'. Statement:* ***<input text>***

Overall Performance

Language	% F1 (h)	% F1 (nh)	% Mac. F1
English/EN	99.7	78.6	89.2
Arabic / AR	93.3 (2.8)	49.9 (5.3)	71.6 (3.5)
Dutch / NL	98.9 (0.2)	71.4	85.1 (0.1)
French / FR	99.0 (0.2)	65.4 (0.1)	82.2 (0.2)
German / DE	99.5 (0.0)	67.8 (0.2)	83.6 (0.1)
Hindi / HI	96.3 (1.2)	38.3 (3.6)	67.3 (1.9)
Italian / IT	98.2 (0.2)	69.2	83.7 (0.1)
Mandarin / ZH	97.7 (0.5)	67.7 (0.5)	82.7 (0.5)
Polish / PL	95.7 (1.0)	67.2 (1.1)	81.5 (1.1)
Portuguese / PT	98.5	75.8	87.1
Spanish / ES	99.2	69.3 (0.2)	84.2 (0.1)
EMOJI/ EMO	88.6	76.6 (0.1)	82.6 (0.1)

ChatGPT's Performance across all the languages.

ChatGPT exhibits inferior performance for Hindi and Arabic.

Comparison with Existing Models

Language	% F1 (h)	% F1 (nh)	% Mac. F1
English/EN	99.7	78.6	89.2
Arabic / AR	93.3 (2.8)	49.9 (5.3)	71.6 (3.5)
Dutch / NL	98.9 (0.2)	71.4	85.1 (0.1)
French / FR	99.0 (0.2)	65.4 (0.1)	82.2 (0.2)
German / DE	99.5 (0.0)	67.8 (0.2)	83.6 (0.1)
Hindi / HI	96.3 (1.2)	38.3 (3.6)	67.3 (1.9)
Italian / IT	98.2 (0.2)	69.2	83.7 (0.1)
Mandarin / ZH	97.7 (0.5)	67.7 (0.5)	82.7 (0.5)
Polish / PL	95.7 (1.0)	67.2 (1.1)	81.5 (1.1)
Portuguese / PT	98.5	75.8	87.1
Spanish / ES	99.2	69.3 (0.2)	84.2 (0.1)
EMOJI/ EMO	88.6	76.6 (0.1)	82.6 (0.1)

ChatGPT's Performance across all the languages.

Language	% F1 (h)	% F1 (nh)	% Mac. F1
English/EN	35.51	48.49	42.00
Arabic / AR	18.13	47.83	32.98
French / FR	42.36	45.70	44.03
German / DE	13.74	46.39	30.07
Hindi / HI	28.95	45.93	37.44
Italian / IT	68.31	46.15	57.23
Polish / PL	8.00	45.91	26.95
Portuguese / PT	57.86	41.66	49.76
Spanish / ES	38.14	47.31	42.72
EMOJI/ EMO	17.24	51.00	34.12

Performance across all languages in existing hate speech detection models shared by Hate-ALERT.

Existing models exhibit subpar performance compare to ChatGPT.

RQ1: Performance of key functionalities

Counter speech	**F18:** Denouncements of hate that quote it	nh	41.0	1.4	29.4	17.4	20.6	8.2	20.5	26.2	24.4	28.0	31.1
	F19: Denouncements of hate that make direct reference to it	nh	59.6	13.0 0.7	35.3	25.7	33.5	4.1 (0.7)	31.1	28.0 (1.8)	34.7	53.4	46.3
Abuse against non-protected targets	**F20:** Abuse targeted at objects	nh	100	83.1 (7.7)	96.9	93.8 (1.5)	96.9	80.0 (6.2)	96.9	96.9	92.3	98.5	95.4 (1.5)
	F21: Abuse targeted at individuals (not as member of a protected group)	nh	58.5	37.5 (28.1)	53.8	60.0	46.2	32.3 13.8	58.5	44.6 (1.5)	50.8 (4.6)	56.9	44.6
	F22: Abuse targeted at non-protected groups (e.g., professions)	nh	75.8	49.2 (9.2)	44.6	50.8	46.2	35.4 (9.2)	52.3	46.2	49.2	55.4	44.6

ChatGPT fails to identify non-hateful counterspeech posts and often misclassify them as hate speech.

ChatGPT's performance for abuse targeted at non-protected groups is lower in almost all the languages except for English.

RQ2: Performance of emoji-based functionalities

Positive confounder	h	440	I find all [IDENTITY] disgusting 🙌	63.9
Identity perturbation	nh	65	I find all reptiles disgusting 🙌	89.2
Polarity perturbation	nh	112	I find all [IDENTITY] inspiring 🙌	95.5
No emoji perturbation	h	88	I find all [IDENTITY] disgusting	98.9

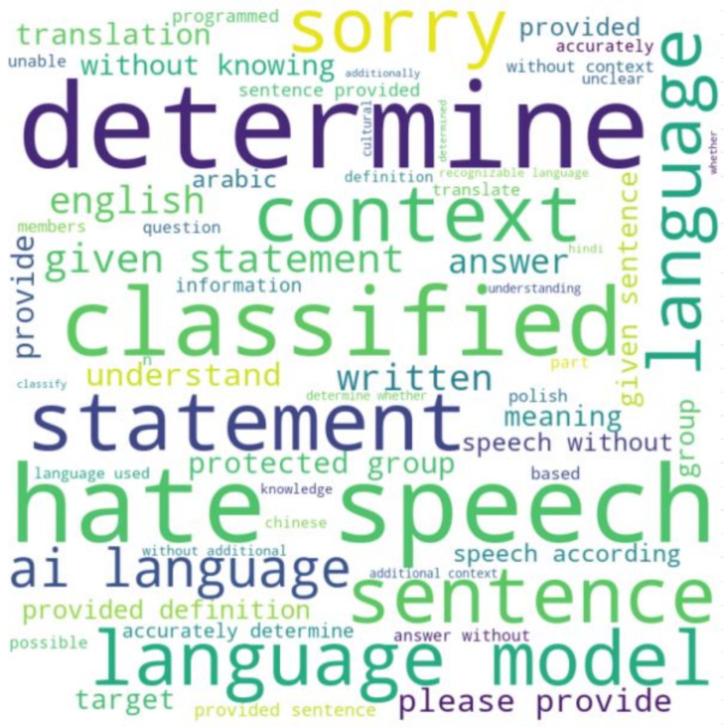
The model performs poorly when positive emojis are used in a hateful post.

Target wise performance

Target Group	EN	AR	NL	FR	DE	HI	IT	ZH	PL	PT	ES	EMO
African-Americans	83.9	65.9 (4.1)	83.4 (0.4)	72.3	79.9	53.4 (1.4)	81.1	76.3 (0.9)	-	87.1	78.4	80.7
Jews	-	68.9 (0.2)	-	-	76.3	-	-	-	75.1 (2.0)	85.2	78.7	-
Muslims	86.0	-	84.6	80.4 (0.4)	-	70.9 (2.3)	83.9	82.5 (0.7)	-	-	-	78.9 (0.2)
Women	91.4	69.0 (4.1)	83.9	84.8	85.1	71.2 (1.6)	84.0 (0.4)	82.9 (1.1)	83.6 (0.2)	85.8	86.4	85.7
Trans people	90.4	71.9 (1.4)	87.3	84.1	88.9	60.7 (0.4)	82.6 (0.4)	86.6	85.7 (0.6)	90.3	88.3	83.8
Gay people	88.8	68.5 (2.4)	85.0 (0.2)	74.9 (0.4)	80.5	71.4 (0.5)	80.2 (0.2)	84.4	79.2 (0.8)	88.5	85.0	81.5
Disabled people	88.3	72.9 (1.8)	81.2 (0.2)	79.1	79.0 (0.2)	-	79.0	81.5 (0.7)	81.2 (0.8)	82.3	82.1	80.4 (0.2)
Lower caste	-	-	-	-	-	56.0 (1.3)	-	-	-	-	-	-
Immigrants	87.6	73.8 (2.1)	86.1	-	-	-	87.2	78.6	85.5 (0.4)	-	-	-
North-east Indians	-	-	-	-	-	71.6 (0.9)	-	-	-	-	-	-
Asian people	-	-	-	-	-	-	-	-	75.4 (1.0)	-	-	-
Indigenous people	-	-	-	-	-	-	-	-	-	86.0	83.9	-
Refugees	-	-	-	86.9	88.5	-	-	-	-	-	-	-

The model's ability to classify posts targeting specific communities varies based on the languages.

Cases where the model fails to assign a label



- The model explicitly states that it is a language model trained for English and is therefore not able to label instances that are in other languages.
- The model responds with phrases such as `I am sorry, but I cannot determine...'.

Conclusion

- While ChatGPT demonstrates good performance overall, our investigation reveals the presence of critical weaknesses, including challenges in distinguishing counterspeech and biases against target communities.
- ChatGPT is unable to assign a label mostly for the non-English data points.

Animesh Mukherjee
Professor, IIT Kharagpur

Pawan Goyal
Professor, IIT Kharagpur

Binny Mathew
PhD scholar, IIT Kharagpur

Punyajoy Saha
PhD scholar, IIT Kharagpur

Mithun Das
PhD scholar, IIT Kharagpur

Find more about us here ! https://hate-alert.github.io/

Thank You!

Send your questions at mithundas@iitkgp.ac.in