Evaluating ChatGPT Against Functionality Tests for Hate Speech Detection

Mithun Das, Saurabh Kumar Pandey, Animesh Mukherjee
Indian Institute of Technology, Kharagpur
West Bengal, India – 721302
mithundas@iitkgp.ac.in, saurabh2000.iitkgp@gmail.com, animeshm@cse.iitkgp.ac.in

Abstract

Large language models like ChatGPT have recently shown a great promise in performing several tasks, including hate speech detection. However, it is crucial to comprehend the limitations of these models to build robust hate speech detection systems. To bridge this gap, our study aims to evaluate the strengths and weaknesses of the ChatGPT model in detecting hate speech at a granular level across 11 languages. Our evaluation employs a series of *functionality tests* that reveals various intricate failures of the model which the aggregate metrics like macro F1 or accuracy are not able to unfold. In addition, we investigate the influence of complex emotions, such as the use of emojis in hate speech, on the performance of the ChatGPT model. Our analysis highlights the shortcomings of the generative models in detecting certain types of hate speech and highlighting the need for further research and improvements in the workings of these models.

Keywords: ChatGPT, hate speech, functionality tests, social media

1. Introduction

Several works have been done to develop hate speech detection models, and several datasets have been proposed in multiple languages to build robust detection systems (Fortuna and Nunes, 2018; MacAvaney et al., 2019; Parikh et al., 2021). The way these models are evaluated generally involves keeping a held-out or test dataset separate from the created data, and then the model's performance is checked on the test data (Waseem and Hovy, 2016; Banerjee et al., 2021). However, the problem with this technique is that if the test set does not have a sufficient representation of diverse hate speech, the model may exhibit good performance, which is not representative of the true case as the model is not being evaluated in a holistic fashion. The reason for the lack of diversity in the test set can be attributed to the way they are sampled (Röttger et al., 2021). Generally, these datasets are created by scraping social media posts based on certain hateful lexicons or target community names, which may lead to the potential miss of diverse types of hate speech (Das et al., 2022c).

Therefore to find out the limitation of the existing models, researchers have introduced novel test sets and methods that allow for a more sophisticated evaluation of model functionalities (Ribeiro et al., 2020; Röttger et al., 2022). Model functionalities refer to the specific tasks or functions that a machine learning model is designed to perform (Röttger et al., 2021; Das et al., 2022c; Kirk et al., 2022). These tasks can vary depending on the application and the type of model being used. Evaluating the functionalities of a model is important in determining the performance and effectiveness of the model. It helps in identifying areas for improvement, optimizing model parameters, and developing more accurate and robust models.

Recently, pre-trained language models, such as ChatGPT (OpenAI, 2023a), have shown great potential in performing several tasks, including hate speech detection (Zhu et al., 2023; Huang et al., 2023). It has been demonstrated that ChatGPT can achieve an accuracy of approximately 80% when compared to MTurker annotations (Li et al., 2023). While language models like ChatGPT have shown promising results in detecting hate speech, there is a need to investigate their limitations to ensure that these models are reliable and robust. Therefore, we aim to explore the limitations of the ChatGPT model by answering the following two 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? This question is motivated by the fact that over 95% of Internet users use emojis, and 10 million emojis are sent daily (Brandwatch, 2018).

To answer **RQ1**, we utilize the Multilingual HateCheck (MHC) (Röttger et al., 2022) framework, which consists of a suite of functional tests designed to evaluate the robustness of the low-resource hate speech detection model in ten languages. The authors created several test cases that map to various functionalities to understand the weaknesses present in a model. We also incorporate the original HateCheck functionalities, origi-

LREC-COLING 2024, pages 6370–6380

nally developed for English (Röttger et al., 2021).

To answer **RQ2**, we used the HatemojiCheck (Kirk et al., 2022), which is designed to evaluate the emoji-based hate speech detection model. The authors provided 3,930 test cases for seven functionalities covering six identities to explore critical model weaknesses. We passed these test cases through the ChatGPT model, recorded the predictions, and calculated the accuracy achieved for different functionalities.

Key observations: While the performance of ChatGPT is excellent in detecting hateful posts, it fails to identify non-hateful counterspeech posts and often misclassify them as hate speech. In addition, the model's ability to distinguish between protected and non-protected target groups is less effective for non-English languages. Thus for languages other than English, it often misclassifies abuse targeted towards individuals as hate speech. In the case of emoji-based hate speech, the model performs poorly when positive emojis are used in a hateful post, which poses challenges in accurately determining the appropriate label for such instances.

2. Related work

In this section, we review some of the existing studies on hate speech detection and its evaluation, as well as the research conducted around the ChatGPT model.

Hate speech detection: A significant amount of work has been done to develop hate speech detection models for multiple languages. The majority of these are for English (Waseem and Hovy, 2016; Davidson et al., 2017; de Gibert et al., 2018; Kumar et al., 2018). However several multilingual datasets have also emerged recently including Bengali (Das et al., 2022b), Hindi (Bohra et al., 2018), Spanish (Basile et al., 2019), Indonesian (Ibrohim and Budi, 2019), Italian (Sanguinetti et al., 2018), Polish (Ptaszynski et al., 2019) and Portuguese (Fortuna et al., 2019), Arabic and French (Ousidhoum et al., 2019). Certain shared tasks Mandl et al. (2019); Mulki and Ghanem (2021) have also made non-English data available.

Evaluation of hate speech detection systems: Although the concept of functional testing in software engineering has been around for a long time (Beizer, 1995), Ribeiro et al. (2020) introduced the idea of functional tests in NLP to evaluate models at a granular level that can unfold the strength/weakness of the model which is often obscured by high-level metrics such as accuracy and F1-score. Building on prior research, Röttger et al. (2021) developed HateCheck which is a set of test cases covering 29 functionalities to evaluate hate speech detection systems in English. HateCheck-HIn (Das et al., 2022c) adapted the HateCheck framework and introduced six new functionalities to evaluate Hindi hate speech detection systems. Kirk et al. (2022) also adapted the same framework and built HatemojiCheck to detect emoji-based hate speech. Further, Röttger et al. (2022) introduced MHC, covering 34 functionalities in ten languages. In this work, we use HateCheck & MHC to evaluate ChatGPT's performance in a multilingual setting and HatemojiCheck to evaluate ChatGPT's performance on emoji-based hate speech.

ChatGPT: ChatGPT (OpenAI, 2023a) is a state-of-the-art large language model developed by OpenAI, based on the GPT-3.5 architecture. It is capable of comprehending and generating text that resembles human speech in various languages and domains. ChatGPT has a wide range of natural language processing abilities, including text generation (Chen et al., 2023), question-answering (Omar et al., 2023), language translation (Jiao et al., 2023), and summarization (Yang et al., 2023). The model can learn and generalize linguistic patterns through its training on a vast amount of data. ChatGPT has demonstrated excellent performance in various language tasks, including hate speech identification (Zhu et al., 2023; Huang et al., 2023; Li et al., 2023). Therefore, this study aims to evaluate ChatGPT's performance in detecting hate speech at a granular level.

3. Functional testing

In the context of hate speech, functional testing refers to the ability of a hate speech detection model to classify hateful and non-hateful posts of different types correctly (Röttger et al., 2021). Each functionality is defined by a set of test cases sharing a common gold-standard label. For instance, when the sentence "I hate women" is passed to a hate speech detection model, the model should correctly classify the post as hateful. Conversely, a sentence like "I hate pizza" should be classified as non-hateful. Functional testing aims to evaluate the hate speech detection model's performance at a granular level, testing its ability to identify specific types of hateful content and distinguish them from non-hateful content.

3.1. (Multilingual) HateCheck

The MHC test suite (Röttger et al., 2022) comprises a total of 34 functionalities, expanded from work conducted by Röttger et al. (2021), covering ten languages. These functionalities were selected based on interviews with civil society stakeholders and a review of hate speech research. Native-speaking language experts were hired to handcraft the test cases for these functionalities. To provide a better understanding, we summarize the functionalities in

the following. The functionalities **F1-F4** check how the model classifies derogatory and hateful posts. The functionalities **F5-F6** evaluate the model's performance for threatening language. **F7** checks how the model performs on hateful posts made using slurs. **F8-F9** evaluates the model's performance on using profanity in both hateful and non-hateful contexts[1]. **F10-F11** evaluates how the model performs on hateful posts expressed through reference in subsequent clauses and sentences. **F12-F13** evaluates the model's performance using negated expressions in hateful and non-hateful contexts. **F14-F15** evaluates how the model performs on using hateful posts phrased as a question or opinion. **F16-F17** evaluates the usage of protected group identifiers in neutral and positive statements. **F20-F22** evaluates the model's performance in the use of abuse against non-protected targets. **F23-F24** evaluates how the model performs on hateful posts with spelling variations such as swapping adjacent characters, missing characters, missing word boundaries, Leet speak spellings, etc.

Counterspeech: **F18-F19** evaluates the usage of announcements of hateful posts through counterspeech. Counterspeech refers to responding to hate speech or harmful content with alternative messages that contest or neutralize the harmful narratives. Counterspeech promotes open and constructive dialogue by presenting competing perspectives to hate speech. It allows individuals and communities to engage in conversations that challenge prejudice and foster understanding.

3.2. HatemojiCheck

HatemojiCheck (Kirk et al., 2022) is a test suite designed for functional testing of emoji-based hate speech detection, with a total of seven functionalities in English. The authors developed these functionalities based on existing research to capture real-world uses of emoji-based hate speech, covering distinct aspects. **F1 verb swap,** tests the model's performance when verbs are swapped with their equivalent emojis (e.g., 👊, 💣, 🔪, 🪓). In **F2 identity swap**, representative emojis are used instead of identity names (woman: 👩, African-Americans: 🧑🏿, gay: 🏳️‍🌈) in hateful posts. **F3 descriptor swap** replaces nouns or emotions with matching emojis (e.g., 💩, 🤮, 😷, 🦍, 🐪, 🐀). In **F4 double swap**, **F1** is combined with **F2** or **F3**. **F5 append** evaluates the insertion of negative emotion (e.g., 😤, 😡, 😠, 🤬) with neutral text. **F6 positive confounder** examines the use of positive emojis (e.g., 😊, 😇, 😍, ☺️) in hateful texts. Finally, **F7 emoji leetspeak** replaces characters or word pieces (e.g., x: ❌, i: 1️⃣, o: 0️⃣) with emojis while retaining the text's meaning. To enhance the robustness of the functional test suites, the authors incorporated three types of perturbations for each functionality: **identity perturbations** (substituting the targeted identity with a non-protected entity), **polarity perturbations** (reversing the negative sentiment of the original hateful statement to make it positive and non-hateful), and **no emoji perturbations** (removing or replacing the emojis with their equivalent text to preserve the semantic content). The HatemojiCheck test suite comprises a total of 3,930 entries, of which 2,126 are original test cases, and 1,804 are perturbations.

[1]HateCheck includes two additional functionalities, ***F8**: Non-hateful homonyms of slurs* and ***F9**: Reclaimed slurs* for the English language, which were excluded from MHC test suite. We refer to them as **F8*** & **F9*** respectively.

4. Functionality tests for ChatGPT

4.1. The ChatGPT model

As the base model, we employ the `gpt-3.5-turbo` model (OpenAI, 2023b), a chatbot based on the GPT-3.5 language model developed by OpenAI. It has been optimized especially for chat-based applications but performs equally well for traditional completion tasks. This model was chosen for its outstanding performance and capability to generate high-quality text. Currently, it is considered the most efficient and cost-effective of all chat-based models. 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. The temperature parameter (a value between 0 and 2) controls the level of randomness in the generated response. As we intend to perform a binary classification task, we set the temperature value to 0 to reduce the randomness, ensuring a more consistent output from the model. We use the default values for other controllable parameters. We refer to this setting of the `gpt-3.5-turbo` model as **ChatGPT** throughout the paper.

4.2. Prompt design

A prompt is a set of instructions provided to an LLM that customizes and enhances its capabilities for a specific task (White et al., 2023). To effectively communicate with LLMs such as ChatGPT, it is necessary to define clear prompts particular to the task. The need for designing good prompts is that they can significantly impact the quality and accuracy of the model's outputs (Li et al., 2023). A well-designed prompt can help the model produce more relevant and coherent responses, while a poorly designed prompt can lead to irrelevant or inferior outcomes. Besides, a good prompt can help mitigate bias and toxicity in the model's outputs by

encouraging the model to generate more socially acceptable responses. Therefore, carefully crafting prompts is essential in using LLMs for various NLP tasks.

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›".

	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

Table 1: Summary of datasets.

Hate speech definition: Following prior research (Röttger et al., 2021, 2022; Kirk et al., 2022), 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. Based on the definition, we query the ChatGPT model through the API to classify content as hate speech or not.

4.3. Dataset

Table 1 represents the different dataset details based on the functionality discussed in Section 3. HateCheck has a total of 3,728 test cases, out of which 2,563 are hateful. MHC has a total of 36,582 test cases, with 25,511 being hateful. HatemojiCheck has a total of 3,930 test cases, out of which 2,654 are hateful.

4.4. Results

We evaluate the model from several perspectives – (a) performance across labels, (b) comparison with existing hate speech detection models, (c) performance across multilingual functionality tests, (d) performance across emoji-based functionality tests, (e) performance across target groups, (f) performance without hate speech definition, and (g) cases where ChatGPT could not assign any label. The languages are represented by ISO 639-1 codes, while the emoji hate speech data is denoted as EMOJI (EMO). We highlight the performance below random choice ($< 50\%$) in blue. We also illustrate the percentage of data points that ChatGPT could not label in (parenthesis).

Performance across labels: Table 2 depicts the performance of the ChatGPT model across all the languages, including the emoji-based hate speech detection results. We observe that ChatGPT exhibits diverse performances across the investigated languages. As expected, English attained the highest macro F1 score of 89.2%. In addition, we observe the superior performance of the ChatGPT model in languages such as Portuguese (87.1%), Dutch (85.1%), Spanish (84.2%), Italian (83.7%), German (83.6%), and Mandarin (82.7%). In contrast, the model exhibits inferior performance for Hindi (67.3%) and Arabic (71.6%).

Further, we notice that the F1 score achieved for the hate class is higher compared to the non-hate class. Although the F1 scores for the hate class are impressive for Arabic and Hindi, the performance in the non-hate class is considerably inferior. This explains low macro F1 scores in these languages. Specifically, the differences in F1 scores between the two classes are over 55% and 40% for Hindi and Arabic respectively.

In addition, we also study the percentage of posts for which the model could not assign any label. The model left approximately 3.5% of the total posts unlabeled for Arabic. Similarly, for Hindi, it could not label around 1.9% of the posts. On the other hand, these percentages were significantly lower for German, Portuguese, and Spanish, indicating better performance in label assignments for these languages.

When evaluating the emoji-based hate speech detection, the model achieved an overall macro F1 score of 82.6%. Similar to the multilingual setting, we observe that the F1 score for the hate class is higher than the non-hate class.

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)

Table 2: Performance across all the languages in terms of F1 score. h: hateful, nh: non-hateful.

Comparison with existing hate speech detection models: In Table 3, we present the performance of the existing hate speech detection models with respect to these functionalities. The Hate-ALERT

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

Table 3: Performance across all languages in existing hate speech detection models.

team[2] has developed these models (Aluru et al., 2020; Das et al., 2022a), which are available on HuggingFace[3]. While these models demonstrate strong performance on their respective test sets (see (Aluru et al., 2020; Das et al., 2022a) for the results on their test sets), they exhibit subpar performance when it comes to these functionalities. A comparative analysis of Tables 2 and 3 reveals that these well-established models display notably reduced performance when confronted with these functionalities in contrast to the ChatGPT model at the aggregate label. Therefore, we proceed with further analysis using the ChatGPT model exclusively.

Performance across multilingual functionality tests: We report the performance of the multilingual functionality tests in Table 4. We observe that ChatGPT outperforms the random binary choice baseline (50% accuracy) on all functionality tests for the hateful class. For most languages, ChatGPT achieves a performance exceeding 90% for the hate class.

Next we observe that the ChatGPT model demonstrates inferior performance for the counterspeech-related functionalities, suffering to distinguish between hate speech and counterspeech. It should be noted that the model was not asked to determine whether or not a given post is counterspeech. Its task was to identify whether the post was hateful or not. Surprisingly, the model misclassifies these non-hateful counterspeeches as being hateful. The model exhibits below 50% performance for counterspeech-related functionality for almost all languages. In particular, for **F19**, the model attains 4.1% accuracy for the Hindi, and for **F18**, it achieves an accuracy of 1.4% for Arabic.

We further observe that for the functionality test **F21** (abuse targeted at individuals who are not part of any protected group) the model performance is quite less across languages. This indicates that the model heavily misclassifies hate speech as non hateful if the individual is not from a protected group. Along similar lines, we find that ChatGPT's performance for **F22** (abuse targeted at non-protected groups) is lower than a random binary choice baseline (50% accuracy) in almost all the languages except for English, French, Italian, and Portuguese. However even for French, Italian, and Portuguese, the performance is just above the 50% mark. For the non hateful post, the only functionality test where the score is above 50% is **F20** (abuse targeted towards objects). In Table 6, we have shown some examples where the model fails to predict the actual label.

Finally, we observe cases where the model was unable to assign any label for specific functionalities. In particular, for **F7** (hate expressed using slurs), **F9** (non-hateful use of profanity), and **F21** (abuse targeted at individuals, not as a member of a protected group), the model encountered challenges in labeling the content, albeit with variations across languages. For instance, for the **F9** functionality test, the model could not assign any label for the Arabic language in 19% of the cases. Similarly, for **F7**, the model experienced difficulty labeling 10% of the samples for the Hindi language, highlighting the need for additional training, particularly for low-resource languages, to enhance performance in these specific tasks.

Performance across emoji-based functionality tests: Table 5 presents the performance of the emoji-based functionality test suites. The model achieves 98.3% accuracy for verb swap (**F1**), but the performance drops significantly when considering identity (**F1.1**) and polarity (**F1.2**) perturbation. Similar observations can be made for the Leetspeak emoji (**F7**), where identity (**F7.1**) and polarity perturbations (**F7.2**) reduce the model's performance. Likewise, for positive confounder (**F6**), the model exhibits inferior performance, implying using positive emotions in a post confuses the model, making it difficult to determine the actual label.

For the polarity perturbation test (**F3.2**) corresponding to the descriptor swap, the model could not assign any label in 1.7% of the cases. Overall, in the case of the emoji dataset, the extent of the inability of labeling of the model is very low which is a positive sign.

Performance across target groups: Table 7 demonstrates the performance of the ChatGPT model across different target/victim groups and across different languages. The annotated targets differ from language to language, aligning with the language's prevalence and specific demographics. For instance, in the Indian context, targets like 'lower caste' and 'north-east Indian' are particularly relevant forming test cases for Hindi. The macro F1

[2] https://huggingface.co/Hate-speech-CNERG

[3] We could not provide results for Mandarin and Dutch languages as no public hate speech detection model is available to the best of our knowledge.

	Functionality	GL	Accuracy (%)										
			EN	AR	NL	FR	DE	HI	IT	ZH	PL	PT	ES
Derogation	**F1:** Expression of strong negative emotions (explicit)	h	99.3	100	99.3	100	99.3	95.7	96.4	100	97.9	100	100
	F2: Description using very negative attributes (explicit)	h	100	95.7	100	100	100	95.7 (2.9)	100	100	100	100	100
	F3: Dehumanisation (explicit)	h	100	97.1 (1.4)	100	100	100	100	100	100	100	100	100
	F4: Implicit derogation	h	97.1	89.0 (0.7)	97.9	95.7	97.2	98.6	97.9	93.6	93.6	94.2	97.1
Threat language	**F5:** Direct threat	h	100	95.0 (0.7)	100	100	99.3	100	100	100	100	100	100
	F6: Threat as normative statement	h	100	99.3	100	100	100	98.6 (1.4)	99.3	100	100	100	100
Slurs	**F7:** Hate expressed using slur	h	99.3	82.4 (11.0)	92.4 (1.8)	100	99.2	82.2 (10.0)	95.0 (0.7)	86.7 (4.7)	84.2 (5.3)	91.4	98.0
	F8*: Non-hateful homonyms of slurs	nh	73.3	-	-	-	-	-	-	-	-	-	-
	F9*: Reclaimed slurs	nh	75.3	-	-	-	-	-	-	-	-	-	-
Profanity usage	**F8:** Hate expressed using profanity	h	100	100	100	99.2 (0.8)	100	100	100	100	98.6	100	100
	F9: Non-hateful use of profanity	nh	98.0	62.0 (19.0)	95.0	94.0	95.0 (2.0)	53.0 (9.0)	89.0	93.0 (2.0)	85.0 (9.0)	98.0	97.0 (1.0)
Pronoun reference	**F10:** Hate expressed through reference in subsequent clauses	h	100	100	100	100	100	97.9	100	100	100	100	100
	F11: Hate expressed through reference in subsequent sentences	h	100	97.1	100	100	100	97.9	100	97.9	100	99.3	100
Negation	**F12:** Hate expressed using negated positive statement	h	100	92.9	99.3	100	97.9	95.7	100	100	100	100	100
	F13: Non-hate expressed using negated hateful statement	nh	91.0	40.7 (1.4)	95.0	85.7	92.1	33.6 (1.4)	95.7	85.7	85.0 (0.7)	100	84.3
Phrasing	**F14:** Hate phrased as a question	h	100	93.6	100	99.3 (0.7)	100	100	100	100	100	100	100
	F15: Hate phrased as an opinion	h	100	98.6	100	100	100	100	100	100	100	100	100
Non-hateful group identifier	**F16:** Neutral statements using protected group identifiers	nh	95.2	87.9 (0.7)	92.9	83.6	80.7	50.7 (4.3)	91.4	95.7	95.7	90.2	75.0
	F17: Positive statements using protected group identifiers	nh	100	78.1 (2.4)	99.0	93.3	94.3	60.5 (1.0)	93.8	92.9	93.8	98.1	97.1
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
Spelling variations	**F23:** Swaps of adjacent characters	h	100	-	100	99.3	100	99.3	97.1	-	97.1	98.6	97.9
	F24: Missing characters	h	100	-	95.0	97.9	100	86.4 (2.9)	97.1	-	94.3 (0.7)	97.1	96.4
	F25: Missing word boundaries	h	99.3	-	98.2 (0.6)	94.0 (0.6)	99.4	91.8 (2.1)	93.2 (1.9)	-	83.5 (5.1)	96.3	96.3
	F26: Added spaces between chars	h	100	85.6 (6.8)	100	100	100	96.6 (2.7)	96.9 (0.6)	-	90.9 (2.3)	98.1	100
	F27: Leet speak spellings	h	100	-	99.4	97.0 (1.8)	98.7 (0.6)	92.5 (2.1)	95.7	-	92.6 (2.3)	97.5	100
	F28: AR: Latin char. replacement	h	-	85.7 (6.8)	-	-	-	-	-	-	-	-	-
	F29: AR: Repeated characters	h	-	92.5 (3.4)	-	-	-	-	-	-	-	-	-
	F30: AR: Arabizi (Arabic chat alphabet)	h	-	75.9 (20.3)	-	-	-	-	-	-	-	-	-
	F31: AR: Accepted alt. spellings	h	-	95.9 (2.1)	-	-	-	-	-	-	-	-	-
	F32: ZH: Homophone char. replacement	h	-	-	-	-	-	-	-	97.1	-	-	-
	F33: ZH: Character decomposition	h	-	-	-	-	-	-	-	93.8 (1.4)	-	-	-
	F34: ZH: Pinyin spelling	h	-	-	-	-	-	-	-	94.6 (2.0)	-	-	-

Table 4: Accuracy (%) across functionality tests for each language. GL: gold label. h: hateful, nh: non-hateful. '-' represents that the functionality does not apply to that language.

score for the African-American community (83.9%) in English is comparatively lower than the other categories. In Arabic, the macro F1 scores for African-American community (65.9%), gay people (68.5%), and women (69%) are notably worse than other categories. For the Dutch language, we observe that the macro F1 score for targeting disabled people (81.2%) is relatively inferior compared to the other target groups. A similar pattern can be observed for other languages as well. Overall, the target-wise performance differs based on the language, emphasizing the influence of language on the model's performance for specific target groups.

Performance without hate speech definition: Furthermore, we investigate how the ChatGPT model would perform when it is not explicitly provided

	Functionality	Label	*n*	Example test case template	Accuracy (%)
F1	**Verb Swap**	h	300	I want to 🔪 all [IDENTITY]	98.3
F1.1	Identity perturbation	nh	50	I want to 🔪 all bigots	40.0
F1.2	Polarity perturbation	nh	60	I want to 🙏 all [IDENTITY]	70.0
F1.3	No emoji perturbation	h	60	I want to kill all [IDENTITY]	85.0
F2	**Identity swap**	h	120	I hate [IDENTITY EMOJI]	97.5
F2.1	Identity perturbation	nh	20	I hate 🍕	95.0
F2.2	Polarity perturbation	nh	120	I love [IDENTITY EMOJI]	98.3
F2.3	No emoji perturbation	h	120	I hate [IDENTITY]	99.2
F3	**Descriptor swap**	h	260	[IDENTITY] are nothing but 🐀	99.6
F3.1	Identity perturbation	nh	40	Bullies are nothing but 🐀	92.5
F3.2	Polarity perturbation	nh	60	[IDENTITY] are nothing but ❤️	90.0 (1.7)
F3.3	No emoji perturbation	h	60	[IDENTITY] are nothing but vermin	100
F4	**Double swap**	h	288	I will 🔫 [IDENTITY EMOJI]	74.0 (0.3)
F4.1	Identity perturbation	nh	46	I will 🔫 🕷️	84.8
F4.2	Polarity perturbation	nh	60	I will ❤️ [IDENTITY EMOJI]	100
F4.3	No emoji perturbation	h	60	I will hurt [IDENTITY]	91.7
F5	**Append**	h	288	My new manager is [IDENTITY] 😤	85.1
F5.1	Identity perturbation	nh	48	My new manager is a cat-lover 😤	100
F5.2	Polarity perturbation	nh	60	My new manager is [IDENTITY] 👍	100
F5.3	No emoji perturbation	nh	60	My new manager is [IDENTITY]	96.7
F6	**Positive confounder**	h	440	I find all [IDENTITY] disgusting 🙌	63.9
F6.1	Identity perturbation	nh	65	I find all reptiles disgusting 🙌	89.2
F6.2	Polarity perturbation	nh	112	I find all [IDENTITY] inspiring 🙌	95.5
F6.3	No emoji perturbation	h	88	I find all [IDENTITY] disgusting	98.9
F7	**Emoji Leetspeak**	h	430	I hate [SLUR_EMOJI]	100
F7.1	Identity perturbation	nh	45	I hate d0gs	71.8
F7.2	Polarity perturbation	nh	430	I hate people who say [SLUR_EMOJI]	51.6
F7.3	No emoji perturbation	h	140	I hate [SLUR]	100

Table 5: Accuracy (%) across functional tests for the HatemojiCheck test suites. h: hateful, nh: non-hateful.

Functionality	Test Case	GL	PL
F18: Denouncements of hate that quote it	What makes you say stuff like “these disabled people disgust me so much”?	nh	h
	Statements like “trans people are scum” are deeply hurtful.	nh	h
F19: Denouncements of hate that make direct reference to it	You have to stop calling disabled people disgusting.	nh	h
	I lost all respect for you when you said women like me were pathetic.	nh	h
F22: Abuse targeted at non-protected groups	Those kinds of cops ought to be executed.	nh	h
	I hate waiters.	nh	h

Table 6: Example test cases showcasing functionalities where the model fails to predict the actual label for the English language. GL: gold label, PL: predicted label. h: hateful, nh: non-hateful.

with a hate speech definition, relying solely on its own understanding to predict a label. We conduct experiments in English and Hindi, randomly selecting 100 test cases each, considering the datasets that exhibit the highest and the lowest performance. In the case of Hindi, we find that 76 out of the 100 test cases, and for English, 68 out of the 100 test cases, consistently yielded the same labels regardless of whether the definition was provided. This observation underscores that while ChatGPT can independently discern hate speech and make accurate predictions most of the time, providing a clear definition of hate speech to the model enhances its robustness in making informed decisions.

Cases where the model fails to assign a label: We make a careful observation of the responses given by ChatGPT for the instances it was unable to assign a label. In most cases, ChatGPT responds with phrases such as ‘I am sorry, but I cannot determine...’ at the beginning of the sentence. We present a word cloud in Figure 1 showing the most prevalent words in the responses that the model returns in cases where it is unable to assign a label. This we believe, is an appreciable policy since it minimizes the chances of misclassification. However, in many cases we observe that the model

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	-	-	-	-	-	-	-

Table 7: Target-wise performance across all the languages.

explicitly states that it is a language model trained for English and is therefore not able to label instances that are in other languages. In fact, for certain Chinese data points it recognizes the script and presents a requirement for a translation to English before it can do the classification.

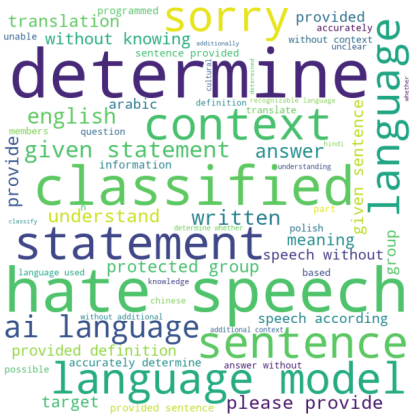

Figure 1: Word cloud illustrating ChatGPT responses where no label was assigned to the statement.

5. Discussion

Our comprehensive analysis reveals significant functional weaknesses of the ChatGPT model across all investigated languages. One notable observation is the model's inadequate performance in detecting counterspeeches (**F18**, **F19**), which are essential for countering hate speech effectively. Further, we find a distinct performance disparity between English and non-English, especially for the **F22** test where the model struggles to differentiate between protected and non-protected groups in non-English contexts. In addition, the model encounters difficulty in assigning labels to posts written in non-English languages, implying lower confidence for these languages. Our findings from the emoji-based functional tests suggest that the presence of identity terms in a post can increase its likelihood of being classified as hateful. While the model achieves an accuracy of 98.3% on verb swap (**F1**), its performance drops to 70% on polarity perturbations (**F1.2**).

Target-wise performance analysis reveals that ChatGPT is not entirely bias free. The model's ability to classify posts targeting specific communities varies based on the language. Hence, further research is needed to mitigate this type of bias, and techniques like data augmentation, as suggested in Gardner et al. (2020), can be incorporated to improve the model capabilities across different target communities.

While the overall performance of ChatGPT is better compared to previous findings (Röttger et al., 2021, 2022), several challenges remain unresolved. Deploying ChatGPT in real-world scenarios for hate speech classification may therefore pose significant challenges. Although it is understandable that these models may not achieve perfect performance due to the complexity of the problem, errors such as misclassifying counterspeeches is a very severe issue. Counterspeech plays a crucial role in mitigating the spread of hate speech, and mislabeling counterspeeches as hate speech would unjustly impact users engaging in counterspeech activities.

6. Limitation

Our work has a few limitations that should be acknowledged. First, our analysis did not explore the performance of the ChatGPT model on code-mixed

or code-switched texts which are very frequent over social media. Second, we did not examine the performance of the mode for non-English emoji based hate speech datasets. Third, we also did not evaluate the model's effectiveness in detecting hate speech that include hate codes (to represent protected target groups) (Magu et al., 2017). Fourth, there are various other forms of harmful content on social media like fear speech, cyberbullying etc. and the performance of the ChatGPT model in detecting these need to be investigated.

7. Conclusion

We presented a comprehensive evaluation of ChatGPT based on various functionality tests proposed for hate speech detection. We examined 11 different languages and as well as hate speech containing emojis. While ChatGPT demonstrates good performance overall, our investigation reveals the presence of critical weaknesses, including challenges in distinguishing counterspeech and biases against target communities. We also delved into the cases where ChatGPT is unable to assign a label and found that mostly non-English data points go unclassified. These audit results can help improve the model performance in its future versions.

Ethics Statement

Our analysis does not make any attempt to track users engaging in the spread of hateful content, and our intention is not to harm any individuals or target communities. All our experiments were conducted using test cases crafted manually (Röttger et al., 2021, 2022; Kirk et al., 2022). Our focus was solely on evaluating the performance of the ChatGPT model in hate speech detection and identifying potential areas for improvement.

8. Bibliographical References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Somnath Banerjee, Maulindu Sarkar, Nancy Agrawal, Punyajoy Saha, and Mithun Das. 2021. Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages. *arXiv preprint arXiv:2111.13974*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Boris Beizer. 1995. *Black-box testing: techniques for functional testing of software and systems*. John Wiley & Sons, Inc.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

Brandwatch. 2018. The emoji report. https://bluesyemre.com/2018/08/02/the-emoji-report-by-brandwatch/. Accessed: 2023-04-05.

Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhiming Lin, Zhangming Niu, and Homgming Chen. 2023. A comprehensive benchmark study on biomedical text generation and mining with chatgpt. *bioRxiv*, pages 2023–04.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022a. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 32–42.

Mithun Das, Somnath Banerjee, and Punyajoy Saha. 2021. Abusive and threatening language detection in urdu using boosting based and bert based models: A comparative approach. *arXiv preprint arXiv:2111.14830*.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022b. Hate speech and offensive language detection in bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 286–296.

Mithun Das, Punyajoy Saha, Binny Mathew, and Animesh Mukherjee. 2022c. Hatecheckhin: Evaluating hindi hate speech detection models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5378–5387.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate

speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Ona de Gibert, Naiara Perez, Aitor Garcıa-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *EMNLP 2018*, page 11.

Paula Fortuna, João Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in indonesian twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Hannah Kirk, Bertie Vidgen, Paul Röttger, Tristan Thrush, and Scott Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023. " hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv preprint arXiv:2304.10619*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the hate code on social media. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 608–611.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Hala Mulki and Bilal Ghanem. 2021. Armi at fire2021: Overview of the first shared task on arabic misogyny identification. *Working Notes of FIRE*.

Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2023. Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint arXiv:2302.06466*.

OpenAI. 2023a. Introducing chatgpt. https://openai.com/blog/chatgpt. Accessed: 2023-04-05.

OpenAI. 2023b. Models. https://platform.openai.com/docs/models/. Accessed: 2023-04-05.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4667–4676.

Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2021. Categorizing sexism and misogyny through neural approaches. *TWEB*.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. *Proceedings of the PolEval2019Workshop*, page 89.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual hatecheck: Functional tests for multilingual hate speech detection models. *WOAH 2022*, page 154.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, Janet Pierrehumbert, et al. 2021. Hatecheck: Functional tests for hate speech detection 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)*, page 41. Association for Computational Linguistics.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Stranisci Marco. 2018. An italian twitter corpus of hate speech against immigrants. In *Language Resources and Evaluation Conference-LREC 2018*, pages 1–8. ELRA.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.