Beyond Hate Speech: NLP's Challenges and Opportunities in Uncovering Dehumanizing Language

Anonymous ACL submission

Abstract

Dehumanization, characterized as a subtle yet harmful manifestation of hate speech, involves denying individuals of their human qualities and often results in violence against marginalized groups. Despite significant progress in Natural Language Processing across various domains, its application in detecting dehumanizing language is limited, largely due to the scarcity of publicly available annotated data for this domain. This paper evaluates the performance of cutting-edge NLP models, including GPT-4, GPT-3.5, and LLAMA-2, in identifying dehumanizing language. Our findings reveal that while these models demonstrate potential, achieving a 70% accuracy rate in distinguishing dehumanizing language from broader hate speech, they also display biases. They are over-sensitive in classifying other forms of hate speech as dehumanization for a specific subset of target groups, while more frequently failing to identify clear cases of dehumanization for other target groups. Moreover, leveraging one of the best-performing models, we automatically annotated a larger dataset for training more accessible models. However, our findings indicate that these models currently do not meet the high-quality data generation threshold necessary for this task.

1 Introduction

Dehumanization, defined as the denial of "humanness" to others (Haslam, 2006), significantly impacts society by fostering conditions that result in extreme and violent behaviors against marginalized groups (Kteily and Landry, 2022). This phenomenon can range from overt derogation, where victims are likened to "dogs" or "monkeys" (Hagan and Rymond-Richmond, 2008), to subtler forms, such as denying the capability of experiencing pain to certain individuals (Deska et al., 2020). The identification of dehumanizing language is crucial for understanding and mitigating its effects on collective violence and the manipulation of public perception in conflicts (Oberschall, 1997).

Despite the importance of detecting dehumanization, this nuanced form of hate speech has been relatively overlooked in natural language processing advancements, primarily due to the lack of publicly available annotated datasets. Annotating dehumanizing language poses unique challenges due to its subjective and abstract nature. However, recent advancements in pretrained models capable of understanding instructions and prompts offer new opportunities to leverage NLP models for this task without the need for extensive fine-tuning.

This study evaluates the capability of prominent pretrained NLP models—specifically, GPT-4 (Achiam et al., 2023), GPT-3.5, and LLAMA-2 (Touvron et al., 2023a)—in accurately identifying dehumanizing language. Through a comprehensive analysis encompassing zero-shot, few-shot, and explainable prompting settings, we evaluate the effectiveness of these models in recognizing dehumanizing content. In zero-shot settings, the models are tested without any prior examples, relying solely on their pre-existing knowledge. In few-shot settings, the models are provided with a limited number of examples to guide their predictions. In explainable prompting settings, apart from the few examples, we also ask the model to explain its results, providing insights into its decision-making process. Our findings reveal that even the best-performing model, GPT-4, has limitations in distinguishing dehumanizing language from other forms of hate speech, achieving only a 70% accuracy rate for this specific task. In addition, our results expose a variable sensitivity across different target groups. GPT models are prone to overclassifying other types of hate speech as dehumanization for certain target groups, such as gay and transgender individuals, while failing to adequately identify dehumanizing language targeting other vulnerable groups, such as immigrants and refugees.

Following our evaluation, we applied the most effective approach, i.e., explainable prompting, to automatically generate annotated data for training smaller open-source models. This phase revealed a notable discrepancy: while the initial results from GP models are promising, the resulting annotations do not meet the expected standards for training high-performing models. Specifically, even the top-performing model only achieved a 61% accuracy rate in distinguishing dehumanizing language from other hate speech types. This finding underscores the fact that annotating nuanced tasks like dehumanization still necessitates the expertise of human annotators.

2 Related Work

Dehumanization has been extensively studied within the realm of social science (Paladino et al., 2002; Haslam et al., 2008; Haslam, 2006; Haslam and Loughnan, 2014; Kteily and Landry, 2022; Harris and Fiske, 2015; Leyens et al., 2000). Recent advancements in NLP techniques present a significant, yet largely unexplored, opportunity to utilize state-of-the-art methodologies and expand traditional dehumanization analysis. These advancements have the potential to identify more comprehensive instances of dehumanization, ultimately contributing to the enhancement of online media safety and enabling a more comprehensive sociological examination of the multifaceted impact of dehumanization on society. However, despite this promising potential, the exploration of dehumanization within the field of NLP has been relatively limited.

The first step in addressing this gap was taken by Mendelsohn et al. (2020) who introduced a computational framework for studying dehumanization with traditional NLP techniques[1], focusing on the analysis of how LGBTQ-related terms were subjected to dehumanization in New York Times articles spanning more than 30 years. Their approach revolved around four key components: (1) Negative Evaluations: assessing the presence of negative judgments directed towards the target group, (2) Denial of Agency: examining instances where the target group's capacity to make decisions or take actions was undermined, (3) Moral Disgust: identifying expressions of moral disgust in the context of the target group, and (4) Use of Vermin Metaphors: detecting the application of metaphors portraying the group as vermin or subhuman.

While Mendelsohn et al. (2020)'s approach was effective in identifying overall trends related to dehumanization, it faces two primary challenges. Firstly, it is challenging to use their proposed approach to pinpoint specific mentions of dehumanization within the text. Secondly, their techniques were less adaptable to shorter texts, such as social media content and comments. In contrast, this paper capitalizes on the capabilities of pretrained NLP techniques to identify specific instances of dehumanization within short input texts.

In addition to Mendelsohn et al. (2020), other researchers have also explored computational analysis of dehumanization. For instance, Friedman et al. (2021) consider dehumanization as a subproblem of moral disengagement and manually annotate 378 examples for both training and evaluation, along with their corresponding entities and relations regarding various forms of moral disengagement. They then utilize a transformer-based model, i.e., a variation of the SpanBERT model (Eberts and Ulges, 2020), to construct a knowledge graph consisting of these entities and relations. The schema of their knowledge graph illustrates entities linked by relationships, with each entity possessing various attributes, including dehumanization, violent, condemned, justified, responsible, and harmed. They report an F_1 score of 50 points on identifying the dehumanization attributes in their dataset. This dataset is not publicly available for incorporation or evaluation in this study.[2]

Similarly, dehumanization is considered as one of the subcategories of hate speech by Vidgen et al. (2021), where they constructed a large-scale dataset of 22K examples of hate speech, among which 906 examples are labeled as dehumanization. We use this subset for the evaluation in this work.

3 Experimental Setup

3.1 Data

As mentioned, we utilize Vidgen et al. (2021)'s publicly available hate speech dataset for our evaluations. This dataset comprises 906 instances of dehumanizing content, enabling us to assess the identification of dehumanization. Additionally, the

[1] E.g., word2vec word embeddings (Mikolov et al., 2013) and connotation frames (Rashkin et al., 2016).

[2] In addition to the mentioned datasets, there are other hate speech datasets that contain a small number of examples, typically fewer than 10, labeled as dehumanization (Calabrese et al., 2022).

inclusion of other hate speech labels, such as animosity or derogation, in this dataset allows us to evaluate the model's ability to distinguish dehumanization from various forms of hate speech.

Another significant advantage of this dataset is that it provides information about the targeted groups. This allows us to analyze whether the model's performance on dehumanizing instances varies depending on the targeted group. For example, we aim to assess whether the model's performance is enhanced when provided with examples from the same targeted group in the input prompts. We devised three evaluation subsets from this dataset for evaluating dehumanization:

Targeted Dehumanization: This evaluation set consists of 42 dehumanization instances aimed at Muslims and an additional 42 randomly chosen samples.[3] This set is designed to assess the model's performance in identifying dehumanization when there is a single, known target a priori.

General Dehumanization: This set contains all 906 dehumanization instances, which may have different targeted groups, along with 906 randomly selected instances from the dataset.[4] This evaluation set assesses the model's performance when the target may vary and is not predetermined.

Dehumanization vs. Hate: It consists of 906 instances of dehumanization as well as 906 randomly selected instances from other hate speech labels, testing the model's ability to distinguish between dehumanization and other forms of hate speech.

3.2 Mendelsohn's Baselines

Other than state-of-the-art models, we also adopt the four linguistic-based components from Mendelsohn et al. (2020)'s framework as our baselines for analyzing the extent to which each of these components can identify instances of dehumanization in our evaluation set.

Negative Evaluation of a Target Group Valence measures how positive or negative the text is, on a scale from completely positive (1) to completely negative (0). We use the NRC VAD lexicon (Mohammad, 2018), which provides scores for valence (positivity or negativity), dominance (control or power), and arousal (excitement or calmness) for various words. To estimate the overall sentiment of a text, we calculate the average valence score of its words using this lexicon.

Additionally, to assess the sentiment directed towards specific target groups, we use the connotation frames lexicon (Rashkin et al., 2015), which assigns scores to 900 English verbs, ranging from very negative (-0.87) to very positive (0.8). We consider a text to be negatively evaluating a target group if its average valence score is below 0.5 and it has a negative perspective score.

Denial of Agency To evaluate agency, which refers to how much control a target group is perceived to have over their actions and decisions, we use the connotation frames for agency (Sap et al., 2017). This method distinguishes between high agency, where entities are seen as having significant control, and low agency, where they are viewed as more passive. The lexicon provides agency levels for over 2000 verbs. We determine a text's overall agency by calculating an aggregate score based on how frequently these verbs appear. If a text predominantly uses verbs that indicate low agency, it is classified as exhibiting a denial of agency. For texts that do not contain any verbs from the lexicon, we apply a default "neutral" label.

Moral Disgust Following the approach by Mendelsohn et al. (2020), we use Graham et al. (2009)'s lexicon to identify instances of moral disgust. This lexicon includes over 30 words and stems associated with moral disgust, such as "disgust", "sin", and "pervert".

To measure moral disgust, we use a vector-based methodology. We calculate the average of the word embeddings for terms linked to moral disgust, with each word's contribution weighted by its frequency in the lexicon.[5] The degree to which an input text is associated with moral disgust is then assessed by computing the cosine distance between the averaged vector of moral disgust terms and the embedding of the input text.

Use of Vermin Metaphors Similar to Mendelsohn et al. (2020), we construct a vector representation for vermin metaphors using keywords such as vermin, rodent(s), rat(s), mice, cockroach(es), termite(s), and bedbug(s). We assess the presence of these metaphors in texts by comparing the input text's embeddings to this vector representation.

[3]The 42 randomly selected examples contain 19 instances of hate speech and 23 non-hate speech labels.

[4]The randomly selected examples contain 414 instances of hate speech and 492 non-hate speech labels.

[5]We use SpaCy's en_core_web_sm model.

3.3 Pretrained Models

In our experiments, we utilize both GPT-4 and GPT-3.5-turbo, which are among the top-performing closed-source NLP models across various tasks. These models are trained on diverse datasets and are capable of understanding and generating human-like text, making them suitable for a wide range of NLP applications. Additionally, we evaluate LLAMA-2-70B[6] (Touvron et al., 2023b), which is one of the leading open-source models in NLP. Similar to the GPT models, LLAMA-2 supports prompt-based usage, allowing it to be applied to our task without the need for additional fine-tuning.

A key determining factor in the success of state-of-the-art pretrained models are their corresponding prompts for each task. A prompt acts as the initial query or instruction, guiding the model to produce the desired output. In this paper, we explore three primary prompting schemes:[7]

Zero-shot: In this setting, the prompt consists of the phrase "Identify target groups and decide if they're dehumanized". This scheme assesses the pretrained model's preexisting knowledge about dehumanization.

Few-shot: We enhance the model's exposure by incorporating five randomly selected instances of dehumanization targeting Muslims into the prompt. This method allows us to evaluate the model's ability to generalize its understanding to other targeted groups, emphasizing the importance of specific examples in improving performance.

Explainable prompting: Building on the few-shot setting, this approach further requires the model to provide explanations for its decisions.

In the zero-shot setting, the model identifies target groups and determines whether the text contains dehumanizing language for that target group. In the few-shot setting, the model goes further by classifying dehumanization within texts as either "blatant" or "subtle". The included few-shot examples with dehumanizing language are labeled as "blatant". The explainable prompting setting mirrors the few-shot approach but adds a requirement for the model to explain its reasoning.

[6]We use the Llama-2-70b-chat-hf model

[7]The prompt templates for each of these settings are included in the appendix.

4 Results

Table 1 presents the results of the models in *zero-shot*, *few-shot*, and *explainable* settings, compared against the four components of *Negative evaluation of a target group*, *Denial of agency*, *Moral Disgust*, and *Use of vermin metaphors*. The *Combination* row shows the results where we consider a text as dehumanization if it contains all four components.

In the *few-shot* setting, a text is flagged for dehumanization if the predicted label for any of the identified targets is true. For *zero-shot* and *explainable*, a text is considered dehumanizing if a "blatant" label is predicted for any of the identified targets. Including both "blatant" and "subtle" labels as dehumanization lowers all models' performance due to reduced precision. The results of this setting are reported in Table 6 in the appendix.

Table 1 reveals the following insights: (1) *Model Performance*: GPT models significantly outperform heuristic components in detecting dehumanizing language, even in the zero-shot setting. In contrast, the LLAMA-2 model tends to overclassify inputs as dehumanizing. (2) *Room for Improvement*: There is substantial room for improvement in this task, as the best accuracy for distinguishing dehumanization from other types of hate speech is only 70%. (3) *Discriminating Dehumanization from Other Hate Speech*: The accuracy of the examined models in recognizing dehumanizing language versus neutral text is higher than their accuracy in discriminating dehumanizing language from other types of hate speech, as indicated by the higher performances in the "general dehumanization" subset. (4) *Impact of Incorporating Targeted Group Details*: Including details about targeted groups in prompts enhances dehumanization detection for the GPT-3.5 model, as shown by improved results in the few-shot and explainable settings within the "Targeted Dehumanization" subset. This improvement is not observed with the GPT-4 model. (5) *Benefits of Explanation*: Requesting explanations for predictions improves the GPT models' ability to differentiate dehumanization from other hate speech, as demonstrated by the higher performance of *explainable* settings over *few-shot* in the "Dehumanization vs. Hate" subset. However, this effect is not observed for LLAMA-2.

5 Analysis

In this section, we analyze the results of GPT models that achieve the highest scores in our evaluation

		Targeted Dehumanization			General Dehumanization			Dehum. vs Hate		
		F_1(other)	F_1(dehum.)	Acc.	F_1(other)	F_1(dehum.)	Acc.	F_1(hate)	F_1(dehum.)	Acc.
Zero-shot	GPT-4	61.76	74.00	69.05	69.73	78.61	74.93	44.80	70.59	61.62
	GPT-3.5	65.75	73.68	70.24	66.36	75.81	71.82	51.90	70.83	63.69
	LLAMA-2	17.45	68.95	54.87	13.92	68.28	53.64	2.84	66.89	50.61
Few-shot	GPT-4	77.33	81.72	79.76	77.09	81.76	79.69	59.41	74.91	68.99
	GPT-3.5	**81.01**	**82.76**	**81.93**	77.13	74.00	75.66	68.01	68.67	68.34
	LLAMA-2	38.38	69.65	59.33	36.77	68.43	57.88	27.87	69.86	57.49
Explainable	GPT-4	73.97	80.00	77.38	**77.38**	**82.02**	**79.97**	59.19	**76.29**	**70.00**
	GPT-3.5	79.07	78.05	78.57	76.15	74.37	75.29	**68.15**	69.96	69.08
	LLAMA-2	13.41	62.83	47.99	33.82	60.56	50.57	32.08	57.74	47.90
Negative Eval.		67.20	4.65	51.19	66.34	5.26	50.33	65.89	5.21	49.83
Agency Denial		64.41	16.00	50.00	62.53	18.56	48.68	63.47	18.84	49.61
Moral Disgust		46.91	50.57	48.81	44.18	46.71	45.47	44.67	46.47	45.58
Vermin Meta.		38.46	46.67	42.86	42.43	45.77	44.15	41.73	45.04	43.43
Combination		67.20	4.65	51.19	66.72	0.44	50.11	66.62	0.44	50.00

Table 1: Comparison of identifying dehumanizing language across different models and settings, focusing on instances explicitly labeled as "blatant" dehumanization. The lower section includes results from Mandelsohn's components. The *Combination* baseline classifies an example as dehumanization if it contains all four components.

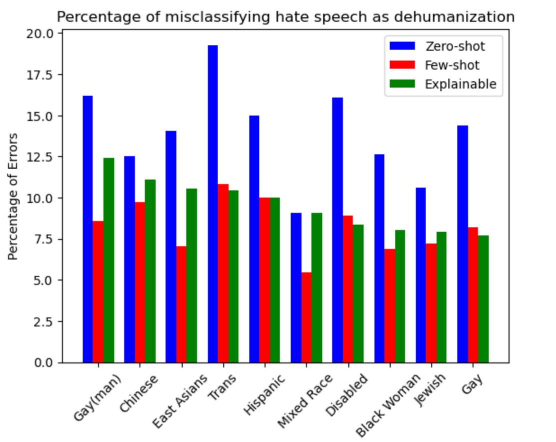

Figure 1: Top 10 target groups with the highest over-sensitivity error ratios for GPT-3.5.

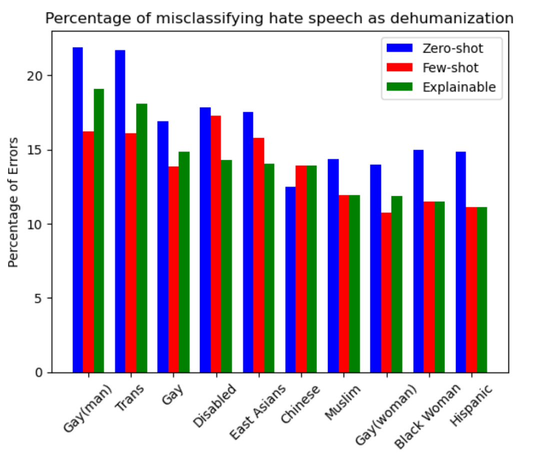

Figure 2: Top 10 target groups with the highest over-sensitivity error ratios for GPT-4.

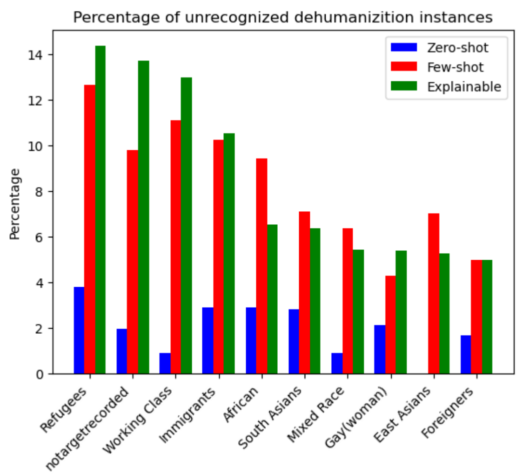

Figure 3: Top 10 target groups with the highest ratio of unrecognized dehumanizition instances for GPT-3.5.

to further explore their shortcomings and strengths.

Does the model perform equally well for different target groups? To address this question, we calculate two types of errors for each target group: (1) over-sensitivity, where the model inaccurately labels less severe hate speech as dehumanization, and (2) recognition blindness, defined by the ratio of instances of dehumanizing language that remain undetected by the model for a specific target group.

We calculate the over-sensitivity error ratio for each target group by dividing the number of instances misclassified as dehumanization by the total number of instances for that group within the "Dehumanization vs. Hate" evaluation set. Figures 1 and 2 show the top 10 target groups with

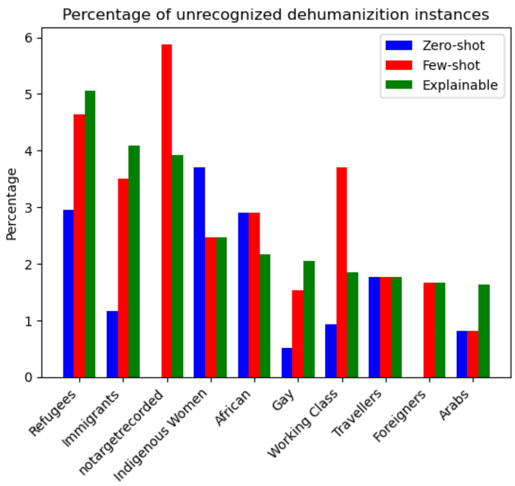

Figure 4: Top 10 target groups with the highest ratio of unrecognized dehumanizition instances for GPT-4.

the highest error ratios for the GPT-3.5 and GPT-4 models, respectively. Target groups are ordered according to their error ratios in the explainable setting, which performs best in the "Dehumanization vs. Hate" subset.

Figures 3 and 4 show the 10 target groups with the highest ratios of recognition blindness, measured by the ratio of instances per target group containing dehumanizing language that the model fails to recognize.

The results reveals that: (1) The error ratio in the zero-shot setting shows significant variability, while few-shot and explainable settings exhibit more consistency in error ratios; (2) The GPT-4 model demonstrates a higher sensitivity to classify hate speech as dehumanization, as evidenced by elevated error ratios among its top 10 target groups in Figure 2; and (3) More importantly, both models exhibit varying levels of sensitivity towards different target groups, overclassifying less severe hate speech as dehumanization for certain groups like gay and transgender individuals, yet more frequently failing to detect dehumanizing language targeting groups such as refugees, immigrants, and the working class. This discrepancy highlights the models' inherent biases towards different target groups.

Table 2 shows some examples in which the zero-shot, few-shot, and explainable settings of GPT-3.5 detect other types of hate speech as dehumanizing language. Similarly, Table 3 shows some examples of dehumanizing language that remained undetected across all three settings of GPT-3.5.

Which types of hate speech does the model frequently confuse with dehumanization? In Figure 5, we present the number of instances from various hate speech categories in the evaluation set that are classified incorrectly by GPT-3.5. Notably, a majority of these errors occur when the model identifies instances of "derogation" as dehumanization.

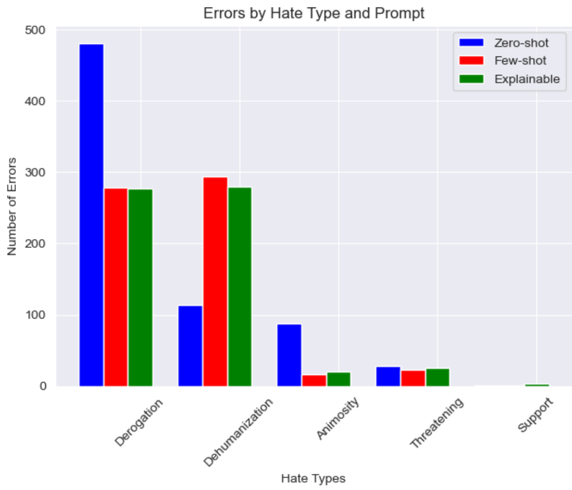

Figure 5: Number of instances from different types of hate speech that are incorrectly classified in "Dehumanization vs. Hate".

6 How Effective is GPT for Annotating Training Data?

The GPT-3.5 and GPT-4 models, being closed-source, might not be accessible to all users. Nevertheless, the task of identifying dehumanizing language has the potential to serve a diverse user base, and its detection shouldn't be limited by users' computational capabilities or the need for paid API access. Inspired by the widespread success of state-of-the-art models, including GPT models, in generating annotated instances for various existing tasks (He et al., 2023; Alizadeh et al., 2023; Latif et al., 2023; Ding et al., 2023; Thapa et al., 2023; Ziems et al., 2023; Gilardi et al., 2023; Törnberg, 2023; Hartvigsen et al., 2022), in this section, we investigate the effectiveness of using GPT-3.5 for generating annotated data to train more accessible (in terms of model size) and open-source models.[8] For this purpose, we employ two different open-source and accessible models: Roberta (Liu et al., 2019)

[8]We use GPT-3.5 because its performance is on-par with that of GPT-4 while it is less expensive.

Text	Label	Zero	Few	Exp.
Why don't [XX] just STFU and stay into the kitchen	Derogation	Dehuman	Blatant	Blatant
It gets to a point where you don't care or even lust for XX anymore. I truly don't think about those deceitful <AA> and it's been like this for years. Never trust one again	Derogation	Dehuman	Subtle	Subtle
It's completely shit that <BB> like him are living and breathing	Derogation	Dehuman	Blatant	Blatant
Next time i see a <CC> person, I'm going to b e a t them up	Threatening	Blatant	Blatant	Blatant
someone who had a strong year and made over 75k in 2018 and who may be struggling currently gets zero but illegals and refugees get $350 mil	Animosity	Dehuman	Blatant	Subtle

Table 2: Examples from GPT-3.5 outputs when the model incorrectly classifies other types of hate speech as dehumanization. Some parts of input texts are masked out to reduce their toxicity. The "Label" column shows the annotated label of the input.

Given that China, Russia and USA are not responsible for the vast majority of war crimes, they should not be the ones responsible
for these swarms of <AA> it's like a plague, the way these <BB> just come spilling over into all corners of the uk
<CC> are the worst kind of dogs and I'm not talking about the animals
<DD> are like puppies, really nice but you get tired of them after some time
The british government is massively fucking up its response to the <EE> crisis, why cant they get rid of this fucking disease?

Table 3: Examples from the evaluation set where the dehumanizing language is not detected by any of the GPT-3.5 variants. Some parts of the input texts are masked out to reduce their toxicity.

Input: For only 20 rupees; [NAME], [NAME], [NAME], [NAME], [NAME], [NAME] and [NAME] publicly assaulted and humiliated Saleem to the point that he ended his life by jumping in front of a train but there won't b any media outrage nor any one would bother to tweet coz d victim is a Muslim **Target:** Saleem **Dehumanization:** Blatant **Explanation:** Saleem is publicly assaulted and humiliated to the point that he ended his life. This is a clear form of dehumanization as his life is disregarded and his suffering is not taken seriously. **Target:** Muslim **Dehumanization:** subtle **Explanation:** The text mentions that there won't be any media outrage or anyone bothering to tweet because the victim is a Muslim. This implies a subtle form of dehumanization, suggesting that the suffering of a Muslim is not as important or worthy of attention as that of others.
Input: This night is very important to all Muslim.May Allah accept everyone's prayers "Ameen" shab e meraj Mubarak to all of you **Target:** Muslim **Dehumanization:** None **Explanation:** The group is mentioned in a respectful and positive manner, with a message of well-wishes and prayers.

Table 4: Samples from the GPT-3.5 output on the collected data. The proper names are anonymized in the input for privacy. Note that each input sentence may have multiple labels per input if it contains more than one target.

and Flan-T5 (Chung et al., 2022).[9] We utilize various model sizes, including Roberta-large (355M parameters), Flan-T5-base (248M), Flan-T5-large (783M), and Flan-T5-XL (3B).

6.1 Automatic Data Annotation

This section outlines the methodology for automatically generating a dataset annotated with instances of dehumanizing language. We use the *Explainable* setting of GPT-3.5 because of its performance in identifying dehumanizing language and distinguishing it from other forms of hate speech. We collected seed data using *snscrape*,[10] a tool for scraping social networking services, to collect 1 million Tweets related to Muslims.[11] From this collection, we randomly selected 20,000 tweets and annotated them by GPT-3.5, yielding 1,208 tweets explicitly marked as instances of blatant dehumanization. Table 4 provides a few examples of the resulting annotated data. These annotations serve as the basis for training smaller models.

6.2 Model Training

We conduct training in two distinct settings to evaluate the impact of dataset size on model performance: (1) using 400 automatically annotated examples, balanced with 200 dehumanization and 200 non-dehumanization texts, and (2) expanding to 2,000 instances, with 1,000 identified by GPT-

[9] Flan-T5 is also a prompt-based approach, for which we use the prompt "Classify this text as either 0 (not dehumanising) or 1 (dehumanising). Text:tweet Answer(0 or 1):"

[10] https://github.com/JustAnotherArchivist/snscrape

[11] Due to recent Twitter policy updates, accessing such data directly is now restricted and subject to the costs associated with the Twitter API.

Train	Model	Dev Acc.	Targeted Dehumanization F_1(other)	Targeted Dehumanization F_1(dehum.)	Targeted Dehumanization Acc.	General Dehumanization F_1(other)	General Dehumanization F_1(dehum.)	General Dehumanization Acc.	Dehum. vs Hate F_1(hate)	Dehum. vs Hate F_1(dehum.)	Dehum. vs Hate Acc.
	Explainable (blatant)	-	79.07	78.05	78.57	76.15	74.37	75.29	68.15	69.96	69.08
400	Roberta-large	79.25	69.85	70.38	70.24	65.88	62.87	64.54	58.82	59.68	59.43
	Flan-base	74.25	53.16	70.59	64.29	52.10	65.08	60.12	45.98	63.20	56.74
	Flan-large	68.75	48.79	69.08	61.51	51.13	66.83	60.58	40.88	63.98	55.32
	Flan-XL	81.25	66.64	71.12	69.05	64.37	66.15	65.29	55.11	62.35	59.05
2K	Roberta-large	80.35	67.71	68.24	68.10	66.95	65.63	66.42	59.42	62.20	**61.02**
	Flan-base	74.00	63.19	68.49	66.67	62.44	63.72	63.71	56.17	61.23	59.59
	Flan-large	75.67	69.31	**74.63**	**72.22**	62.41	**69.66**	66.43	51.52	**65.77**	59.88
	Flan-XL	76.63	64.40	71.66	68.45	63.79	68.91	**66.67**	53.79	65.11	60.40

Table 5: Investigating the impact of training smaller, open-source models with annotations generated by GPT-3.5 using 'Explainable' prompting, comparing the effects of fine-tuning the model with (1) 400 annotated examples and (2) 2,000 annotated examples.

3.5 as dehumanizing. This approach allows us to explore how varying amounts of automatically annotated data influence performance outcomes. While automatic annotation is less expensive than using human experts, it still requires a paid API. Therefore, it is important to examine the impact of the number of training instances on performance.

Table 5 shows the results of this experiment. The "Explainable" row displays the results of the GPT-3.5 model, which is used for annotating additional data. The findings indicate a general trend: all models fine-tuned with the automatically annotated data underperform compared to the GPT-3.5 annotation model. However, an increase in training data volume correlates with improved accuracy across models.[12] Overall, the FLAN-large model achieves higher accuracy and F_1 scores in detecting dehumanization in various evaluation sets. Using FLAN-XL instead of FLAN-large does not show a significant advantage.

7 Conclusion

The automatic identification of dehumanizing language is a crucial task, given its role in spreading subtle and harmful hate speech with severe consequences, especially for marginalized communities. In this paper, we explored the use of state-of-the-art NLP models to identify dehumanizing language. While our findings show considerable promise, there are still various directions for future research. These include enhancing the models' ability to distinguish dehumanizing language from other forms of hate speech. More importantly, our analysis highlights potential disparities in the models' effectiveness across different target groups. This raises caution about relying on these models for broad conclusions when analyzing large-scale data for social research on different target groups.

[12]The exception is the results of the Roberta-large model on the "Targeted Dehumanization" evaluation set. However, due to the small size of this evaluation set, the differences may not be significant.

8 Limitations

We have exclusively relied on Vidgen et al. (2021)'s dataset for dehumanization evaluation. The study could have yielded deeper insights with access to a dataset that categorizes dehumanization into blatant and subtle instances. Additionally, our evaluation of the impact of automatic annotation is limited to Twitter data related to a single target group. A more comprehensive assessment would involve multiple target groups. It's important to acknowledge that recent restrictions on social media APIs have posed challenges in this field of study. Moreover, there is a potential risk in deploying NLP models to detect dehumanizing language: the inadvertent reinforcement of biases. Models may disproportionately flag or overlook certain groups' speech, sustaining inequality and suppressing free expression.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*.

Agostina Calabrese, Björn Ross, and Mirella Lapata. 2022. Explainable abuse detection as intent classification and slot filling. *Transactions of the Association for Computational Linguistics*, 10:1440–1454.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jason C Deska, Jonathan Kunstman, E Paige Lloyd, Steven M Almaraz, Michael J Bernstein, JP Gonzales, and Kurt Hugenberg. 2020. Race-based biases in judgments of social pain. *Journal of Experimental Social Psychology*, 88:103964.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *24th European Conference on Artificial Intelligence - ECAI 2020*, pages 2006–2013. IOS Press.

Scott E Friedman, Ian Magnusson, Sonja Schmer-Galunder, Ruta Wheelock, Jeremy Gottlieb, Christopher Miller, et al. 2021. Toward transformer-based nlp for extracting psychosocial indicators of moral disengagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

John Hagan and Wenona Rymond-Richmond. 2008. The collective dynamics of racial dehumanization and genocidal victimization in darfur. *American Sociological Review*, 73(6):875–902.

Lasana T Harris and Susan T Fiske. 2015. Dehumanized perception. *Zeitschrift für Psychologie*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and social psychology review*, 10(3):252–264.

Nick Haslam, Yoshihisa Kashima, Stephen Loughnan, Junqi Shi, and Caterina Suitner. 2008. Subhuman, inhuman, and superhuman: Contrasting humans with nonhumans in three cultures. *Social cognition*, 26(2):248–258.

Nick Haslam and Steve Loughnan. 2014. Dehumanization and infrahumanization. *Annual review of psychology*, 65:399–423.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Nour S. Kteily and Alexander P. Landry. 2022. Dehumanization: trends, insights, and challenges. *Trends in Cognitive Sciences*, 26(3):222–240.

Siddique Latif, Muhammad Usama, Mohammad Ibrahim Malik, and Björn W Schuller. 2023. Can large language models aid in annotating speech emotional data? uncovering new frontiers. *arXiv preprint arXiv:2307.06090*.

Jacques-Philippe Leyens, Paola M Paladino, Ramon Rodriguez-Torres, Jeroen Vaes, Stephanie Demoulin, Armando Rodriguez-Perez, and Ruth Gaunt. 2000. The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Personality and social psychology review*, 4(2):186–197.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184.

Anthony Oberschall. 1997. Vojislav Seselj's nationalist propaganda: contents, techniques, aims and impacts, 1990-1994. How mass media propaganda impacts on ordinary people's acceptance and participation in collective violence, and how Seselj's nationalist propaganda promoted and justified coercion and violence by the Serbs against non-Serbs.

Maria-Paola Paladino, Jacques-Philippe Leyens, Ramon Rodriguez, Armando Rodriguez, Ruth Gaunt, and Stéphanie Demoulin. 2002. Differential association of uniquely and non uniquely human emotions with the ingroup and the outgroup. *Group Processes & Intergroup Relations*, 5(2):105–117.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2015. Connotation frames: A data-driven investigation. *arXiv preprint arXiv:1506.02739*.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2329–2334.

Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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)*, pages 1667–1682, Online. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

A Which model identifies more instances as dehumanization?

To determine which model most frequently identifies instances of dehumanization, we examine the recall and precision values for these models in detecting dehumanization labels across different evaluation subsets. These results are reported in Tables 7 and 8.

We find that, with the exception of the "targeted dehumanization" subset, the zero-shot settings detect more instances of dehumanization compared to few-shot and explainable settings. This is particularly evident in the "Dehumanization vs. Hate" subset for the GPT-3.5 model, indicating that without specific examples, GPT-3.5 is prone to misclassifying hateful content as dehumanization. Table 2 provides examples of such misclassifications from the evaluation set. Additionally, considering both "blatant" and "subtle" labels as indicators of dehumanization significantly increases recall but reduces precision.

B Prompt templates

Zero-shot Prompt `You are an expert in psychology.\nText: {text}.\nFirst, Identify target groups in the text\nDecide whether each target is dehumanised by the author.\nOnly response in the example output format that can be parsed by json:\n{{\n "Targets": ["group_name1", "group_name2", "..."],\n "Dehumanisation": [\n ["target1", true],\n ["target2", false], \n ["...", true]\n]\n}}\n`

Few-shot Prompt `You are an expert in psychology and NLP.Examples of dehumanization:\n{examples}\n\nAnalyze the following text and identify all`

		Targeted Dehumanization			General Dehumanization			Dehum. vs Hate		
		F_1(other)	F_1(dehum.)	Acc.	F_1(other)	F_1(dehum.)	Acc.	F_1(hate)	F_1(dehum.)	Acc.
Few-shot	GPT-4	52.46	72.90	65.48	57.45	76.14	69.43	19.32	68.93	55.14
	GPT-3.5	**69.44**	76.59	73.49	**74.27**	77.99	**76.27**	**57.26**	**70.88**	**65.36**
	LLAMA-2	8.70	66.07	50.52	15.87	67.93	53.56	8.20	69.80	54.55
Explainable	GPT-4	48.28	72.73	64.29	52.04	74.78	66.94	15.22	70.64	56.39
	GPT-3.5	70.42	**78.35**	**75.00**	70.73	**78.20**	75.01	48.48	70.66	62.61
	LLAMA-2	5.00	62.30	46.02	13.47	61.42	46.64	8.36	61.28	45.56

Table 6: Comparison of identifying dehumanizing language across different models and settings, when both "blatant" and "subtle" predictions are classified as dehumanization.

	Targeted Dehumanization			General Dehumanization			Dehum. vs Hate		
	Recall	Precision	F_1(dehum.)	Recall	Precision	F_1(dehum.)	Recall	Precision	F_1(dehum.)
Zero-shot	83.33	66.03	73.68	88.18	66.47	75.81	88.18	59.18	70.83
Few-shot (blatant)	87.80	78.20	**82.76**	69.32	**79.34**	74.00	69.31	68.03	68.67
Explainable (blatant)	76.19	**80.00**	78.05	71.80	77.11	74.37	71.85	**68.16**	69.96
Few-shot (+subtle)	87.80	67.92	76.60	84.16	72.66	77.99	84.22	61.19	**70.88**
Explainable (+subtle)	**90.48**	69.09	78.35	**89.79**	69.26	**78.20**	**89.85**	58.23	70.66

Table 7: The Recall, Precision and F_1 values of GPT-3.5 model for recognizing the dehumanization label in the "Dehumanization vs. Hate" evaluation set in the *zero-shot*, *few-shot*, and *explainable* setting when the "blatant" predictions are considered as dehumanization.

	Targeted Dehumanization			General Dehumanization			Dehum. vs Hate		
	Recall	Precision	F_1(dehum.)	Recall	Precision	F_1(dehum.)	Recall	Precision	F_1(dehum.)
Zero-shot	88.10	63.79	74.00	92.05	68.59	78.61	92.05	57.24	70.59
Few-shot (blatant)	90.48	**74.51**	**81.72**	91.06	74.19	81.76	91.06	63.62	74.91
Explainable (blatant)	90.48	71.70	80.00	91.39	**74.39**	**82.02**	91.39	**65.46**	**76.29**
Few-shot (+subtle)	92.86	60.00	72.90	97.57	62.43	76.14	97.57	53.29	68.93
Explainable (+subtle)	**95.24**	58.82	72.73	**98.01**	60.45	74.78	**98.01**	55.22	70.64

Table 8: The Recall, Precision and F_1 values of GPT-4 model for recognizing the dehumanization label in the "Dehumanization vs. Hate" evaluation set in the *zero-shot*, *few-shot*, and *explainable* setting when the "blatant" predictions are considered as dehumanization.

the groups targeted in it. For each group, assess whether they are subjected to any form of dehumanization, be it blatant, subtle, or none at all.\n\nText: \n{text}\n\nMake sure target name exist and exactly the same in given text. Only response in the example output json format:\n [\n"target_name":target name in text, "dehumanize": "None" or "Subtle" or "Blatant"\n],...]\n

Explainabel Prompt You are an expert in psychology and NLP.Examples of dehumanization:\n{examples}\n\nAnalyze the following text and identify all the groups targeted in it. For each group, assess whether they are subjected to any form of dehumanization, be it blatant, subtle, or none at all.\n\nText: \n{text}\n\nMake sure target name exist and exactly the same in given text. Only response in the example output json format:\n [\n"target_name":target name in text, "dehumanize": "None" or "Subtle" or "Blatant", "explanation":explanation\n],...]\n

C Evaluating the Impact of Examples in the Prompts

We conducted an additional evaluation using two sets of cases from 42 dehumanizing texts targeting Muslims, subjectively chosen by the authors. One set of texts was closer to dehumanization (Example set 1), while the other set was closer to generalized hate speech (Example set 2). We repeated the experiments using GPT-3.5 with these two different example sets as the few-shot examples in both few-shot and explainable settings. The results are reported in Table 9. We observe that including more subtle forms of dehumanization as few-shot examples increases the F_1 score for detecting dehumanization. However, this also leads the model to misclassify more instances of other types of hate speech as dehumanization, resulting in lower performance in those categories. However, the overall accuracy does not change significantly by using different examples in the prompt.

Evaluation Sets This section will list two sets of 5 examples specifically chosen to differentiate between dehumanization and generalized hate speech. The selected examples aim to provide insight into the nuanced differences and challenges in classifying such texts, without further analysis.

D Analyzing Biases in Smaller Models

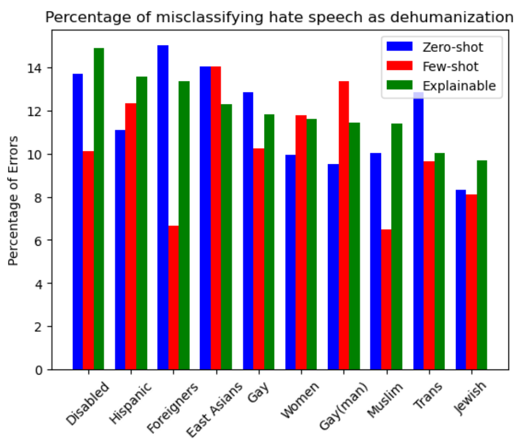

Figure 6: Top 10 target groups with the highest oversensitivity error ratios for Roberta.

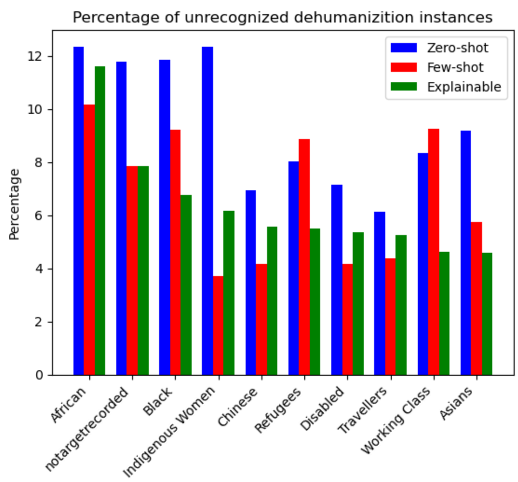

Figure 7: Top 10 target groups with the highest ratio of unrecognized dehumanizition instances for Roberta.

Figures 6 and 7 display the top 10 target groups with the highest error ratios and recognition blindness for the fine-tuned RoBERTa-large model with 400 annotated examples. We observe some overlap between the biases of the examined model and the annotator model.

E Hyperparameters and Training Configuration

This appendix outlines the key hyperparameters used for the Flan T5 model training:

- **Batch Size**: The batch size for both training and evaluation is set to 8.
- **Learning Rate**: The learning rate for the model is configured at 5×10^{-5}.
- **Number of Training Epochs**: The model training is set to run for 15 epochs.

F Example of output

	Targeted Dehumanization			General Dehumanization			Dehum. vs Hate		
	F_1(other)	F_1(dehum.)	Acc.	F_1(other)	F_1(dehum.)	Acc.	F_1(hate)	F_1(dehum.)	Acc.
Few-shot (Example 1)	71.15	70.13	72.29	77.31	72.27	75.04	70.58	68.35	69.5
Few-shot (Example 2)	77.5	79.07	78.31	75.93	76.97	76.46	65.4	71.59	68.8
Explainable (Example 1)	74.15	70.89	72.61	76.69	73.23	75.08	68.42	68	68.21
Explainable (Example 2)	74.66	79.57	77.38	74.43	77.26	75.92	63.73	72.24	68.55
Few-shot (Example 1+subtle)	68.42	73.33	71.08	74.06	77.63	75.98	58.89	71.21	66.13
Few-shot (Example 2+subtle)	71.43	79.17	75.9	70.71	77.72	74.7	52.15	70.93	63.83
Explainable (Example 2+subtle)	73.23	80.41	77.38	71.35	78.88	75.69	47.13	70.7	62.3
Explainable (Example 2+subtle)	59.38	75	69.04	65.83	77.45	72.83	38.62	69.62	59.36

Table 9: Evaluating the impact of few-shot examples in the prompt for the GPT-3.5 model when (1) the examples are selected from clear, blatant cases of dehumanization (Example 1), and (2) the examples are selected from more subtle dehumanization (Example 2).

Input: For only 20 rupees; [NAME], [NAME], [NAME], [NAME], [NAME], [NAME] and [NAME] publicly assaulted and humiliated Saleem to the point that he ended his life by jumping in front of a train but there won't b any media outrage nor any one would bother to tweet coz d victim is a Muslim
Target: Saleem
Dehumanization: Blatant
Explanation: Saleem is publicly assaulted and humiliated to the point that he ended his life. This is a clear form of dehumanization as his life is disregarded and his suffering is not taken seriously.
Target: Muslim
Dehumanization: subtle
Explanation: The text mentions that there won't be any media outrage or anyone bothering to tweet because the victim is a Muslim. This implies a subtle
form of dehumanization, suggesting that the suffering of a Muslim is not as important or worthy of attention as that of others.

Input: This night is very important to all Muslim.May Allah accept everyone's prayers "Ameen" shab e meraj Mubarak to all of you
Target: Muslim
Dehumanization: None
Explanation: The group is mentioned in a respectful and positive manner, with a message of well-wishes and prayers.

Input: In Karnataka, Muslim Couple Who Came To Watch Kantara In Theatre Assaulted By MobReportedly some youth on the premises warned them against watching the movie.
Target: Muslim Couple
Dehumanization: Subtle
Explanation: The Muslim couple is subjected to subtle dehumanization through the use of the word 'reportedly', which suggests that the incident may not have actually occurred or that the couple's account may not be reliable.

Input: These videos should be watched by everyone on the planet. NAME1, who admitted to killing a pregnant Muslim woman and several others, has been acquitted. In contrast, NAME2, who advocates for peace, is imprisoned. Dear Universe, THIS IS INDIA.
Target: Muslims
Dehumanization: Blatant
Explanation: The text refers to a pregnant Muslim woman being killed and implies that the acquittal of the perpetrator is unjust. This suggests a dehumanization of Muslims as their lives are not valued or protected.
Target: NAME2
Dehumanization: None
Explanation: The text states that NAME2 advocates for peace and is imprisoned, but it does not contain any dehumanizing language or implications.

Table 10: Examples from the results of the GPT-3.5 output on the collected data. The proper names are anonymized in the input for privacy. Note that each input sentence may have multiple labels per input if it contains more than one target.

Input: just watched news and one of the news is black man killing a white man and got charged well a dog biting a man is not news. It’s what us niggers go through all the time
Target: Black man
Dehumanization: Subtle
Explanation: The comparison of a black man’s actions to a dog biting a man subtly dehumanizes black individuals by equating their actions to those of an animal, suggesting that their behavior is expected to be violent or criminal, which is a form of dehumanization.
Target: niggers
Dehumanization: Blatant
Explanation: The use of the term ’niggers’ is a blatant form of dehumanization. It is a derogatory term that has historically been used to dehumanize and demean black people, reducing them to a status less than human.

Input: I have a lot of muslim students enrolled on my course. This is going to be a looooong semester trying to get those mozzies to understand anything
Target: Muslim students
Dehumanization: Subtle
Explanation: The term ’mozzies’ is used pejoratively to refer to Muslim students, which is a form of subtle dehumanization. The statement implies a generalized and negative stereotype about the intelligence or capability of Muslim students, suggesting they will be difficult to teach due to their religious or cultural background.

Input: Widely predicted to have his sights on the Labour leadership following [NAME1] troubled tenure in office, [NAME2] is quickly raising to the leadership of the Labour Party, which would result in the official opposition to the British government falling under Muslim control and making the Muslim takeover of the British government a nightmarish spectre.
Target: Khan
Dehumanization: None
Explanation: The text does not dehumanize Khan directly but rather discusses his potential political rise.
Target: Labour Party
Dehumanization: None
Explanation: The Labour Party is mentioned in a political context without any dehumanizing language directed towards it.
Target: Muslim
Dehumanization: Blatant
Explanation: The phrase ’Muslim takeover of the British government’ implies a hostile, dehumanizing view of Muslims, suggesting they are a monolithic group with a singular, threatening agenda towards the British government.

Table 11: Examples from the results of the GPT-4 output on the evaluation data. Note that each input sentence may have multiple labels per input if it contains more than one target.