Hate Speech Detection and Reclaimed Language: Mitigating False Positives and Compounded Discrimination

Eszter Zsisku[1], Arkaitz Zubiaga[1], Haim Dubossarsky[1,2,3]
[1] Queen Mary University of London, [2] University of Cambridge, [3] The Alan Turing Institute
ezsisku@gmail.com,{a.zubiaga,h.dubossarsky}@qmul.ac.uk

ABSTRACT

While minimising false negatives in hate speech classification remains an important goal in order to reduce discrimination and increase fairness for online communities, there is a growing need to produce models that are sensitive to nuanced language use. This is particularly true for terms that may be considered hateful in certain contexts, but not others. The LGBTQ+ community has long faced stigmatisation and hate, which continues to be the case online. There has been a rise in appreciation and understanding of this community's use of "mock impoliteness" and the reclaiming of language that has traditionally been used derogatorily against them. Reclaimed language in particular presents a challenge in the field of hate-speech detection. As a first-of-its-kind study looking into the impact of reclaimed language on hate speech detection models, we create a novel dataset, Reclaimed Hate Speech Dataset (RHSD), which enables investigation into the phenomenon. Through the use of a state-of-the-art hate speech detection model, we demonstrate that models may inadvertently discriminate against the LGBTQ+ community's reclaimed language use through misclassifying such content as hateful. As a result, there is a risk of compounding discrimination against this population through restricting their language use and self-expression. In response to this issue, we produce a fine-tuned hate-speech detection model which aims to minimise false positive classifications of reclaimed language. By creating and publishing the first dataset that focuses on reclaimed language and investigating its impact on hate speech detection models, our research highlights the importance of semantically aware approaches to hate-speech detection that are not overly reliant on individual words or phrases associated with hate. We thus establish a benchmark methodology for further investigation into reclaimed language, that promises to support marginalised groups, taking into account the intersectional nature of their discourse.

NB: Readers should be advised that this paper contains use of and reference to racial, homophobic and transphobic slurs which they may find triggering. References to sentences containing such slurs are purely for illustration purposes and in no way reflect the author's attitudes or opinions.

WEBSCI '24, May 21–24, 2024, Stuttgart, Germany

ACM ISBN 979-8-4007-0334-8/24/05
https://doi.org/10.1145/3614419.3644025

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Philosophical/theoretical foundations of artificial intelligence**; **Discourse, dialogue and pragmatics**.

KEYWORDS

Discrimination, Hate speech classification, LGBTQ+, Natural language processing, NLP, Reclaimed language, Semantic understanding

ACM Reference Format:
Eszter Zsisku[1], Arkaitz Zubiaga[1], Haim Dubossarsky[1,2,3]. 2024. Hate Speech Detection and Reclaimed Language: Mitigating False Positives and Compounded Discrimination. In *ACM Web Science Conference (WEBSCI '24), May 21–24, 2024, Stuttgart, Germany*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3614419.3644025

1 INTRODUCTION

With the worldwide adoption and popularity of social media, online communication has given rise to several challenges. The ability to reach wide audiences and in some cases maintain anonymity has led to a rise in prevalence of toxic, abusive and hateful behaviour in online communities such as Reddit and Twitter [24].

Researchers define hate speech as language which threatens or offends groups based on characteristics such as race, religion, ethnicity, gender, sexuality or disability [4, 8, 18]. Language that incites violence or promotes prejudice or stereotypical views of protected groups can also be defined as hateful [5] and some authors highlight that more subtle language that implies disrespect or hostility should also be included under the umbrella of hate speech [21]. For online communities, the challenge becomes being able to balance user base desires to express opinions and share ideas, whilst safeguarding individuals with protected characteristics from discrimination and abuse, thus creating a cohesive and fair set of rules for online spaces [6, 7]. Efforts have therefore increased to moderate such content, both in terms of manual flagging and reporting of hateful content, and the deployment of Machine Learning algorithms which aim to accurately detect hate speech [14].

The efficacy and generalisability of computational models for hate speech detection, which are essentially classifiers from a Machine Learning point of view, is highly dependent on the quality and diversity of their training datasets [27]. For labellers providing the ground truth for these datasets, challenges arise due to the presence of ambiguous slang words, sarcasm, and profanity [28]. These complexities are in turn mirrored by a classifier's performance: it must gain a semantic understanding of the relationship between tokens in each sentence in order to perform optimally. To achieve this goal, efforts have been directed towards extracting meaningful features for hate speech identification [3]. More recently, Vidgen

et al. [20] produced a large corpus on which to train a RoBERTa classifier. This dataset includes four rounds of increasingly sophisticated methods of delivering hate-speech, as well as adversarial, non hateful examples. The nuanced use of semantic context in this dataset reflects the nature of language used online: highly diverse, prone to misinterpretation, and often context-dependent.

An area of research which highlights the importance of context examines the phenomenon of language reclamation, where traditionally hateful slurs are embraced and repurposed by the communities they were initially used to disparage [19]. An example of reclaimed language is the use of the word "bitch", a historically derogatory word used towards women. "Bitch" is popularly used in the drag community (a subcommunity of the LGBTQ+ domain) in a neutral, exclamatory or even endearing sense. This demonstrates how the typically negative stereotype of gay males being effeminate is reclaimed and celebrated by the drag community.

We anticipate that reclaimed language is one such case where relying solely on individual words or phrases for hate speech classification does not suffice, as the meaning and intent behind such reclaimed language may differ significantly from their original derogatory usage. As a result, computational models for hate speech detection might not be able to differentiate between hate speech and reclaimed language, with significant societal implications and the potential for compounded discrimination of already marginalised groups such as the LGBTQ+ community.

Our research makes the first such effort in investigating the impact of reclaimed language on hate speech detection, and highlights the need for capturing more nuanced elements of language, such as the cultural context in which it is used [14]. To enable this research, we create the first hate speech dataset that incorporates reclaimed language labelled as such. We first assess whether a state-of-the-art hate speech detection model can effectively handle reclaimed language, where we observe significant deficiencies owing to high rate of false positives. Then, we address this challenge through fine-tuning the previously described hate speech classifier [20] using a curated dataset that is composed of both "standard" hate speech and reclaimed language. This Reclaimed Hate Speech (RHS) model is then tested on both standard hate speech and reclaimed language. Additionally, we test the model's performance on a hate speech test suite, and an unseen "hateful" target word to assess its robustness and generalisability. We demonstrate that the RHS model can retain an overall competitive performance detecting hate speech while improving on detection of reclaimed language, which in turn highlights the need for considering the latter when building hate speech detection models. Our research contributes with insights into a novel direction of research looking into reclaimed language, highlighting the need for further exploration into this understudied phenomenon.

Our work makes the following novel contributions:

- We collect Reclaimed Hate Speech Dataset (RHSD), the first hate speech dataset that is dedicated to investigating reclaimed language, and provides a labelled subset of instances of reclaimed language. The dataset will be released upon publication.
- Using a state-of-the-art hate speech detection model, we investigate the impact of reclaimed language on the model, showing the model's tendency of falsely predicting cases of reclaimed language as hate speech. This demonstrates the inequality that marginalised groups experience from current models.
- We further fine-tune a baseline model on samples including reclaimed language, creating the Reclaimed Hate Speech (RHS) model, demonstrating and stressing the need for incorporating this knowledge into hate speech detection models for a broader and more generalisable handling of hate speech in the presence of reclaimed language.
- This study quantifies and highlights the need for further research and more careful investigation into ways of mitigating the impact of reclaimed language on hate speech detection models, which we further enable by establishing a benchmark methodology for the research. It demonstrates not only the shortcomings of current computational approaches in providing inclusive environment, but also the potential to alleviate some of the ethical challenges modern computational approaches are faced with, improving fairness in modern AI tools.

2 RELATED WORK

Hate speech detection is typically approached as a classification task, which is however challenging due to the need of understanding the meaning of a sentence at a semantic level to determine if it is hateful [23]. Protected characteristics may not be outrightly insulted: instead, the propagation of harmful stereotypes, recounting of negative encounters, or sarcastic mockery of behaviour or culture are all examples of hate speech, which are much more difficult to be detected. As such, bag-of-word approaches to hate speech classification tend to have poor precision due to their tendency to associate profane words with hate speech, leading to high false positive rates [2, 11]. Even the presence of slurs with typically hateful connotations may not suffice: the reclamation of racial slurs by the African-American community illustrates the importance of a semantic level analysis of sentences [26]. Even with the emergence of more sophisticated Neural Network models, like transformer, research has found that their predictions are heavily impacted by the presence of keywords that can be abusive in certain contexts [28].

The high false positive rate is not only evident with automated models, but also with inexperienced human annotators [3, 4], who have a tendency of labelling texts as abusive based on the presence of certain keywords [22]. Furthermore, those without understanding or appreciation of the context in which language is used may also falsely label certain texts as hateful, producing annotator bias which can, for example, introduce racial bias into hate speech detection models [17]. Sap et al. [17] demonstrate such bias with example usage of the word "n*gga" or the suffix "-ass", which the authors explain are considered harmless in African American English (AAE) and yet without knowing the ethnicity of the writer, are interpreted as hateful. Xia et al. [26] addressed the presence of bias in classification of text containing AAE through an adversarial debiasing procedure that was able to reduce false positive rates of hate classification for AAE text whilst maintaining comparable accuracy for other forms of hate-speech.

Similar biases exist for another protected community. Thiago et al. [19] identified that Perspective API's toxicity detection tool appeared to over-classify text from American drag queens as toxic without regard for the social function of the language they use. In a significant number of cases, drag queen accounts were found to have higher perceived toxicity levels than the politically controversial Donald Trump, as well as white nationalist users. Further word-level analysis uncovered words used in an LGBTQ+ context (e.g. "drag", "gay") and common swear words (e.g., "bitch", "ass") contributed highly to toxicity ratings. "Bitch" was highlighted as appearing the most frequently in tweets with high toxicity (70%) which is problematic considering the word is often used by drag queens in a neutral or positive sense [19].

Similarly, specific words traditionally viewed as derogatory and hateful towards the LGBTQ+ community have in recent years been "reclaimed" by some members of this population. Words such as "fag", "dyke" or "tranny" may now be used in a joking or ironic context with the aim of removing the power of such language from heterosexual or cisgender adversaries. As explained by Thiago et al. [19] such reclamation of slurs can convey solidarity against attackers as well as aim to desensitise the queer community against such language being used against them. "Mock impoliteness" has also been identified as a communication style employed by the queer community as a method for preparing for hostility or "building a thick skin to face a hostile environment" [12]. Such communication may appear toxic or hateful, but to those within the queer community, it is generally evaluated positively. These are prime examples of language use that are often not understood by hate speech detection models and thus result in falsely labelling such texts as hateful. This is not only inaccurate, but extremely harmful to an already vulnerable community who is consequently doubly victimised in the online realm – both by hate speech perpetrators and by tools intended to prevent spread of such hate. Our work hence aims to research this understudied phenomenon of reclaimed language and its impact on hate speech detection models.

In an effort to keep up with the ever-evolving slang language used to insult protected demographics and to generalise hate speech detection models, Vidgen et al. [20] demonstrated the use of a dynamic process whereby human annotators worked to iteratively train a human-in-the-loop RoBERTa model pre-trained on a large corpus of hate speech, by introducing increasingly complex examples of hate speech for the model to learn from. The process incorporated not only original entries of synthesised hate speech, inspired by real-life examples, but also perturbations and adversarial examples, which resulted in increased robustness and generalisation of the model by the final round of training. Although they provide a compelling methodology for a human-in-the-loop approach to improving hate-speech classification, their process involved significant manpower (including expert annotators overseeing the process), time and resources. Their research did not look at the presence and impact of reclaimed language, which is the focus of our study by building on the efforts of Vidgen et al. [20].

3 PROBLEM QUANTIFICATION

As a first step towards studying the ability of current hate speech detection models to handle reclaimed language, we evaluated whether an existing state-of-the-art model for hate speech detection struggles with cases of reclaimed language, and to what extent. For that purpose we compare the prevalence of hate speech in tweets of two TV series, one whose content is related to the LGBTQ+ community and a second that has other social media context to serve as control.

The series were *RuPaul's Drag Race*, a drag queen competition reality show, and the popular reality TV dating series *Love Island*. The assumption was that discourse about the two series would hold distinct features in terms of language choices and dialect: tweets referencing *Drag Race* were predicted to contain more reclaimed language and mock impoliteness, which would in turn be reflected by higher rates of false hate speech classification (i.e., false positive).

Hashtags "#DragRace" and "#LoveIsland" were used to collect a random sample of 5000 tweets for each series between January 1st, 2023 and June 2nd, 2023. These tweets were then classified using Vidgen et al. [20], a state-of-the-art hate speech classification model that was pretrained on a hate speech dataset and served as our baseline model (see Section 5 for further details). Proportions of hate speech predictions were recorded. All tweets of both series were assumed to have a ground-truth of "No Hate" based on not being removed by Twitter's own hate speech identification tools. Consequently, model predictions labelled as hate speech were considered as false positives.

Analysis of classification proportions showed that tweets from threads discussing *RuPal's Drag Race* were classified as hate speech at almost triple the rate of tweets discussing *Love Island* (Figure 1).

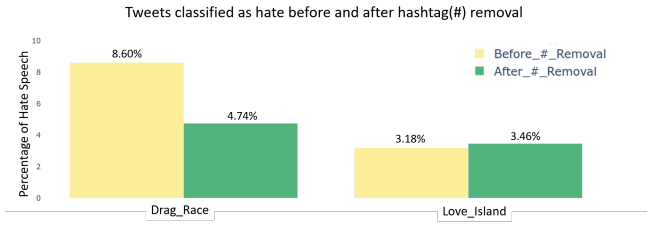

Figure 1: Hate speech classification rates for #LoveIsland and #DragRace tweets before and after removal of these hashtags.

In order to better understand which phrases were contributing to these classifications, we used the Captum model interpretability library [9] to extract the highest contributing words in each sentence (Table 1). This was done as an attempt to identify which words were causing higher false positive rates in the *Drag Race* Tweets compared to the *Love Island* Tweets.

Table 1: Top 10 highest contributing tokens in hate-speech classification of tweets containing #DragRace, and the percentage of sentences with that token.

Drag	Race	</s>	bitch	I	fucking	race	I	ass
30.5	6.05	5.12	4.19	2.56	2.33	2.09	2.09	1.63

The results indicate that the "#DragRace" hashtag is largely responsible for the disparity between the hate speech rates in the two datasets. The token "Drag" was the highest contributing word

to the prediction in 30.5% of sentences, whilst "Race" contributed most highly in 6.05% of cases.

Following these findings, the keywords "#DragRace" or "#LoveIsland" were removed to confirm no over-reliance on specific words was causing the difference in misclassification rates. A re-classification using the same baseline Vidgen model after these removal indeed produced lower rates of hate speech predictions for *Drag Race* tweets. However, hate speech predictions for *Drag Race* were still 50% higher than for *Love Island*, demonstrating that handling reclaimed language is a significant and persistent problem for NLP models, and strengthening the rationale for exploring ways for its mitigation. Additional results that demonstrate poor performance of reclaimed language and further support this finding are presented in Figures 2 and 3. In what follows we describe our approach for the problem, which relies on a specially curated dataset and model fine-tuning.

4 THE RECLAIM HATE SPEECH DATASET (RHSD)

With the insights obtained by the preliminary analysis above, four different sources were used to compile our experimental dataset in order to train and test models to better handle reclaimed language. Creating a unified dataset was motivated by the need to produce a robust model that could retain its ability to classify general hateful content, whilst being optimised for the task of recognising reclaimed language use, and not misclassifying it as hateful. This dual objective was made clear following a pilot study we conducted (see § 5). Therefore, the RHSD unified dataset contains items of derogatory, non-derogatory, and reclaimed language usage, with the first labelled as hate speech, and the last two as non-hate for the purpose of training our model.[1] Further statistics of the RHSD dataset can be found below in § 4.6.

4.1 Transcripts of Full Episodes

To provide a large corpus with examples of reclaimed language use, a total of 18,772 sentences were scraped from Seasons 11-14 of the drag queen competition television series *RuPaul's Drag Race*.[2] It was assumed that these sentences would contain no genuine examples of hate speech due to their broadcasting on national television, and would contain only cases of reclaimed language because these were spoken by drag queens and their allies. Under this assumption, we used the same Vidgen model used in § 3 to obtain ground truth labels for the scraped sentences. As a result, 941 sentences (about 5%) were perceived as reclaimed language (sentences that were classified as hate speech by the model that were interpreted as false positives), and included in the RHSD dataset. The reminder non derogatory sentences were discarded. An additional three episodes were used as held out set for evaluation (see 4.7).

4.2 Target Word Selection

Following an exploration of the transcripts above, 12 target words were identified which were hypothesised to be examples of reclaimed language that may contribute to misclassification of sentences that contain them. These target words were: "bitch", "drag", "fag", "gay", "homo", "puss", "queen", "queer", "sissy", "slay", "slut", and "whore". Definitions of these words in terms of their hateful connotations as well as their reclaimed counterparts can be found in Appendix A. Regular expressions were used so that variations that contained these words would also be included in their group (e.g. "bitchy" = "bitch", "faggot" = "fag", "pussy" = "puss"). In the case of the word "whore", variants were defined separately as being "hoe", "ho", or "ho's" and were also categorised under "whore". Sentences which contained none of these target words were classed as "other".

4.3 Hate Speech and Offensive Language Dataset

The selected target words were used as search terms in the "Hate Speech and Offensive Language" Dataset [16]. This dataset consists of hateful or offensive tweets. Sentences were selected which contained these target words. This dataset labelled sentences based on whether they were a) hateful b) contained offensive language c) neither hateful nor offensive. Mapping these labels onto the gold labels was done as follows: sentences which were identified as hateful were given the gold label "DEG"; sentences which were neither hateful nor offensive were labelled "NDG". All other sentences were excluded due to the lack of consistency as to whether sentences containing offensive language were also hateful, or not. In total, 2488 sentences were obtained from this dataset.

4.4 Slur Corpus

We used a subset of the Slur Corpus dataset [10], consisting of 13,332 texts each containing the homophobic slur "faggot" used in different contexts, and used the existing gold labels of the dataset. The appropriative (APR) subcategory as defined by Kurrek et al. [10] was consistent with our definition of reclaimed language, as was labelled accordingly.

4.5 Synthetic Slur Dataset

A small dataset of 154 synthetic sentences was created by feeding prompts to GPT-4 in order to produce examples with more nuanced usage of reclaimed language. 32.5% were clear out-group uses of the target word, and given the label "DEG". The remaining 67.5% of the sentences were examples of reclaimed language use in various contexts. In some sentences, the writer clearly defined their in-group membership to the drag community. In others, they defined their membership implicitly through using other vocabulary typical in the drag community. The sentiment of these texts were mixed: some were positive towards others or the drag community, others were hostile towards individuals. However, in terms of language profile, these sentences still adhere to the definition of reclaimed language. For further information on prompts given to GPT-4 and example sentences from each category in the synthetic dataset, refer to Appendix B.

4.6 RHSD Dataset Statistics

After merging all four data sources described above, our unified RHSD dataset contains 16,868 items, 58.4% are labelled as hate and the remaining 41.6% as nothate. Of the nothate items 33.5% were standard non derogatory, and 8% were cases of reclaimed language.

[1] Dataset is available on https://github.com/haimdub/RHSD
[2] We used https://subslikescript.com/ for scraping the show's substitles.

The contribution of each of the sources used to curate the dataset were as follows: 78.8% from Kurrek et al. [10], 14.7% from Samoshyn [16], 5.57% from the *Drag Race* Transcript, and the remainder of 0.91% from the Synthetic Dataset. The same proportions of class labels were maintained (within 1 d.p.) in the subset of RHSD used for testing.

The target word "fag" is by far the most common word present in RHSD (found in 67.8% of texts). This is followed by "bitch" (12.6%) and gay (8.37%). The remaining target words are present in smaller proportions (<3%). See Appendix A for further details of the characteristics and proportions of target words within the dataset.

4.7 Evaluation Datasets

RHSD. We used a held-out set of RHSD as our first evaluation set, comprising 15% of all the samples (70% and 15% were used for training and validation, respectively). This was complemented with three episodes from Season 8 of *RuPaul's Drag Race* that were not scraped previously and preprocessed in the same way as the training set. In order to have a balanced split for training, validation and testing, we stratified that each label group was represented proportionally, and that the splits also maintained the original distribution of the hate/nothate labels within each group.

HateCheck Validation. A suite of functional tests for hate speech detection models [15] was used as our second evaluation set. It is composed of 29 tests, where 18 consist of distinct types of hate expression and 11 consist of non-hateful contrasts. The aim of HATECHECK is to uncover specific weaknesses in models, and consists of 3,728 labelled entries (69% Hate, 31% Not Hate). We followed Vidgen et al. [20] in using this dataset who achieved substantially higher scores than models previously tested [15].

Unseen Target Word dataset. Using the same approach for the *Slur Corpus* above (§ 4.4), an additional subset of 13,332 texts each containing the homophobic slur "tranny", previously unseen by the model, was used as a test set. We remove any instances of this target word from the entire training set, so the model had not been fine-tuned on any reclaimed uses of this word.

The purpose of using these three datasets is to investigate the differences of models dealing with a dataset that contain reclaimed language, RHSD, and two datasets that do not, HateCheck and Unseen Target Word.

5 THE RECLAIMED HATE SPEECH (RHS) MODEL

The baseline Vidgen model [20] is a RoBERTa model that was initially trained on a large corpus of English language hate speech and toxicity datasets. The model was fine-tuned on a dataset of approximately 40,000 synthetic examples of sentences created by annotators, including over 15,000 examples of challenging perturbations aimed to trick the model into causing misclassifications. The dataset comprises 54% hateful entries, significantly higher than comparable datasets. Training was conducted over four rounds with increasingly sophisticated perturbations introduced in each round. The initial model (M1) was the most easily tricked with 54.7% of entries being misclassified, including 64.6% of Hate and 49.2% of Not Hate. The final model (M4) was tricked only by 27.7% of content, including 23.7% of Hate and 31.7% of Not Hate. The final model (M4) achieved a Macro-F1 score of 72.93±0.56 on the final round (R4) test set, compared to the first model (M1) only achieving 63.44±0.26 Macro-F1 on this test set.

The fine-tuning process was executed using the Hugging Face Transformers library's Trainer class [25]. Fine-tuning was done for a total of 3 epochs, with a batch size of 16, and employed the Adam optimizer with a learning rate of 0.0005. The model was trained with early stopping enabled, and a patience of 2 epochs for the validation set to prevent overfitting. All capitalisation and punctuation was preserved as it was assumed to contribute to semantic information.

Using Vidgen et al. [20] as a base model, we first fine-tuned it on a subset of the RHSD that focused on cases of reclaimed language, using only the transcripts examples from *Drag Race* (see § 4.1). This, however, led to very poor performance on standard hate speech detection, as evaluated on Vidgen's own test split , with F-Scores of 40.1 relative to 72.93 for Vidgen. These preliminary results further emphasised the need for our RHSD as a balanced dataset. We then develop our model called **"Reclaimed Hate Speech" (RHS)**, which is fine-tuned on the full RHSD dataset (see § 4). This additional fine-tuning is performed to assess the importance of incorporating reclaimed language, together with standard hate speech language, into the model.

6 RESULTS AND ANALYSES

Overall performance of our RHS model was compared against the Vidgen model on four datasets, and further analyses were performed for the different class labels, datasets, and target words.

6.1 Overall performance

Accuracy and F1 scores were computed on our TestSet from our unified model, HATECHECK Test Suit, and our Unseen Target Word Dataset (see § 4.7). Results show (Table 2) that RHS performs markedly better than Vidgen's model on our TestSet, with 20% and 10% gains in the accuracy and F-Scores, respectively. An opposite trend is observed in the HateCheck and Unseen Target Word datasets, where RHS shows a slight diminished performance than Vidgen's model. These reduced performance is significantly smaller in size than its gains on the TestSet (about 5% and 3%, respectively). These results demonstrate that RMS is able to maintain high performance on standard hate speech cases (HateCheck), despite being trained on a much more challenging dataset than contains cases of reclaimed language. The ability or RMS to actually handle reclaimed language is investigated in what follows, where we provide fine-grained analysis on the items in our TestSet.

	RHSD		**HateCheck**		**Unseen Word**	
	Acc	F1	Acc	F1	Acc	F1
RHS	86.3	88.0	91.5	91.6	76.3	77.0
Vidgen	72.7	80.4	95.6	95.6	79.1	82.7

Table 2: Model accuracy and F1-scores across all datasets.

6.2 Class Level Analysis

A breakdown of the results on our RHSD reveals the source of the improved results compared to Vidgen's model (Figure 2). A significant large improvement was seen in the RHS model's accuracy in classifying reclaimed language (REC: 84.7% vs 17.5%), as well as for non-derogatory language (NDG: 86.1% vs. 45.9%). On the other hand, a slight degradation in the model's performance in identifying derogatory language was observed (DEG: 86.5% vs 95.5%). Overall, these results demonstrate that RHS is able to perform far better that Vidgen in handling reclaimed language, while still maintaining good performance on standard hate speech detection.

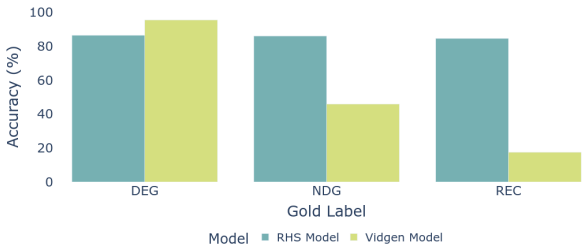

Figure 2: Class level model performance in RHSD.

6.3 Dataset Level Analysis

A breakdown of the individual sources that comprise our RHSD dataset reveals similar pattern of results (Figure 3). Substantial improvement in RHS ability to accurately classify reclaimed language from the *Drag Race* transcript relative to Vidgen's model is observed, an accuracy of 93.9% compared to 15.0% for the latter. This is not surprising given that these dataset contain only cases of reclaimed language (see § 4). Overall accuracy also improved for the Slur Corpus [10] that contain derogatory and non-derogatory usages of the word *faggot*, but without known cases of reclaimed language. Performance on the Samoshyn dataset which was primarily composed of derogatory sentences shows a slight degradation, in-keeping with the slight drop in performance observed in the derogatory class above (Figure 2). Here too, findings suggest that RHS is able to maintain high level of performance on standard hate speech detection, and also obtains dramatic improvements in handling reclaimed language, compared with previous model.

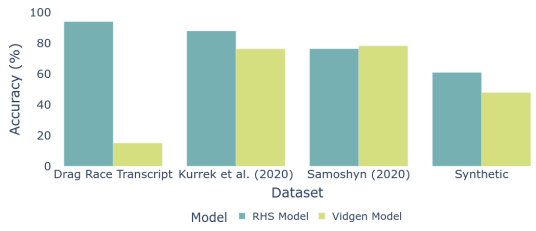

Figure 3: Dateset level model performance in RHSD.

6.4 Word Level Analysis

For the word level analysis of our RHSD, all target words showed improvement in accuracy rates on the test dataset except "sissy" and "slut" (Figure 4) which were present in very small proportions. The target word "fag" comprised the majority of target words in the dataset (appearing in over half of texts), and showed substantial improvement in accuracy (87.5% vs 76.2%).

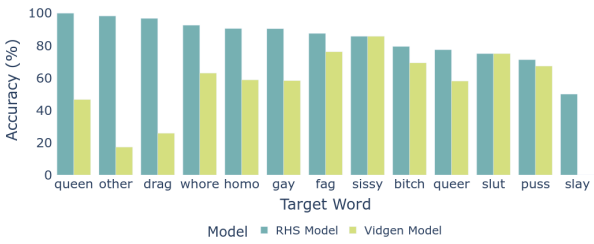

Figure 4: Word level model performance in RHSD.

6.5 Performance on Unseen Word Dataset

The RHS model performed relatively well on the Unseen Target Word dataset, achieving accuracy and F1-scores over 75% (Table 2). However, performance was slightly lower compared to the Vidgen's baseline model. Notably, the proportion of reclaimed language in this test dataset was lower than in the dataset the RHS model was trained upon (2.5% compared to 7%). Looking at the different classes, RHS performed worse for derogatory (DEG) language, but was substantially better for reclaimed language (REC; 62.3% vs 46.5%) and non-derogatory (NDG; (86.9% vs 69.4%) categories (Figure 5). However, compared to the RHSD dataset, RHS's accuracy for reclaimed language was lower in this class category. These results solidify our conclusion that RHS can maintain high performance on standard hate speech task (with notable fluctuations of course), and simultaneously improve on reclaim language.

7 DISCUSSION

7.1 Overview

The results of this study show promise in the ability of hate speech classification models to be robustly optimised in order to prevent compounded discrimination of marginalised communities' use of reclaimed language. The RHS model showed substantial improvement in classifying reclaimed language, whilst maintaining performance

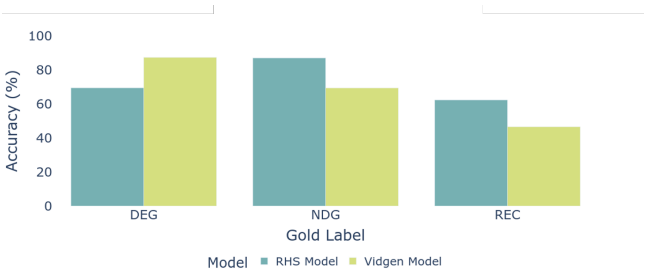

Figure 5: Models performance in Unseen Word Dataset.

on standard hate speech detection tasks, though some challenges remain. This work underscores the importance of semantically-driven approaches when training hate speech classification models.

7.2 Key Findings

7.2.1 Improvement in Reclaimed Language Classification. The RHS model exhibited a marked improvement in classifying reclaimed language. The model's ability to differentiate reclaimed language by learning from drag queen speech is a significant advancement towards building hate speech classifiers that are sensitive to the unique language dynamics within the LGBTQ+ community. The importance of using a diverse range of training data, including directly from the groups intended to be protected by hate speech classifiers, is also highlighted.

7.2.2 Improved Accuracy In Different Evaluation Sources. By showing either improvement in, or maintenance of, accuracy rates across the four sources used to create our dataset, the RHS model's performance is shown to hold generalised applicability across different contexts and sources. The lack of improvement for the Samoshyn dataset, which was composed mainly of "standard" derogatory sentences, is consistent with the RHS model's slight decrease in accuracy in classifying these type of derogatory sentences, demonstrating a trade-off between sensitivity towards reclaimed language use versus "standard" hate speech identification.

7.2.3 Target-Word Level Improvement. The improvements at the target-word level demonstrates the RHS model's ability to recognize and adapt to different semantic meanings of words. This aligns with the goal of teaching the model to not be overly reliant on the presence of certain key words when making hateful predictions.

7.3 Limitations and Future Work

While the study has achieved its primary goal of developing a special dataset to facilitate the handling of reclaimed language, and fine-tuned a model on it in order to demonstrate its necessity, some limitations must be acknowledged.

7.3.1 Performance on HateCheck. The decrease in performance on the HateCheck test suite validation dataset also shows a compromise in model robustness when fine-tuned to a highly specific context. Future research may focus on addressing these limitations by focusing on subsections of the test suite where the RHS model performed most poorly and ensuring the training dataset includes a diverse range of hateful examples to learn from.

7.3.2 Generalisability and Community Engagement. As the study focuses specifically on reclaimed language used in the drag community, further research must assess whether results, and our approach in general, are generalisable to other reclaimed language use, and also to the LGBTQ+ community as a whole. Engaging directly with members of the LGBTQ+ community to understand and represent their viewpoints on reclaimed language use is essential in creating hate speech detection models that hold the highest real-world applicability and value. Notably, the use of reclaimed language may not be deemed acceptable by all LGBTQ+ members and so labelling reclaimed language as inoffensive may be an unjustified assumption.

7.3.3 Binary Hate Speech Classification. Although the aim of this work was to separate the use of slurs in a reclaimed context from their use in hate speech, in real world applications this may not be sufficient. Instances could arise where slurs are used by an in-group member, but the sentence overall is still derogatory and insulting to another member of the in-group. Bullying within an in-group is very much a possibility, and so a blanket classification of reclaimed language as "nothate" may be inappropriate. Future work may focus on training models with an additional third "offensive" classification option, as done by researchers such as Davidson et al. [3].

7.3.4 Meaning Change Over Time. It is well known that the meaning of words may change over time, and taboo and slang words are particularly susceptible to such changes (i.e., they change their meanings more frequently) [13]. Therefore, any word that is classified today in the RHSD as derogatory or reclaimed could soon be outdated. While there is no apparent solution to this problem, which also affects Machine Learning models in general [1], being aware of this limitation that stems from the temporal dynamics of word meaning and their fluidity is essential. However, putting hate speech in the spotlight, either in its derogatory or reclaimed meaning, could facilitate research in this domain which is relatively lacking, and contribute to the identification of the underlying conditions that drive taboo and slang words to change more than other words.

7.3.5 Increor reclamied meaning, asing Potential for Synthetic Dataset. The small size of the synthetic dataset results in a modest contribution to the fine-tuned model's performance. Further research may expand the synthetic dataset in order to produce sufficient data for fine-tuning and assessment of a model's ability to perform in situations where in-group members are using reclaimed language whilst being derogatory to another in-group member. Additionally, the concept that in-group members will not always "reveal" their group membership presents an additional challenge. In these cases, a model may need to learn to identify an entire in-group vocabulary rather than only specific reclaimed words. This problem may warrant investigation of more sophisticated methodology like the adversarial debiasing procedure by Xia et al. [26] for African American English, applied to drag queen or LGBTQ+ dialect. Furthermore, the rarity of some target words (e.g. "sissy") which were not balanced with derogatory examples, may mean the model does not generalise well to derogatory uses of such words. This presents an opporuntity to utilise generative models to engineer more sophisticated instances of reclaimed language use that can improve the balance of future datasets.

7.3.6 Unseen Target Word Performance. Overall, the RHS model did not outperform the Vidgen model on an unseen target word. This may be due to the difference in proportions of reclaimed language in the test dataset compared to the Unseen Target Word dataset, as well as the lack of drag-queen specific reclaimed language use in the unseen target word dataset. This could imply that the RHS model has been optimised too specifically to the drag race population's use of reclaimed language, and lacks the generalisability to other forms of reclaimed language use. Future work should assess the homogeneity of reclaimed language to assess whether further subdivision is needed.

7.3.7 Implications For the LGBTQ+ Community. The findings of this study have significant implications for the LGBTQ+ community. An improved ability to recognise reclaimed language reduces the risk of compounding discrimination against this population's self-expression and their goal of gaining ownership over language historically used derogatorily against them. This is a vital step towards ensuring that online spaces are inclusive and respectful of different communities.

7.3.8 For Hate Speech Detection. This research contributes to the ongoing effort to develop sophisticated hate speech classification models. By highlighting the importance of understanding context and cultural nuances of language, it sets a precedent for future research to be semantically informed.

8 CONCLUSION

Our research is a pioneering investigation of the impact of reclaimed language on hate speech detection. We formulate the problem and the data collection methodology to enable research on reclaimed language, which leads to the collection of the first such dataset that focuses exclusively on reclaimed language, i.e. Reclaimed Hate Speech Dataset (RHSD). After assessing the impact of reclaimed language on a state-of-the-art hate speech detection model, we propose an alternative solution through a strategically fine-tuned model by introducing our "reclaimed Hate Speech" (RHS) model. We observe that a model specifically fine-tuned on reclaimed language can effectively keep a competitive performance on general hate speech, while significantly improving performance on cases of reclaimed language. The results of our model also show the importance of including language from diverse populations in training hate speech classifiers that are more inclusive and contextually aware. While the results indicate potential for progress, the observed trade-offs between reclaimed and derogatory uses of slurs emphasise that this is a nuanced and multifaceted field which requires careful balancing of the sensitivity and robustness of models. Our research provides a dataset and establishes a benchmark methodology to enable research in hate speech detection involving reclaimed language. We suggest that future research should ideally address the dearth of datasets involving reclaimed language to enable better generalisability, and avoid potential overfitting.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their insightful comments which improved this manuscript, and provides us with food for thought for future research. This research was partly funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021).

REFERENCES

[1] Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2023. Building for tomorrow: Assessing the temporal persistence of text classifiers. *Information Processing & Management* 60, 2 (2023), 103200.

[2] Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet* 7, 2 (2015), 223–242.

[3] 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*. 512–515.

[4] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, Vol. 12.

[5] Katharine Gelber and Luke McNamara. 2016. Evidencing the harms of hate speech. *Social Identities* 22, 3 (2016), 324–341.

[6] Y Gerrard. 2020. Book review: Behind the Screen: Content Moderation in the Shadows of Social Media. *New Media & Society* 22, 3 (2020), 579–582.

[7] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[8] Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media* 27 (2022), 100182.

[9] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896* (2020).

[10] Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 138–149.

[11] Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1621–1622.

[12] Sean McKinnon. 2017. "Building a thick skin for each other" The use of 'reading'as an interactional practice of mock impoliteness in drag queen backstage talk. *Journal of Language and Sexuality* 6, 1 (2017), 90–127.

[13] April M. S. McMahon. 1994. *Understanding Language Change.* Cambridge University Press.

[14] Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon. 2021. Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE Access* 9 (2021), 88364–88376.

[15] Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 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)*. 41–58.

[16] Andrii Samoshyn. 2020. Hate Speech and Offensive Language Dataset. https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset. [Online; accessed 20-June-2023].

[17] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.

[18] Sanjana Sharma, Saksham Agrawal, and Manish Shrivastava. 2018. Degree based Classification of Harmful Speech using Twitter Data. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. 106–112.

[19] Dias Oliva Thiago, Antonialli Dennys Marcelo, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & culture* 25, 2 (2021), 700–732.

[20] 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)*. 1667–1682.

[21] Jeremy Waldron. 2012. *The harm in hate speech.* Harvard University Press.

[22] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142.

[23] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access* 6 (2018), 13825–13835.

[24] Matthew Leighton Williams and Olivia Pearson. 2016. Hate crime and bullying in the age of social media. (2016).

[25] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[26] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting Racial Bias in Hate Speech Detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*. 7–14.

[27] Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science* 7 (2021), e598.

[28] Wenjie Yin and Arkaitz Zubiaga. 2022. Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media. *Online Social Networks and Media* 30 (2022), 100210.

Target Word	Proportion Hateful (%)	Proportion Reclaimed (%)	Proportion of Dataset Containing Word (%)
fag	59.3	2.20	67.8
bitch	65.2	14.7	12.6
gay	28.8	5.77	8.37
puss	72.3	3.72	2.92
homo	20.0	1.88	2.89
other	0.26	99.7	1.88
queer	29.8	8.26	1.08
drag	13.3	74.7	0.82
whore	63.0	18.1	0.68
queen	33.3	55.2	0.47
slut	55.4	8.93	0.28
sissy	44.8	3.45	0.14
slay	16.7	66.7	0.06

Figure 6: Table showing (left to right): Target word; Proportion of each target word with ground truth "hate"; proportion of each target word classed as reclaimed; proportion of sentences in the dataset containing each target word (NB: sentences may contain more than one target word)

Target Word	Hateful Definition	Reclaimed Definition*
fag	"*Extremely Disparaging and Offensive.* A contemptuous term used to refer to a gay person, especially a gay man."	**
bitch	"*Offensive*, slang. A malicious, spiteful, or coarse woman."	"A fierce woman. A friend. Used as a term of endearment among drag queens."
gay	"*Usually Offensive.* A person, especially a man, who is sexually or romantically attracted to people of the same sex or gender."	"Male homosexuality."
pussy	"*Slang: Disparaging and Offensive.* A timid, passive person, especially a man."	In the context of "*Pussy on Fire*": "An expression meaning to work hard and excel. To be awesome or outstanding. To work so hard that your vagina catches on fire."
homo	"*Slang: Extremely Disparaging and Offensive.* A contemptuous term used to refer to a gay person, especially a gay man."	**
queer	"*Often Disparaging and Offensive.* (The noun "a queer" is often used with disparaging intent and considered offensive even among those who approve of the corresponding adjective "queer." The plural "queers" is less likely to offend than the singular "a queer.") a) A gay or lesbian person. b) A person whose sexual orientation or gender identity falls outside the heterosexual mainstream or the gender binary."	**

Figure 7: Definitions of target word in hateful and reclaimed contexts (Part A)

A FURTHER INFORMATION ABOUT TARGET WORDS

Please see Figures 6, 7 and 8 for more detail with the definitinos and distributions of the target words used in our study.

B DESCRIPTION OF PROCESS FOR GENERATING SYNTHETIC TEST DATASET

GPT-4 was asked to produce sentences written as "young adults on social media" and so included artefacts such as hashtags. All sentences were created with the placeholder "Drangle" and substituted for the slur "faggot" after creation. This was in order to meet GPT-4's terms of use policy which prohibits generation of hateful content. The DEG category contained hateful examples of the slur, while the REC category contained positive or neutral examples of the slur used in a reclaimed context.

Target Word	Hateful Definition	Reclaimed Definition*
drag	(in the context of "drag queen": "*Slang.* A performer, typically one who was assigned male at birth, whose act involves a stylized and exaggerated interpretation of femininity that plays with stereotypical gender themes."	**. In the context of "Dragulation": "The process of being given a drag makeover by the "Dragulator" computer."
whore	"*Disparaging and Offensive.* A person who is sexually promiscuous."	** Example use: "You can't keep a good ho down" - RuPaul
queen	"*Slang: Usually Disparaging and Offensive.* A term used to refer to a gay man, especially one who is flamboyantly campy or effeminate."	People who dress up in clothes of the opposite gender usually. Slang term used for transvestites.
slut	"*Disparaging and Offensive.* A person, especially a woman, who is sexually promiscuous."	** Example use:"I'm not a slut, I'm a whore. Whores get paid" —Willam Belli (Drag Queen)
sissy	"*Disparaging and Offensive.* An effeminate boy or man."	In the context of "Sissy That Walk": "To walk with sass. Usually a way of serving up fish. often told to do by RuPaul."

Figure 8: Definitions of target word in hateful and reclaimed contexts (Part B)

Gold Label of Synthetic Dataset Item	Example Sentence
DEG	"gays trying to act like they're the arbiters of cool is the ultimate faggot illusion. #GetOffYourHighHorse"
REC	"Mismatched socks? gays take it to the next level! We mix patterns and colors to create a faggot fashion statement that screams confidence and uniqueness! "
REC	"Doesn't it annoy anyone else among my fags how Dave, one of our own, can't keep a secret? #LooseLips"

Figure 9: Example sentences from chatGPT-4 after reinstating slur words instead of the placeholder "drangle"