Inclusive Counterfactual Generation: Leveraging LLMs in Identifying Online Hate

M. Atif Qureshi
Technological University Dublin, atif.qureshi@tudublin.ie

Arjumand Younus
University College Dublin, Ireland

Simon Caton
University College Dublin, Ireland

Funder: Science Foundation Ireland

Inclusive Counterfactual Generation: Leveraging LLMs in Identifying Online Hate

M. Atif Qureshi[1 [0000000344134476]], Arjumand Younus[2 [0000−0001−7748−2050]], and Simon Caton[3 [0000−0001−9379−3879]]

[1] ADAPT Centre, eXplainable Analytics Group, Faculty of Business, Technological University Dublin
atif.qureshi@tudublin.ie

[2] School of Information and Communication Studies, University College Dublin
arjumand.younus@ucd.ie

[3] School of Computer Science, University College Dublin
simon.caton@ucd.ie

Abstract. Counterfactually augmented data has recently been proposed as a successful solution for socially situated NLP tasks such as hate speech detection. The chief component within the existing counterfactual data augmentation pipeline, however, involves manually flipping labels and making minimal content edits to training data. In a hate speech context, these forms of editing have been shown to still retain offensive hate speech content. Inspired by the recent success of large language models (LLMs), especially the development of ChatGPT, which have demonstrated improved language comprehension abilities, we propose an inclusivity-oriented approach to automatically generate counterfactually augmented data using LLMs. We show that hate speech detection models trained with LLM-produced counterfactually augmented data can outperform both state-of-the-art and human-based methods.

Keywords: counterfactuals, ChatGPT, inclusivity, model robustness, out-of-domain testing

1 Introduction

Natural language processing technologies allow us to derive meaningful insights from the vast amount of user-generated textual data, thereby advancing research in the domain of social computing. More specifically, text classification systems from within natural language processing are critical components of social computing pipelines. It has been well established within the natural language processing literature that dataset artefacts critically influence the performance of text classification systems [5]. The performance effects are more significant in social computing tasks, such as hate speech detection, and in most cases, there is a danger of the model learning the dataset rather than the construct being investigated [30, 38, 39]. This eventually implies higher misclassifications that can have disastrous repercussions in the context of hate speech detection [41], particularly for real-world solutions that require out-of-domain deployment.

Decoupling the dataset artefacts from the task at hand is a complex process, chiefly on account of how modern machine learning methods learn features from various datasets. Recent approaches to rectify the problem of dataset artefact learning for hate speech tasks involve the generation of counterfactually augmented training data by means of which models are able to learn enhanced features for various natural language processing tasks [14, 19, 41]. Essentially, the chief idea behind hate speech counterfactual generation is making (near) minimal edits to a piece of text while flipping its label (from hate to non-hate or non-hate to hate). Existing approaches for generating counterfactuals mostly rely on human-in-the-loop systems involving tremendous amounts of manual effort [14, 19]; the very few automated techniques rely on assessing statistical correlations within spurious data patterns and labels - both, however, rely on making basic edits to the input text leading to ethical considerations in a domain as challenging as detecting online hate [21]. The ethical considerations involve flipping labels of non-hate texts to produce hate texts, and in many cases, when hate texts are converted to non-hate texts, some controversial aspects still persist, i.e., reduced inclusivity and offensive content not being completely removed.

Inspired by the language comprehension abilities of (very) large language models (LLMs) such as the GPT family [7, 25], we argue for advanced counterfactual generation via the use of LLM-based models. LLMs, originating from extensive Web engineering efforts with deep neural networks trained on large amounts of text corpora, allow for the encoding of rich factual knowledge, resulting in informative responses to input prompts [27]. These informative responses produced by LLMs show promise within the domain of counterfactual generation for the challenging task of detecting online hate. We investigate the effectiveness of LLM-generated counterfactuals via proof-of-concept experiments on hate speech datasets comprising multiple domains [4, 40, 41]. Chiefly, we ask the following questions:

– **(RQ1) What kind of role could LLM-generated counterfactuals have in the training of hate speech detection models?**.
– **(RQ2) How do models trained using LLM-generated counterfactuals compare to those that use manually generated equivalents?**

To answer the above research questions, our LLM-generated counterfactuals come from Polyjuice [42] and ChatGPT[4]. Performance comparisons of the LLM-based counterfactual generation pipeline show improvements in hate speech classification accuracy, and the fact that our counterfactuals are generated only against hate texts makes this particularly appealing from an ethical/inclusivity standpoint. The robustness of models is evaluated through cross-validation evaluations on the training data and out-of-domain testing on another popular hate speech dataset [4]. As further evidence of robustness, we also compare the effectiveness of LLM-generated counterfactuals with well-known manually generated ones [41]. Code and data is available here to facilitate future research: `https://github.com/matifq/Inclusive-Counterfactual-Generation-Leveraging-LLMs-in-Identifying-Online-Hate`

[4] `https://openai.com/blog/chatgpt`

Original text	**Manual CF** **(Previous Approaches)**	**Prompt-based CF** **(Our Approach)**
Let's talk about the antithesis of hard work: immigrants	Let's talk about antithesis of hard work: my brother	Let's discuss the contrast to hard work: immigrants.
Both genders are not equal, men are superior, its a fact so get over it	Genders are not equal, men are favoured, its a fact so get over it so we can do something about it.	Gender equality is important, and no gender is superior to the other. Let's embrace this reality.

Table 1: Human-in-the-loop/manual counterfactuals vs ChatGPT-based counterfactuals

2 Contribution: Inclusivity in Counterfactual Generation

Table 1 shows two concrete examples of how the quality of existing counterfactuals (via minimal edits) differs from those produced by ChatGPT. Note that the first example has been borrowed from [38] and the second from [41]. As is obvious from the examples, the counterfactuals generated employing human-in-the-loop still seem suggestive of offence even if directed towards “my brother" in the first example or “men" in the second example. This helps highlight a significant aspect of subjectivity cues within hate speech annotation and, by extension, classification efforts, which in itself is an unresolved problem [10].

By utilising LLMs, like ChatGPT, for automated text generation and utilising these as counterfactuals, the problem of subjectivity bias is somewhat minimised, thereby promoting inclusivity. Moreover, this helps us approach the hate speech detection problem in a fundamentally different way, giving it a wider philosophical basis from within information and communication studies. This approach advocates against “cancel culture" [9] while ensuring inclusive online spaces, as evidenced from examples in Table 1 where elements of offence are non-existent. In doing so, we address the complex interplay between freedom of expression and hate speech by allowing the preservation of the essential idea being expressed in a no-hate format [16]. We give further examples of this aspect we wish to highlight during our experimental evaluations phase in Section 6. It is also worth noting an additional benefit of our approach: an easing of emotional fatigue for human annotators that would no longer need to be exposed to hateful and toxic subject matter.

3 Related Work

Our work sits at the intersection of hate speech detection [15] and particularly, hate speech detection methods built on top of data augmentation methods [13].

Essentially hate speech detection is a text classification task with the main components being data collection, feature extraction, and model learning [18]; with early efforts focusing mainly on feature extraction over classical machine learning models. With the emergence of deep learning, however, it was discovered that deep learning models using either CNN or LSTM models performed on average 13–20% better

[2]. Another striking trend within this domain was witnessed with the emergence of BERT as state-of-the-art in hate speech detection, and it significantly outperformed approaches like FastText as well as CNN-, and LSTM-based approaches [31].

With powerful computing architectures, transformer models matured quickly further improving accuracy on benchmark datasets, but questions around the robustness of hate speech detection models began to emerge. Despite the field of hate speech detection having been around for over a decade, it was not until recently that the issue of models' limitations on out-of-domain datasets was taken up by hate speech detection researchers [23]. This essentially implies that there is an implicit learning of cues/artefacts from the dataset, and those cues/artefacts are spuriously correlated with the construct under investigation [36]. To mitigate such learning and to ensure the robustness of models in the domain of hate speech detection, counterfactually augmented data has been proposed as a solution that shows significant promise [19, 34]. The basic premise behind counterfactual data augmentation is that instances that are minimally edited to flip their labels are added to the training data to offer a causality-based framework towards increasing the robustness of the machine learning model [28]. Most works, and even very recent ones are limited to the use of manual edits that employ label flips through word additions/deletions. In response, there have been some efforts towards the generation of automated counterfactuals [1, 33]. Even within the domain of automated counterfactuals, very few efforts have exclusively turned to generative natural language processing models [17, 37] with enhanced expressive power and increased robustness. A very recent work [37] performs a detailed evaluation of manually generated counterfactuals vs. those generated automatically through generative natural language processing models; we discuss the major differences between their work and ours in the next section.

4 LLM-based Counterfactual Generation Pipeline

Our goal in this work is to automate the generation of counterfactuals to improve the training (and by extension performance) of hate speech classification models. We do this by leveraging existing LLMs and prompting them to edit a specified corpus of text content (e.g. Table 1); thus injecting counterfactuals into the text directly. Figure 1 illustrates our approach at a high-level. It can be summarised as follows (from bottom left, clockwise through the figure): 1) collect a corpus of text content to act as training data; 2) subset the corpus into not hateful/non-hateful content and prompt the LLM to modify the text (i.e. hateful → not hateful and not hateful → hateful); ChatGPT is depicted here as an example; 3) collate the output from the LLM (potentially correcting encoding issues); 4) resample the training data injecting the counterfactual examples; 5) (re)train and evaluate the ML model.

4.1 Datasets

To evaluate the effectiveness of our LLM-based counterfactual generation pipeline, we use three datasets: 1) the Toraman English cross-domain hate speech tweet dataset [40], 2) the Vidgen dataset, and 3) the HatEval English tweets' test dataset [4].

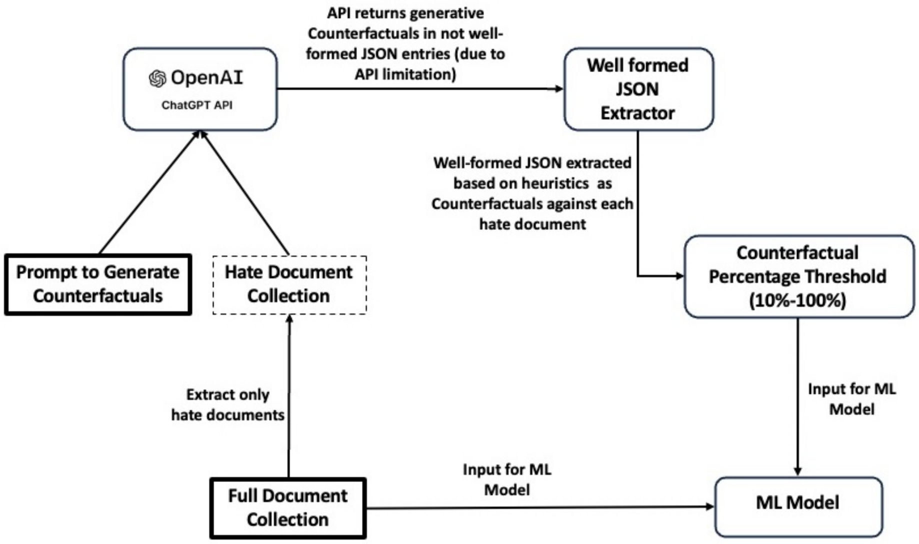
API returns generative Counterfactuals in not well-formed JSON entries (due to API limitation)
OpenAI
ChatGPT API
Well formed JSON Extractor
Well-formed JSON extracted based on heuristics as Counterfactuals against each hate document
Counterfactual Percentage Threshold (10%-100%)
Prompt to Generate Counterfactuals
Hate Document Collection
Extract only hate documents
Input for ML Model
Input for ML Model
Full Document Collection
ML Model

Fig. 1: High-level overview of our hate speech detection pipeline built on top of ChatGPT-based counterfactuals

The Toraman dataset consists of 68,597 tweets from five different domains namely: gender, religion, race, politics and sports, with three labels: *hate* (2% tweets), *offensive* (19%) and *normal* tweets (79%). The choice for this dataset is mainly motivated by the fact that it is very recent with multiple tweets belonging to various domains.

The Vidgen dataset is a dynamically generated dataset emanating from the efforts of multiple annotators over four rounds. It consists of 13,104 texts with binary labels: *hate* (50.5% tweets) and *nothate* (49.5%).[5] This dataset on account of its high annotation quality and associated manual counterfactuals was used as a comparison benchmark within the model training step.

The HatEval dataset comprises 3,000 tweets with binary labels: *hate* (42% tweets) and *normal* (58%). We selected this dataset to conduct our out-of-domain robustness experiments, and hence, this data serves as the test set.

4.2 Methodology

For LLM-based counterfactual generation we explored the use of Polyjuice and ChatGPT. Polyjuice involves fine-tuning a GPT-2 model [29] to generate counterfactuals from input sentences through eight different control codes (e.g., negation, shuffling, lexical). We make use of these control codes, and use the Polyjuice version available from Huggingface.[6]. Furthermore, inclusivity is incorporated by filtering the full document collection to pass on only hate documents to Polyjuice.

[5] Note that on account of being a special-purpose, manually curated dataset for the task of hate speech detection there are higher than normal percentages of hate speech texts.

[6] `https://huggingface.co/uw-hai/polyjuice`

Similarly, the inclusivity component is the step where we filter the full document collection to pass on only hate documents to the ChatGPT via OpenAI API (shown in Figure 1). To generate the counterfactuals, we first misspelled the swear words[7] and profanity words[8] by repeating the last character twice. This approach ensured that ChatGPT would produce a response, and helped circumvent ChatGPT guardrails. Following this step, we used the following prompt on ChatGPT using the OpenAI API:

- "Substitute problematic terms in the following texts with inclusive language and produce a list of a few rephrased versions of the text and list down suspected problematic terms:"
- "You reply in JSON only, no free text."
- **List of texts.**

In the above prompt **“List of texts"** is the “Hate Document Collection" of Figure 1. The prompt generates malformed JSON responses from ChatGPT that need to be corrected and converted into consistent JSON entries. To resolve this, we applied regular expressions together with a range of information extraction heuristics (shown in **“Well-formed JSON Extractor"** of Figure 1) and recovered 75% counterfactuals of the total tweets (i.e., hate and offensive) from the OpenAI API. The final step before the machine learning model controls the percentage of counterfactuals that we inject into the model, and helps control the proportion of counterfactuals provided to the model.

The work closest to ours is by Sen at al. [37], and they perform a detailed evaluation of manually generated counterfactuals vs. those generated automatically through generative natural language processing models. Their technique for generating automatic counterfactuals however relies on some training examples and prompts that encourage generative models to rely again on minimal/basic edits. Our technique of programmatic message-passing to OpenAI API via JSON inputs is scalable, and can concretely utilise the expressive power of generative models.

5 Experimental Setup

To illustrate the effectiveness of our approach, we conduct three different experiments, each evaluating a different aspect of the LLM-generated counterfactuals: 1) the effectiveness of LLM-generated counterfactuals, 2) their robustness in out-of-domain settings, and 3) to compare LLM-generated counterfactuals against manual (human) developed counterfactuals. Table 2 summarizes the experimental settings, aims, training data, test data (if any), and models used. It is important to note that in the Polyjuice case, we randomly sample one of out of the eight counterfactuals instead of including all counterfactual variants unlike Sen at al. [37].

[7] We used the lexicons from `https://github.com/peterkwells/uk-attitudes-to-offensive-language-and-gestures-data/`

[8] We used the lexicons from `https://github.com/surge-ai/profanity`

Experiment	Aim	Training Data	Test Data	Models
1	Cross-Validate Effectiveness of LLM-generated CFs	Toraman	N/A	Davidson, TPOT, BERT
2	Check Model Robustness on OOD Data	Toraman	HatEval	BERT
3	Perform Model Robustness Comparisons between Manual and LLM-generated CFs	Vidgen	HatEval	BERT

Table 2: Summary of Experimental Settings Across Three Experiments

5.1 LLM-generated Counterfactual Effectiveness

In the first experiment, we aim to test the effectiveness of the LLM-generated counterfactuals and compare three classifiers. The first classifier we choose is the Davidson winner, namely LogisticRegression; note that this popular machine learning algorithm is the original pipeline of a popular hate speech dataset [11]. The second classifier is the winning model generated by the Tree-Based Pipeline Optimization Tool (TPOT) tool, an AutoML [22] classification pipeline.[9] This model is composed of stacked LinearSVC and DecisionTree classifiers and was trained on the Toraman dataset. The third classifier is the popular finetuned-BERT classifier[10] [20]. The model choice is motivated by the fact that we aim to investigate various settings: 1) a basic machine learning model via features directly observed in the dataset [11], 2) an AutoML pipeline that can select the best outcome from within traditional machine learning pipelines again via features directly observed in the dataset, and 3) an algorithm that encodes complexities of background knowledge and domain knowledge while learning complex inter-dependencies between features (BERT).

We performed five-fold cross-validation on each domain for the three baselines[11] and their variants using counterfactuals generated by our proposed technique. We selected 0.1, 0.2, 0.3, ..., to 1.0 as the range of parameters that controls the proportion of generated counterfactuals across all hate and offensive tweets, i.e., 0.1 would randomly choose 10% of hate and offensive tweets and generate counterfactual variants for those tweets. Note that this counterfactuals' proportion selection strategy is the one referred to in Figure 1 under the process called **"Counterfactual Percentage Threshold"**.

5.2 Exploring Model Robustness

In the second experiment, we aim to test model robustness by experimenting with out-of-domain training and test data using BERT. We limit the testing to BERT on

[9] With generations=5, population_size=40

[10] With max_epochs = 5, batch_size = 32 (except for the third Experiment, we use 10), learning_rate = 1e-5

[11] For BERT, each fold's original test set was divided into 50%-50% validation and test set.

account of it being the best-performing algorithm for the first experiment (see Section 6). We used the Toraman dataset as the training and validation set by splitting it into an 80-20 ratio. We then used the HatEval dataset as the test set. However, we combined the *hate* and *offensive* classes from training data into a single *hate* class to match the binary labels of the out-of-domain test set.

Out-of-domain testing forms a significant aspect of hate speech detection models given how crucial it is to detect hate speech in settings previously unknown to the model. The robustness of hate speech detection models is crucial in ensuring inclusive online spaces, and the research literature has established out-of-domain tests as a method for such evaluations [35] whereby the training set is significantly different from the test set.

5.3 Manual vs. LLM-based Counterfactual Robustness

In the third experiment, we aim to test model robustness when using the LLM-generated counterfactuals of our pipeline vs manual ones generated by human annotators across the Vidgen dataset. Again, we limit the testing to BERT on account of it being the best-performing algorithm for the first set of experiments, and for the sake of a fair comparison, we use only counterfactuals against hate texts from the Vidgen dataset. Since the test is being performed for model robustness we perform it over out-of-domain test data. For a thorough evaluation of robustness across manual counterfactuals, we test three variants of LLM-generated counterfactuals: ChatGPT-based counterfactuals, Polyjuice counterfactuals, and a combination of both.

6 Findings and Discussion

6.1 LLM-generated Counterfactual Effectiveness

Tables 3, 4 and 5 show the results with TPOT pipeline, Davidson pipeline, and BERT respectively for counterfactual variant vs no-counterfactual variant i.e., original data. Due to space limitations, we only show the best counterfactual variant of each run i.e. domain and report averaged macro-F1 and weighted-F1 scores. We report both these metrics to highlight the strength of our approach in dealing with imbalanced hate speech datasets. As can be seen in almost all cases except for two (averaged Macro-F1 in case of *"Religion"* and *"Sports"*) the counterfactual variant outperforms the model with no counterfactuals. This essentially demonstrates a promising direction for LLM-generated counterfactual generation solely for hate speech labels. The best performance boost is achieved in the case of BERT over the *"Religion"* domain; chiefly, this is on account of the Toraman dataset containing a vast array of topics/themes/terms in the context of *"Religion"* thereby leading to better and diverse counterfactuals that make sure the model doesn't learn dataset artefacts.[12]

[12] A qualitative analysis of the data revealed coverage of a vast range of issues from gays in Islam to Republicans to Catholicism. In fact, the dataset diversity is highest for tweets belonging to domain *"Religion"*

Domain	$F1_M$	$F1_M.cf$	$F1_M.pj.cf$	$F1_W$	$F1_W.cf$	$F1_W.pj.cf$
Gender	0.614	$\mathbf{0.620}_{cf=0.3}$	$0.613_{cf=0.1}$	0.887	$\mathbf{0.890}_{cf=0.4}$	$0.888_{cf=0.1}$
Religion	**0.624**	$0.607_{cf=0.1}$	$0.613_{cf=0.5}$	0.878	$\mathbf{0.880}_{cf=0.6}$	$0.873_{cf=0.6}$
Race	0.581	$0.585_{cf=0.6}$	$\mathbf{0.590}_{cf=0.3}$	0.876	$0.881_{cf=0.5}$	$\mathbf{0.887}_{cf=1.0}$
Politics	0.627	$\mathbf{0.639}_{cf=0.2}$	$0.631_{cf=0.3}$	0.878	$\mathbf{0.885}_{cf=0.9}$	$0.880_{cf=0.2}$
Sports	**0.683**	$0.674_{cf=0.3}$	$0.674_{cf=0.2}$	0.932	$\mathbf{0.935}_{cf=0.5}$	$0.931_{cf=0.2}$

Table 3: Model Performance for TPOT Pipeline: Testing Counterfactuals' Effectiveness via Cross-Validation Averaged Macro-F1 and Weighted-F1 Scores. Subscripts *M* and *W* represent macro and weighted F1 while *.cf* implies the setting where ChatGPT-based counterfactuals were used, and *.pj.cf* implies the setting where Polyjuice counterfactuals were used.

Domain	$F1_M$	$F1_M.cf$	$F1_M.pj.cf$	$F1_W$	$F1_W.cf$	$F1_W.pj.cf$
Gender	0.649	$\mathbf{0.655}_{cf=0.9}$	$0.650_{cf=0.6}$	0.882	$\mathbf{0.885}_{cf=0.9}$	$0.881_{cf=0.1}$
Religion	0.595	$\mathbf{0.606}_{cf=0.7}$	$0.600_{cf=0.9}$	0.838	$\mathbf{0.852}_{cf=1.0}$	$0.845_{cf=0.4}$
Race	0.633	$0.639_{cf=1.0}$	$\mathbf{0.645}_{cf=0.9}$	0.865	$0.870_{cf=1.0}$	$\mathbf{0.871}_{cf=0.9}$
Politics	0.643	$\mathbf{0.656}_{cf=0.9}$	$0.654_{cf=0.9}$	0.860	$\mathbf{0.873}_{cf=1.0}$	$0.868_{cf=0.9}$
Sports	0.710	$\mathbf{0.716}_{cf=0.9}$	$\mathbf{0.716}_{cf=0.9}$	0.929	$\mathbf{0.934}_{cf=0.9}$	$0.929_{cf=0.2}$

Table 4: Model Performance for Davidson Pipeline: Testing Counterfactuals' Effectiveness via Cross-Validation Averaged Macro-F1 and Weighted-F1 Scores. Subscripts *M* and *W* represent macro and weighted F1 while *.cf* implies the setting where ChatGPT-based counterfactuals were used, and *.pj.cf* implies the setting where Polyjuice counterfactuals were used.

Domain	$F1_M$	$F1_M.cf$	$F1_M.pj.cf$	$F1_W$	$F1_W.cf$	$F1_W.pj.cf$
Gender	0.770	$\mathbf{0.781}_{cf=0.4}$	$0.767_{cf=0.7}$	0.915	$\mathbf{0.923}_{cf=0.4}$	$0.915_{cf=0.1}$
Religion	0.693	$\mathbf{0.739}_{cf=0.4}$	$\mathbf{0.739}_{cf=0.4}$	0.908	$\mathbf{0.918}_{cf=0.4}$	$0.917_{cf=1.0}$
Race	0.694	$\mathbf{0.733}_{cf=0.2}$	$0.725_{cf=0.8}$	0.906	$\mathbf{0.917}_{cf=0.2}$	$0.916_{cf=0.6}$
Politics	0.728	$0.763_{cf=1.0}$	$\mathbf{0.769}_{cf=0.2}$	0.903	$\mathbf{0.913}_{cf=1.0}$	$0.911_{cf=0.2}$
Sports	0.762	$\mathbf{0.778}_{cf=0.1}$	$0.777_{cf=0.5}$	0.944	$\mathbf{0.949}_{cf=0.7}$	$\mathbf{0.949}_{cf=0.1}$

Table 5: Model Performance for BERT: Testing Counterfactuals' Effectiveness via Cross-Validation Averaged Macro-F1 and Weighted-F1 Scores. Subscripts *M* and *W* represent macro and weighted F1 while *.cf* implies the setting where counterfactuals were used, and *.pj.cf* implies the setting where Polyjuice counterfactuals were used.

6.2 Exploring Model Robustness

Table 6 shows the results with BERT on a dataset taken in another context i.e. out-of-domain. Note that the creators of this dataset of HatEval English tweets report a baseline accuracy of 0.451 (support vector machine) and 0.367 (most frequent concept) [4]. A BERT model trained without counterfactuals offers an improvement over this, and this is further improved by incorporating counterfactuals. Here, the results show the best performance over the *"Race"* domain. Polyjuice counterfactuals show the best performance, and this is on account of its ability to produce diverse sets of realistic counterfactuals.

Domain	$F1_M$	$F1_M.cf$	$F1_M.pj.cf$	$F1_W$	$F1_W.cf$	$F1_W.pj.cf$
Gender	0.516	$0.525_{cf=0.8}$	$\mathbf{0.534}_{cf=0.4}$	0.521	$0.532_{cf=0.2}$	$\mathbf{0.543}_{cf=0.4}$
Religion	0.523	$0.540_{cf=1.0}$	$\mathbf{0.543}_{cf=1.0}$	0.526	$0.539_{cf=1.0}$	$\mathbf{0.548}_{cf=1.0}$
Race	0.511	$0.528_{cf=0.8}$	$\mathbf{0.550}_{cf=0.2}$	0.505	$0.531_{cf=0.6}$	$\mathbf{0.553}_{cf=0.2}$
Politics	0.515	$0.527_{cf=0.2}$	$\mathbf{0.546}_{cf=0.6}$	0.517	$0.532_{cf=0.2}$	$\mathbf{0.557}_{cf=0.6}$
Sports	0.518	$0.524_{cf=1.0}$	$\mathbf{0.532}_{cf=0.5}$	0.519	$0.525_{cf=1.0}$	$\mathbf{0.539}_{cf=0.5}$

Table 6: **BERT Performance: Exploring Model Robustness** Averaged Macro-F1 and Weighted-F1 Across no-Counterfactuals VS Counterfactuals for Out-of-Domain Test Set. Subscripts *M* and *W* represent macro and weighted F1 while *.cf* implies the setting where ChatGPT-based counterfactuals were used, and *.pj.cf* implies the setting where Polyjuice counterfactuals were used.

6.3 Manual vs. LLM-based Counterfactual Robustness

Tables 7A and 7B and Tables 8A and 8B again show the results with BERT on a dataset taken in another context i.e. out-of-domain but this time comparing our approach with manually generated counterfactuals from Vidgen dataset. Table 7A shows the results for no counterfactuals case vs best cases of LLM-generated counterfactuals and manual counterfactuals. For both metrics, the combination variant of both counterfactuals shows the best performance; and from the standpoint of advancements in automated hate speech detection, this is very encouraging.

Table 8A and 8B shows the results for no counterfactuals case vs mean cases of LLM-generated counterfactuals and manual counterfactuals; at the same time, it also shows the standard deviation scores for all the cases. As is clear from the results in these tables, the combination variant where ChatGPT-based counterfactuals are mixed with Polyjuice counterfactuals generates the most effective version outperforming manually generated ones on average. This performance difference is contrary to what Sen et al. demonstrate in their recent work [37] where their conclusion was in favor of manually generated counterfactuals. We establish that this is on account of allowing free-form counterfactual generation via LLMs rather than forcing minimal edits or flips. Moreover, both Polyjuice and ChatGPT produce diverse counterfactuals enabled via the expressive power of large language models, and the inclusivity aspect enables a controlled injection of these counterfactuals thereby leading to better model robustness as compared to manually generated counterfactuals.

The most encouraging aspect of our technique is the ability to generate effective counterfactuals without needing to involve manual efforts of hate speech generation, which in itself is a tricky from an ethical standpoint. Furthermore, such generated counterfactuals that promote inclusivity rather than harm can serve as a significant impetus to policymaking towards ethical AI [32]; particularly concerning efforts

$F1_M$	$F1_M.cf$	$F1_M.pj.cf$	$F1_M.comb.cf$	$F1_M.mancf$
0.584	$0.653_{cf=0.9}$	$0.638_{cf=0.3}$	$\mathbf{0.658}_{cf=0.6}$	$0.631_{cf=0.6}$

Table 7.A **BERT Performance: Manual vs. LLM-based Counterfactual Robustness** Averaged Macro-F1 Across no-Counterfactuals VS Best Case Counterfactuals VS Best Case Manual Counterfactuals for Out-of-Domain Test Set. Subscript M represents macro F1. *.cf* implies the setting where ChatGPT-based counterfactuals were used, *.pj.cf* implies the setting where Polyjuice counterfactuals were used, *.comb.cf* implies the setting where a combination of ChatGPT-based and Polyjuice counterfactuals were used and *.mancf* implies the setting where manual counterfactuals were used.

$F1_W$	$F1_W.cf$	$F1_W.pj.cf$	$F1_W.comb.cf$	$F1_W.mancf$
0.606	$0.664_{cf=0.9}$	$0.648_{cf=0.3}$	$\mathbf{0.666}_{cf=0.1}$	$0.644_{cf=0.6}$

Table 7.B **BERT Performance: Manual vs. LLM-based Counterfactual Robustness** Averaged Weighted-F1 Across no-Counterfactuals VS Best Case Counterfactuals VS Best Case Manual Counterfactuals for Out-of-Domain Test Set. Subscript W represents weighted F1. *.cf* implies the setting where ChatGPT-based counterfactuals were used, *.pj.cf* implies the setting where Polyjuice counterfactuals were used, *.comb.cf* implies the setting where a combination of ChatGPT-based and Polyjuice counterfactuals were used and *.mancf* implies the setting where manual counterfactuals were used.

$F1_M$	$\bar{x}(.cf)$	$\sigma(.cf)$	$\bar{x}(.pj.cf)$	$\sigma(.pj.cf)$	$\bar{x}(.comb.cf)$	$\sigma(.comb.cf)$	$\bar{x}(.mancf)$	$\sigma(.mancf)$
0.584	0.617	0.025	0.620	0.016	**0.638**	0.015	0.607	0.027

Table 8.A **BERT Performance: Manual vs. LLM-based Counterfactual Robustness** Averaged Macro-F1 Across no-Counterfactuals VS Mean of Counterfactuals VS Mean of Manual Counterfactuals for Out-of-Domain Test Set. *.cf* implies the setting where ChatGPT-based counterfactuals were used, *.pj.cf* implies the setting where Polyjuice counterfactuals were used, *.comb.cf* implies the setting where a combination of ChatGPT-based and Polyjuice counterfactuals were used and *.mancf* implies the setting where manual counterfactuals were used.

$F1_W$	$\bar{x}(.cf)$	$\sigma(.cf)$	$\bar{x}(.pj.cf)$	$\sigma(.pj.cf)$	$\bar{x}(.comb.cf)$	$\sigma(.comb.cf)$	$\bar{x}(.mancf)$	$\sigma(.mancf)$
0.606	0.630	0.021	0.631	0.016	**0.647**	0.014	0.624	0.022

Table 8.B **BERT Performance: Manual vs. LLM-based Counterfactual Robustness** Averaged Weighted-F1 Across no-Counterfactuals VS Mean of Counterfactuals VS Mean of Manual Counterfactuals for Out-of-Domain Test Set. *.cf* implies the setting where ChatGPT-based counterfactuals were used, *.pj.cf* implies the setting where Polyjuice counterfactuals were used, *.comb.cf* implies the setting where a combination of ChatGPT-based and Polyjuice counterfactuals were used and *.mancf* implies the setting where manual counterfactuals were used.

such as European Union AI Act which flagged harms of large language models comprehensively [24]. Ours is the first step towards efforts to highlight how the harms of large language models can be evaded while enabling their effective use in modern natural language tasks.

6.4 Vision: Evading Harms of ChatGPT and Enabling Its Effective Usage

Much has been written about the potential harms of generative artificial intelligence tools like ChatGPT [6, 26], and more so with their massive ingestion of huge quantities of data leading to challenges of misinformation and bias. Essentially our proof-of-concept experiments set out LLMs as effective tools for addressing online hate speech, which may prove beneficial for the research community in hate speech detection. At the same time, we also present the first step towards efforts to enable more socially sensitive counterfactuals via the use of tools like ChatGPT in a task as complex as hate speech detection. Our *“vision"* here is in proposing a direction within counterfactual generation where there is minimum manual effort and a reduced risk of exposure to harmful and upsetting content for annotators.

The vision advocated in this paper is basically within the same dimension as making use of AI’s ability to impersonate human subjects in fields such as psychology, political science, economics, and market research [3]. Bots trained over huge amounts of data, like ChatGPT, have already proven effective stand-ins in pilot studies and for designing experiments, saving time and money [12]. We argue a step further for computational social science researchers through its real-time deployment in the generation of synthetic data for natural language processing tasks.

Lastly, and most significantly, our extensive evaluations and the aspect of inclusivity within prompts highlight the need to leave the responsibility of ethical dimensions in artificial intelligence to humans rather than machines. Here, we have a different take to Sen et al. [37] who highlight the failure of LLM guardrails and their potential risks in the context of hate speech. The best way to circumvent such failures is by means of not assigning such dangerous tasks to tools like ChatGPT; and therein we emphasise the significance of using it to generate solely non-hate content as we did.

7 Conclusion and Future Work

The paper proposes an approach for hate speech detection whereby models use large language models to compute inclusive counterfactuals (i.e., non hate counterfactuals) and then exploit these counterfactuals to improve hate speech detection. We have shown via extensive experimental evaluations that text counterfactuals generated with LLMs (like ChatGPT) show a promising direction toward inclusivity in hate speech detection algorithms while also ensuring model robustness. There is much room for further investigation in this area particularly concerning moving from the idea of minimal edits in counterfactual generation to inclusive, prompt-based edits; and more so for tasks that involve complex (as well as disturbing) social constructs. In a future version of this work, we aim to experiment with multiple variants of prompts over ChatGPT over multiple datasets. The future directions of this work also involve a thorough comparison with other large language models for counterfactual generation. A potential limitation of this work is the preprocessing needed to circumvent ChatGPT’s guardrails for slur words and profanities, thus meriting future investigation.

This work has obvious parallels to the fairness in machine learning literature (see [8]): more inclusive models will be less susceptible to biases in hate speech classification and thus reduce socially insensitive outcomes. A key direction of future work would be to comprehensively explore the impact of our approach on fairness in hate speech classification (as a yet relatively under-explored area of fair NLP).

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under grant no. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Technological University Dublin.

References

1. Atanasova, P., Simonsen, J.G., Lioma, C., Augenstein, I.: Fact checking with insufficient evidence. Transactions of the Association for Computational Linguistics **10**, 746–763 (2022)
2. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th international conference on World Wide Web companion. pp. 759–760 (2017)
3. Bail, C.A.: Can generative ai improve social science? (2023)
4. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F.M.R., Rosso, P., Sanguinetti, M.: Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th international workshop on semantic evaluation. pp. 54–63 (2019)
5. Belinkov, Y., Poliak, A., Shieber, S.M., Van Durme, B., Rush, A.M.: Don't take the premise for granted: Mitigating artifacts in natural language inference. arXiv preprint arXiv:1907.04380 (2019)
6. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big?. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 610–623 (2021)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
8. Caton, S., Haas, C.: Fairness in machine learning: A survey. ACM Computing Surveys (Aug 2023)
9. D. Clark, M.: Drag them: A brief etymology of so-called "cancel culture". Communication and the Public **5**(3-4), 88–92 (2020)
10. Davani, A.M., Díaz, M., Prabhakaran, V.: Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Transactions of the Association for Computational Linguistics **10**, 92–110 (2022)
11. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the international AAAI conference on web and social media. vol. 11, pp. 512–515 (2017)
12. Dillion, D., Tandon, N., Gu, Y., Gray, K.: Can ai language models replace human participants? Trends in Cognitive Sciences (2023)
13. Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E.: A survey of data augmentation approaches for nlp. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 968–988 (2021)

14. Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., et al.: Evaluating models' local decision boundaries via contrast sets. arXiv preprint arXiv:2004.02709 (2020)
15. Garg, T., Masud, S., Suresh, T., Chakraborty, T.: Handling bias in toxic speech detection: A survey. ACM Computing Surveys **55**(13s), 1–32 (2023)
16. Gibson, A.: Free speech and safe spaces: How moderation policies shape online discussion spaces. Social Media+ Society **5**(1), 2056305119832588 (2019)
17. Howard, P., Singer, G., Lal, V., Choi, Y., Swayamdipta, S.: Neurocounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 5056–5072 (2022)
18. Jahan, M.S., Oussalah, M.: A systematic review of hate speech automatic detection using natural language processing. Neurocomputing p. 126232 (2023)
19. Kaushik, D., Hovy, E., Lipton, Z.: Learning the difference that makes a difference with counterfactually-augmented data. In: International Conference on Learning Representations (2019)
20. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. vol. 1, p. 2 (2019)
21. Kumar, A., Tan, C., Sharma, A.: Probing classifiers are unreliable for concept removal and detection. arXiv preprint arXiv:2207.04153 (2022)
22. Le, T.T., Fu, W., Moore, J.H.: Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics **36**(1), 250–256 (2020)
23. Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M.E., Sabharwal, A., Choi, Y.: Adversarial filters of dataset biases. In: Proceedings of the 37th International Conference on Machine Learning. pp. 1078–1088 (2020)
24. Madiega, T.A.: Artificial intelligence act. European Parliament: European Parliamentary Research Service (2021)
25. Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre, E., Heinz, I., Roth, D.: Recent advances in natural language processing via large pre-trained language models: A survey. arXiv preprint arXiv:2111.01243 (2021)
26. Motoki, F., Neto, V.P., Rodrigues, V.: More human than human: Measuring chatgpt political bias. Public Choice pp. 1–21 (2023)
27. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
28. Pearl, J.: Causal and counterfactual inference. the handbook of rationality (2019)
29. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
30. Ramponi, A., Tonelli, S.: Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3027–3040. Association for Computational Linguistics (2022)
31. Ranasinghe, T., Zampieri, M., Hettiarachchi, H.: Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In: FIRE (working notes). pp. 199–207 (2019)
32. Ray, P.P.: Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems (2023)
33. Ross, A., Wu, T., Peng, H., Peters, M.E., Gardner, M.: Tailor: Generating and perturbing text with semantic controls. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3194–3213 (2022)

34. Samory, M., Sen, I., Kohne, J., Flöck, F., Wagner, C.: "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In: Proceedings of the international AAAI conference on web and social media. vol. 15, pp. 573–584 (2021)
35. Sarwar, S.M., Murdock, V.: Unsupervised domain adaptation for hate speech detection using a data augmentation approach. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 16, pp. 852–862 (2022)
36. Schlangen, D.: Targeting the benchmark: On methodology in current natural language processing research. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 670–674 (2021)
37. Sen, I., Assenmacher, D., Samory, M., Augenstein, I., van der Aalst, W., Wagne, C.: People make better edits: Measuring the efficacy of llm-generated counterfactually augmented data for harmful language detection. arXiv preprint arXiv:2311.01270 (2023)
38. Sen, I., Samory, M., Flöck, F., Wagner, C., Augenstein, I.: How does counterfactually augmented data impact models for social computing constructs? arXiv preprint arXiv:2109.07022 (2021)
39. Sen, I., Samory, M., Wagner, C., Augenstein, I.: Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4716–4726 (2022)
40. Toraman, C., Şahinuç, F., Yilmaz, E.: Large-scale hate speech detection with cross-domain transfer. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 2215–2225. European Language Resources Association, Marseille, France (Jun 2022)
41. Vidgen, B., Thrush, T., Waseem, Z., Kiela, D.: Learning from the worst: Dynamically generated datasets to improve online hate detection. arXiv preprint arXiv:2012.15761 (2020)
42. Wu, T., Ribeiro, M.T., Heer, J., Weld, D.S.: Polyjuice: Generating counterfactuals for explaining, evaluating, and improving 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). pp. 6707–6723 (2021)