

Towards Fairness Assessment of Dutch Hate Speech Detection

Julie Bauer¹, Rishabh Kaushal^{1,2}, Thales Bertaglia³, Adriana Iamnitchi¹

¹Maastricht University

²Indira Gandhi Delhi Technical University for Women ³Utrecht University

Correspondence: jts.g.bauer@student.maastrichtuniversity.nl

Abstract

Numerous studies have proposed computational methods to detect hate speech online, yet most focus on the English language and emphasize model development. In this study, we evaluate the counterfactual fairness of hate speech detection models in the Dutch language, specifically examining the performance and fairness of transformer-based models. We make the following key contributions. First, we curate a list of Dutch Social Group Terms that reflect social context. Second, we generate counterfactual data for Dutch hate speech using LLMs and established strategies like Manual Group Substitution (MGS) and Sentence Log-Likelihood (SLL). Through qualitative evaluation, we highlight the challenges of generating realistic counterfactuals, particularly with Dutch grammar and contextual coherence. Third, we fine-tune baseline transformer-based models with counterfactual data and evaluate their performance in detecting hate speech. Fourth, we assess the fairness of these models using Counterfactual Token Fairness (CTF) and group fairness metrics, including equality of odds and demographic parity. Our analysis shows that models perform better in terms of hate speech detection, average counterfactual fairness and group fairness. This work addresses a significant gap in the literature on counterfactual fairness for hate speech detection in Dutch and provides practical insights and recommendations for improving both model performance and fairness.

1 Introduction

While the ease of expressing oneself on social media platforms has led to creative and meaningful interactions, it has also amplified the spread of hate speech – particularly content targeting specific groups based on ethnicity, gender, sexual orientation, and similar characteristics. To address this issue, researchers have developed numerous computational methods for detecting hate

speech (Alkomah and Ma, 2022; Yin and Zubiaga, 2021; MacAvaney et al., 2019). However, the majority of these efforts focus on the English language. Regardless of language, a critical question remains: *Are these detection models fair?* In other words, do they detect hate speech targeted at all social groups with similar accuracy? If not, the models are unfair (Mehrabi et al., 2021; Pessach and Shmueli, 2022). Unfair models risk perpetuating biases, which can exacerbate existing issues and erode users’ trust in social media platforms. This work evaluates the fairness of hate speech detection models for the Dutch language. Consider the following two sentences: S1: “*All Moroccans are troublemakers.*” and S2: “*All Dutch are troublemakers.*”. A fair hate speech detection model should classify both sentences as equally hateful since they share the same structure and level of negativity. However, if the model predicts S1 to be 98% likely to be hateful but S2 only 10%, this disparity indicates unfairness. This kind of bias often arises when sensitive identity attributes, such as nationality, ethnicity, sexuality, or religion, disproportionately influence model predictions (Garg et al., 2019). This example highlights the importance of *counterfactual fairness*: the principle that a model’s decision should remain consistent if sensitive attributes in the input data are changed. For instance, S2 is the counterfactual version of S1 (and vice versa). If a model evaluates these sentences differently, it fails to meet the standard of counterfactual fairness (Kusner et al., 2017). This challenge is particularly critical in hate speech detection, where sensitive attributes often appear in potentially biased contexts. More formally, a model M is considered counterfactually fair if it produces the same predictions for all possible values of a sensitive attribute A . Mathematically, this can be expressed as: $\Pr(\hat{Y} | A = a_1) = \Pr(\hat{Y} | A = a_2)$, where $A \in \{a_1, a_2\}$ represents different values of the sensitive attribute.

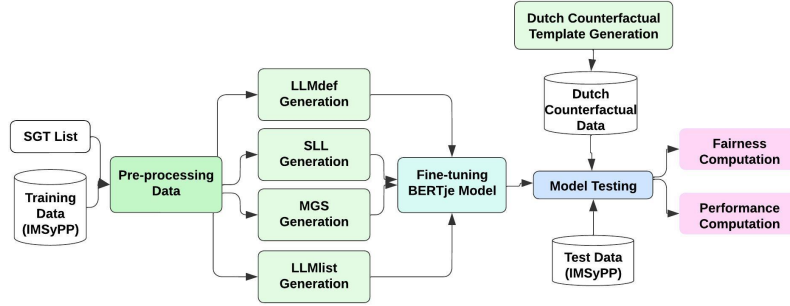


Figure 1: Proposed methodology outlining key steps. SGT list is curated. Training data forms input to the four counterfactual data generation methods, namely, LLMlist, LLMdef, SLL, and MGS. BERTje model is fine-tuned with counterfactual data. The model is tested on test data and Dutch counterfactual data generated using templates. Finally, performance evaluation and fairness computations are performed.

Despite progress in counterfactual fairness research for the English language, not much work has been done in European languages such as Dutch. Several studies on Dutch hate speech detection have been conducted using pre-trained language models and mainly focused on creating datasets and models for hate speech detection (Caselli et al., 2021; Caselli and Van Der Veen, 2023; Lemmens et al., 2021; Markov and Daelemans, 2022; Markov et al., 2022; Ruitenbeek et al., 2022; Theodoridis and Caselli, 2022). This work aims to bridge the gap in the existing literature by exploring the counterfactual fairness of pre-trained models in detecting hate speech in Dutch on social media platforms. We aim to answer the following research questions: *What are the methods to generate counterfactual data for the Dutch language? Are hate speech detection models in the Dutch language counterfactually fair? What is the impact of generated counterfactual data on the performance and fairness of the hate detection model in the Dutch language?*

Figure 1 presents the key steps in our methodology. We begin by curating a list of Dutch Social Group Terms and generating counterfactual sentences. To do this, we use Large Language Models (LLMs) and other techniques like Manual Group Substitution (MGS) and Sentence Log-Likelihood (SLL) originally proposed for English. Through qualitative evaluation, we identify challenges in generating realistic counterfactual sentences that conform to the rules of Dutch grammar. Next, we fine-tune transformer-based hate speech detection models using the generated counterfactual data. We then evaluate the performance of these models in detecting hate speech in Dutch. Finally, we assess counterfactual fairness using Counterfactual Token

Fairness (CTF). We also compute group fairness metrics, namely, equality of odds and demographic parity.

2 Related Work

Hate speech detection has been extensively studied (MacAvaney et al., 2019; Bertaglia et al., 2021; Mullah and Zainon, 2021; Yin and Zubiaga, 2021; Alkomah and Ma, 2022; Subramanian et al., 2023; Rawat et al., 2024; Gandhi et al., 2024). Defining hate speech is inherently challenging, as it is a complex phenomenon influenced by interpretation (Hietanen and Eddebo, 2023). Fortuna et al. (Fortuna and Nunes, 2018) proposed a comprehensive definition of hate speech: *“Hate speech is language that attacks or diminishes, incites violence or hate against groups based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, or others. It can manifest in various linguistic styles, including subtle forms or even through humour.”* This definition captures the diverse ways in which hateful language can be expressed.

Although most of the research in this domain has focused primarily on the English language, some studies have explored hate speech detection in multiple languages (Corazza et al., 2020). Notable examples include investigations into Italian (Del Vigna et al., 2017), Danish (Sigurbergsson and Derczynski, 2020), and Spanish (Plaza-del Arco et al., 2021), among others. We focus on hate speech detection in the Dutch language. Among earlier works, Tulkens et al. (Tulkens et al., 2016a,b) performed a dictionary-based approach for the detection of racist discourse in Dutch using automated

means. Markov et al. (Markov et al., 2021) explored features based on emotions and style for cross-domain hate speech detection in multiple languages including Dutch. Caselli et al. (Caselli et al., 2021) introduced a new dataset, Dutch Abusive Language Corpus (DALC v1.0), which comprises manually annotated tweets for abusive language. Ruitenbeek et al. (Ruitenbeek et al., 2022) curated a corpus containing more than 11k posts on Twitter in Dutch which are abusive and offensive. Hilte et al. (Hilte et al., 2023) investigated the demographics of authors who spread hate speech in Dutch and found that older men indulge in more hate speech. Vries (Vries, 2024) used a BERT-CNN based model for detecting the targets against whom hate is triggered on the X platform in Dutch. However, none of these works address the important issue of fairness of Dutch hate speech detection models.

Fairness is becoming increasingly important in the context of hate speech classification. A model is considered fair when it (1) does not use sensitive attributes in making decisions and (2) treats individuals with the same sensitive attributes similarly (Mehrabi et al., 2021). Approaches to increase fairness are predominantly based on sensitive attributes that point to (un)privileged groups, which are disproportionately likely to be positively classified as hateful by a model (Caton and Haas, 2024). Although different notions of fairness exist (Verma and Rubin, 2018), we focus on causality-based notions because evaluating causal relationships provides a more comprehensive evaluation of model fairness that can uncover model bias. More specifically, we apply counterfactual fairness (Kusner et al., 2017; Garg et al., 2019) which is defined as *“the intuition that a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group”*. In the context of fairness of hate speech models (Davani et al., 2021), this means that changing the sensitive attribute in a sentence should not impact the outcome of classification. The sentences “some people are gay” and “some people are straight” should, therefore, receive a similar toxicity prediction by fair models. While most previous work focuses on English, we focus on the Dutch language and evaluate hate speech detection models from a fairness perspective.

3 Counterfactual Data Generation

Evaluating the counterfactual fairness of a model requires access to counterfactual sentences. Counterfactual generation is a data augmentation strategy that creates such sentences by modifying sensitive identity terms while preserving the original meaning. This additional data is then used to fine-tune the model with the goal of improving both performance and fairness. We apply three methods for generating counterfactuals: Large Language Model (LLM), Sentence Log-Likelihood (SLL), and Manual Group Substitution (MGS).

3.1 Dataset Augmentation with Social Groups

We selected the IMSyPP Dutch hate speech dataset¹, which is publicly available and was curated as part of a project funded by the European Commission to tackle online hate speech through prevention, awareness, and regulation. The dataset contains comments from several Dutch social media platforms and forums, such as Twitter or Dumpert, posted from January 2018 to October 2020. It consists of a training set with 25,720 posts and an evaluation set with 2,858 posts. The records in the dataset include the website URL that the post originates from, the text of the hate speech post, the hate speech target and the type of hate speech (appropriate, inappropriate, offensive, or violent). Of these labels, the ‘offensive’ and ‘violent’ classes are typically regarded as hate speech. The target categories are one of the following: racism, migrants, islamophobia, antisemitism, religion, homophobia, sexism, ideology, media, politics, individual and other (Novak et al., 2021). Because this data set is focused on targeted identity groups, it is particularly suitable for this study. However, what is missing from this dataset for our objective is the identification of the social groups for each of the target categories in these hate speech examples.

To address this limitation, we manually curated a list of Dutch Social Group Terms (SGT) following the approach used by Davani et al. (Davani et al., 2021) for English. A naive approach of simply translating the English language SGT into Dutch would not suffice. For example, a social group ‘Moroccan’, which is a prevalent minority group in the Netherlands, would not be a relevant social group in the United States. At the same time, some other social groups such as ‘Sikhs’ are irrelevant for the hate speech context of the Dutch society and thus

¹<http://imsypp.ijs.si/>

are excluded. Additionally, due to complexities of Dutch grammar, we add variants of the SGTs to the list because nouns tend to have an adjective form in Dutch and vice versa. For example, the adjective ‘Nederlands’ means ‘Dutch’, while the word ‘Nederlander’ is a noun which means a ‘Dutch person’. Also, since the Dutch language conjugates adjectives, we added one conjugation form to the SGT list, for instance, only ‘Marokkaans’ and not ‘Marokkaanse’, to keep it simple. We did not consider plural nouns for this reason. Following this approach, we curated a list of 85 SGTs, which we refer to as *the Dutch SGT List*. The full list is available in Appendix A. After identifying the SGTs, we filtered the dataset and found that 2,649 posts contained at least one SGT from the Dutch SGT list, referred to as *baseline data*, which we use for the next steps. In addition, we perform standard preprocessing tasks, including removing emojis and deleting extra spaces, special signs, commas, and full stops.

3.2 LLM-based Counterfactual Generation

Inspired by recent work (Sen et al., 2023; Mishra et al., 2024), we introduce two distinct prompting approaches for generating counterfactual sentences using large language models (LLMs); both approaches take posts from the baseline dataset as input. The first approach, denoted as *LLMdef*, operates implicitly by instructing the LLM to modify social group terms in a given input post. These modifications are based on various identity attributes, such as gender, race, class, sexuality, political affiliation, religion, education level, age, and profession, among others. In this approach, the model is expected to identify the relevant social group term within the input post and replace it with another term from the same category. The selection of the replacement term is left to the discretion of the LLM, allowing for a dynamic and context-aware generation of counterfactuals. This method generated 15,175 counterfactual posts.

The second approach, denoted as *LLMlist*, builds upon the first method but introduces an explicit mechanism for social group term substitution. Instead of relying solely on the LLM’s internal decision-making, we provide a predefined list of social group terms, which we have carefully curated and documented in Appendix A. The model is then directed to generate counterfactual posts by substituting the social group term in the input sentence with an alternative from this predefined

list. This explicit specification ensures greater control over the counterfactual generation process and enables more systematic and interpretable modifications. By employing these two approaches, we aim to explore the capabilities of LLMs in generating counterfactual statements that reflect variations in social identity attributes, facilitating a deeper understanding of biases, fairness, and representation in language models. This method generated 21,562 counterfactual posts.

3.3 Sentence Log-Likelihood (SLL)

Following previous work (Nadeem et al., 2021; Davani et al., 2021), we generate counterfactuals considering the log-likelihood of the sentence, denoted as *SLL*. Equation (1) quantifies the log-likelihood ($\lg(P(x))$) of a sentence, where $x_0, x_1, \dots, x_{i-1}, x_i$ refer to the words in a sentence x .

$$f(x) = \log(P(x)) = \sum_{i=1}^n \log P(x_i | x_0, \dots, x_{i-1}) \quad (1)$$

Consider, for example, the sentence, “all Moroccans should go back to their country.” The SLL method assumes that the word ‘Moroccans’ is more likely to be replaced by ‘Turks’ than ‘Germans’. Turks is more likely to occur linguistically because, unlike Germans, it is a minority group that is often discriminated against in this way in Dutch society. For each of 2,649 comments, we substitute the SGT in the comment with each of the other SGTs in the Dutch SGT list to obtain potential counterfactual comments. We employ the pre-trained GPT-2 model to compute log-likelihood for each of these potential counterfactual substitutions. We consider only those counterfactual comments that have a higher or equal log-likelihood than the original comment. This method generated 49,104 counterfactual comments.

3.4 Manual Group Substitution (MGS)

In line with previous works (Yang et al., 2020; Madaan et al., 2021), we perturb SGTs based on the specific identity group and grammatical function in the sentences, specifically for Dutch. This method works with dictionaries, substituting an SGT with other SGTs that are present in the dictionary. The substitution process is mathematically defined as:

$$x_{cf} = \{x' | x' \in substitute(x, D)\} \quad (2)$$

In this equation, x_{cf} refers to the set with correctly generated counterfactuals. The equation uses

Table 1: Dataset Statistics (cnt: count; len: average length of sentence; ent: entropy of social group terms).

Category	Baseline			LLMdef			LLMlist			SLL			MGS		
	cnt	len	ent	cnt	len	ent	cnt	len	ent	cnt	len	ent	cnt	len	ent
Appropriate	1011	34.9	4.4	5801	29.9	4.8	8273	28.8	6.8	15624	31.2	5.8	6580	26.2	4.6
Inappropriate	260	37.1	4.8	1527	32.2	5.2	2211	29.7	7.2	6850	33.7	5.9	2863	27.1	3.9
Offensive	1327	39.8	4.9	7549	36.4	5.3	10713	34.2	7.2	25663	38.6	5.9	10481	39.3	3.6
Violent	46	37.3	3.5	298	34.5	4.0	365	36.1	5.4	967	29.8	5.8	469	30.5	3.6

$substitute(x, D)$ to replace the original SGT in a sentence with a counterfactual SGT based on its location in the dictionary. We curate this dictionary that contains several lists of SGTs based on whether the SGT is an adjective or noun and whether it belongs to one of the following identity groups: Nationality, skin colour, migrants, gender, sexuality, religion, age and ideology. For example, the word ‘woman’ would belong to the gender/noun category in the defined dictionary and would, therefore, be replaced only with other gender/noun terms, ‘transgender’ and ‘man’ in this case. In this manner, the MGS method creates grammatically correct and likely counterfactual sentences. Following this process, we generate 20,393 counterfactual comments.

Table 1 describes these datasets in terms of count, average sentence length, and entropy of SGTs.

Table 2: Performance Metrics for Evaluated Models (Accuracy, F1 Score, Precision and Recall)

Model	Acc	Prec	Rec	F1
Baseline	0.75	0.78	0.75	0.75
BERTje + LLMdef	0.79	0.61	0.62	0.61
BERTje + LLMlist	0.77	0.65	0.52	0.61
BERTje + SLL	0.79	0.79	0.79	0.79
BERTje + MGS	0.79	0.79	0.79	0.79

4 Qualitative Assessment

Next, we perform qualitative evaluation to assess whether *realistic* counterfactuals are created.

4.1 Sentence Log-likelihood Evaluation

SLL methods do not create realistic counterfactuals. Firstly, some counterfactuals do not adhere to the Dutch grammar rules. Table 13 in Appendix D shows an incorrectly generated counterfactual that uses “jong” as an adjective when Dutch grammar rules dictate that it should be “jonge” (because it belongs to a plural noun). This problem does not occur in English because adjectives are not conjugated based on the matching noun in this language. Therefore, this problem is specific to the Dutch context. In addition, Table 13 demonstrates the problem of substituting the adjective “Turkish” when

a noun is expected, according to Dutch grammar rules. Sentences that do not adhere to Dutch grammar are expected to get lower sentence likelihood since faulty grammar is expected to occur less often than correct grammar. However, this is not the case for the SLL method, suggesting the suboptimality of this method in the Dutch language regarding grammar adherence. Secondly, the SLL method does not always generate counterfactuals that adhere to the sentence context. Table 13 shows the original sentence “coming out as female”. A correctly generated example substitutes “female” with “gay”, as this is a usual expression, retaining the sentence’s meaning. However, the sentence likelihood method also substitutes the word “female” with “young”, which is not a usual expression and, therefore, unrealistic. The counterfactual with “young” would have been expected to receive a lower log-likelihood than the original sentence, but this is not the case. This suggests that the method fails to properly process sentence context in Dutch.

4.2 Manual Group Substitution Evaluation

As compared to SLL, this method tends to adhere better to the Dutch grammar rules than the SLL method but still makes occasional mistakes. This happens especially when a word can be used as both an adjective and a noun, which is the case for several Dutch SGTs. Table 14 in Appendix D shows a sentence in which this is the case. The original sentence contains the word “Chinese”. This word can be both an adjective and a noun in Dutch, resulting in the generation of grammatically incorrect sentences. In this case, when a noun is treated as an adjective and vice-versa, the method produces grammatically incorrect counterfactuals. The method is thus sub-optimal in creating realistic counterfactuals. Moreover, the question remains whether MGS captures the context of a sentence adequately. Even when sentences adhere to Dutch grammar rules, the original meaning of the sentence is not fully captured by every counterfactual generated. Table 14 shows an example. The original sentence “speak Dutch” requires substituting

Table 3: Performance results of classification models. All models are the pretrained BERTje Dutch language model finetuned with the counterfactual datasets generated as described in Section 3. The baseline is the original BERTje model from (De Vries et al., 2019).

Category	Baseline			LLMdef			LLMlist			SSL			MGS		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Appropriate	0.91	0.72	0.81	0.87	0.83	0.85	0.87	0.81	0.84	0.86	0.83	0.85	0.86	0.84	0.85
Inappropriate	0.54	0.28	0.37	0.38	0.38	0.38	0.32	0.42	0.36	0.33	0.36	0.34	0.33	0.36	0.35
Offensive	0.63	0.91	0.75	0.76	0.81	0.78	0.76	0.80	0.78	0.77	0.81	0.79	0.79	0.80	0.79
Violent	0.66	0.44	0.53	0.42	0.44	0.43	0.43	0.47	0.45	0.58	0.44	0.50	0.55	0.49	0.52

words from the adjective list. The sentence correctly substitutes an adjective, but “Latina” is not a language, creating an unlikely sentence.

4.3 LLM Evaluation

Generally, we see that both LLM methods can solve issues that arise in the other generation methods. MGS or SLL generation methods can create sentences like “a black (een zwart)”, that are incorrect in the Dutch language. An LLM can work around this issue, by substituting “a black person”, making the counterfactual sentences grammatically correct. However, LLMlist (Table 15 in Appendix D) still makes grammatical mistakes, usually based on already existing mistakes in the training data. The method creates sentences like “marokkaan broeders” rather than “marokkaanse broeders” from the original sentence “mocro broeders”, which is grammatically incorrect. Additionally, this method creates unlikely counterfactuals and makes interpretation mistakes. They change, for example, “iemand zwart maken” (expression in Dutch, which means to discredit someone) to “iemand wit maken” which is incoherent. “Black” here is incorrectly detected as an SGT in this context. It seems that the LLMlist method strictly adheres to the specified group terms in the prompt when detecting SGTs and creating counterfactuals. This is further demonstrated by the creation of counterfactuals with a wide range of SGTs, despite being unlikely. For example, it will create a sentence like “coming out as old” rather than coming out as a specific gender or sexuality. These examples demonstrate the limits of using a pre-defined list within LLMs in the realm of counterfactual generation. In contrast, LLMdef (Table 16 in Appendix D) creates more realistic counterfactuals that fall within a specific social group. It will, for example, retain the gender aspect in a sentence like “coming out as a woman”, creating only counterfactuals with a gender group term. This results in more likely counterfactuals. The qualitative analysis also shows a wider variety

of social group terms than present in the SGT list, but can go beyond that, for example, by changing stereotypically Dutch names into Arabic-sounding or English sounding names. This strength can, however, work counterproductively. Too freely interchanging terms, creates counterfactuals that are too dissimilar from the original sentence. For example, LLMdef substituted the colors of a Dutch football club with “education” group terms, creating an unlikely counterfactual. Such *free* substitution can result in either a mismatch between the generated sentence and the original label or an unlikely synthetic sentence that does not resemble human speech.

5 Performance Evaluation

We are interested in understanding the counterfactual fairness of pre-trained models for Dutch hate speech detection and evaluating methods for improving it. To this end, we focus on the BERTje model, a monolingual pre-trained model used in previous research in Dutch hate speech detection (Markov and Daelemans, 2022; Novak et al., 2021). This model categorizes hate speech into four classes: ‘acceptable’, ‘inappropriate’, ‘offensive’, and ‘violent’, corresponding to the categories in the IMSyPP dataset. The BERTje model was finetuned specifically for the IMSyPP dataset (De Vries et al., 2019), making it a natural choice for evaluating fairness in this context. Additionally, we further fine-tune the baseline BERTje model with counterfactual sentences generated using SLL, MGS and LLM-based methods. This results in five models: (1) Baseline BERTje, (2) BERTje + SLL, (3) BERTje + MGS, (4) BERTje + LLMdef, and (5) BERTje + LLMlist, which will be evaluated on (counterfactual) fairness and performance. We tested all of these models on the test data from the IMSyPP dataset with stratified sampling.

As shown in Table 2, the best-performing models are BERTje+SLL and BERTje+MGS, both achieving 79% accuracy with better precision, recall, and

F1-scores. LLMdef and LLMlist do not give better precision and recall, suggesting that counterfactual data generated by LLMs might introduce noise.

Analyzing the performance per class label (Table 3) gives more information on the model performance for specific hate speech classes. Interestingly, performance scores differ greatly between classes. The “inappropriate” and “violent” classes generally have lower precision, recall, and F1 scores than the “appropriate” and “offensive” classes, indicating class imbalances in the training data. The fine-tuned models show a slightly greater discrepancy than the baseline model.

6 Fairness Assessment

We also evaluate the fairness of these models using individual (counterfactual) and group fairness metrics. For fairness evaluation, we have categorized the SGTs into seven groups based on nationality, skincolor, gender, sexuality, religion, age, and ideology (see Table 7 in Appendix A). For an evaluation that focuses on fairness, we create a set of similar counterfactual data, which we refer to as *Dutch counterfactual evaluation dataset*. We manually curated this dataset with synthetic non-stereotypical instances created from a given set of templates. These templates consist of short toxic and non-toxic phrases like “hug [identity term]” or “destroy [identity term]”, in which SGTs are substituted. The idea is to emphasize adjectives without explicitly connecting them to the substituted SGTs, thereby assessing the model’s fairness based on adjective interpretation. All sentences with four tokens or fewer were selected from the dataset of Dixon et al. (Dixon et al., 2018) and manually translated to Dutch. In this process, template sentences with equal translations were only added once. For instance, the words ‘terrible’ and ‘horrible’ both translate to ‘verschrikkelijk’ in Dutch. This resulted in 17 toxic and 17 non-toxic synthetic templates. Consequently, the Dutch SGTs were substituted into these sentence templates, resulting in this Dutch counterfactual dataset with 2,890 sentences.

Note that the models are trained on four labels (appropriate, inappropriate, offensive or violent), and are being evaluated on the Dutch counterfactual dataset which has toxic and non-toxic labels. Because of the discrepancy between the predicted labels and the original labels, we consider the ‘appropriate’ and ‘inappropriate’ labels as non-toxic

and the ‘offensive’ and ‘violent’ as toxic.

Counterfactual Fairness: We use Counterfactual Token Fairness (CTF) (Garg et al., 2019) as a metric to assess the counterfactual fairness of our models. CTF aims to measure the fairness of the model outputs by assessing how much those outputs change when the inputs are altered with counterfactual examples. The CTF is quantified using the following equation:

$$CTF(X, X_{cf}) = \sum_{x \in X} \sum_{x' \in X_{cf}} |g(x) - g(x')| \quad (3)$$

In this equation, X represents the set of original input instances, which have at least one of the SGTs. Each instance x in X has a corresponding counterfactual instance x' in the set X_{cf} . The functions $g(x)$ and $g(x')$ compute the labels for the original sentence and the corresponding counterfactual sentence, respectively. The absolute difference $|g(x) - g(x')|$ captures how much the model’s predictions change between an original instance x and its counterfactual x' . A lower CTF value indicates that the model’s outputs for original sentences and their counterfactual versions are more similar. A lower CTF is desired because it suggests that the model treats counterfactual variants of the same input fairly.

Analyzing and comparing the CTF of the models clarifies how the specified counterfactual generation methods affect counterfactual fairness. As shown in Table 4, the baseline model has 0.24 as an overall. All counterfactual models improve fairness for non-toxic class but their performance deteriorates for toxic class. Among the counterfactual models, SLL is the most counterfactually fair both on average and also for the toxic templates. The baseline model gives more biased predictions for non-toxic templates compared to toxic templates.

Table 4: Counterfactual fairness results. Lower values mean more fair model.

Model	Toxic	Non-Toxic	Average
Baseline	0.11	0.36	0.24
BERTje+LLMdef	0.26	0.011	0.13
BERTje+LLMlist	0.32	0.001	0.16
BERTje+SLL	0.20	0.001	0.10
BERTje+MGS	0.28	0.003	0.14

Table 5 presents CTF scores for each SGT category for different models. As evident, all counterfactual models perform well for non-toxic templates for each of the SGT categories. For the toxic

Table 5: CTF Scores per Social Category. Lower values mean more fair model.

Model	Baseline		LLMdef		LLMlist		SLL		MGS	
	Tox	NonTox	Tox	NonTox	Tox	NonTox	Tox	NonTox	Tox	NonTox
Religion	0.02	0.38	0.20	0.00	0.23	0.00	0.22	0.00	0.25	0.00
Skin Color	0.31	0.26	0.14	0.00	0.25	0.00	0.11	0.00	0.27	0.00
Nationality	0.10	0.37	0.25	0.02	0.33	0.00	0.20	0.00	0.24	0.00
Ideology	0.01	0.46	0.39	0.01	0.27	0.01	0.19	0.01	0.33	0.01
Age	0.07	0.03	0.08	0.00	0.17	0.00	0.10	0.00	0.10	0.00
Gender	0.10	0.07	0.21	0.00	0.23	0.00	0.22	0.00	0.25	0.00
Sexuality	0.05	0.43	0.17	0.00	0.23	0.00	0.18	0.00	0.29	0.02

templates, counterfactual models match the fairness with base model only for the age category. In all remaining categories, fairness performance of counterfactual models underperform when compared to the baseline model.

Table 6: Fairness Metrics for fine-tuned models (based on groups rather than group terms). We compute demographic parity difference (DPD) and equalized odds difference (EOD). Lower values mean more fair model.

Model	DPD	EOD
Baseline	0.38	0.53
BERTje+LLMdef	0.09	0.18
BERTje+LLMlist	0.13	0.25
BERTje+SLL	0.06	0.11
BERTje+MGS	0.18	0.36

Group Fairness: Comparing the group fairness (Hardt et al., 2016) results gives insight into how the specified counterfactual generation methods impact a model’s group fairness (Table 6). The baseline model has the highest demographic parity difference (DPD) of 0.38, indicating that the model’s predicted positive rate differs greatly among identity groups. All counterfactual models perform better than the baseline. MGS improves the score to 0.18, while the SLL method reaches the lowest DPD of 0.06. Both MGS and SLL perform better than LLM-generated counterfactuals. Furthermore, the baseline model scores poorly for the equalized odds difference (EOD) with 0.53 as maximum score. The LLMdef and LLMlist methods improve this score to 0.18 and 0.25, respectively. MGS method moves it to 0.36, while SLL performs the best with the lowest EOD score of 0.11, indicating same trend as DPD. Although these scores are still considered sub-optimal, they signify an improvement with respect to the baseline model. Overall, both counterfactual generation models significantly improve in DPD and EOD, indicating fairer, less biased models. The SLL method outperforms MGS for both DPD and EOD, implying superior group fairness.

7 Limitations

Generating counterfactuals in Dutch for hate speech detection presents several challenges. We observe that some counterfactuals produced using the SLL and MGS methods do not adhere to proper Dutch grammar and occasionally result in sentences that lack semantic coherence. This issue is particularly evident with the SLL method, which relies on GPT-2, a model not optimally suited for the Dutch language. In contrast, counterfactuals generated by large language models (LLMs) tend to be more grammatically accurate and semantically plausible in Dutch.

Another limitation of our work lies in model performance. Specifically, we find that finetuning the Dutch-language BERTje model on counterfactual datasets generated via SLL and MGS methods improves fairness metrics. However, finetuning on datasets generated by LLMs does not yield similar improvements, especially for the toxic class. This is a counterintuitive outcome, suggesting that the counterfactual generation process may inadvertently introduce or reinforce biases, negatively impacting fairness for certain classes. One possible explanation is that unrealistic or noisy counterfactuals degrade the quality of the training data.

Additionally, Dutch social group terms were generated manually, and there is a risk that some relevant terms may have been omitted. Our prompt design approach also relies on baseline strategies; more sophisticated techniques, including ablation studies, are not explored in this work. Finally, we do not analyze how the number of counterfactuals affects model performance and fairness.

8 Conclusions and Future Work

This paper generates and evaluates the effect of counterfactual Dutch-language datasets for hate speech detection. The research finds that counterfactual generation methods introduce notable fairness improvements while maintaining performance

in the Dutch context. The resulting datasets are publicly available for research ², being the first such datasets for the Dutch language. Our empirical evaluations indicate improved overall counterfactual fairness and classification performance. The counterfactual models, in general, slightly improve on F1 score, precision and recall. Performance improvements may be partially due to increased training data for the counterfactual models, yet sometimes increased training data also exacerbates class imbalances. However, performing ablation studies with varying size of training data and fine-tuning strategies can be done in the future to study the impact of data variations on performance and fairness. In addition, extending this work with state-of-the-art LLMs can provide different results and insights.

References

- Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.
- Thales Bertaglia, Andreea Grigoriu, Michel Dumontier, and Gijs van Dijck. 2021. [Abusive language on social media through the legal looking glass](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 191–200, Online. Association for Computational Linguistics.
- Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke Van Der Veen, Gerben Timmerman, and Malvina Nissim. 2021. Dalc: the Dutch abusive language corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harm*, pages 54–66. Association for Computational Linguistics (ACL).
- Tommaso Caselli and Hylke Van Der Veen. 2023. Benchmarking offensive and abusive language in Dutch tweets. In *7th Workshop on Online Abuse and Harms, WOAHA 2023, co-located with ACL 2023*, pages 69–84. Association for Computational Linguistics, ACL Anthology.
- Simon Caton and Christian Haas. 2024. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38.
- Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.
- Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2021. Improving counterfactual generation for fair hate speech detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 92–101.
- Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.
- Fabio Del Vigna, Andrea Cimino, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.
- Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.
- Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41(8):e13562.
- Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.
- Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.
- Mika Hietanen and Johan Eddebo. 2023. Towards a definition of hate speech—with a focus on online contexts. *Journal of Communication Inquiry*, 47(4):440–458.
- Lisa Hilde, Ilija Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2023. Who are the haters? a corpus-based demographic analysis of authors of hate speech. *Frontiers in artificial intelligence*, 6:986890.
- Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.
- Jens Lemmens, Ilija Markov, and Walter Daelemans. 2021. Improving hate speech type and target detection with hateful metaphor features. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 7–16.
- Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLoS one*, 14(8):e0221152.

²<https://github.com/Juulba/Dutch-counterfactual-fairness>

- Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Dip-tikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13516–13524.
- Iliia Markov and Walter Daelemans. 2022. The role of context in detecting the target of hate speech. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, October, Gyeongju, Republic of Korea, pages 37–42.
- Iliia Markov, Ine Gevers, and Walter Daelemans. 2022. An ensemble approach for Dutch cross-domain hate speech detection. In *International conference on applications of natural language to information systems*, pages 3–15. Springer.
- Iliia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylistic and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.
- Ashish Mishra, Gyanaranjan Nayak, Suparna Bhat-tacharya, Tarun Kumar, Arpit Shah, and Martin Foltin. 2024. Llm-guided counterfactual data generation for fairer ai. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1538–1545.
- 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*, 9:88364–88376.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language 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)*, pages 5356–5371.
- Petra Kralj Novak, Igor Mozetič, Guy De Pauw, and Matteo Cinelli. 2021. [Hate speech detection and trends - imsyp multilingual hate speech database](#). Accessed: 2024-09-04.
- Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44.
- Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Urena-López, and M Teresa Martín-Valdivia. 2021. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.
- Anchal Rawat, Santosh Kumar, and Surender Singh Samant. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(2):e1648.
- Ward Ruitenbeek, Victor Zwart, Robin Van Der Noord, Zhenja Gnezdilov, and Tommaso Caselli. 2022. “zo grof!”: A comprehensive corpus for offensive and abusive language in Dutch. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 40–56.
- Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil van der Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of llm-generated counterfactually augmented data for harmful language detection. *arXiv preprint arXiv:2311.01270*.
- Guðbjartur Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for danish. In *Proceedings of the International Conference on Language Resources and Evaluation: LREC 2020*, pages 3498–3508. European Language Resources Association.
- Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G Deepalakshmi, Jaehyuk Cho, and G Manikandan. 2023. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, 80:110–121.
- Dion Theodoridis and Tommaso Caselli. 2022. All that glitters is not gold: Transfer-learning for offensive language detection in Dutch. *Computational Linguistics in the Netherlands Journal*, 12:141–164.
- Stéphan Tulkens, Lisa Hilde, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016a. The automated detection of racist discourse in Dutch social media. *Computational linguistics in the Netherlands journal*, 6:3–20.
- Stéphan Tulkens, Lisa Hilde, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016b. A dictionary-based approach to racism detection in Dutch social media. In *Workshop Programme*, page 11.
- Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7.
- R Vries. 2024. Analysing targets of hate speech on x in the netherlands using bert-cnn. B.S. thesis, University of Twente.
- Linyi Yang, Eoin Kenny, Tin Lok Ng, Yi Yang, Barry Smith, and Ruihai Dong. 2020. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *COLING 2020-28th International Conference on Computational Linguistics, Proceedings of the Conference*.
- Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

A Categories of Social Group Terms in the Dutch Context

Table 7 maps the social group terms from the Dutch context into seven categories.

Table 7: Categorization of Social Group Terms.

Category	Values
nationality	turk, marokkaan, mocro, nederlander, hollander, indo, suri, surinamer, antiliaan, syrier, chinees, duitser, belg, afrikaan, aziaat, amerikaan, arabier, latina, latino, turks, marokkaans, mocro, nederlands, hollands, indonesisch, indo, suri, surinaams, antilliaans, syrisch, chinees, duits, belgisch, afrikaans, engels, aziatisch, Amerikaanse, arabisch, latina, latino, immigrant, migrant, asielzoeker, vluchteling, autochtoon, allochtoon
skincolor	wit, blank, zwart, donker
gender	vrouw, man, transgender, non-binair, vrouwelijk, mannelijk, masculien, intersekse
sexuality	homo, hetero, heteroseksueel, homoseksueel, trans, bi, lesbisch, biseksueel
religion	christelijk, katholiek, moslim, protestant, jood, joods
age	oudere, jongere, volwassene, student, oud, jong, volwassen
ideology	marxist, populist, feminist, liberaal, boer, communist, rechts, links, marxistisch, populistisch, feministisch, liberaal, communistisch

B LLM Prompts

We use the following two LLM prompts for generating counterfactual sentences.

LLMdef: *Your task is to generate five counterfactual sentences for a given input sentence in Dutch. Identify social group terms in the input sentence and modify the input sentence by changing the social group(s) with other social group terms. Each counterfactual sentence should differ from the original while preserving the original meaning. Social group terms are terms that signify social*

groups based on characteristics like gender, race, class, sexuality, political leaning, religion, education, age, profession etc. Please generate five distinct counterfactual sentences for the input sentence: `input_sentence` Your response should be a JSON list of five dictionaries where each dictionary contains: - `'input_sentence'`: the original sentence. - `'counterfactual sentence'`: the generated counterfactual sentence.

LLMlist: *Your task is to generate seven counterfactual sentences for a given input sentence in Dutch language. Identify and modify the input sentence by changing the social group term(s) with other social group term(s) and change other words without changing the level of toxicity expressed in input sentence based on `input_sentence`. The generated sentence should make contextual sense. Some social group terms in Dutch context are: "heteroseksueel", "hetero", "homoseksueel", "homo", "trans", "transgender", "non-binair", "vrouw", "vrouwelijk", "man", "mannelijk", "masculien", "biseksueel", "lesbisch", "bi", "intersekse", "turk", "turks", "marokkaan", "marokkaans", "mocro", "nederlander", "nederlands", "hollander", "hollands", "indonesisch", "indo", "suri", "surinamer", "surinaams", "antiliaan", "antilliaans", "zwart", "wit", "blank", "syrisch", "syrier", "chinees", "duits", "duitser", "belg", "belgisch", "afrikaans", "afrikaan", "jood", "joods", "asielzoeker", "vluchteling", "autochtoon", "allochtoon", "engels", "aziatisch", "aziaat", "amerikaan", "amerikaans", "arabier", "arabisch", "latina", "latino", "immigrant", "migrant", "donker", "christelijk", "katholiek", "moslim", "protestant", "rechts", "links", "marxist", "marxistisch", "populist", "populistisch", "feminist", "feministisch", "liberaal", "boer", "communist", "communistisch", "oudere", "oud", "jongere", "jong", "volwassene", "volwassen", and "student". Input sentence is: `input_sentence`. Give your response as a dictionary of "counterfactual sentences" with consistent formatting.*

C Dataset Statistics

In Tables 8, 9, 10, 11, and 12, we provide distribution of categories of social group terms in each of the four labels in all datasets.

D Qualitative Assessment

Table 8: Baseline Dataset.

label	count	age	gender	ideology	nationality	religion	sexuality	skincolor
Appropriate	1057	98	507	120	134	46	25	118
Inappropriate	289	17	72	16	83	36	20	41
Offensive	1436	66	503	259	298	86	64	156
Violent	47	3	19	5	11	2	3	4
	2825	184	1101	400	526	170	112	319

Table 9: LLMdef generated Dataset.

label	count	age	gender	ideology	nationality	religion	sexuality	skincolor
Appropriate	6110	500	2574	633	1359	265	105	619
Inappropriate	1702	97	340	73	691	185	91	200
Offensive	8249	383	2464	1458	2388	459	262	800
Violent	304	16	98	46	99	10	15	20
	16305	996	5476	2210	4537	919	473	1639

Table 10: LLMlist generated Dataset.

label	count	age	gender	ideology	nationality	religion	sexuality	skincolor
Appropriate	7965	1066	1971	636	2088	524	1022	580
Inappropriate	2277	189	429	144	793	168	323	189
Offensive	10747	965	2346	1483	3107	678	1251	787
Violent	360	29	109	26	98	24	55	17
	21160	2249	4855	2289	6086	1394	2651	1573

Table 11: SLL generated Dataset.

label	count	age	gender	ideology	nationality	religion	sexuality	skincolor
Appropriate	15624	1652	1572	2586	5547	543	1956	1514
Inappropriate	6850	723	676	1111	2653	277	736	555
Offensive	25663	2731	2690	4110	9509	948	2922	2322
Violent	967	106	102	156	342	33	116	92
	49104	5212	5040	7963	18051	1801	5730	4483

Table 12: MGS generated Dataset.

label	count	age	gender	ideology	nationality	religion	sexuality	skincolor
Appropriate	6580	198	894	518	4532	117	133	188
Inappropriate	2863	28	110	48	2418	96	105	58
Offensive	10481	133	818	1302	7506	213	287	222
Violent	469	4	36	33	361	6	21	8
	20393	363	1858	1901	14817	432	546	476

Table 13: Examples of incorrect counterfactual sentences generated by SLL method. English translation is within brackets.

<i>incorrect adjective conjugation</i>	
Original sentence	volwassen mensen (adult people)
Incorrect counterfactual	jong mensen (young people)
Correct counterfactual	jongere mensen (younger people)
<i>wrong noun adjective substitution</i>	
Original sentence	zielige allochtoon (pitiful immigrant)
Incorrect counterfactual	zielige Turks (pitiful Turkish)
Correct counterfactual	zielige Turk (pitiful Turk)
<i>unlikely counterfactual generation</i>	
Original sentence	uit de kast komen als vrouw (coming out as female)
Incorrect counterfactual	uit de kast komen als jong (coming out as young)
Correct counterfactual	uit de kast komen als homo (coming out as gay)

Table 14: Examples of incorrect counterfactual sentences generated by MGS method. English translation is within brackets.

<i>wrong noun adjective substitution</i>	
Original sentence	Chinees gehaald (taken out Chinese)
Incorrect counterfactual	Turk gehaald (taken out Turk)
Correct counterfactual	Turks gehaald (taken out Turkish)
<i>unlikely counterfactual generation</i>	
Original sentence	spreek Nederlands (speak Dutch)
Incorrect counterfactual	spreek Latina (speak Latina)
Correct counterfactual	spreek Belgisch (speak Belgian)

Table 15: Examples of incorrect counterfactual sentences generated by LLMlist. English translation is within brackets.

<i>incorrect adjective conjugation LLMlist</i>	
Original sentence	Mocro broeders (Moroccan brothers)
Incorrect counterfactual	Marokkaan broeders (Moroccan person brothers)
Correct counterfactual	Marokkaanse broeders (Moroccan brothers)
<i>unlikely counterfactual generation</i>	
Original sentence	iemand zwart maken (discredit someone)
Incorrect counterfactual	iemand wit maken (to make someone white)
<i>unlikely counterfactual generation</i>	
Original sentence	uit de kast komen als vrouw (coming out as female)
Incorrect counterfactual	uit de kast komen als jong (coming out as old)
Correct counterfactual	uit de kast komen als man (coming out as male)

Table 16: Example of incorrect counterfactual sentence generated by LLMdef. English translation is within brackets.

<i>unlikely counterfactual generation</i>	
Original sentence	rood, zwart en groen (red, black and green)
Incorrect counterfactual	student, docent en directeur (student, teacher and principal)