Detecting Microaggressions Using LLM-Generated Counterfactuals: Transformer Cross-Domain Evaluation and LLM Classification

Nura Hossainzadeh, Neeharika Kotte, Carlos Schrupp

Abstract

Microaggressions are subtle forms of hate speech. Because they are not directly or explicitly offensive in nature, they require a nuanced approach to identify accurately, particularly in text. Our study presents both transformer and LLM (GPT-5.1) models that are able to detect microaggressive speech with a high degree of accuracy. The key to achieving this result was the development of a novel dataset, used to train our transformer models and evaluate our LLM model, consisting of microaggressive statements from a well-known corpus paired with LLM-generated non-microaggressive counterparts that matched their linguistic and stylistic properties while differing in semantic meaning. We found that our LLM model achieved a 90% accuracy on a 100-entry sample of our generated dataset. Of the transformer models we tested, deBERTa exhibited the best performance, achieving high accuracy (81.87%) even when cross-evaluated on a previously unseen workplace microaggressions dataset on which it had not been trained.

1 Introduction

The term “microaggression” was coined by African American psychiatrist and Harvard professor Chester Pierce in 1970 to describe “subtle” and “often automatic” offenses directed against people of color. Since then, the concept has been extended to other marginalized groups. Microaggressions often implicitly demean or stereotype a person’s heritage or identity (for example, by telling a person of color “you are a credit to your race”), or they may exclude or minimize the thoughts and experiences of a marginalized person (for example, telling a stay-at-home mother “life must be breeze when you don’t have to work”) (Sue and Spanierman, 2020). Because they are so subtle, the responsibility of raising awareness about microaggressions often falls on the shoulders of marginalized groups. Having to experience and simultaneously prove the existence of microaggressions has been found to have long-term adverse physical and emotional effects on victims; in fact, this double burden can have a more debilitating effect than explicitly aggressive acts (Sue, 2010; Nadal et. al., 2014).

With the explosive growth of social media and online communication, hate speech and microaggressions have become pervasive in digital text. Therefore, it has become increasingly important to develop methods of identifying microaggressions, a task for which modern natural language processing models can be leveraged. Previous work has mostly focused on explicit hate (see, for example, Miqdadi, et. al., 2024; Benítez-Andrades et. al., 2022; and Lee, et. al., 2022), but much less research has been completed on how and whether natural language processing can be used to classify microaggressions. While hate speech is overtly violent, microaggressions are linguistically softer, and violent meaning emerges from non-violent words. The difficulty of classifying microaggressions is compounded by the fact that contextual factors, such as cultural knowledge, body language, and tone, may not be accessible to the model. For instance, a seemingly innocuous joke about "turning someone in" may be deeply offensive given recent instances of police violence and brutality. Therefore, these complex subtleties, combined with the quick, implicit cadences of microaggressions, present fundamental challenges for machine learning algorithms.

Our work addresses these challenges by composing BERT and LLM model-based classifiers, trained on a novel microaggressions dataset, that have learned to effectively differentiate microaggressions and non-microaggressions. We confirm this with strong cross-evaluation performance on a second smaller dataset composed of microaggressions expressed in a workplace context unseen during training.

Our second contribution is an engineered, balanced dataset composed of, first, microaggressions drawn from a well known microaggression corpus, and second, LLM-generated non-microaggressions that mirror the syntax and style of the microaggression. This larger, balanced corpus of microaggressive and non-microaggressive statements differ primarily in meaning rather than in surface-level linguistic or stylistic features, allowing models trained on this dataset to focus on learning what distinguishes microaggressions semantically.

Beyond its scholarly and scientific relevance, our work has practical relevance. Hate speech—microaggressive or not—is an impediment to social justice, insofar as it mutes and intimidates minority voices and undermines their well-being. Our hope is that these models can be used to detect microaggressive speech in online fora to moderate content and create safer spaces for minorities to exist and express themselves. We consider our approach successful if, first, our performance on the evaluation datasets yields an overall accuracy of at least ten percentage points above our baseline, and second, if we observe high precision on the microaggressions class (and consequently a high F1-score); this would indicate that our classifier has learned what fundamentally characterizes microaggressive statements.

2 Background

Because microaggressive speech often goes beyond the realm of the explicitly linguistic, the literature on this topic is sparse. Two early papers studied microaggressions by focusing on data analysis and without employing complex machine learning architectures. First, a widely-cited study was done in 2019 (Breitfeller et. al.) that focused mostly on the question of how to collect and accurately label microaggressive comments, creating the selfMA dataset, which we also draw from in our study. An early paper (Ali et. al, 2020) used feature selection and classical machine learning models (such as logistic regression and Naïve Bayes) to classify whether or not statements were racial microaggressions. We go beyond this study by using more recent deep learning architectures, aiming to capture the semantic, contextual meaning of expressions rather than relying on shallow lexical features such as word frequency.

In addition to studying microaggressions, researchers have studied implicit hate speech, which is hard to differentiate from microaggressive speech; these studies, we would hold, are in fact about microaggressions even if authors do not use this term. Sasse et. al (2025) study how to detect dog whistles, which are terms that appear innocuous to the general public but are known to be disparaging by certain "in-groups." They use vector databases (with the help of an LLM filter) to detect dog whistles, which are generally semantically similar to the disparaging terms and ideas they imply. The approach we follow here instead is to use transformer and LLM models rather than vector databases to detect semantic similarity of subtle or indirect terms.

In the same vein as dog whistles, Wiegand and Ruppenhofer (2024) study depictions of minority groups deviating from the norm, which is implicitly disparaging ("gays like to sprinkle flour on their gardens"); they find that deBERTa fine-tuned on LLM-generated data is able to detect this "othering" accurately. Like these authors, we use a dataset that is in part LLM-generated, and we also find deBERTa to be a highly effective transformer to use for this classification task. However, our interest goes beyond detecting statements that "other" and thereby implicitly disparage. We are interested in the full spectrum of subtle disparaging comments that may be directed against members of marginalized communities (stereotyping, assertions of power/authority, etc.).

Finally, Hartvigsen et. al. (2022) create Toxigen, a machine-generated database composed of both implicit hate speech and benign speech; the authors cite a 93% accuracy on a human-validated Toxigen test split after fine-tuning transformers trained on explicit hate speech (hateBERT and ToxDectRoBERTa) on Toxigen. Unlike Toxigen, our primary dataset is composed of a mix of machine-generated statements and other statements pulled directly from the internet. In this way, we avoid a potential pitfall of over-reliance on machine-generated text that differs in substantial ways from human-written texts found in the wild.

3 Methods

3.1 Proposed Approach and Balancing Challenges

Our proposed approach involves training a convolutional neural network as a baseline, various transformer models (BERT-base-cased, RoBERTa, HateBERT, and deBERTa), and a LLM-based model on both microaggressive and non-microaggressive statements, and then evaluating on a separate, smaller workplace microaggressions dataset, drawn from a different source and context, for cross-evaluation.

Currently, there are no well-established or referenced balanced microaggressions datasets for model training and evaluation. The largest, most well-known microaggressions dataset is selfMA (Self-Annotated Microaggressions), created by Breitfeller et. al (2019). It is composed of self-reported microaggressions drawn from "The Microaggressions Project" Tumblr website (www.microaggressions.com), which was created in 2010 by two Columbia University students. It serves as a platform for individuals to share their experiences with microaggressions and bring awareness to how these microaggressions perpetuate harm across multiple settings, including at work, in school, or in social settings. A key challenge with selfMA, however, is that it only contains microaggression examples, resulting in a severe class imbalance that prevents effective binary classification (microaggression versus normal text) tasks. We address this challenge futher below, in Section 3.2.

We also leverage the Microaggressions in the Workplace (workplace MA) dataset as our secondary corpus for cross-evaluation, as we are limited by its small sample size (171 entries) and domain-specific (workplace) data. Unlike selfMA, this dataset is balanced, with roughly equal numbers of microaggressive and non-microaggressive statements written by the dataset author. The limited number of entries in this dataset allowed us to thoroughly review the dataset's content, ensuring that microaggressions and non-microaggressions were accurately labeled.

3.2 Balancing the Data: Non-Microaggressions Pair Generation

To address the key issue of class imbalance with selfMA, we first attempted to balance it with external datasets. We used two datasets for this task: Toxigen (the large-scaled machine-generated hate speech database described in the Background section) and the Social Bias Inference Corpus (SBIC), a corpus of social media posts including both offensive and non-offensive statements. However, while initial experiments resulted in high overall accuracy on test splits, there was significant degradation in cross-evaluation performance on workplace MA. Readability assessments also revealed substantial differences between selfMA and these external datasets; therefore, this approach was unsuccessful likely because the model capitalized on stylistic differences unique to the datasets, rather than the semantic differences between microaggressions and nonmicroaggressions (see more in Appendix Section 8.4).

Given these challenges, our final approach was to generate non-microaggression equivalents of the 1,300 microaggression entries in selfMA. We provided a large language model with an explicit definition of microaggressions and a few-shot prompt, asking it to produce an alternative version of the microaggression (see more in Appendix Section 8.1). For example, this microaggression entry: "You're pretty for an Asian girl" would be balanced with a structurally equivalent, benign text: "You're pretty." We aimed to maintain as similar a sentence structure and semantic meaning as possible, with pairs differing only in the inclusion of phrase(s) with microaggression qualities. This construction pairs each microaggression with a closely matched non-microaggressive control, encouraging models to focus on the presence or absence of microaggressive content rather than on dataset-specific artifacts. We combine and shuffle original statements and counterfactuals to obtain a balanced binary dataset, which we now refer to as `selfMA_generated`.

To ensure that this dataset did not exhibit the same flaw as the previous datasets we had created—significant readability score differences between the text of microaggressions and non-microaggressions—we again performed the same readability tests on LLM-generated non-microaggressions and the microaggressions drawn from selfMA. We found that readability score differences between these statements were much less drastic than they had been before, indicating that our model was unlikely to rely on superficial linguistic differences. Microaggressive statements drawn from selfMA in `selfMA_generated` had a Flesch-Kincaid Grade Level score of 3.6, while non-microaggressive statements generated by the LLM had a slightly higher grade level score of 5.5. Similarly, while selfMA had a Gunning Fog index of 6.1, LLM-generated statements had an index of 7.5. Finally, the average sentence length for each dataset was hardly differentiable: 9.7 for selfMA and 9.6 for the generated entries.

To establish an acceptable difference threshold, we used readability score differences between classes in workplace MA as our benchmark. Since workplace MA contained entries from both classes, any readability differences could not be attributed to differences in dataset authorship. To apply the benchmark, we calculated the ratio of the readability scores between microaggressions and non-

Metric	**Workplace Ratio (Nonmicro / Micro)**	**Original Generated Ratio (Generated / Original)**
Flesch reading ease	0.93	0.85
Flesch Kincaid grade	1.24	1.48
Gunning Fog	1.00	1.23
Smog index	1.01	1.23
Coleman Liau index	1.37	1.53
Automated readability index	1.18	1.41
Dale Chall readability	1.45	1.18
Linsear Write	1.07	1.13
Difficult words	1.60	1.20
Avg sentence length	1.06	1.00
Avg syllables per word	1.04	1.12
Word count	1.05	0.71
Char count	1.08	0.77

Table 1: A comparison of readability score ratios (between classes) in Workplace MA vs. in `selfMA_generated`, showing no major differences.

microaggressions in workplace MA and compared them to the ratios of readability scores between components of `selfMA_generated` (see Table 1). We found that overall, ratios were quite similar, indicating that the microaggressive and non-microaggressive statements of `selfMA_generated` had readability differences that were comparable to microaggressive and non-microaggressive statements of workplace MA, a dataset that had not been generated and included entries from both classes from a single source. This result strongly indicated that the model would not rely on stylistic linguistic differences to differentiate between microaggressions and non-microaggressions in `selfMA_generated`.

Beyond readability, we also conducted an unsupervised proxy audit on the `selfMA_generated` training split to test whether simple stylistic or lexical patterns could recover the labels. We extracted three feature views—style-only indicators (such as length, punctuation, capitalization), identity terms such as race, gender, religion and nationality, and fastText-based sentence embeddings—and applied KMeans with k=2, treating the resulting cluster assignments as predicted labels. Cluster–label agreement, measured with Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), was near zero across all three views (e.g., ARI $\in$ [-0.000, 0.019], NMI $\in$ [0.000, 0.041]), indicating no more alignment than would be expected by chance (see Table 2). Taken together with the readability analysis, these results suggest that the `selfMA_generated` dataset reduces trivial non-semantic shortcuts; models must rely on more substantive semantic cues, rather than superficial stylistic differences, to distinguish microaggressions from non-microaggressions.

Feature view	**ARI**	**NMI**
Style-only	0.02	0.04
Lexical-proxy	0.00	0.00
fastText semantic	0.00	0.01

Table 2: Unsupervised proxy audit on the selfMA-generated train split. The table reports the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) between KMeans clusters and gold labels for three feature views (style-only, lexical-proxy, and fastText semantic), showing that all unsupervised clusterings are only weakly aligned with the true labels.

3.3 Experimental Design, Evaluation Strategy, and Success Criteria

To determine the most effective model architecture for microaggression detection, we conducted multiple experimental approaches in addition to our baseline CNN model, with `selfMA_generated` as our main training corpus.

We selected a CNN as our main deep-learning baseline because it is a standard, compute-efficient architecture that has shown strong performance for racism, xenophobia, and subtle toxicity detection on short texts in prior work (e.g., Benítez-Andrades et al., 2022; Gilda et al., 2022; El Miqdadi et al., 2024; Ledalla et al., 2023).

Our main set of experiments trained on `selfMA_generated` across multiple transformer architectures (BERT, RoBERTa, HateBERT, DeBERTa) to identify which model best captures the subtle semantic distinctions between microaggressions and non-microaggressions. We started with BERT as a basic transformer and then experimented with RoBERTa since it was trained on a larger corpus than BERT. We also experimented with HateBERT (created by Caselli et. al., 2021) since it is already trained on hate speech and exhibits high accuracy with identifying not only explicit hate speech but, as demonstrated by Hartvigsen et. al. (2022) on their Toxigen dataset, hate speech that is subtle. Finally, we chose to test DeBERTa because it is a model commonly used to detect implicit speech; for example, Farha et. al (2022) uses it to detect sarcasm and Wiegand et. al (2024) for "othering." Across these transformer models, we utilize the following hyperparameters for training: batch_size = 16, num_epochs = 3, learning_rate = 2e-5. We then performed cross-evaluation with workplace MA to assess the models' ability to generalize across contexts and identify microaggressions in a completely independent and unseen corpus. This dual evaluation strategy tests both whether models learn meaningful semantic features and whether they avoid overfitting to source-specific patterns.

Lastly, we performed a definition-based prompting evaluation with GPT-5.1, supplying the model with the following classification rules: ***not_microaggressive****: The message is neutral, factual, or generally benign without any underlying bias or aggression;* ***microaggressive****: The message contains subtle, indirect, or unintentional expressions of bias, insensitivity, or invalidation. Look for backhanded compliments, questioning remarks, assumptions based on identity, or microinvalidations (e.g., "Where are you really from?", "You're so articulate for a [minority group]", "I'm not racist, but...").*

The LLM model was evaluated on a random 100-entry sample drawn from the entire `selfMA_generated` dataset. To make the classification task more challenging for the LLM model, we shuffled the dataset beforehand to ensure that non-microaggressive statements and the original microaggressive statement from which they were generated were not likely to both be in the 100-entry sample (perhaps the model would too easily be able to pick up on the difference between a microaggression and its exact negation).

Our first success metric across all experiments was overall accuracy on the `selfMA_generated` test split. In the case of transformer models in particular, our second metric was cross-evaluation accuracy on workplace MA; since this was a completely unseen dataset, it was the more important success metric for our transformer models. We also examined precision and F1-scores for the microaggressions class in workplace MA as secondary success metrics. We focused on precision in particular, as high precision indicates that the classifier reliably identifies true microaggressions with minimal mischaracterizations of benign statements. Strong F1-scores additionally demonstrate more balanced performance, ensuring that the model achieves high precision without being overly conservative and missing too many of these true microaggressions.

4 Results

4.1 Baseline

Our baseline model is a CNN consisting of an embedding layer followed by multiple parallel convolutional filters with different kernel sizes, global max-pooling, and two dense layers, and we train it on our labeled microaggression corpus and evaluate it on a held-out test split. On this test set, the CNN baseline achieves the accuracy and F1 scores reported in Table 3; the drop in performance on the workplace MA dataset shows that the CNN primarily learns dataset-specific patterns rather than fully generalizable semantic representations of microaggression. This reinforces the need for more complex transformer-based models and the cross-dataset generalization experiments described later.

4.2 BERT Models Results

We evaluated four transformer-based architectures trained directly on our `selfMA_generated` dataset, which showed improved performance compared to the baseline. All models (BERT-base-cased, hateBERT, roBERTa-base, deBERTa-v3-base) achieved strong overall performance on the unseen test split of this dataset, with accuracies ranging from 87.69% with hateBERT to 92.69% with deBERTa-v3-base, as shown in the table below. Cross-evaluation on the workplace MA dataset, however, while strong, showed varying levels of generalization across these models. deBERTa-v3-base had the highest cross-evaluation overall accuracy, at 81.87% (approximately 10 percentage point drop from `selfMA_generated` test set), followed closely by BERT-base-cased at 80.70% (approximately 8 percentage point drop) and hateBERT at 80.12% (approximately 7 percentage point drop), while roBERTa-base showed the weakest cross-dataset generalization, achieving only 76.61% accuracy on workplace MA (approximately 14 percentage point drop). This is expected, however, as deBERTa-v3-base was developed to outperform roBERTa-base and BERT models by incorporating a disentangled attention mechanism and an enhanced mask decoder (He, Liu, et al., 2020). Still, because the majority of these models were able to achieve high accuracies on an unseen dataset, this indicates that they have likely learned meaningful microaggression patterns rather than relying on superficial, text-based cues. Similarly, Hartvigsen et. al (2022) also cross-evaluate hateBERT and RoBERTa fine-tuned on their entirely LLM-generated dataset (Toxigen) on several unseen publicly available human-written implicit hate speech datasets, but observe much larger drops in cross-evaluation accuracy than we did (in most cases, close to 20 percentage points or more).

Examining the confusion matrices and precision and recall metrics across all models when evaluated on workplace MA reveal a consistent pattern: high precision but moderate recall for microaggressions. For instance, deBERTa-v3-base achieved perfect precision on identifying microaggressions but only 63% recall, resulting in an F1-score of 77%. All the models, however, very successfully identify non-aggressive or normal texts (ranging from 93-100% recall). Therefore, the models' classification behavior is quite conservative in that it would rather incur false negatives than false positives.

4.3 LLM Results

The LLM model proved to be effective at distinguishing microaggressions from non-microaggressions in the 100-entry sample drawn from `selfMA_generated`; it achieved an overall accuracy of 90%. Like our transformer models, the LLM exhibited similar pattern: high precision with identifying microaggressions and lower recall. However, both precision and recall were still high compared to BERT models—its precision on microaggression identification was 97.6% and its recall was 82%, resulting in an F1-score of 89%. Given that the LLM was given a very short prompt (text included in Section 3.3), its ability to identify microaggressions was notable.

The accuracy of the LLM model can in part be attributed to its access to the contexts that surround microaggressions, as scholars have noted (see, for example, Guo et. al). While our transformer models studied the microaggressive statements with limited knowledge of the context in which they were spoken, LLMs are potentially more aware of this broader context. Since context is so crucial to understanding microaggressions (asking a person of color "where are you from?" at an International students' fair is not offensive, while in most other contexts it would be), this knowledge likely was particularly helpful for microaggression detection. In addition, the precise definition we provided the LLM in the prompt likely helped it to identify microaggressions. Numerous studies demonstrate the impact of prompt content on model performance, showing that detailed prompts are key, and providing the models with examples of statements that fall into each category often (but not always) help performance (see, for example, Kumarage et. al. and Weber et. al.). Our prompt, and the microaggression examples embedded in the prompt, likely helped the model to achieve such high accuracy.

Model	**Eval. acc./F1** **on `selfMA_generated`**	**Cross eval. acc./F1** **on workplaceMA**
CNN (baseline)	78.46% / 78.45%	69.01% / 66.75%
BERT-base-cased	89.62% / 90.00%	80.70% / 80.00%
hateBERT	87.69% / 88.00%	80.12% / 80.00%
roBERTa-base	91.54% / 92.00%	76.61% / 75.00%
deBERTa-v3-base	92.69% / 93.00%	81.87% / 81.00%
GPT-5.1 (definition-based prompting)	90.00% / 89.94%	N/A

Table 3: Cross-domain performance of baseline and transformer models trained on `selfMA_generated`. All models except GPT-5.1 are fine-tuned on the `selfMA_generated` training split and evaluated on its test split (in-domain) and on the workplaceMA corpus (cross-domain); GPT-5.1 is evaluated on a 100-example sample with a definiton-bases prompt and is not fine-tuned.

At the same time, these results should not be taken to imply that large proprietary LLMs such as GPT-5.1 are a straightforward solution for microaggression detection. First, they are opaque, non-deterministic systems whose decision boundaries cannot be audited or reproduced in the same way as fixed transformer models fine-tuned on a known dataset, which raises concerns about transparency and accountability. Second, because such models are trained on broad, largely undocumented web corpora, they may encode and reproduce societal biases in ways that are difficult to anticipate, potentially over-flagging the language of marginalized speakers or under-flagging harmful language that aligns with majority norms. Third, relying on an external, general-purpose API for classification introduces additional risks around data governance, privacy, and long-term reproducibility of results, particularly when the inputs involve sensitive or identifiable accounts of discrimination. For these reasons, we view GPT-5.1's performance in our study as evidence of the capabilities of contemporary LLMs rather than an endorsement of their direct, uncritical use as production microaggression detectors.

5 Conclusion

Human communication is perhaps just as much implicit as it is explicit, and microaggressions are a form of implicit communication that often escapes the notice of humans (at least those who are not their victims) and machine-learning algorithms alike. In our work, we have shown that machine learning models can effectively detect microaggressions, despite their implicit qualities, perhaps taking some of the burden off of marginalized individuals who must convince others that they exist, they are widespread, and they are deeply harmful. We have shown that LLM-generated counterfactual balancing is crucial for training models to distinguish microaggressions from non-microaggressions, with our BERT models—particularly DeBERTa—and our LLM classifier achieving high performance on both our generated dataset and, in the case of the transformer models, an independently authored workplace microaggressions dataset, with all but one BERT model and the LLM surpassing our threshold of at least ten percentage points above baseline. Future studies could incorporate human feedback on our dataset, asking for multiple perspectives on whether each statement is correctly labeled, and even asking participants to share their own examples of microaggressions and corresponding non-microaggressions. Models trained and tested on text that reflects human understandings of aggression and hate become better able to recognize the very words and sentiments that have historically produced such harm.

6 Author Contributions

Nura Hossainzadeh: Led data curation, literature review, and transformer and LLM evaluation experiments.
Neeharika Kotte: Performed initial exploratory data analysis, implemented and tuned transformer models and conducted sequential transfer learning experiments.
Carlos Schrupp: Designed and implemented the dataset construction pipelines, developed the CNN baseline, and maintained the experimental infrastructure.
All team members composed the final report together.

7 References

Code repository

- Project Group. (2025). *Detecting Microaggressions Using LLM-Generated Counterfactuals: Transformer Cross-Domain Evaluation and LLM Classification* [Code repository]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha

Project notebooks

Main analysis notebooks

- Project Group. (2025). *Dataset Balancing via LLM Counterfactual Generation for selfMA-generated* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/1_1_dataset_balance_selfMA_counterfactual_generation.ipynb

- Project Group. (2025). *Baseline CNN Model Training on selfMA-generated (v2)* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/1_2_baseline_CNN_SelfMA_generated_v2.ipynb

- Project Group. (2025). *BERT Model Training and Evaluation on selfMA-generated* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/1_3_bert_selfMA_generated.ipynb

- Project Group. (2025). *LLM-based Classification Experiments on selfMA-generated* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/project_Carlos_Nura_Neeha/blob/main/Notebooks/1_4_llm_SelfMA_generated.ipynb

- Project Group. (2025). *Readability Metrics for selfMA-generated and workplaceMA* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/1_5_readability_metrics_selfMA_generated_workplaceMA.ipynb

- Project Group. (2025). *Proxy Audit on Counterfactual selfMA-generated Dataset* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/1_6_audit_proxy_counterfactual_generated.ipynb

Additional notebooks

- Project Group. (2025). *Data Scraping and Collection for Microaggressions* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/2_0_scraping_code_microaggressions.ipynb

- Project Group. (2025). *Initial Exploratory Data Analysis of Microaggression Datasets (Neeha)* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/3_1_EDA_Neeha.ipynb

- Project Group. (2025). *EDA of selfMA and Toxigen Datasets (Carlos)* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/3_2_EDA_Carlos_SelfMA_Toxigen.ipynb

- Project Group. (2025). *EDA of workplaceMA Dataset (Nura)* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/3_3__EDA_Nura_workplaceMA.ipynb

- Project Group. (2025). *EDA of workplaceMA Dataset (Carlos)* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/3_4_EDA__Carlos_workplaceMA.ipynb

- Project Group. (2025). *Exploratory Data Analysis of the selfMA-generated Dataset* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/3_5_EDA_SelfMA_generated.ipynb

- Project Group. (2025). *Baseline CNN Training on selfMA + SBIC Balanced Dataset* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/4_1_baseline_CNN_SelfMA_SBIC.ipynb

- Project Group. (2025). *Baseline CNN Training on selfMA + Toxigen Balanced Dataset* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/4_2_baseline_CNN_SelfMA_Toxigen.ipynb

- Project Group. (2025). *Sequential BERT-base Training Pipeline with workplaceMA* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/5_1_sequential_bert_base_cased_workplaceMA.ipynb

- Project Group. (2025). *Sequential BERT-base Training on selfMA + SBIC* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/5_2_sequential_bert_base_cased_selfMA_SBIC.ipynb

- Project Group. (2025). *Sequential BERT-base Training on ISHate, iSarcasm, and selfMA* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/5_3_sequential_bert_base_cased_ISHate_Isarcasm_selfMA.ipynb

- Project Group. (2025). *Sequential BERT-base Training: ISHate → iSarcasm → selfMA → workplaceMA* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/5_4_sequential_bert_base_cased_ISHate_Isarcasm_selfMA_workplaceMA.ipynb

- Project Group. (2025). *Sequential BERT-base Training: selfMA-generated, workplaceMA, and iSarcasm* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/5_5_sequential_bert_base_cased_selfMA_generated_workplaceMA_iSarcasmt.ipynb

- Project Group. (2025). *RoBERTa Model Training on the ISHate Dataset* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/6_1_roBERTa_ISHate.ipynb

- Project Group. (2025). *HateBERT Model Training on the ISHate Dataset* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/6_2_hateBERT_IShate.ipynb

- Project Group. (2025). *BERT Model Training on the selfMA Dataset* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/6_3_bert_selfMA.ipynb

- Project Group. (2025). *ModernBERT Model Training on the selfMA Dataset* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/6_4_modernbert_selfMA.ipynb

- Project Group. (2025). *BERT Model Training on the selfMA and Toxigen Datasets* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/6_5_bert_selfMA_toxigen.ipynb

- Project Group. (2025). *LLM-based Experiments with BART and DeBERTa on selfMA* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/7_1_llm_bart_deberta_selfMA.ipynb

- Project Group. (2025). *GPT-5.1 Evaluation on selfMA-generated* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/7_2_llm_gpt_5_1_selfMA_generated.ipynb

- Project Group. (2025). *Readability Analysis of selfMA and Toxigen* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/8_1_readability_selfMA_Toxigen.ipynb

- Project Group. (2025). *Readability Analysis of selfMA-generated* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/8_2_readability_selfMA_generated.ipynb

- Project Group. (2025). *Readability Analysis of the Toxigen Dataset* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/8_3_readability_toxigen.ipynb

- Project Group. (2025). *Comparative Readability Analysis: Toxigen, selfMA, SBIC, and workplaceMA* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/8_4_readability_toxigen_selfMA_sbic_selfMA_workplaceMA.ipynb

- Project Group. (2025). *Dataset Balancing Pipeline: selfMA + SBIC* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/9_1_dataset_balance_selfMA_SBIC.ipynb

- Project Group. (2025). *Dataset Balancing Pipeline: selfMA + Toxigen* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/9_2_dataset_balance_selfMA_toxigen.ipynb

- Project Group. (2025). *Dataset Loading and Integration Utilities* [Jupyter notebook]. GitHub. https://github.com/neeharika-kotte/w266_final_project_Carlos_Nura_Neeha/blob/main/Notebooks/9_3_datasets_load.ipynb

Papers

- Abu Farha, I., Oprea, S. V., Wilson, S., & Magdy, W. (2022). *SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic*. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022* (pp. 802–814). Association for Computational Linguistics. https://aclanthology.org/2022.semeval-1.111/

- Ali, O., Scheidt, N., Gegov, A. E., Haig, E., Adda, M., & Aziz, B. (2020). *Automated detection of racial microaggressions using machine learning*. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI* (pp. 2477–2484). IEEE. https://doi.org/10.1109/SSCI47803.2020.9308569

- Benítez-Andrades, J. A., González-Jiménez, Á., López-Brea, Á., Aveleira-Mata, J., Alija-Pérez, J.-M., & García-Ordás, M. T. (2022). *Detecting racism and xenophobia using deep learning models on Twitter data: CNN, LSTM and BERT*. *PeerJ Computer Science, 8*, e906. https://doi.org/10.7717/peerj-cs.906

- Bhat, M. M., Hosseini, S., Hassan, A., Bennett, P., & Li, W. (2021, November). *Say 'YES' to positivity: Detecting toxic language in workplace communications*. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2017–2029). https://aclanthology.org/2021.findings-emnlp.173/

- Breitfeller, L., Ahn, E., Jurgens, D., & Tsvetkov, Y. (2019). *Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP* (pp. 1664–1674). Association for Computational Linguistics. https://aclanthology.org/D19-1176/

- Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). *HateBERT: Retraining BERT for abusive language detection in English*. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021* (pp. 17–25). https://aclanthology.org/2021.woah-1.3/

- Cornett, K. E. (2024). *BCC'ing AI: Using modern natural language processing to detect micro and macro e-ggressions in workplace emails* (Master's thesis, Virginia Polytechnic Institute and State University). VTechWorks. https://vtechworks.lib.vt.edu/items/b55a8c11-f4de-4fe9-aef9-99a42f7df5b4

- El Miqdadi, I., Hourri, S., El Idrysy, F. Z., Hayati, A., Namir, Y., Nikolov, N. S., & Kharroubi, J. (2024). *Enhancing racism classification: An automatic multilingual data annotation system using self-training and CNN*. *Data Mining and Knowledge Discovery, 38*(6), 3805–3830. https://doi.org/10.1007/s10618-024-01059-2

- ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). *Latent hatred: A benchmark for understanding implicit hate speech*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 345–363). Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.29/

- Fersini, E., Gasparini, F., Rizzi, G., Saibene, A., Chulvi, B., Rosso, P., Lees, A., & Sorensen, J. (2022). *SemEval-2022 Task 5: Multimedia automatic misogyny identification*. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022* (pp. 533–549). Association for Computational Linguistics. https://aclanthology.org/2022.semeval-1.74/

- Gilda, S., Silva, M., Giovanini, L., & Oliveira, D. (2022). *Predicting different types of subtle toxicity in unhealthy online conversations*. *Procedia Computer Science, 198*, 360–366. https://doi.org/10.1016/j.procs.2021.12.254

- Gunturi, U. S., Kumar, A., & Rho, E. H. (2024). *Linguistically differentiating acts and recalls of racial microaggressions on social media*. *Proceedings of the ACM on Human-Computer Interaction, 8*(CSCW1), 1–36. https://dl.acm.org/doi/10.1145/3637366

- Guo, K., Hu, A., Mu, J., Shi, Z., Zhao, Z., Vishwamitra, N., & Hu, H. (2023). *An investigation of large language models for real-world hate speech detection*. In *2023 International Conference on Machine Learning and Applications (ICMLA* (pp. 1568–1573). IEEE. https://ieeexplore.ieee.org/abstract/document/10459901

- Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). *TOXIGEN: A large-scale machine-generated dataset for adversarial and implicit hate speech detection*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers*. Association for

Computational Linguistics. https://aclanthology.org/2022.acl-long.234/

- He, P., Liu, X., Gao, J., & Chen, W. (2020). *DeBERTa: Decoding-enhanced BERT with Disentangled Attention.* In *Proceedings of the International Conference on Learning Representations (ICLR 2021.* https://arxiv.org/abs/2006.03654
- Howard, P., Singer, G., Lal, V., Choi, Y., & Swayamdipta, S. (2022). *NeuroCounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation.* In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 5056–5072). Association for Computational Linguistics. https://aclanthology.org/2022.findings-emnlp.371/
- Khodak, M., Saunshi, N., & Vodrahalli, K. (2018). *A large self-annotated corpus for sarcasm.* In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018.* European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/pdf/160.pdf
- Kirk, H. R., Yin, W., & Röttger, P. (2023). *SemEval-2023 Task 10: Explainable detection of online sexism.* https://aclanthology.org/2023.semeval-1.305/
- Kumarage, T., Bhattacharjee, A., & Garland, J. (2024). *Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection.* https://arxiv.org/abs/2403.08035
- Ledalla, S., J., A., Sathwik E., A., Reddy M., S., & Kumar N., H. (2023). *Racism detection using deep learning techniques. E3S Web of Conferences, 391*, 01052. https://doi.org/10.1051/e3sconf/202339101052
- Lee, E., Rustam, F., Washington, P. B., El Barakaz, F., Aljedaani, W., & Ashraf, I. (2022). *Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble GCR-NN model. IEEE Access, 10*, 9717–9728. https://doi.org/10.1109/ACCESS.2022.3144266
- Mendelsohn, J., Le Bras, R., Choi, Y., & Sap, M. (2023). *From dogwhistles to bullhorns: Unveiling coded rhetoric with language models.* In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers* (pp. 15162–15180). Association for Computational Linguistics. https://aclanthology.org/2023.acl-long.845/
- Nadal, K. L., Griffin, K. E., Wong, Y., Hamit, S., & Rasmus, M. (2014). *The impact of racial microaggressions on mental health: Counseling implications for clients of color. Journal of Counseling & Development, 92*(1), 57–66. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1556-6676.2014.00130.x
- Nadeem, M., Bethke, A., & Reddy, S. (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* (pp. 5356–5371). Association for Computational Linguistics. https://aclanthology.org/2021.acl-long.416/
- Ocampo, N. B. (2025). *Unmasking implicit and subtle hate speech: NLP approaches for detecting and countering online harm* (PhD thesis, Université Côte d'Azur). HAL Open Archive. https://hal.science/tel-05247463/
- Ògúnremí, T., Sabri, N., Basile, V., & Caselli, T. (2021). *Leveraging bias in pre-trained word embeddings for unsupervised microaggression detection.* In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021.* https://journals.openedition.org/ijcol/1066
- Parihar, A. S., Thapa, S., & Mishra, S. (2021). *Hate speech detection using natural language processing: Applications and challenges.* In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI* (pp. 1302–1308). IEEE. https://doi.org/10.1109/ICOEI52431.2021.9452882

- Perez Almendros, C., & Camacho-Collados, J. (2024). *Do large language models understand mansplaining? Well, actually. . . .* In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024* (pp. 5235–5246). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.466/

- Pierce, C. (1970). *Offensive mechanisms*. In F. B. Barbour (Ed.), *The Black seventies* (pp. 265–282). Porter Sargent.

- Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). *HateCheck: Functional tests for hate speech detection models*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers* (pp. 43–61). Association for Computational Linguistics. https://aclanthology.org/2021.acl-long.4/

- Sasse, K., Aguirre, C. A., Cachola, I., Levy, S., & Dredze, M. (2025, July). *Making FETCH! Happen: Finding Emergent Dog Whistles Through Common Habitats*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers* (pp. 5687–5709). https://aclanthology.org/2025.acl-long.284/

- Sue, D. W., & Spanierman, L. (2020). *Microaggressions in everyday life*. John Wiley & Sons.

- Wang, S., Zhou, J., Sun, C., Ye, J., Gui, T., Zhang, Q., & Huang, X. (2022). *Causal intervention improves implicit sentiment analysis*. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022* (pp. 6966–6977). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.607.pdf

- Wang, Z., & Potts, C. (2019). *TalkDown: A corpus for condescension detection in context*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP* (pp. 3711–3719). Association for Computational Linguistics. https://arxiv.org/abs/1909.11272

- Weber, M., Huber, M., Auch, M., Döschl, A., Keller, M. E., & Mandl, P. (2025). *Digital Guardians: Can GPT-4, Perspective API, and Moderation API reliably detect hate speech in reader comments of German online newspapers?* https://arxiv.org/abs/2501.01256

- Wiegand, M., & Ruppenhofer, J. (2024, November). *Oddballs and Misfits: Detecting implicit abuse in which identity groups are depicted as deviating from the norm*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 2200–2218). https://aclanthology.org/2024.emnlp-main.132/

Datasets

- Abu Farha, I., Oprea, S. V., Wilson, S., & Magdy, W. (2022). *iSarcasmEval: Intended Sarcasm Detection in English and Arabic* [Data set]. GitHub. https://github.com/iabufarha/iSarcasmEval

- Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., & Granitzer, M. (2020). *AbuseEval v1.0: Implicit/explicit messages in offensive and abusive language* [Data set]. GitHub. https://github.com/tommasoc80/AbuseEval

- Conversation AI. (2022). *Unhealthy conversations dataset* [Data set]. GitHub. https://github.com/conversationai/unhealthy-conversations

- Curry, A., Tausczik, Y. R., & De Choudhury, M. (2021). *ConvAbuse: A dataset for understanding abusive language in online conversations* [Data set]. GitHub. https://github.com/amandacurry/convabuse

- Davidson, T. (2017). *Hate speech and offensive language dataset* [Data set]. Hugging Face. https://huggingface.co/datasets/tdavidson/hate_speech_offensive

- Derczynski, L. (n.d.). *Hate speech data* [Data set]. GitHub. Retrieved November 17, 2025, from https://github.com/leondz/hatespeechdata

- ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). *ImplicitHate* [Data set]. Hugging Face. https://huggingface.co/datasets/SALT-NLP/ImplicitHate

- Jigsaw. (2019). *Jigsaw unintended bias in toxicity classification* [Data set]. Kaggle. https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data

- Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Jr., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., Hoover, J., Azatian, A., Hussain, A., Lara, A., G., O., Al Omary, A., Park, C. G., Wang, C., Wang, X., Zhang, Y., & Dehghani, M. (2018). *The Gab hate corpus: A collection of 27k posts annotated for hate speech* [Data set]. Open Science Framework. https://osf.io/edua3/

- khanak27. (n.d.). *Microaggressions in the Workplace* [Data set]. Hugging Face. Retrieved November 17, 2025, from https://huggingface.co/spaces/khanak27/microaggressionsdetector/blob/main/micro_agg.csv

- Khodak, M., Saunshi, N., & Vodrahalli, K. (2018). *Self annotated Reddit corpus (SARC)* [Data set]. GitHub. https://github.com/NLPrinceton/SARC

- Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2020). *HateXplain: A benchmark dataset for explainable hate speech detection* [Data set]. GitHub. https://github.com/hate-alert/HateXplain

- Microaggressions. (n.d.). *Microaggressions* [Website]. Retrieved November 17, 2025, from http://www.microaggressions.com

- Nadeem, M., Bethke, A., & Reddy, S. (2021). *StereoSet* [Data set]. ACL Anthology. https://aclanthology.org/2021.acl-long.416/

- Ocampo, N. B. (2023). *ISHate: An in-depth analysis of implicit and subtle hate speech messages* [Data set]. Hugging Face. https://huggingface.co/datasets/BenjaminOcampo/ISHate

- Sap, M., Gabriel, S., Qin, L, Jurafsky, D., Smith, N. A., & Choi, Y. (2019). *Social bias frames* [Data set]. Hugging Face. https://huggingface.co/datasets/allenai/social_bias_frames

- SarcasmNet. (n.d.). *Sarcasm detection dataset* [Data set]. Hugging Face. Retrieved November 17, 2025, from https://huggingface.co/datasets/SarcasmNet/sarcasm

- Shang_49102. (n.d.). *Microaggression detection research at Areto Labs* [Web page]. Medium. Retrieved November 17, 2025, from https://medium.com/@shang_49102/microaggression-detection-research-at-areto-labs-6d608b215583

8 Appendix

8.1 SelfMA non-microaggression pair generation

We generate the non-microaggression pairs with the following prompt:
Transform the given microaggressive text into a non-microaggressive equivalent. The non-microaggressive equivalent should remove any bias, insensitivity, or invalidation present in the original text, while retaining the core positive or neutral sentiment if applicable. Do not repeat the original text in the answer. Do not use labels in the answer. Produce the non-microaggressive equivalent phrase as the answer.
Examples:

> **Microaggressive:** "You're very articulate for someone like you."
> **Non-Microaggressive:** "Your presentation was very articulate."
> **Microaggressive:** "Where are you really from?"
> **Non-Microaggressive:** "Where did you grow up?"
> **Microaggressive:** "You must be good at math since you're Asian."
> **Non-Microaggressive:** "You are really good at math."
> **Microaggressive:** "You're too pretty to be a software engineer."
> **Non-Microaggressive:** "You are pretty."
> **Microaggressive:** "Are you sure you

want to lead this project?"
Non-Microaggressive: "Can you lead this project?"
Microaggressive: "You're surprisingly smart for someone like you."
Non-Microaggressive: "You are really smart."
Microaggressive: "I didn't expect someone like you to be this good at coding."
Non-Microaggressive: "You are good at coding."
Microaggressive: "Are you the diversity hire?"
Non-Microaggressive: "Are you the new hire?"
Microaggressive: "You look too feminine to be a scientist."
Non-Microaggressive: "You are a scientist."
Microaggressive: "You're lucky you're cute."
Non-Microaggressive: "You're cute."
Microaggressive: "{microaggressive_text}"
Non-Microaggressive: ""

We show the non-microaggression pairs generation process in Figure 1.

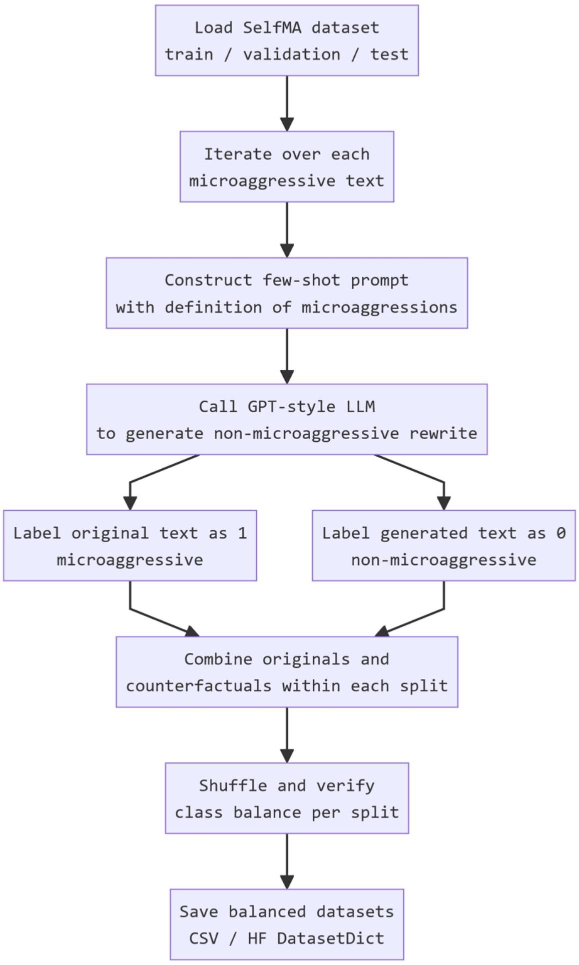

Figure 1: Counterfactual dataset balancing.

8.2 Exploratory Data Analysis on workplace MA and `selfMA_generated`

In our exploratory analysis of the Microaggressions in the Workplace (workplaceMA) corpus, we first confirmed that the dataset is small but reasonably balanced: it contains 171 entries with 87 labeled as non-microaggressive (label 0) and 84 as microaggressive (label 1). The schema is simple, with a short "speech" field and a binary label, and there are no missing values in either column. Text-length statistics indicate that workplaceMA examples are short and focused, with character lengths ranging from 17 to 66 (mean 37.7) and word counts from 3 to 13 (mean 6.7), consistent with brief workplace comments rather than long narratives. Microaggressive statements were on average 6.4 words long, while non-microaggressive statements were 6.9 words long. And since average word-lengths between microaggressions and non-microaggressions were so similar, the model likely could not use these basic markers as shortcuts to label the statements.

For the `selfMA_generated` dataset, we conducted analogous EDA to characterize the training distribution produced by our counterfactual balancing procedure. The resulting corpus is substantially larger and exactly balanced across splits, with 1,040 microaggressive (label 1) and 1,040 non-microaggressive (label 0) texts in the training set and 130 per class in both validation and test sets. By construction, each microaggressive selfMA incident is paired with a generated non-microaggressive rewrite that preserves topic and context, and our EDA confirms that this pairing is reflected in the lexical statistics: unigram and bigram analyses show that microaggressive examples contain more identity-and evaluation-related phrases (e.g., "gay people", "you're Black"), while the non-microaggressive counterparts reuse much of the same vocabulary but with the harmful framing removed. Readability analysis further indicates that both originals and counterfactuals remain short, relatively simple sentences (e.g., Flesch Reading Ease means of 85 for originals and 72 for generated texts, corresponding to easy-to-read prose), which is consistent with the short workplace utterances we ultimately evaluate on. Together, these EDA findings (charts shown in Figure 2) support our design choice: `selfMA_generated` provides a balanced, lexically diverse training set that is structurally similar to workplaceMA while focusing the label signal on the presence or absence of microaggressive content rather than on differences in length, domain, or style.

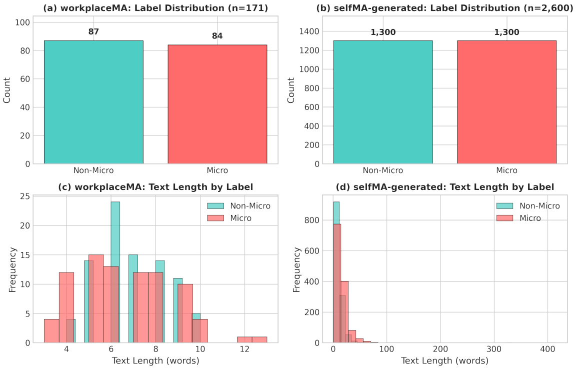

Figure 2: Exploratory data analysis on `selfMA_generated` and workplace MA.

8.3 Limitations of `selfMA_generated`

While our findings suggest that counterfactual balancing on selfMA is a promising approach for proxy-aware microaggression detection, several limitations point to directions for future work. First, both selfMA and workplace MA are relatively small and English-only, so a natural extension is to scale our evaluation to larger, noisier, and more diverse domains, including additional workplaces and online communities. Second, our assessment of LLM-generated counterfactuals is primarily indirect, via readability and proxy audits; a more rigorous evaluation would incorporate human judgments of fluency, naturalness, and the absence of residual microaggressive content. Third, although we explicitly examine non-semantic shortcuts, we do not yet perform systematic fairness or subgroup analyses—for example, by target group (race, gender, sexual orientation) or by type of microaggression—which would be essential for deployment in real-world settings.

8.4 Initial Balancing and Modeling Attempts: What Did Not Work

To address the class imbalance in selfMA, we explored balancing the dataset with other external datasets which were proven effective for hate speech and toxicity detection. One such approach was to combine microaggressions from selfMA with non-aggressive (normal text) entries from the Social Bias Inference Corpus (SBIC).

Additionally, we hypothesized that training models on related tasks, such as hate speech, sarcasm, indirectness, and/or subtlety would better capture the nuanced linguistic nature of microaggressions. To test this hypothesis, we utilized two additional datasets: ISHate for hate speech detection, referenced by Ocampo et al (2023), and iSarcasmEval, referenced by Abu Farha et al (2022), for sarcasm detection. ISHate includes both hate speech and non-hate speech drawn from seven different social media websites, where hate speech was labeled as explicit, implicit, subtle, and/or non-subtle. According to the authors of this dataset, implicit hate speech involves the use of figurative language, such as irony, metaphor, exaggeration, rhetorical question and sarcasm, such that hateful words are not directly used (“Are you sure Islam is a peaceful religion?). On the other hand, subtle speech is literal but indirect (“I’m either in North Florida or Nigeria; sometimes I can’t tell the difference”) (Ocampo et. al., 2023). iSarcasmEval is a collection of human-written and annotated sarcastic and non-sarcastic statements, where the statements had originally been posted on Twitter.

Using a sequential transfer learning approach with selfMA balanced with SBIC (training first on the ISHate dataset, followed by the iSarcasm Evaluation dataset, and then finally on this combined, balanced selfMA + SBIC dataset), we achieved an overall accuracy of 97.44%. However, when we performed model cross-evaluation on the workplace MA dataset, the resulting accuracy was only 52.05% (a significant drop of approximately 45 percentage points). This sharp decline in accuracy revealed a fundamental overfitting problem: the model was likely learning to distinguish unique, stylistic text patterns in selfMA compared to SBIC, rather than learning true microaggression detection.

We then attempted a three-class approach (non-aggressive, aggressive, and micro-aggressive statements) by balancing selfMA microaggressions with benign and toxic entries from the Toxigen dataset. However, this approach proved to be ineffective as well, as readability tests revealed that the Toxigen dataset was far more textually complex. For example, while selfMA had a Flesch-Kincaid Grade Level score of 3.8 (almost 4th grade reading level),

Experimental Approach	**Eval. acc.**	**Cross eval.**
ISHate → iSarc → selfMA	91.92%	77.19%
iSarc → selfMA	92.69%	75.44%

Table 4: Sequential training results.

Toxigen had a score of 8.0, more than four grade levels apart. Similarly, while the Toxigen dataset had a Gunning Fog Index of 10.2, selfMA had a score of 6.1, a difference of about four grade levels. Furthermore, the average sentence length for selfMA entries was 9.7 words, whereas for Toxigen it was 16.8. As before, this indicated that the model could exploit surface-level stylistic artifacts of the separate datasets, rather than semantic and linguistic differences between microaggressions, aggressions and non-aggressions.

We also explored sequential transfer learning with `selfMA_generated`. Specifically, we tested a three-stage approach (ISHate → iSarcasm → `selfMA_generated`) and a two-stage approach (iSarcasm → `selfMA_generated`). The BERT model included the following hyperparameters for training: batch_size = 16, num_epochs = 3, learning_rate = 2e-5. As shown in the results in Table 4, while these models do achieve strong accuracy on the `selfMA_generated` test split, both experimental approaches (three-stage and two-stage) also show the highest degradation in overall cross-evaluation accuracy with workplace MA. These results suggest that the benefits of sequential transfer learning from hate speech and sarcasm detection were minimal and perhaps contributed to unhelpful and excessive noise when the microaggressions-trained classifier was evaluated in a separate context.