Rape Culture Language Detection

Rebecca Bargiachi
University of California, Berkeley | Fall, 2024
Rebecca.Bargiachi@berkeley.edu

Abstract

This study explores the development of a machine learning model to identify language that perpetuates rape culture, focusing on nuanced textual cues such as victim-blaming, euphemistic language, and minimization of sexual violence. Using a dataset of over 4,800 unique sentences annotated for their perpetuation of rape culture, the study implements a baseline Bag of Words (BoW) model and iteratively improves performance through the application of a transformer-based model (BERT). To address the dataset's class imbalance, GPT-generated synthetic sentences were incorporated into the training data. While the addition of synthetic data improved precision in some cases, its impact on recall and F1-scores varied, particularly in the BERT-based model. The BERT model without synthetic data demonstrated the highest overall performance, achieving a validation F1-score of 0.82 and a test F1-score of 0.64 for the minority class. Error analysis revealed that all models struggled with implicit and context-dependent cues, such as euphemistic expressions and language that minimized perpetrators' responsibility. These findings highlight the potential of transformer-based models for identifying rape culture language at scale, while also underscoring the need for more robust datasets and further advancements in handling nuanced linguistic harms.

1. Introduction

Over half of women and transgender individuals in the United States experience sexual violence in their lifetime, with many survivors retraumatized by harmful myths and narratives that hinder justice and perpetuate stigma (Cooper et al., 2023). Rape culture is a societal framework characterized by attitudes, norms, and language that normalize, trivialize, or excuse sexual violence. Despite increased awareness, it remains deeply embedded in society and its effects are far-reaching, reinforcing systemic inequities and silencing survivors. Self-reported rates of sexual violence more than doubled from 2017 to 2018, yet reporting to law enforcement dropped from 40% to a mere 25% over the same period (National Sexual Violence Resource Center, 2018). This decline in reporting underscores the critical role of public discourse, media narratives, and cultural attitudes in shaping survivors' willingness to come forward.

Language, as both a reflection and reinforcer of societal norms, is at the heart of this issue. Researchers have found a significant increase in victim-blaming language in printed media over the past decade, highlighting the persistent and evolving ways in which rape culture manifests (Layman, 2020). This troubling trend demands innovative approaches to identify and counteract harmful rhetoric in both traditional and digital media landscapes.

This study aims to contribute to broader efforts to challenge harmful narratives, empower survivors, and promote more inclusive standards in reporting by developing a machine learning model capable of detecting language that reinforces rape culture. The model focuses on identifying specific linguistic patterns and rhetorical devices (e.g., victim-blaming, trivialization, euphemistic language, etc.) that perpetuate rape culture. The intent of this research is to contribute to potential interventions by providing a scalable and systematic approach to analyzing harmful language in sexual violence reporting.

2. Background

Rape Culture & Lingustics

Research on rape culture has identified various linguistic patterns that perpetuate harmful narratives, including victim-blaming, trivialization of assault, and excusing perpetrator behavior (Pritchard, 2014). Burt (1980) introduced the concept of rape myths, defined as "prejudicial, stereotyped, or false beliefs about rape, rape victims, and rapists" that foster a

hostile environment for survivors (Burt, 1980; Maiorano et al., 2023). These myths are often reinforced through media and everyday interactions.

Critical discourse analysis has revealed how euphemistic language and gendered stereotypes in media coverage obscure the severity of sexual violence and shift blame onto survivors (Aroustamian, 2020). For example, euphemisms like "had his way with her" or "took liberties" minimize the gravity of sexual assault, while focusing on a survivor's behavior rather than the perpetrator's actions fosters victim-blaming. This research underscores the need for a systematic approach to identifying such language. Despite growing urgency, there has been limited work in developing computational tools that can automatically detect language reinforcing rape culture.

NLP & Harmful Language Classification

Natural Language Processing (NLP) has been widely used to detect harmful language, including hate speech (Sun et al., 2019), online harassment (Van Royen et al., 2022), and gender bias (Sun et al., 2018). Schmidt and Wiegand (2017) provide a comprehensive survey of hate speech detection techniques, emphasizing the challenges of capturing implicit and context-dependent language. Some recent work has targeted specific aspects of rape culture-related language, such as Suvarna & Bhalla's use of LSTM- and CNN-based architectures to classify victim-blaming tweets following the #MeToo Movement. However, the use of transformer-based architectures to detect rape culture language remains largely unexplored.

Bag of Words and Early NLP Models

Traditional NLP approaches, such as Bag of Words models, have been effective for basic text classification tasks like sentiment analysis and topic modeling (Langen, 2024). Studies have shown that simple models incorporating character n-grams can achieve strong performance in text classification tasks (Li, 2020). However, BoW models are limited in their ability to detect nuanced language because they ignore the semantic and contextual relationships between words. This shortcoming is particularly problematic when identifying rape culture language, which often relies on implicit cues and subtle framing in addition to more overtly harmful expressions (Pritchard, 2014).

Contextual Language Models (BERT and Beyond)

Recent advancements in pre-trained language models, such as BERT have significantly improved NLP by capturing the contextual relationships between words (Devlin et al., 2019). These models excel at tasks requiring a deep understanding of linguistic nuance, such as hate speech detection. For instance, Jigsaw's "Perspective" tool leverages BERT-based architectures to analyze "toxicity" in online comments, demonstrating the potential for such models in addressing implicit harmful language.

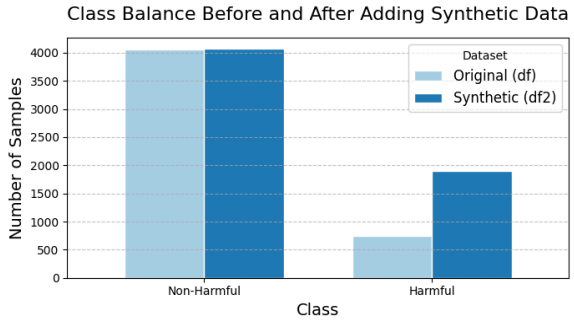
While prior research has focused on detecting overtly harmful language, such as hate speech or explicit victim-blaming, this study addresses a more nuanced problem: identifying both overt and subtle linguistic patterns that perpetuate rape culture. Furthermore, much of the prior research has focused on social media datasets. This study takes a novel approach by training NLP models on news articles reporting on sexual violence, providing a different and crucial perspective on how rape culture is perpetuated. By leveraging both traditional (BoW) and modern (BERT) approaches, I seek to fill a gap in the literature, offering insights into the effectiveness of computational tools for identifying and analyzing rape culture language in news media.

3. Data

The Rape Culture Language dataset is from a content analysis conducted by The Representation Project's research team. The dataset consists of 210 news articles about sexual violence from 2000-2024, parsed into sentences. A team of researchers achieved interrater reliability on a small codebook of variables and hand-coded each sentence for the presence or absence of language that perpetuated rape culture (0-No, 1-Yes). Language that perpetuates rape culture might blame the survivor for the assault, excuse the perpetrator's behavior, minimize the severity of sexual violence, reinforces gender stereotypes and power imbalances, and/or silence survivors and discourage reporting.

To prepare the data for analysis, I first pre-processed the text for the baseline BoW model. This involved converting all sentences to lowercase for compatibility and removing duplicate sentences to avoid inflating the representation of specific language patterns. After preprocessing, the dataset contained 4,803 unique sentences.

An exploratory analysis revealed a substantial class imbalance, with significantly fewer sentences classified as harmful (1) compared to non-harmful (0). To address this imbalance, I generated synthetic sentences classified as harmful using GPT prompts. Prompt engineering took multiple iterations in order to circumvent OpenAI's policy around using terms related to rape and sexual violence. In order to generate sentences demonstrating six different categories of rape culture language, I split the queries up into six separate prompts, generating 1,200 new sentences in total. I then added them to the training dataset after the original data had been split into training, validation, and test sets. Importantly, GPT-generated data was only used in the training set for the "enhanced" models to preserve the integrity of validation and test set performance metrics.



4. Methods

Baseline: Bag of Words

To establish a baseline, I implemented a BoW model, which is a widely used approach in text classification tasks and serves as a valuable starting point for understanding fundamental patterns in textual data (Langen, 2024). I used logistic regression as the classifier, with a focus on key performance metrics: accuracy, precision, recall, and F1-score. Given the imbalanced nature of the dataset, the F1-score was prioritized as the primary evaluation metric, as it balances precision and recall. The use of accuracy as a supplementary metric provides a general sense of overall model performance but is not sufficient alone for imbalanced datasets.

Transformer-Based Model

Next I chose to use BERT due to its track-record for understanding contextual information, which is critical for nuanced tasks like identifying language that perpetuates rape culture.

The BERT model was fine-tuned using the Hugging Face Transformers library. I initialized the model with the pre-trained bert-base-cased checkpoint and set it up for a binary classification task. Training was performed using the AdamW optimizer with a learning rate of 2x10^-5 and a linear learning rate scheduler. Hyperparameters included a batch size of 16 for both training and evaluation, a weight decay of 0.01 to reduce overfitting, and a warm-up ratio of 0.1 for learning rate stabilization. I trained the model for 3 epochs, saving the best-performing model based on F1-score using early stopping. Dropout layers with a rate of 0.1 were applied to reduce overfitting as well. Like the BoW model, both BERT models were evaluated using accuracy, precision, recall, and F1-score on both the validation and test datasets.

5. Results

The results of my analysis are summarized in the table comparing the baseline BoW model, the BoW model with the enhanced dataset, the BERT model, and the BERT model with enhanced data.

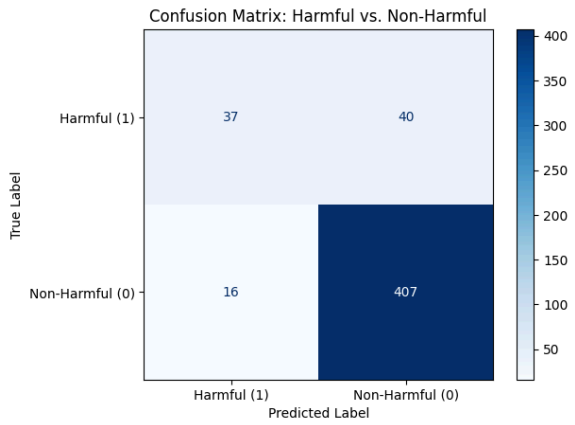
Metric	BoW	BoW + Synth. Data	BERT	BERT + Synt. Data
Val Accuracy	91%	91%	94.2%	88.6%
Val F1	0.72	0.66	0.82	0.69
Test Accuracy	87%	89%	89.8%	88.4%
Test F1	0.58	0.56	0.64	0.52
Test Recall	0.58	0.45	0.60	0.42
Test Precision	0.58	0.71	0.70	0.71

The baseline Bag of Words model achieved a validation accuracy of 91% and a test accuracy of 87%. Precision and recall for harmful language detection (Class 1) were balanced, but relatively low at 0.58. This highlights the model's difficulty in detecting harmful language, which is likely due to the lack of context and semantic understanding inherent in BoW methods.

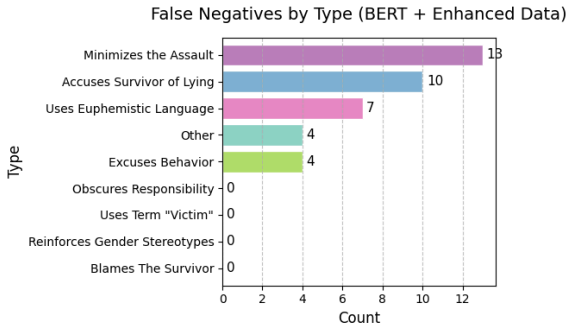
Adding synthetic data to the training set slightly improved the baseline model's test accuracy to 89% and increased precision for Class 1 to 0.71. However, recall for Class 1 dropped to 0.45, resulting in a lower F1-score of 0.56. This indicates that while the enhanced dataset improved the model's ability to correctly identify harmful sentences, it struggled to recall all instances of harmful language. This suggests a trade-off between precision and recall introduced by the synthetic data.

The BERT model achieved the highest validation accuracy (94.2%) and the best overall test accuracy (89.8%). For the harmful class, it outperformed the BoW models with an F1-score of 0.64 and a recall of 0.60, indicating a better ability to detect nuance. However, at .70 the model's precision for Class 1 was only marginally better than the BoW model with synthetic data.

Introducing synthetic data to the BERT model led to a noticeable and surprising drop in performance. Test accuracy declined to 88.4%, and the F1-score dropped to 0.52. Precision slightly increased to 0.71, but recall for Class 1 decreased to 0.42.



The confusion matrix shows a greater number of false negatives compared to the BERT model with unenhanced data, suggesting that while synthetic data improved precision, it disrupted the model's ability to generalize across the broader dataset.



As shown in the visualization of false negatives by type, the majority of errors occurred for sentences that minimized the assault (13 false negatives), accused survivors of lying (10 false negatives), or used euphemistic language (7 false negatives). These categories especially highlight the nuanced nature of harmful language and the challenges faced by all models in capturing these subtleties. Below are examples of false negatives belonging to these three sub-types, categorized as either “too nuanced” (where the model failed to recognize implicit harm) or “mislabeled” (suggesting possible over-labeling in the training data or ambiguity in definitions).

Accuses Survivor of Lying

Sentences in this category often involve skepticism of survivors’ accounts, accusations of false reporting, or language that undermines credibility. The model struggled to differentiate between nuanced harmful language and legitimate reporting of false claims. Sentences such as “Credibility problems also surface in rape reports” subtly perpetuate the rape myth that false accusations are common without saying it explicitly enough for the model to pick up on it.

Too Nuanced:

- "Credibility problems also surface in rape reports from neighborhoods frequented by prostitutes and crack addicts."
- "...Brooks said: 'For the last two months, I have been hassled to no end with threats, lies, and tragic tales of what my future

would be if I did not write a check for many millions of dollars.'"
- "But public defender Lisa Wayne, who has represented about 20 men accused of rape, contends that many rape reports are wrong."

Potentially Mislabeled:

- "Kim Pursley, 30, pleaded no contest earlier this month to a misdemeanor charge of false reporting."
- "The police later learned that the girl had lied to cover her tracks."

Minimizes the Assault

This category includes language that downplays the severity of sexual violence. The model often failed to classify sentences where harm was described indirectly or in euphemistic terms. In one example, the language describes the legal definition of rape while avoiding the term altogether.

Too Nuanced:

- "According to an arrest warrant, Artis had sex with the student against her will in February 2016 when he should have known she was 'mentally incapacitated and physically helpless.'"
- "Even the marching band was hit with unflattering news this week."

Potentially Mislabeled:

- "...Charges have been dismissed against a North Carolina football player for misdemeanor sexual battery and assault on a female student."
- "The move comes in the wake of the sexual abuse scandal involving former sports doctor Larry Nassar."

Uses Euphemistic Language

This category includes the use of softened language to describe sexual violence, such as “allegations,” “accused,” or other terms that downplay the seriousness of the assault." Interestingly, the model often misclassified sentences with euphemistic language as non-harmful. This highlights a broader limitation in detecting linguistic patterns that rely on implicit tone or framing rather than explicit statements. Additionally, the presence of certain legal jargon in sentences, such as “the suit alleges,” may have led the model to incorrectly categorize them as non-harmful, despite their broader context reinforcing rape culture.

Too Nuanced:

- "This publication contacted Civeo for comment, but the company declined because of the nature of the allegations."
- "The suit alleges sexual assault, battery and gender violence."

Potentially Mislabeled:

- "Lisa Wayne defends accused rapists."

The error analysis suggests that the model’s reliance on explicit patterns might have contributed to its failure to detect implicit or context-dependent harms across all three of these categories. The analysis also revealed that some false negatives seem to result from labeling issues with the dataset. For example, sentences with direct quotes that perpetuate rape culture might also subtly cue to the reader that the quoted party’s take is problematic. Despite improving precision in some areas, the introduction of synthetic data may have introduced noise, potentially affecting the model’s ability to generalize to more nuanced harmful language.

6. Conclusions

This study demonstrates both the promise and challenges of applying machine learning to identify language that perpetuates rape culture. The results reveal that while BERT achieved better performance in capturing context-dependent cues, particularly on the validation set, the inclusion of synthetic data did not yield consistent improvements across all metrics. For instance, while synthetic data improved precision in the BoW models, it had mixed effects on recall and F1-scores for BERT, particularly in identifying nuanced forms of harmful language. Error analysis highlighted the difficulty of detecting implicit harms like victim-blaming, minimization, and euphemistic

language, further illustrating the limitations of models in addressing deeply contextual linguistic phenomena. While the synthetic data improved recall in certain cases by exposing the model to more harmful examples, it also introduced variability that may have confused the model even more during training.

Future research should prioritize several areas to build on this work. While transformer-based models have the potential to augment human efforts to flag harmful content, their effectiveness relies on robust and nuanced training data. There is a need for larger, more diverse datasets that can train models to generalize well across various contexts and linguistic styles. Efforts should include annotating sentences for nuanced subtypes of harm, which could improve the models' ability to distinguish subtle harmful language. When it comes to the use of GPT-generated synthetic data, careful curation and weighting strategies are necessary to avoid potential trade-offs in model precision. Experimenting with more advanced architectures, such as domain-adapted or task-specific versions of BERT, could enhance a model's ability to detect rape culture language.

Ultimately, this work serves as a foundation for leveraging machine learning to address the pervasive issue of rape culture in media and society. These findings carry broader implications for automated content moderation and awareness-building initiatives. In contexts such as news reporting or online moderation, these tools could serve as a first line of defense, flagging potentially harmful language for review. By refining these tools, researchers, practitioners, and activists can better identify and challenge harmful media narratives and contribute to cultural change.

References:

1. Aroustamian, C. (2020). Time's up: Recognizing sexual violence as a public policy issue: A qualitative content analysis of sexual violence cases and the media. *Aggression and Violent Behavior, 50*, 1359–1789. https://doi.org/[DOI pending].
2. Burt M. R. (1980). Cultural myths and supports for rape. Journal of Personality and Social Psychology, 38(2), 217. 10.1037/0022-3514.38.2.217
3. Cooper, R., & Heldman, C. (2023). Ten Rape Myths in Media: A Quantitative Analysis of Sexual Violence in Film & TV. *The Representation Project.*
4. Khalid, O., & Srinivasan, P. (2020). Style matters! Investigating linguistic style in online communities. *Proceedings of the International AAAI Conference on Web and Social Media, 14*, 360–369.
5. Katherin Layman (2020). The Representation of Rape and Sexual Assault Within News Media, Senior Thesis. *Portland State University.*
6. Langen, H. (2024). The impact of the #MeToo movement on language at court: A text-based causal inference approach. *PLoS ONE, 19*(5), e0302827. https://doi.org/10.1371/journal.pone.0302827.
7. Li, T. (2020). An empirical evaluation of text classification using character n-grams. Retrieved from https://arxiv.org/pdf/2312.11504.
8. Lovell, R., Gruber, A., & Johnson, T. (2023). Using machine learning to assess rape reports: Sentiment analysis detection of officers' "signaling" about victims' credibility. *National Criminal Justice Reference Service.* Retrieved from https://www.ojp.gov/ncjrs/virtual-library/abstracts/using-machine-learning-assess-rape-reports-sentiment-analysis.
9. Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2018). Gender bias in sentiment analysis. Retrieved from https://arxiv.org/pdf/1807.11714.
10. Maiorano, N., Travers, Á., & Vallières, F. (2023). The relationship between rape myths, revictimization by law enforcement, and well-being for victims of sexual assault. *Violence Against Women, 29*(14), 2873-2890. https://doi.org/10.1177/10778012231196056.

11. National Sexual Violence Resource Center. (2018). Statistics: In Depth. Retrieved from https://www.nsvrc.org/statistics/statistics-depth.
12. Pritchard, A. (2014). Changing Conversations About Sexual Assault. *Grand Valley State University.* Retrieved from https://www.gvsu.edu/cms4/asset/903124DF-BD7F-3286-FE3330AA44F994DE/changing_conversations_about_sexual_assault.pdf.
13. Rho, E.H.R., Mark, G., & Mazmanian, M. (2018). Fostering civil discourse online: Linguistic behavior in comments of #MeToo articles across political perspectives. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW), 1–28. https://doi.org/10.1145/3274416.
14. Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection Using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media,* 1–10. Retrieved from https://aclanthology.org/W17-1101.pdf.
15. Stevens, H., Acic, I., & Taylor, L.D. (2021). Uncivil Reactions to Sexual Assault Online: Linguistic Features of News Reports Predict Discourse Incivility. *Cyberpsychology, Behavior, and Social Networking, 24*(12), 815-821. https://doi.org/10.1089/cyber.2021.0075.
16. Sun, T., Gaut, A., Tang, S., Huang, Y., Wei, J., & Qian, S. (2019). Mitigating Gender Bias in Natural Language Processing: A Case Study of Corpus Selection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 1536–1546. Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1159/.
17. Suvarna, A., & Bhalla, G. (2020). #NotAWhore! A computational linguistic perspective of rape culture and victimization on social media. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop,* 328-334.
18. Van Royen, K., Milosevic, T., & Davis, B. (2022). Artificial Intelligence to Address Cyberbullying, Harassment and Abuse: New Directions in the Midst of Complexity. *International Journal of Bullying Prevention, 4*(1), 1-5. https://doi.org/10.1007/s42380-022-00117-x.