8-2024

Offensive Content Detection in Online Social Platforms

Ebuka Okpala
eokpala@clemson.edu

Offensive Content Detection in Online Social Platforms

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Computer Science

---

by
Ebuka Johnbosco Okpala
August 2024

Accepted by:
Dr. Long Cheng, Committee Chair
Dr. Feng Luo
Dr. Nianyi Li
Dr. Matthew Costello

Abstract

Online social platforms enable users to connect with large, diverse audiences and the ability for a message or content to flow from one user to another user, user to followers, followers to user, and followers to followers. Of course, the advantages of this are apparent, and the dangers are also clearly obvious. The user-generated content could be abusive, offensive, or hateful to other users, possibly leading to adverse health effects or offline harm. As more of society's public discourse and interaction move online and these platforms grow and increase their reach, it is inherently important to protect the safety of the users of these platforms. Platforms ensure safety by enforcing rules on the type of content allowed that, when violated, could lead to a warning, user suspension or the removal of the user-generated content before or after the content is published. Monitoring and removing policy-violating content is labor and resource-intensive. Recently, the growth of machine learning, specifically deep learning-based natural language processing, has made it possible to detect offensive content and flag it for review automatically. The automatic detection of offensive content is non-trivial because of its subjectivity, as what is considered offensive in one country is not in another, its nuances, and the constant evolution of public discussions around political or social issues across different cultures. Detecting and understanding offensive content during political or social issues offers an understanding of how platforms can improve safety and the dynamics of offensive content in public discourse. However, of equal importance is the fairness of the deep learning systems used in detecting offensive content, which can propagate bias in their training datasets and could lead to unequal treatment of minority groups by the platform. The study of offensive content intersects

social science and is becoming an emerging socially relevant cybersecurity problem. The works investigated in this dissertation are composed of our efforts to enable healthy online discourse by detecting and analyzing offensive content, understanding and mitigating bias in deep learning-based offensive content detection models, and developing teaching materials and an experiential learning lab to engage students in AI-cyberharassment education.

With regards to the detection and analysis of offensive content, the first study conducts a large-scale analysis of the emotions expressed, the extent of offensive discussions, and the role of emotions in offensive discussions; as anger is the emotion of epistemic injustice, the characteristics of users who generated offensive content and users who received offensive content, and the topics discussed in the Black Lives Matter (BLM) related discussions on social media after the death of George Floyd in 2020, and the protests that followed. To examine offensive language and emotion, we first develop a classifier that uses sentiment representation to aid offensive language detection. We then develop an emotion classifier based on deep attention fusion with sentiment features to classify emotions. The offensive and emotion classifiers were used to detect offensive content and classify emotions in over 20 million tweets. Finally, topic modeling was used to analyze the topics of the offensive and no-offensive tweets.

Regarding bias in offensive content detection models, in the second study, we looked at how offensive language datasets contain bias that offensive content detection models propagate. When these models classify tweets written in African American English (AAE), they predict AAE tweets as a negative class at a higher rate than tweets written in Standard American English (SAE). This study assessed bias in language models fine-tuned for offensive content detection and the effectiveness of adversarial learning in reducing such bias. We introduce AAEBERT, a pre-trained language model for African American English obtained by re-training BERT-base on AAE tweets. The representation of tweets from AAEBERT is fused with the representation of tweets from the offensive content classifier and used as input to an adversarial network to perform debiasing. We then compared the effects of adversarial debiasing in language models before and after debiasing.

Artificial intelligence (AI) is becoming increasingly popular and is being used to complete tasks in our daily activities. The third work extends the second work by exploring the implications of using large language models (LLMs), such as the version of the generative pre-training (GPT) model, GPT-4, in annotating offensive language datasets used in fine-tuning downstream models for detecting offensive content. We used different prompting techniques to annotate several offensive language datasets and fine-tuned models on the LLM-annotated datasets. Then, we assess racial bias towards AAE tweets in the models fine-tuned on LLM-annotated datasets compared to models fine-tuned on human-annotated offensive language datasets, and the rate of false positives in the models fine-tuned on LLM-annotated datasets towards AAE tweets. We also explore whether using dialect priming in the prompt techniques explored helps reduce racial bias in LLM annotation of offensive language datasets.

Finally, the popularity of AI calls for creating an AI-ready workforce across academic disciplines and professions. Most AI education research focuses on developing curricula for computing and engineering students while paying little attention to non-computing students. In the fourth work, given the interdisciplinary nature of this emerging social cybersecurity problem, engaging non-computing students without prior knowledge of AI in AI can be challenging. We take the first step to develop educational materials and a hands-on lab that introduces AI to non-computing students and how AI can be used for socially relevant cybersecurity, like offensive content detection.

Dedication

This dissertation is dedicated to my parents - Chief Ojukwu Raphael Okpala and Mrs Akuada Rose Okpala, and my siblings - Chinyere, Nonye, Obiageliaku, Nkechi, Chidozie, Izundu, Kelechi, and Amaka for their unwavering support and encouragement.

And to the memory of Oduraa Quartey and Abena Ofori, whose Ph.D. journey at Clemson University was cut short in the Summer of 2023.

Acknowledgments

I want to thank my advisor, Dr. Long Cheng, for his support, guidance, and encouragement throughout my Ph.D. journey. I would also like to thank my committee members, Dr. Feng Luo, Dr. Nianyi Li, and Dr. Matthew Costello, for their comments and suggestions.

I also want to express my sincere gratitude to Clemson Online for supporting and funding me throughout my Ph.D. studies at Clemson University. To my previous managers at Clemson Online, Anne Marie Rogers and David Bassett, thank you for your patience, encouragement, support, guidance, and reassurance.

Finally, I would like to thank Mr. Louis Lacio and Mrs LaNiece Lacio for their words of encouragement, support, and warm welcome when I arrived at Clemson. To my friends Dr. Seun Oti-Aina, Gugu Selela, Patricia Ng'ethe, Kehinde Elelu, Damilola Aiyetigbo, Nachiappan Chockalingam, Shivam Pandit, Chidi Igbelina, Dr. Simeon Babatunde, Dr. Prosper Anyidoho, Dr. Emmanuel Adjei, Dr. Camilius Amevorku, Nkemjikanma Ohanyere, Chibuzor Frank Obi, Henrique Ferreira, and Marco Soto thank you all for making this path memorable.

Table of Contents

Title Page . . . . . **i**

Abstract . . . . . **ii**

Dedication . . . . . **v**

Acknowledgments . . . . . **vi**

List of Tables . . . . . **ix**

List of Figures . . . . . **xii**

1 Introduction **1**
1.1 Background 1
1.2 Research Contributions 3
1.3 Dissertation Outline 6

2 Literature Review **8**

3 Analyzing Offensive Content and Emotional Dynamics in Black Lives Matter Discourse on Twitter **12**
3.1 Abstract 12
3.2 Introduction 13
3.3 Related Work 15
3.4 Methodology 16
3.5 Results and Discussion 34
3.6 Implications 50
3.7 Limitations 51
3.8 Conclusions and Future Work 52

4 AAEBERT: Debiasing BERT-based Hate Speech Detection Models via Adversarial Learning **54**
4.1 Abstract 54
4.2 Introduction 55
4.3 Related Work 57
4.4 Methodology 58
4.5 Experiments 62

4.6 Results . . . 65
4.7 Conclusion . . . 69

5 Large Language Model Annotation Bias in Hate Speech Detection . . . 71
5.1 Abstract . . . 71
5.2 Introduction . . . 72
5.3 Related Work . . . 75
5.4 Methodology . . . 77
5.5 Results . . . 89
5.6 Broader Perspectives . . . 99
5.7 Limitations . . . 100
5.8 Conclusions and Future Work . . . 100

6 AI-Cybersecurity Education Through Designing AI-based Cyberharassment Detection Lab . . . 114
6.1 Abstract . . . 114
6.2 Introduction . . . 115
6.3 Related Work . . . 118
6.4 Design & Development of AI Socially Relevant Cyberharassment Lab . . . 120
6.5 Methods . . . 126
6.6 Results and Discussion . . . 127
6.7 Limitations . . . 136
6.8 Conclusion and Future Work . . . 137
6.9 Acknowledgment . . . 137

7 Conclusion and Future Work . . . 138
7.1 Conclusions . . . 138
7.2 Future Recommendations . . . 140

References . . . 141

List of Tables

3.1 Performance of the three language models used in sentiment classification. . 22
3.2 Performance of the sentiment model for the negative and positive classes. Evaluation metrics are macro averages. 22
3.3 Performance of the offensive detection model for the offensive and non-offensive classes. Evaluation metrics are macro averages. 24
3.4 Performance of the offensive detection model compared to state-of-the-art methods. 27
3.5 Model generalization results after training on a specific dataset and testing on another dataset. The evaluation metric shown is macro F1 score. 28
3.6 Performance of our emotion model on 11 emotions. F1-macro: 55.8%, F1-micro: 68.70%. 31
3.7 Statistics of predicted offensive and non-offensive tweets. 34
3.8 Statistics of the offensive reply network. 37
3.9 The topics discovered by topic modeling in the 2020 offensive and non-offensive tweets without the representative tokens in the topics. 41
3.10 The topics discovered by topic modeling and the most representative tokens in the 2020 offensive tweets. 42
3.11 The topics discovered by topic modeling and the most representative tokens in the 2020 non-offensive tweets. 43
3.12 The topics discovered by topic modeling and the most representative tokens in the 2021 offensive tweets. 46
3.13 The topics discovered by topic modeling and the most representative tokens in the 2021 non-offensive tweets. 47
3.14 The topics discovered by topic modeling in the 2021 and 2022 offensive tweets without the representative tokens in the topics. The highlighted topics are some of the topics in the 2021 offensive tweets that persisted in 2022. 48
3.15 The topics discovered by topic modeling in the 2021 and 2022 non-offensive tweets without the representative tokens in the topics. The highlighted topics are some of the topics in the 2021 non-offensive tweets that persisted in 2022. After 2020, topics related to Floyd, Breonna, and Riots/Protests are still being discussed. 48
3.16 The topics discovered by topic modeling and the most representative tokens in the 2022 offensive tweets. 49
3.17 The topics discovered by topic modeling and the most representative tokens in the 2022 non-offensive tweets. 50

4.1 Datasets used in our work . . . 62
4.2 Evaluation results of fine-tuned models on each hate speech dataset without applying adversarial debiasing . . . 66
4.3 Racial bias analysis of fine-tuned BERT models. Showing result with and without adversarial debiasing . . . 67
4.4 Racial bias analysis of fine-tuned BERTweet models. Showing result with and without adversarial debiasing . . . 68
4.5 Racial bias analysis of fine-tuned HateBERT models. Showing result with and without adversarial debiasing . . . 68

5.1 Statistics of the datasets. . . . 81
5.2 Performance of the dialect classification model for the AAE and non-AAE (SAE) classes. Evaluation metrics are macro averages. . . . 84
5.3 Performance of the dialect classification model on the dataset of users who self reported their race/ethnicity (AA and White). Evaluation metrics are macro averages. . . . 84
5.4 Annotation prompt samples from the Davidson and Golbeck datasets for the three prompting strategies. . . . 102
5.5 Classifier performance after fine-tuning on each GPT-annotated dataset with multi-class labels for each prompting strategy. Evaluation metrics are macro averages. . . . 103
5.6 Classifier performance after fine-tuning on each GPT-annotated dataset with binary labels for each prompting strategy. Evaluation metrics are macro averages. . . . 104
5.7 Classifier performance after fine-tuning on each human-annotated dataset with multi-class labels for each prompting strategy. Evaluation metrics are macro averages. . . . 105
5.8 Classifier performance after fine-tuning on each human-annotated dataset with multi-class labels for each prompting strategy. Evaluation metrics are macro averages. . . . 106
5.9 Performance of human vs GPT-4 general prompt annotation. Evaluation metrics are macro averages. . . . 106
5.10 Performance of human vs GPT-4 general prompt annotation with datasets conditioned on dialect. Evaluation metrics are macro averages. . . . 107
5.11 Performance of human vs GPT-4 few-shot learning prompt annotation. Evaluation metrics are macro averages. . . . 107
5.12 Performance of human vs GPT-4 few-shot learning prompt annotation with datasets conditioned on dialect. Evaluation metrics are macro averages. . . . 107
5.13 Performance of human vs GPT-4 chain-of-thought prompt annotation. Evaluation metrics are macro averages. . . . 108
5.14 Performance of human vs GPT-4 chain-of-thought prompt annotation with datasets conditioned on dialect. Evaluation metrics are macro averages. . . . 108
5.15 Racial bias analysis of fine-tuned BERT model on human annotated datasets (left) and GPT-4 (right) annotated datasets using general prompt annotation. 108

5.16 Racial bias analysis of fine-tuned BERT model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right). 109
5.17 Racial bias analysis of fine-tuned BERTweet model on human annotated datasets (left) and GPT-4 (right) annotated datasets using general prompt annotation. 109
5.18 Racial bias analysis of fine-tuned BERTweet model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right). 110
5.19 Racial bias analysis of fine-tuned HateBERT model on human annotated datasets (left) and GPT-4 (right) annotated datasets using general prompt annotation. 110
5.20 Racial bias analysis of fine-tuned HateBERT model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right). 111
5.21 Racial bias analysis of fine-tuned BERT model on GPT-4 annotated datasets using general prompt annotation with dialect priming. 111
5.22 Racial bias analysis of fine-tuned BERT model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right) with dialect priming. 111
5.23 Racial bias analysis of fine-tuned BERTweet model on GPT-4 annotated datasets using general prompt annotation with dialect priming. 112
5.24 Racial bias analysis of fine-tuned BERTweet model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right) with dialect priming. 112
5.25 Racial bias analysis of fine-tuned HateBERT model on GPT-4 annotated datasets using general prompt annotation with dialect priming. 112
5.26 Racial bias analysis of fine-tuned HateBERT model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right) with dialect priming. 113

6.1 Demographics of students who participated in the Social Statistics 1 course in the Spring and Fall 2022 semesters. 125
6.2 Spring 2022 semester survey results. 127
6.3 Fall 2022 semester survey results. 128

List of Figures

3.1 Annotation decision tree used in data labeling. 20
3.2 An overview of the proposed offensive language classification model with fused representation of the input text from the sentiment and offensive models. 23
3.3 An overview of the proposed Emotion classification model with deep attention fusion. FC indicates a fully connected layer. 29
3.4 The model precision of the topic models. Higher is better. The red line within the boxplot represents the median. 33
3.5 (A) Daily tweet count in log scale. (B) Temporal evolution by emotions for offensive tweets in log-scale. (C) Temporal evolution by emotions for non-offensive tweets in log-scale (D) Emotion distribution of 2020 tweets. (E) Emotion distribution of 2021 tweets. (F) Emotion distribution of 2022 tweets. The temporal evolution of emotions is based on smoothed weekly averages for visual clarity. Anger and disgust correlate with a Pearson correlation score of 0.99 (p-value $\ll 0.0001$). The emotion distribution shows a significant increase in anger, disgust, and fear in 2020. The gray vertical lines signify points with significant changes in the number of daily tweets and emotional distribution. 35
3.6 Emotion dynamics of offenders. (A) Temporal evolution by emotions for offensive tweets in log-scale. (B) Temporal evolution by emotions for non-offensive tweets in log-scale (C) Emotion distribution of 2020 tweets. (D) Emotion distribution of 2021 tweets. (E) Emotion distribution of 2022 tweets. The temporal evolution of emotions is based on smoothed weekly averages for visual clarity. Anger and disgust correlate with a Pearson correlation score of 0.99 (p-value $\ll 0.0001$). The gray vertical lines signify points with significant changes in emotional distribution. 39
3.7 Emotion dynamics of recipients. (A) Temporal evolution by emotions for offensive tweets in log-scale. (B) Temporal evolution by emotions for non-offensive tweets in log-scale (C) Emotion distribution of 2020 tweets. (D) Emotion distribution of 2021 tweets. (E) Emotion distribution of 2022 tweets. The temporal evolution of emotions is based on smoothed weekly averages for visual clarity. Anger and disgust correlate with a Pearson correlation score of 0.99 (p-value $\ll 0.0001$). The gray vertical lines signify points with significant changes in emotional distribution. 40

4.1 Illustration of the proposed debiasing architecture 59

5.1 An overview of the white-aligned and black-aligned datasets creation from the race dataset. 79
5.2 Confusion matrix of human annotation and GPT-4 general prompt annotation on full (training and testing) datasets. 90
5.3 Dialect-wise results for BPSN AUC on general prompt annotated test datasets. 95
5.4 Dialect-wise results for BPSN AUC using few-shot annotation. 96
5.5 Dialect-wise results for BPSN AUC using chain-of-thought annotation. . . . 97

6.1 A screenshot illustrating the lab interface 122
6.2 A screenshot illustrating Lab 1 instruction manual 123
6.3 The visual aid for model explanation in Lab 1 124
6.4 One example of a lab activity 124
6.5 One example of Lab 1's discussion questions 125

Chapter 1

Introduction

1.1 Background

What is offensive language and how does it differ from hate speech? The problem with offensive language and its related concepts such as abuse [36], hate speech [43, 50, 113], toxicity [166, 233], aggression [109], and cyberbullying [40, 236] is that they are often difficult to differentiate. These related concepts are specific types of offensive language [239]. Despite there relatedness, multiple definitions have been used in the literature. While there is no formal definition of hate speech, there is a general consensus that it is any "*communication that disparages a target group of people based on some characteristic such as color, race, ethnicity, gender, sexual orientation, nationality, religion or other characteristic*" [45]. Even though hate speech is protected under the First Amendment in the United States, it is prohibited by law in countries such as the United Kingdom, France, and Canada. Due to increased criticism from users that online platforms are too permissive and to comply with laws in the national jurisdictions in which they operate. Social media companies such as Twitter and Facebook have invested in infrastructures such as algorithms to remove content, the ability for users to flag content for reviews by human moderators, and have developed policies to guide the use of their platforms in promoting content that attacks people based on protected characteristics such as race, ethnicity, gender, and sexual orientation.

Hate speech have also been defined differently in the literature [43, 61, 183]. Davidson *et al.* [43] defined hate speech as "*language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group*". Including the intention to cause harm or to incite violence. Salminen *et al.* [183] define online hate as "*the use of language that contains either hate speech targeted toward individuals or groups, profanity, offensive language, or toxicity - in other words, comments that are rude, disrespectful, and can result in negative online and offline consequences for the individual, community, and society at large*".

Offensive language has been defined as posts that include "insults, threats, and posts containing any form of untargeted profanity" [239]. Posts containing targeted (individual, group or others i.e an organization, a situation, an event or an issue) profanity are also considered offensive. In addressing remaining issues in [239], [36] defines abusive language as "*hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions*". Hate speech, derogatory language, profanity, toxic comments, racists and sexist statements are included in this definition.

Toxic comment is defined in [166, 233] as a "rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion".

Going by the definitions in [43, 233, 239], one can begin to see the similarities, yet the differences, that make the separation of these phenomenons difficult and the extent of their subjectivity. This difficulty stems from differing opinions on what can be considered offensive or hate as people have different tolerance levels, lived experiences, and political views [75, 108]. A rude, disrespectful, or unreasonable comment can also be derogatory or humiliating irrespective of it being targeted or not. The similarities and differences between these phenomena have been studied in the literature [225, 230]. In [225], they argue that offensive language can be reduced into two factors:

- Is the language directed towards a specific individual or entity or a generalized group?

- Is the content explicit or implicit?

Definitions based on these two factors may be a better way to better differentiate between various related phenomenons (hate speech, offensive language, cyberbullying) and reconcile definitions and achieve a consensus in the community [36].

Social media platforms are increasingly used to share opinions and engage in discussions online [94]. While these engagements have their benefits, the topics and contents of discussions can become offensive, hateful or abusive due to the anonymity that some platforms provide. Surveys have indicated that 40% of internet users have experienced online abuse, with members of minority groups targeted more often [53, 83, 115]. To help mitigate this problem and encourage healthy online discussions, the Natural Language Processing (NLP) community has employed machine learning (ML) and deep learning (DL) approaches to enable the automatic identification of various forms of offensive language to facilitate content moderation as the amount of data (over 500 million) shared per day is manually impossible to track and moderate [91, 98, 152].

1.2 Research Contributions

The offensive language literature has focused on designing classifiers to detect offensive content. Little attention has been paid to the analysis of offensive content to understand the nature of offensive language on social media. More specifically, little work has been done in characterizing offensive language during large-scale global events such as the COVID-19 pandemic and social movements such as #GamerGate [38], #MeToo [143], and #BlackLivesMatter [63] seeking social justice in a range of social and political concerns and the issue of racial bias in offensive content detection models. The global COVID-19 pandemic in 2020 and the following lockdown drove the rise of COVID-19-related discussions on social media. The discussions, which turned political, were also fueled with misinformation about the origins of the virus, leading to attacks both online and offline on people of Asian origin, especially people of Chinese descent. Due to the prevalence of hateful and offensive

language towards Asians, researchers studied the evolution of anti-Asian hate [250], level of aggression, target, and type of hate speech [149], dynamics of hate speech and hateful communities [210], hate-related keywords [218], and offensive content analysis and their targets [124].

In this work, we first perform a large-scale analysis of offensive content during the Black Lives Matter (BLM) movement in 2020. In the summer of 2020, massive protests erupted around the world after the death of George Floyd. Protesters marched in solidarity with the BLM movement, demanding justice for the death of George Floyd and an end to systemic injustice. While the BLM movement is well-researched in the literature, very few works have studied the nature of offensive content in the BLM movement, especially during the global protests in 2020 and beyond on social media. The death of George Floyd and the protests that followed triggered an increase in discussion about BLM and issues of racial injustice on social media. To contribute to the literature on online safety, fairness and justice through the lens of social movements, we conduct a large-scale study on the prevalence of offensive content and the emotional dynamics in BLM-related discussions on Twitter during the BLM global movement in 2020. As anger is the emotion of epistemic injustice and the fact that people express different emotions during political protests [17, 213], we study the effect of emotions on offensive content in the BLM movement. We develop a BERT-based model with sentiment representation fusion to detect and analyze offensive content. To analyze emotions, we develop a BERT-based model with deep attention fusion with sentiment representations to classify the emotions expressed in offensive and non-offensive content. We created an offensive reply graph and analyzed the nature of the relationship between offensive content authors and the receivers of offensive content. Using a topic model BERTopic [77], we qualitatively analyze the important topics discussed in the offensive and non-offensive BLM-related discussions and the persistence of topics beyond 2020.

Second, researchers have recently shown that ML models exhibit unfairness or bias in its predictions. As the quality of the learned model depends on the quality of the training data, if a model is learned using poor data, the learned model will be poor. Since

most datasets are human-generated and labeled by humans, human biases can lead to biased labels due to human stereotypical bias based on sensitive attribute such as race or gender. Another source of bias is skewed training data distribution. In most cases, the learned model will perform less accurately for the underrepresented group of people or users than the general population or perform unfairly due to the dependency between some sensitive or protected attributes in the data and the class label. As ML models are used in real world to detect offensive language used in targeting groups or individuals belonging to protected categories, if these models are biased, they will not perform effectively in protecting the groups they were designed to protect, defeating its goal. The second work focuses on racial bias in hate speech and offensive language detection datasets, assessed bias in offensive language detection models based on large language models (LLMs), specifically, bidirectional encoder representations from transformers (BERT) [47] based models. We fine-tuned BERT-based models on each of the Twitter datasets analyzed, used a dataset of African American English (AAE) and Standard American English (SAE) to compare how the fine-tuned models performed on each race dataset, and used a hypothesis based method to calculate the percentage of tweets assigned to each class. To mitigate bias, we introduce AAEBERT, a retrained BERT language model and used the AAE representation of tweets from AAEBERT in an adversarial learning setting to debias the fine-tuned models. Results indicate that the fine-tuned models assigns tweets written in AAE to the negative classes at a higher rate than tweets written in SAE and that adversarial learning is effective in reducing racial bias with a trade-off in classification performance.

Third, larger LLMs, such as the generative pre-trained (GPT) models like ChatGPT, GPT-3, GPT-4, and LLaMA, have significantly improved performance in various NLP tasks. Due to their performance, they are used in various ways, such as in data annotation tasks, a costly and time-consuming process in the ML pipeline. In the offensive language detection task, using larger LLMs for data annotation is attractive due to the health implications of annotation offensive content may have on human annotators [216]. While using larger LLMs for annotating offensive language datasets may be beneficial, the downstream implications

of using LLM-annotated data in downstream models may propagate racial bias. In this work, we extend the second work by analyzing the effects of using larger LLMs, specifically GPT-4, in annotating offensive language datasets. We perform extensive experiments using three prompting strategies (general, few-shot learning, and chain-of-thought reasoning) to annotate seven offensive language datasets. Then we fine-tuned three models (BERT, BERTweet, and HateBERT) on each of the datasets and assessed racial bias towards AAE and SAE tweets using a hypothesis-based metric to determine the rate of racial bias and an AUC-based metric to determine the rate at which each model reduces false positives.

Finally, the need to develop an AI-ready workforce has been discussed by global organizations and governments and has been recognized as a topic that should be explored across disciplines [136, 153]. In a report published by the National Academics of Sciences, Engineering, and Medicine (NASEM) [147], the importance of learning AI by students of all backgrounds, disciplines and professional goals was highlighted, indicating the need to educate across disciplines. Most AI education research has focused on developing curricula for computing and engineering students, focusing less on humanities students. We fill this gap by developing experiential learning education materials to engage students in AI-driven, socially relevant cybersecurity curricular modules. Our educational materials which includes a hands-on lab introduces non-computing students to AI and how AI can be leveraged to solve social issues such as the detection of offensive content.

1.3 Dissertation Outline

Chapter 2 consists of a consolidated literature review on offensive language detection and efforts in assessing and mitigating racial bias in offensive language detection models. Chapter 3 presents the study on analyzing offensive content and the emotional dynamics in BLM-related discussions. Chapter 4 presents the results of assessing racial bias in offensive language detection models and mitigating racial bias using adversarial debiasing. Chapter 5 studies the implications of using larger LLMs, such as GPT-4, in annotating datasets for

offensive language detection. Chapter 6 introduces our educational materials designed to introduce non-computing students to artificial intelligence (AI) and how AI can be used for socially relevant cybersecurity. Chapter 7 presents the conclusion of this study and recommendations for future directions.

Chapter 2

Literature Review

A consolidated literature review of this study is presented in this chapter, summarizing works in offensive language detection on social platforms and racial bias and fairness in offensive language detection models. Each of the studies in Chapters 3, 4, 5, and 6 have their literature reviews, respectively.

Offensive content detection on online social platforms has recently been studied extensively. Researchers have introduced varying definitions [43, 141, 152, 233, 241], developed datasets annotated to tackle different aspects of offensive content from the binary classification [20, 71, 239], multi-class classification [36, 43, 62, 224, 226], to the detection of targets and types of offensive content [241] and created competition for the detection of offensive language [39, 241]. The different machine learning and deep learning algorithms and features utilized in offensive content detection in a text have been well summarized [15, 61]. As offensive content can occur in images and text and in videos and text, multimodal offensive content detection datasets [41, 72, 99, 181] have been introduced. Multimodal methods [72, 99, 181, 238, 247] developed as well as challenges [52] designed to encourage the development of new methods to tackle this multi-dimensional problem. These previous works have primarily focused on developing models to detect various forms of offensive content. The understanding and the characterization of the dynamics of offensive content [235, 242, 243], their targets [124, 194, 219] and the authors who disseminate them have

also been explored [107, 196]. While these works have offered insights into offensive content on online social platforms, little attention has been paid to understanding offensive content during social or global movements and the issues of fairness, equity and society in offensive content detection models.

Fairness and justice are essential aspects of AI and ethics research [229]. Social movements have always been a part of social justice and fairness, as recognizing collective injustice motivates participation in social movements, collective protests, and political rebellions [209]. In the analysis of the #GamerGate [133] controversy, which turned into a social justice, sexism, and feminism issue, [38] collected a dataset of 1.6M tweets written by 340K users and analyzed the difference between these users, their tweets, and that of random Twitter users. They find that GamerGaters are more engaged than random users and that their tweets are aggressive, hateful and less joyful. The #MeToo movement, which was an offline movement created by a woman of color [78, 157] to enable the safe discussion of sexual harassment and violence, has been criticized for under-representing the voices of women of color and the contributions to feminist movement [120, 143, 159, 231]. In a large-scale analysis of the #MeToo movement on Twitter, [143] examined the users who tweeted about the movement using topic modeling, the stories they shared and how it differed across different inferred demographic groups (gender, race, and ethnicity) and intersectionality, and the dynamics of the hashtags across demographic groups. They find that when compared to other demographics, white women authored more tweets and were overrepresented, matching the criticism of unequal representation. They also found that tweets by black women were emotionally supportive and critical of the difference in treatment in the justice system and the police. Recently, the #StopAsianHate hashtag has been used in the Stop Asian Hate movement following the COVID-19 pandemic, which may have contributed to the increase in hate towards Asians [116, 177, 237]. This led to studies analyzing offensive content towards Asians during the COVID-19 pandemic [124, 138].

Social movements played a critical role in anti-discriminatory practices and a broad range of political and social concerns [206]. As previous works have indicated above, online

social platforms have given minority communities the power to express their struggles and quickly disseminate their call for social justice. In promoting social justice, online social platforms mustn't use unfair or biased offensive content detection models to moderate content. While content moderation is pertinent, it must be performed cautiously. Research has shown that offensive content detection models are biased towards African-American English (AAE), a variation of Standard English common in the African-American community. Researchers in [42] show that offensive content detection models trained on biased hate speech detection datasets inherit and propagate the racial bias in the datasets by assigning tweets written in AAE into negative classes (hate, offensive, abuse, etc.) at a higher rate than tweets written in Standard American English (SAE). Different methods have been developed to mitigate this racial bias towards AAE in offensive content detection models based on traditional deep learning architectures such as LSTM and more recently pre-trained language models based on the transformer-based architectures [214] such as BERT [46]. Using a two-phased training approach in a model consisting of an LSTM with an attention mechanism encoder, a multilayer perceptron binary classifier, and a multilayer perceptron adversarial network, [234] reduced the false positive rate for AAE tweets. A regularization-based technique based on a re-weighting mechanism was used in [142] to reduce the rate at which a BERT-based model classifies AAE tweets into negative classes. While previous works [42, 142, 234] have studied the racial bias towards AAE tweets, they do not perform extensive experiments on multiple classifiers, datasets and might not extend to transformer based architectures. The work of [42] focused only on one traditional ML classifier (regularized logistic regression). While [234] used an adversarial network in a two-phased training of an LSTM classifier, they did not explore its effect on transformed-based architectures. While [142] explored racial bias in a transformer-based architecture (BERT_base), their regularization-based method that depends on re-weighting samples using high-frequency 2-grams might not scale to higher n-grams such as 3-grams. As AI systems become widespread in the case of even bigger large language models (LLMs) such as ChatGPT [130], researchers have explored the effectiveness of ChatGPT-like models

in detecting offensive content [79, 82, 104, 217], its ability to generate offensive content [66, 189], debiasing such models [189], and for annotating offensive content datasets [88], dialect prejudice [85], and the risk of using LLMs for annotation [204].

With the ubiquitousness of artificial intelligence (AI) in our society today, AI literacy is needed to better equip students and future researchers with the skills and knowledge to prepare for a changing and new workforce needs [31, 33, 95, 127, 131, 199, 201]. Research in AI education have focused on the creation of general AI curriculum for kindergarten, middle school, high school, and university students [95], undergraduate and graduate students [], and the integration of AI curriculum across all disciplines in a university [197] with little attention being paid to AI-based socially-relevant cybersecurity education such as offensive content detection.

Chapter 3

Analyzing Offensive Content and Emotional Dynamics in Black Lives Matter Discourse on Twitter

This work has been accepted at the International AAAI Conference on Web and Social Media (ICWSM), 2025.

3.1 Abstract

The Black Lives Matter (BLM) movement seeks to spread awareness and fight against social and racial injustice. In 2020, BLM-related discussions surged on social media after the death of George Floyd and the protests that followed. Previous works have qualitatively analyzed the scaling, dynamics, and topics of BLM discussions on social media. However, very few works have studied the offensive content, the emotions expressed, and the topics of offensive discussions in BLM-related discussions. In this measurement study, to examine offensive language and emotion, we conduct a large-scale study of BLM discussions on Twitter. We first develop a classifier that uses sentiment representation to aid offensive language detection. We then develop an emotion classifier based on deep atten-

tion fusion with sentiment features to classify emotions. We further use topic modeling to analyze the topics of offensive tweets. Our analysis of over 20 million tweets revealed that offensive tweets peeked in the weeks following George Floyd's death and rapidly decreased but remained stable. The analysis further revealed that negative emotions were the most expressed emotions. Offensive reply network analysis reveals that most offensive replies are unidirectional. Our contribution in this work is five-fold: (1) We identify offensive content during BLM protests; (2) we identify online emotions that were significant in the offensive and non-offensive content during the protests; (3) we assess the characteristics of users who replied offensively and those who are the recipients of offensive content; (4) we assess emotion dynamics across offenders and recipients; (5) we identify the hot topics that most drove the offensive content on Twitter. Our work offers important implications for content moderation and the conscious and unconscious attitudes towards the black/African American community.

3.2 Introduction

Digital tools such as social media platforms have significantly increased the number of online discussions among users, particularly around topics related to social and political issues. Black Lives Matter (BLM) is an activist organization that seeks to raise awareness of racial injustice and police brutality [203] and utilized social media as an essential tool in broadening the organization's impact dating back to July 2013 when the hashtag "#Blacklivesmatter" was created on Twitter by Black Lives Matter activist founders. At the time of BLM's creation, the use of "#Blacklivesmatter" in discussions was low until it spiked in the fall of 2014 due to its use in the context of the Ferguson, Missouri protests after the shooting of Michael Brown [63]. A similar rise in BLM movement-related discussions was observed after the killing of George Floyd by a Minneapolis police officer on May 25th, 2020 [12]. George Floyd's death initiated large protests organized by BLM, leading to discussions about George Floyd's death, police brutality and racism, and other related events such as

the death of Breonna Taylor and Ahmaud Arbery [151].

While the movement has drawn researchers to study its different aspects [89, 167, 208] few attempts have been made to study offensive language in BLM discussions. Less is known about the content of offensive language and what topics were discussed in these contents. Examining offensive content in the BLM movement is critical to effecting change through content moderation. It can encourage individuals, especially those affected by the issues the movement seeks to highlight and their allies, to engage in healthy online conversations about the movement. Studies have revealed that both the physical and psychological health of individuals can be affected by police brutality, especially among Black Americans, as shown in the high levels of depression among Black Americans after Floyd's death was widely shared on social media [56]. The findings relate to the findings that offensive language has adverse health effects that can lead to suicide [84].

Online emotions may have played a significant role in the rise of offensive content on Twitter during the BLM protests. Sociology and political science research suggest that emotions play an essential role in social movements and protests [213**?**]. On emotions of protests, protesters experience negative emotions such as anger and fear when interacting with opponents and positive emotions such as joy when interacting with other activists in the movement [213]. People can experience emotions without being directly confronted by the triggering situation [213]. Due to the role of emotions in protests, we investigate the emotional dynamics of offensive and non-offensive content on Twitter during the BLM protest.

Distinguishing from existing works, our paper presents the first study analyzing offensive content in the BLM online movement on Twitter. We aim to answer the following research questions.

- **RQ1:** What was the extent of offensive content during the 2020 BLM movement and the years (2021 and 2022) after? Did offensive content increase during the 2020 movement, and was it sustained after? What is the nature of the relationship between offensive content authors and the recipients of offensive content?

- **RQ2:** What emotions were expressed during these periods, and how does the emotion of offensive tweets differ from the non-offensive tweets? How does emotions vary across the authors and recipients of offensive content?

- **RQ3:** What are the offensive and non-offensive topics discussed in 2020 during the BLM protests sustained in 2021 and 2022?

3.3 Related Work

Emotion is one of the most complex affective concepts and is defined as reactions attributed to stimuli (response to situational events) [244]. Emotions play a significant role in online behavior, as found in the dynamics of retweets [102, 200] and online consumption [179]. Social media users primarily rely on affective rather than cognitive information processing, as suggested by psychology research [128], making emotions essential drivers of online behavior [30, 92, 148, 179]. Motivated by previous studies, we expect that unique emotional dynamics characterize the BLM protests and online discussions.

In the past, very few works have attempted to analyze offensive content in online social movements, particularly the Black Lives Matter social movement. The work of [110] is close to our work regarding offensive content detection in the BLM movement on social media. They use deep-learning models to classify collected tweets into hate and non-hate classes. Other works have used classical machine learning [61], and deep learning techniques [16], to study offensive content on social platforms. Recently, large-scale pre-trained language models have been used in offensive language detection. In SemEval-2018 Task 6, subtask A category, Liu *et al.* [126] obtained first place by fine-tuning the BERT [46] model. Hate speech and offensive language [7, 10, 11, 101, 149, 210] were intensively studied during the COVID-19 pandemic. Following the COVID-19 outbreak, Liao *et al.* [124] investigated the ebb and flow of offensive tweets and their targets. Schild *et al.* [190] studied Sinophobic content on Twitter and 4chan. Vishwamitra *et al.* [218] used a BERT attention model to discover hate-related keywords, and Li *et al.* [123] developed COVID-HateBERT,

a domain-specific model for COVID-19 hate detection.

Researchers have also considered sentiment features by including these as features in supervised machine learning and deep learning approaches [191]. For instance, a multitask framework is developed in [174] that uses emotions to inform and improve abusive language detection. To our knowledge, this paper is the first time large-scale topic modeling, emotion analysis, and user and network analysis of offensive tweets and users have been conducted in BLM-related discussions. Our analysis has revealed new insights into the topics discussed in the offensive tweets, and the emotions expressed, and the nature of users in BLM-related discussions. Previous work in the BLM movement has focused on a variety of themes, including the dynamics of user behavior [89], #AllLivesMatter [64], the scaling of the movement [144], the resurgence of Anonymous during BLM protests [93], the common and different topics in BLM and Stop Asian Hate movements [208], the social media engagement in the movement over time and the relationship between online engagement and offline activities [44], and the analysis of sentiment and emotions during the movement [60, 168].

3.4 Methodology

We aim to understand offensive language during BLM-related online social movements and protests using machine learning models and computational methods. The qualitative analysis performed with these models and methods provides details on offensive language in online social movements and helps answer our RQs. Our process is detailed below.

3.4.1 Data Collection

To identify offensive content during the BLM-related online social movements and protests, we first collected a large sample of BLM-related tweets.

We collected three years (2020, 2021, and 2022) of English public tweets contained the hashtags and keywords #BLM, #BlackLivesMatter, #AtlantaProtests, #KenoshaProtest,

#MinneapolisProtest, #ChangeTheSystem, #JusticeForGeorgeFloyd, #GeorgeFloyd, #Floyd, #BreonnaTaylor, #JusticeForBreonnaTaylor, #Breonna, #JusticeForJacobBlake, #JacobBlake, #JusticeForAhmaud, #AhmaudArbery, #Ahmaud, Black Lives Matter, George Floyd, Breonna Taylor, and Ahmaud Arbery in both lowercase and uppercase. Data retrieval was from May 1, 2020, to December 31, 2020, from January 1, 2021, to December 31, 2021, and from January 1, 2022, to October 27, 2022. These hashtags and keywords were chosen after surveying media reports covering the movement and protests. Part of the hashtags was also selected from a previous work that introduced TweetBLM, a Black Lives Matter-related hate speech dataset [110]. After post-processing by removing tweets with less than four words and removing duplicates, our dataset consists of 21,596,115 tweets, 16,592,382 tweets in 2020, 3,615,913 tweets in 2021, and 1,387,820 tweets in 2022. Data was collected using Twarc[1], a Python library for retrieving and archiving Twitter JSON data via the Twitter API.

3.4.2 Data Annotation

We aim to identify offensive tweets in our dataset using an offensive language classifier. To train the classifier, we annotated a random sample of our dataset, given the dataset size, annotation cost, and the few occurrences of offensive language as the proportion of offensive tweets are generally low [191]. Therefore, before post-processing, we sampled 100,000 tweets from the 21M tweets and used Google's Perspective API[2] to identify potentially offensive tweets to annotate. Researchers have used Perspective API to detect toxic comments in YouTube [156], to understand behaviors of toxic account on Reddit [107], and to filter potentially offensive tweets in COVID-19 dataset [124]. Following [124], we use the Perspective API to filter potentially offensive tweets for labeling. Perspective API assigns a toxicity probability score between 0 and 1 to a text, with higher values indicating high perceived toxicity. A threshold of 0.7 was used to filter the sampled tweets after assigning a toxicity score to each tweet using Perspective. The Perspective API suggests using a

[1]https://twarc-project.readthedocs.io/en/latest/
[2]https://perspectiveapi.com/

threshold value between 0.7 - 0.95 to filter potentially toxic content. We decided to use the lower value of that range (0.7) as the threshold because the use of 0.9 and 0.8 produced a small number of potentially toxic tweets (442 and 1612, respectively).

We note that the Perspective API was not used in our work to determine the final offensiveness of tweets. Instead, it was used to select potentially offensive tweets, which were relabeled and used to train our offensive model. Offensive language detection models have been shown to propagate bias, especially racial bias, in the training data they were trained on [158]. In particular, Perspective API is biased towards African American English (AAE) [186]. Due to the potential bias of the Perspective API [186, 188, 228], relabeling tweets helps mitigate bias. While we do not provide dialect or race priming to our data annotators [186], annotators were informed to consider context and the possible race/ethnicity of a tweet's author during annotation. Specifically, all annotators were made aware by the most knowledgeable annotator familiar with African American English (AAE) that some lexical markers of AAE are reclaimed offensive slurs that are used safely and are not particularly offensive [188].

We selected tweets with a score greater than or equal to the threshold, obtaining 3,492 tweets. We trained a sentiment classification model discussed in Section 3.4.4 and used this model to classify each of the 3,492 tweets into two classes - positive and negative sentiment. Then, we selected 2,482 tweets classified as negative as the potentially offensive tweets that were annotated using our annotation guideline and offensive language definition.

In the offensive language detection literature, researchers have adopted different definitions of offensive language [61]. These definitions can be attributed to the similarity between offensive/abusive language and other related tasks - hate speech, toxicity, abuse, cyberbullying, aggression, etc. In this work, we group these commonalities into a single umbrella, offensive language, and use our labeling strategy to capture the differences. The comprehensive definition in [36] considers the similarity between offensive language and other related phenomena such as hate speech and abusive language and context to understand offensive content. We adopt the definition and thus define offensive language

as: *"language that insults or offends or attacks a person or group based on their social or personal characteristics such as race, sexual orientation, gender, national origin, religion, disability, occupational status, opinions, statements, or actions"*. Additionally, *"language that promotes/incites violence, harass with/without racial epithet, or expresses inferiority is considered offensive"*. Context is instrumental in determining offensive content because a text can be offensive without having offensive words. In this case, we ensure the content is directed to a person or group before labeling a text offensive. Texts that do not belong in our definition or contain offensive words not directed to a person, a group, or other (i.e., an organization, an event, an issue, or a situation) [239] are considered non-offensive. In line with this definition, a tweet is labeled as one of two categories: offensive or non-offensive.

Three internal annotators and one of the authors of this paper, who are native English speakers, labeled the 2,482 tweets in three stages. They were given our definition and example tweets used to explain the definition in detail further. They were instructed to pay attention to the context before labeling a tweet as offensive, even in the presence of a particular word, as it does not indicate that a tweet is offensive. In the first stage, the four annotators labeled 100 tweets, and a Fleiss' Kappa = 0.43 was measured, indicating moderate agreement [48, 114]. Then, they discussed the annotations, modified the guideline accordingly, and used the revised guideline to re-label the 100 tweets with a Fleiss' Kappa = 0.65, indicating a substantial agreement [36, 114]. The modified guideline is formulated as a decision tree as shown in Fig 3.1 and is used in stages 2 and 3; If a tweet contains an offensive word and explicitly refers to a person, group, or other, and the tweet simply expresses emotion (e.g., @USER I miss you bitch!!!), as often done in social media, it is labeled as non-offensive. Otherwise, it is labeled offensive. If the tweet does not explicitly refer to a person, group, or other, and a person, group, or other can be easily inferred through context, it is labeled offensive. Otherwise, it is labeled non-offensive. If a tweet does not contain an offensive word but is offensive because it is implied (i.e., implicit), it is labeled offensive.

In stage 2, all annotators labeled a new batch of 100 tweets, and a Fleiss' Kappa

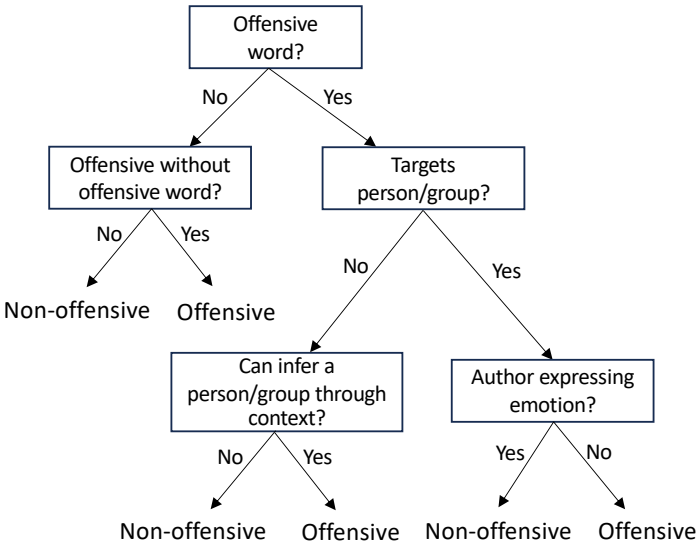


Figure 3.1: Annotation decision tree used in data labeling.

= 0.66 was measured. In stage 3, annotators labeled a set of 2,282 tweets, and the inter-annotator agreement score provided by Fleiss' kappa is 0.4, a moderate agreement. The results of the annotation are in line with similar work [48], achieved a moderate agreement strength [114], and demonstrate the difficulty in annotating offensive content due to high level of subjectivity. Additionally, recent work has argued that a low agreement score does not necessarily imply poor-quality annotation [119].

We used a majority decision to assign a final label to a tweet. When there is a tie, the most knowledgeable annotator in offensive language, one of this paper's authors, breaks the tie. Our annotated dataset consists of 2,465 tweets after removing duplicates and tweets of less than four words.

3.4.3 Ethical Statement

We reflect on the ethical and privacy implications of our work due to the sensitive nature of this study. First, when data was collected, no deleted, protected, or suspended accounts were included in our dataset as we adhered to Twitter's Standard API terms and Conditions (Twitter, Inc, 2022). Furthermore, before our analyses (in 2023), we performed a non-compliance (e.g., deleted tweets or from suspended accounts) check of tweet IDs to

ensure they are still compliant with Twitter rules. All non-compliant tweets are excluded from our analyses. Second, example tweet quotations have been modified to protect the identity of the original author. Some content is offensive and sensitive, so readers should read content cautiously. Lastly, following abusive research guidelines [216], researchers and data annotators were encouraged to pace themselves, take frequent breaks, and were provided a sense of purpose to prevent mental health and emotional problems that could arise through vicarious trauma[3]. Our institution's instutional review board (IRB) approved this study.

3.4.4 Sentiment Classification

Having identified potentially offensive tweets after using the Perspective API to score each of the randomly sampled 100K tweets from the 21M tweets in our dataset, we further filtered the 3,492 tweets that Perspective identified as potentially offensive. To do this, we used pre-trained language models and fine-tuned them on the Twitter Sentiment140 dataset [70]. The dataset comprises 1.6 million tweets labeled into three categories; positive, neutral, and negative sentiment. We randomly split the dataset in a 90:10 ratio to obtain the train ($n = 1400000$) and test ($n = 160000$) sets. We formulated the sentiment classification task as a binary classification task by dropping the neutral class. We experimented with three pre-trained language models: BERT [46], DistilBERT [185], and BERTweet [150] (vinai/betweet-base on HuggingFace) for sentiment classification. These language models are fine-tuned by replacing their pre-training head with a randomly initialized classification head. Then fine-tuning is performed by training the models on classification examples while minimizing the cross-entropy loss to learn the randomly initialized parameters. Before fine-tuning, we pre-processed the dataset by replacing web links with URL, user mentions with @USER, numbers with NUMBER, removing the # sign contained in hashtags, and removing platform-specific tokens like "RT" (retweets on Twitter). We trained the models on the dataset using Adam optimizer, a learning rate of $10^{-}5$, five epochs, and a batch

[3]https://firstdraftnews.org/wp-content/uploads/2017/04/vicarioustrauma.pdf

Language - Model	F_1	Precision	Recall
BERT	0.870	0.870	0.870
DistilBERT	0.863	0.863	0.863
BERTweet	**0.888**	**0.888**	**0.888**

Table 3.1: Performance of the three language models used in sentiment classification.

size of 256. The BERT, DistilBERT, and BERTweet models obtained 0.872, 0.866, and 0.889 F1 scores, respectively. Hence, we used the BERTweet [150](vinai/betweet-base on HuggingFace) model to classify the potentially offensive tweets identified by Perspective and all the 21M tweets in our dataset. Our fine-tuned BERTweet model is slightly better than the fine-tuned BERT [46] model (0.87) used in [168] and is competitive when compared to the adapted BERTweet + SVM model (0.905) used in [19] in terms of F1 scores as shown in Table 3.1. The per class performance of the sentiment model is shown in Table 3.2.

Target	F1	Precision	Recall
Negative	0.890	0.899	0.882
Positive	0.889	0.880	0.897

Table 3.2: Performance of the sentiment model for the negative and positive classes. Evaluation metrics are macro averages.

3.4.5 Detecting Offensive Language

To detect offensive language during BLM-related events and protests (in answer to RQ1), we used the 2,465 tweets we annotated to identify offensive content, of which 1,110 were non-offensive, and 1,355 were offensive. We randomly split the 2,465 tweets in a 90:10 ratio to obtain the train ($n = 2218$) and test ($n = 247$) sets used in training and testing our offensive language detection model. The train split contained ($n = 1215$) offensive tweets

and ($n = 1003$) non-offensive tweets. In contrast, the test split included ($n = 140$) offensive tweets and ($n = 107$) non-offensive tweets.

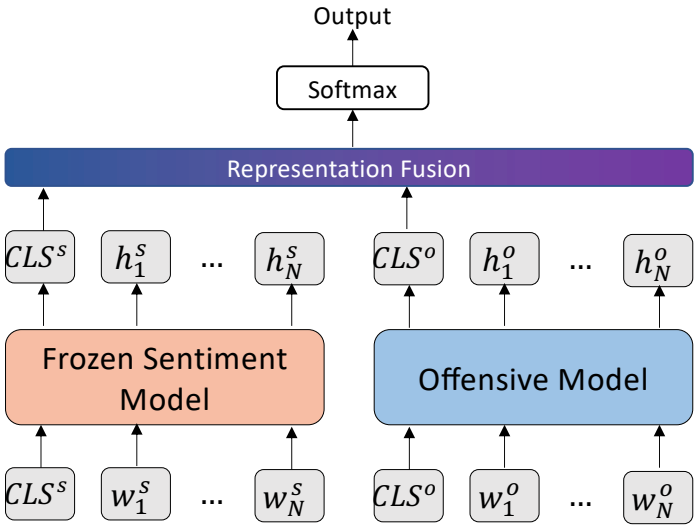


Figure 3.2: An overview of the proposed offensive language classification model with fused representation of the input text from the sentiment and offensive models.

We implement our offensive model by fine-tuning BERTweet [150] (vinai/betweet-base on HuggingFace). During fine-tuning, the latent representations of the input text from our fine-tuned sentiment model and the offensive model being fine-tuned are fused to obtain a joint representation used in detection. Offensive language and sentiment analysis are closely related, and it can be safely assumed that negative sentiment is likely related to a text that is offensive [191]. The architecture of our model is shown in Fig. 3.2.

From Fig. 3.2, we freeze the weights of the fine-tuned sentiment model and use the model to extract representations of the input text by taking the special classification token (CLS) output of the penultimate layer. The CLS token is added to every input sequence as a special classification token. The corresponding token in the penultimate layer can be regarded as the aggregate representation of the input sequence [46]. The sentiment model is frozen to prevent wights from being updated during the fine-tuning of the offensive model. We denote this representation by vector $\mathbf{S}$ (with dimension 768). During fine-tuning, we also extract the CLS representation of the input text from the penultimate layer of the offensive

model denoted by vector **O** (with dimension 768). The concatenation of the sentiment and offensive model representations (i.e., $\mathbf{S} \oplus \mathbf{O}$) of the input text is fed to an output layer (number of neurons = the number of classes in the labeled dataset). The offensive language detection task is formulated as a binary classification task, where the label:0 corresponds to the non-offensive class and the label:1 corresponds to the offensive class. To train, we use the Adam optimizer with the learning rate initialized at 10^{-5}, five epochs, batch size of 16, and max sequence length of 128. Our offensive detection model achieved a macro F1, macro precision, and macro recall of 0.814, 0.815, and 0.813, respectively. Without using sentiment features, the model obtained 0.791, 0.798, and 0.787 in macro F1, macro precision, and macro recall, respectively. Additionally, without the sentiment features, the model achieved a macro F1 score of 0.829 for the offensive class and a macro F1 score of 0.756 for the non-offensive class, indicating that including sentiment features helps performance. After training, the fine-tuned model is used to classify each of the 21M tweets in our dataset. Table 3.3 summarizes the classifier's performance.

Target	F1	Precision	Recall
Non-offensive	0.787	0.798	0.776
Offensive	0.841	0.832	0.850

Table 3.3: Performance of the offensive detection model for the offensive and non-offensive classes. Evaluation metrics are macro averages.

We perform robustness checks to validate our results. (1) We chose to split the BLM dataset into a 90:10 ratio because it had a better performance when compared to the model trained on splitting the dataset into an 80:20 ratio to obtain the train (n=1972) and test (493) sets. The dataset obtained using the 80:20 split ratio contained (n=1077) offensive tweets and (n = 895) non-offensive tweets. The test split contained (n=278) offensive and (n=215) non-offensive tweets. The model achieved 0.785, 0.791, and 0.782 macro F1, precision, and recall, respectively. Per class, the non-offensive class achieved 0.747, 0.792,

and 0.707 macro F1, precision, and recall, respectively. The offensive class achieved 0.822, 0.791, and 0.858 macro F1, precision, and recall, respectively. The model obtained from the 90:10 split ratio, as discussed in Section 3.4.5, outperforms the 80:20 split ratio model in both overall F1 and per class F1 scores. We further validate this result by using bootstrap confidence intervals [55, 175, 176]. We randomly draw n samples with the replacement of k tweets from the test dataset of each split, where $n = 1000$, $k = 247$ for the 90:10 split, and $k = 493$ for the 80:20 split. We calculate the 95% confidence intervals (CI) of the macro F1 score and AUROC of each split repeated over 1000 bootstrap iterations. We obtained a 95% CI of 0.744-0.863 and 0.828-0.911 for the F1 scores of the 90:10 and 80:20 splits, respectively, and a 95% CI of 0.817-0.906 and 0.887-0.952 for AUROC scores of the 90:10 and 80:20 splits respectively. There are overlaps between the F1 and AUROC scores of the two splits, indicating no significant difference in performance between the two splits. We repeat the experiment on the difference between the F1 scores and AUROC scores of the two splits obtaining 95% CI of [0.0, 0.411] for the F1 scores and [0.0, -0.645] for the AUROC scores. Since both CIs contain zero there is no significant difference in performance of both splits. (2) We retrained our offensive model using 10-fold cross-validation with and without the sentiment features. The offensive model with the sentiment features obtained 0.779 and 0.857 average F1 score and AUROC score, respectively, and the offensive model without the sentiment features obtained 0.769 and 0.851 mean F1 score and AUROC score, respectively. We observed that there is significant difference (p-value < 0.05) in AUROC scores using paired t-test indicating the sentiment features helps in improving the offensive model's ability to separate between offensive and non-offensive tweets which is why we used this model in this study.

While our goal is not to advance the state-of-the-art, we compare our model to state-of-the-art methods and show that our model is competitive and has good generalization performance. Tables 3.4 and 3.5 contain the performance details of our model compared to other state-of-the-art models, and the cross-dataset generalization performance when compared to other methods revealing that our model is competitive and has good generalization

ability. Table 3.4 shows the comparison of our model to [126], [34], and [90], which are the models that performed the best in the classification of offensive language using the OffensEval dataset [241], abusive language [36] dataset, and hate speech [20] dataset. We also show the results of fine-tuning BERT, BERTweet, and HateBERT [34] on our annotated dataset.

The OffensEval [239] dataset was shared as part of SemEval 2019: Task 6 evaluation exercise about identifying offensive language in social media (sub-task A). The dataset contains 14,100 tweets, with 13,240 tweets in the training set and 860 tweets in the test set. The BERT-based model of [126] performed the best in the sub-task A category. As part of the SemEval 2019: Task 5 evaluation exercise about detecting hate speech against immigrants and women, the HatEval dataset [20] was shared. The dataset's English portion contains 13,000 tweets, with 10,000 tweets in the training set and 3,000 tweets in the test set. The AbusEval dataset [36] was developed by adding an extra layer to the OffensEval [239] and re-annotating for implicit and explicit abuse. The dataset is the same size as the OffensEval dataset and differs in the distribution of the positive class in the training and test sets. The BLM dataset is the in-house dataset we collected and annotated as described in Section 3.4.2.

From Table 3.4, our model (Ours) is very competitive to other methods when fine-tuned on various datasets. On the BLM dataset, we fine-tuned BERT, BERTweet, and HateBERT on our annotated dataset. Fine-tuning of BERT and BERTweet uses the same hyperparameters we used in our model as described in Section 3.4.5, fine-tuning of HateBERT uses the fine-tuning hyperparameter specified in the original work [34]. As shown in Table 3.4, our model outperforms other models on the BLM dataset.

We further validated our model by evaluating how well it generalizes to other data sets. To estimate this, we train our model on a data set (e.g., OffensEval) and test the trained model on another data set (e.g., AbusEval) as shown in Table 3.5. Also, we com-

Dataset	Method	Macro F1
OffensEval	BERT [126]	**.829**
	HateBERT [34]	.809
	Ours	.803
AbusEval	HateBERT [34]	**.765**
	[36]	.716
	Ours	.743
HatEval	HateBERT [34]	.516
	[90]	**.651**
	Ours	.548
BLM	BERT	.679
	HateBERT [34]	.710
	BERTweet	.791
	Ours	**.814**

Table 3.4: Performance of the offensive detection model compared to state-of-the-art methods.

pare our model's performance to the model developed by [16] as corrected in [14], the model developed by [8] as corrected in [14], fine-tuned BERT [46] (bert-base-uncased on HuggingFace) using the same hyperparameter used in our model as described in Section 3.4.5, and HateBERT [34]. The results of [16] and [8] are obtained from [14] as they fixed the problems with [16] and [8]. The Waseem & Hovy data set [226] is a public data set for hate speech detection annotated into three classes - "sexism", "racism", and "non-hate". In our experiments; we convert the classes into a binary class - "sexism" and "racism" classes are converted into "hate" for consistency with the SemEval data sets. From Table 3.5, we can observe that our model is very competitive when compared to other models and generalizes very well when trained on Waseem & Hovy [226] and HatEval [20] datasets. Furthermore, when trained on our dataset (BLM), our model outperforms other models in generalization, indicating that our model performs well in classifying offensive language, abusive language, and hate speech. The generalization experiment also validates our data annotation scheme.

Train		Test		
		OffensEval	AbusEval	HatEval
Waseem & Hovy	[16]	-	-	.475
	[8]	-	-	.472
	Ours	-	-	**.651**
OffensEval	BERT	-	**.824**	**.628**
	HateBERT [34]	-	.750	.547
	Ours	-	.800	.628
AbusEval	BERT	**.821**	-	.632
	HateBERT [34]	.759	-	.622
	Ours	.801	-	**.641**
HatEval	BERT	**.529**	.568	-
	HateBERT [34]	.512	.565	-
	Ours	.517	**.570**	-
BLM	BERT	.547	.526	.608
	HateBERT [34]	.587	.595	.563
	Ours	**.626**	**.648**	**.655**

Table 3.5: Model generalization results after training on a specific dataset and testing on another dataset. The evaluation metric shown is macro F1 score.

3.4.6 Emotion Classification

To understand the emotions expressed during the BLM-related events and protests (in answer to RQ2), we conducted emotion classification on our dataset. This approach enables us to identify the different emotional states of users during the protests.

Emotion analysis differs from sentiment analysis as it is a fine-grained classification of text based on emotional categories. Six basic emotions (anger, fear, sadness, enjoyment, disgust, and surprise) are defined by [57], and most emotion analysis studies focus on these emotions. In [57], the author further argued that these emotions can differ in antecedent events, behavioral response, physiology, etc. To classify emotions, following [249], we used the Semeval-2018 Twitter dataset [140] with 11 emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust) to fine-tune a pre-trained language model using joint representation from our sentiment model. We formulate this task as a multi-class classification problem and use the standardized training ($n = 6838$)

and test ($n = 3259$) for training and testing our model. We perform the same pre-processing as in our sentiment analysis.

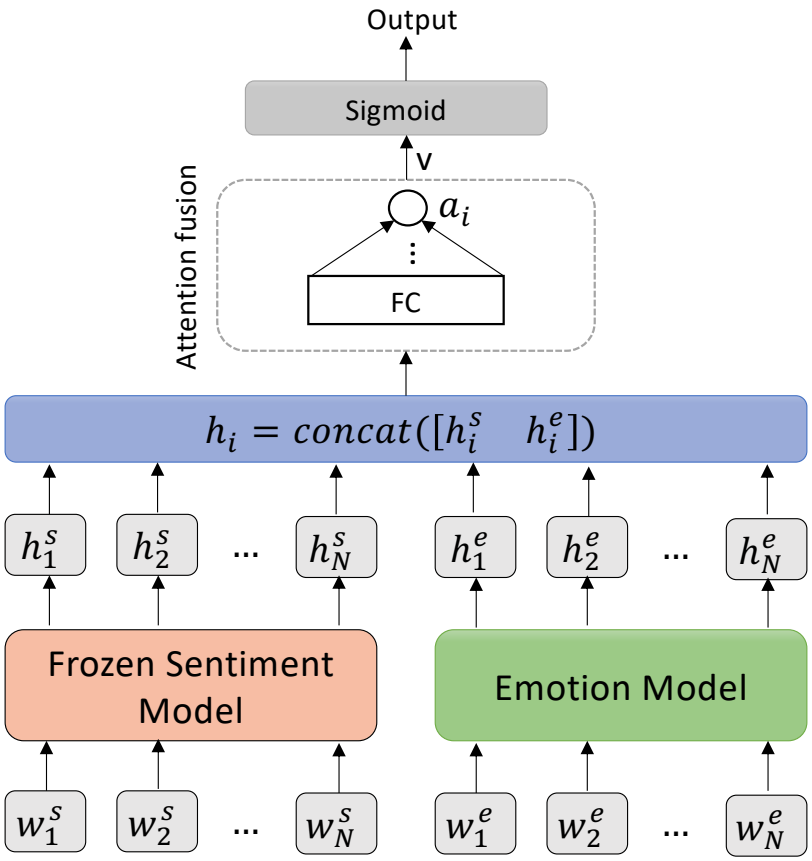


Figure 3.3: An overview of the proposed Emotion classification model with deep attention fusion. FC indicates a fully connected layer.

The emotion classification model is similar to the bidirectional LSTM (BiLSTM) architecture with a deep self-attention mechanism introduced by [21] and used in [249] to study emotion during COVID-19 pandemic. In our model, we use BERTweet [150](vinai/betweet-base on HuggingFace) and used our fine-tuned sentiment model from Section 3.4.4 to improve the performance of the emotion model by performing deep attention fusion of the representation of the words from the last encoder layer of each of the models using five hidden fully connected (FC) layers in the attention module. The architecture of our emotion model is shown in Fig. 3.3. From Fig. 3.3, the network consists of the fine-tuned sentiment model with weights frozen, the pre-trained language model (BERTweet) being

fine-tuned, an attention fusion layer, and an output layer (neurons = number of labels) with a sigmoid activation function. The input to the network is a tweet represented as a sequence of N words including the SEP token. Let $x_s = [w_1^s, ..., w_N^s]$ represent the input tweet to the sentiment model and let $x_e = [w_1^e, ..., w_N^e]$ represent the same input tweet to the emotion model. The sentiment model encodes the input and produces word representations $h_1^s, ..., h_N^s$ for each word in x_s from the last layer. Similarly the emotion model encodes the input and produces word representations $h_1^e, ..., h_N^e$ for each of the words in x_e from the last layer. We obtain the final representation for each word h_i by concatenating the representations h_i^s and h_i^e from both models, $h_i = [h_i^s h_i^e]$. Each representation $h_i \in \mathbb{R}^{2D}$ is a vector, where D is the size (768) of each word representation. In the attention layer, we use a multilayer perceptron (MLP) in place of the self-attention mechanism [165], [21] to amplify the influence of each word:

$$a_i = \frac{\exp(tanh(W_a h_i))}{\sum_{j=1}^{N} \exp(tanh(W_a h_j))} \tag{3.1}$$

$$v = \sum_{i=1}^{N} a_i h_i \tag{3.2}$$

where a_i is the attention weight that measures the importance of the current word i, W_a is the weight to be learned, and v is the final feature representation of the input tweet. The MLP is composed of $l = 4$ hidden layers (768, 768, 768, 256 neurons) with *tanh* activation function and an output layer. We use the Adam optimizer with a learning rate initialized at 10^{-5}, batch size of 8, a max sequence length of 128, minimize binary cross entropy loss, and applied early stopping to stop training when the loss value stops improving for seven consecutive epochs to avoid overfitting. Our emotion model achieved a macro F1 score of 55.8% (5.1% improvement when compared to [249]) and a micro F1 score of 68.7%. Following [249], we focus our analysis on emotions (anger, disgust, fear, joy, and optimism) with F1-scores above 0.7. The performance of our emotion model on each of the 11 emotions is shown in Table 3.6.

Emotion	Precision	Recall	F1-Score
Anger	0.79	0.77	0.78
Anticipation	0.36	0.22	0.27
Disgust	0.75	0.71	0.73
Fear	0.69	0.75	0.72
Joy	0.86	0.82	0.84
Love	0.64	0.56	0.59
Optimism	0.69	0.71	0.70
Pessimism	0.48	0.34	0.40
Sadness	0.76	0.64	0.69
Surprise	0.40	0.18	0.25
Trust	0.19	0.14	0.16

Table 3.6: Performance of our emotion model on 11 emotions. F1-macro: 55.8%, F1-micro: 68.70%.

3.4.7 Topic Analysis

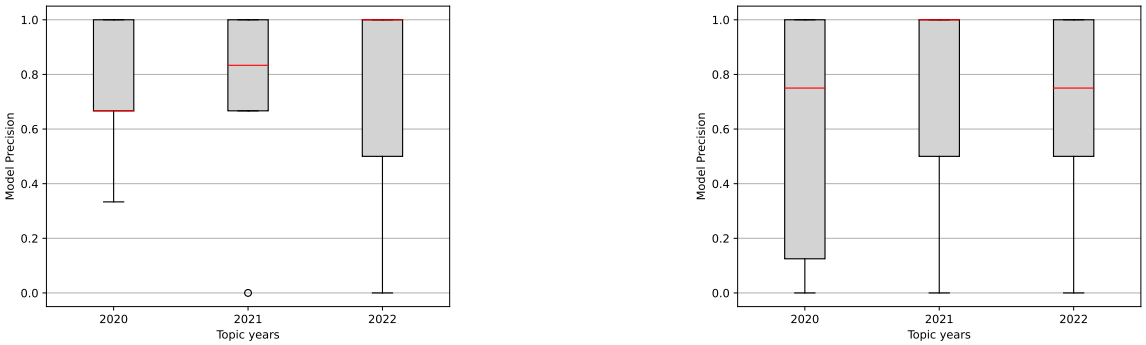
In order to analyze the offensive and non-offensive discussions in BLM-related online social movements, we conducted topic modeling on the predicted offensive and non-offensive tweets (in answer to RQ3). This approach enabled us to identify the important topics that engaged users during the movement.

Topic models help discover latent topics in a collection of documents. In this work, we used BERTopic [77], a topic model that generates coherent topics using a class-based TF-IDF (term frequency and inverse document frequency) and pre-trained transformer-based language models. BERTopic follows the clustering approach of topic modeling; the model leverages semantic relationships among words by using pre-trained language models to generate document embedding. The dimensions of the generated embeddings are reduced, and clusters of similar documents representing distinct topics are created. Finally, a class-based TF-IDF is used to extract topic representation from each topic. We use the following configuration for each of the main steps for topic modeling with BERTopic [77] most of which are BERTopic defaults, embedding (sentence transformer - all-MiniLM-L6-

v2), dimensionality reduction (UMAP with a seed parameter for reproducibility), clustering (HDBSCAN), vectorizer (CountVector), and class-based TF-IDF for topic representation. Unlike Latent Dirichlet Allocation (LDA) [24], which requires a user to specify the number of topics k, BERTopic does not require this specification. Instead, the minimum number of topics to be generated can be set, defaulting to 10 if not set explicitly. In this work, we set the minimum number of topics to 100 if the number of documents is between 1M and 1.5M and 500 if the number of documents is greater or equal to 1.5M. The smaller the value, the more topics are generated. We fitted distinct BERTopic models to all 2020, 2021, and 2022 offensive and non-offensive tweets. For 2020 and 2021 non-offensive tweets, we randomly sampled 2 million tweets from each dataset and fitted BERTopic on each. We sampled 2020 and 2021 non-offensive tweets due to computational resource constraint[4]. A manual examination was conducted on the top terms in each of the top 9 topics, the tweets associated with the topic, and the topic labeled according to the subject the terms likely represented. Topics that do not have coherent semantic groupings are excluded from our results (hence, the numbered topics presented in our results are not in chronological order). Non-coherent groups were found by analyzing the top 10 terms in a topic and the most representative documents in the topic as generated by BERTopic. We qualitatively merged similar topics (e.g., topics discussing the movement using #BLM and topics discussing the movement without the hashtag). The topic labels and example documents in each topic were analyzed qualitatively.

To validate our topic model, especially how well the inferred topics correspond with human concepts, we utilized word intrusion [37]. Word intrusion measures the semantic cohesiveness of the topics inferred by a topic model and verifies that the topics correspond to natural groupings by humans using model precision (the fraction of the subjects that agrees with the model). In this task, a subject is given six randomly ordered words, and the subject is tasked with identifying a word that is out of place or does not belong with the others, i.e., the intruder. To select a set of words given to the user, we randomly choose a topic

[4]We use a shared resource that terminates a job after three days, and it takes more than three days to fit more than 2M tweets to BERTopic.



(a) Offensive

(b) Non-offensive

Figure 3.4: The model precision of the topic models. Higher is better. The red line within the boxplot represents the median.

from the model. Then, select the top five words in the topic with the highest probability. An intruder word is randomly chosen from a randomly selected topic's five most probable words. All six words are shuffled and given to the subject. As stated earlier in this section, BERTopic does not require the specification for the number of topics; therefore, for this task, we restrict our analysis to the top 50 topics produced by BERTopic. We chose 50 because the percentage of documents assigned to each topic reduces to less than 25% of the documents after 50. Three internal subjects completed this task; they were instructed on the task, i.e., finding an intruder word in a set of words. No specialized training was offered to the subjects. Each subject was presented with ten sets of this task for each year in our study. The results are shown as a boxplot in Fig 3.4. From Fig 3.4, we observe that in each year, the level of agreement is good, indicating that the inferred topics are semantically meaningful.

3.4.8 Network Analaysis

To understand the interaction between offensive tweet authors and the recipients of offensive tweets, we study the reply graph of users who interacted with each other directly [107]. We construct a directed weighted graph $G = (V, E, w)$ for each year in our study where V are Twitter users, E are edges, a user u is directed to a user v through the edge

(u, v) if u tweeted offensively to v. And w represents the number of offensive tweets from u to v.

3.5 Results and Discussion

In this section, we discuss the results of our study examining offensive language in BLM-related discussions.

3.5.1 Offensive and Non-offensive Content (RQ1)

In answer to RQ1, we examined the presence and increase of offensive content in BLM tweets. Table 3.7 summarizes the statistics of offensive and non-offensive tweets in our dataset. From Table 3.7, 2.5M tweets were predicted to be offensive, and 18.8M tweets were predicted to be non-offensive. The year 2020 had the highest number (1.7M or 71%) of total offensive tweets when compared to 2021 (500k or 20%) and 2022 (235K or 10%). A similar trend is observed for the non-offensive tweets. With a total of 2.5M offensive tweets, approximately 12% of the total tweets, it shows that **BLM-related discussions had a considerable amount of offensive tweets**.

Year	Offensive	Non-offensive
2020	1,766,491	14,621,431
2021	500,039	3,077,946
2022	235,503	1,124,404

Table 3.7: Statistics of predicted offensive and non-offensive tweets.

To further examine the presence of offensive and non-offensive tweets, we looked at the number of offensive and non-offensive tweets created per day across our study period (Fig. 3.5). We used the Pruned Exact Linear Time (PELT) [100] algorithm to detect change points in the number of tweets per day and possibly the likely events around the change point that caused the change. We indicate possible events with the letter "E" in Fig. 3.5. From Fig. 3.5A, we find a notable uptick on May 31, 2020 (E1). The noteworthy increase in

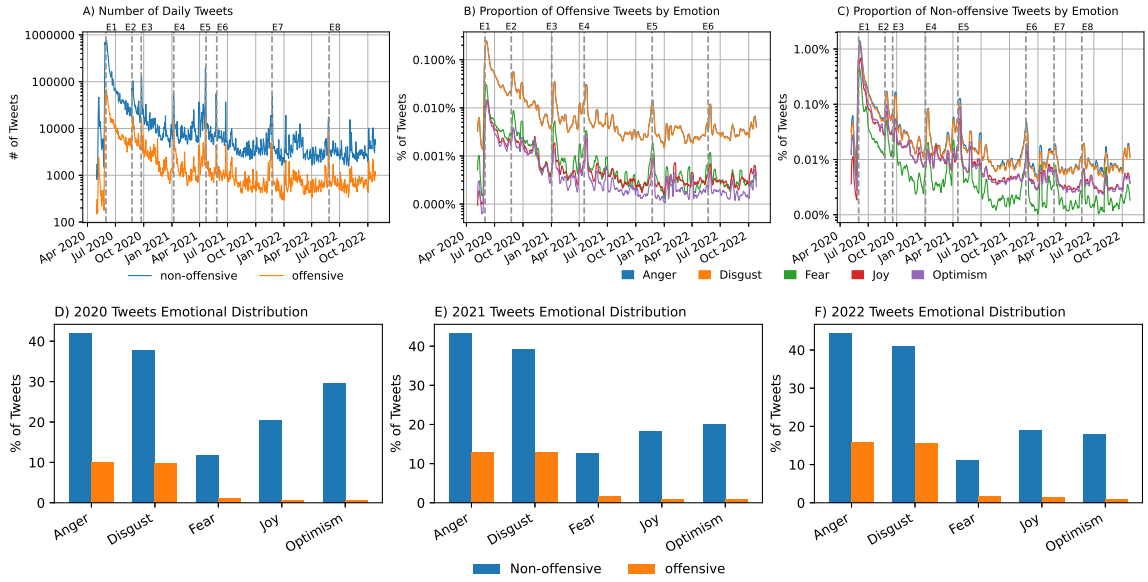


Figure 3.5: (A) Daily tweet count in log scale. (B) Temporal evolution by emotions for offensive tweets in log-scale. (C) Temporal evolution by emotions for non-offensive tweets in log-scale (D) Emotion distribution of 2020 tweets. (E) Emotion distribution of 2021 tweets. (F) Emotion distribution of 2022 tweets. The temporal evolution of emotions is based on smoothed weekly averages for visual clarity. Anger and disgust correlate with a Pearson correlation score of 0.99 (p-value $\ll 0.0001$). The emotion distribution shows a significant increase in anger, disgust, and fear in 2020. The gray vertical lines signify points with significant changes in the number of daily tweets and emotional distribution.

offensive and non-offensive tweets accounts for approximately 2.4% of the offensive tweets and 3.6% of the non-offensive tweets, respectively. E1 (May 31, 2020) corresponds to a day that protests continued in large cities across the United States following George Floyd's death and when President Trump announced his plans to designate Antifa as a terrorist organization. After May 31, 2020 (E1), there was a noticeable decline and stability in the number of offensive and non-offensive tweets, with the number of offensive tweets decreasing the most. We observe other upticks over time. E2 (August 24, 2020) potentially corresponds to the shooting of Jacob Blake in Kenosha, Wisconsin, and the clash between the police and those protesting Jacob Blake's shooting. E3 (September 23, 2020) potentially corresponds to protests over Breonna Taylor's death and the indictment of a Louisville, Kentucky, officer for firing into the apartment of Breonna Taylor's neighbor. E4 (January 7, 2021) potentially

corresponds to a day after the January 6, 2021, United States Capitol attack by pro-Trump protesters. E5 (April 22, 2021) potentially corresponds to the aftermath of the shooting of Daunte Wright and the shooting and Killing of Ma'Khia Bryant by a Columbus police officer. E6 (May 27, 2021) potentially corresponds to one year after George Floyd's death, the trial period of Derek Chauvin, the Minneapolis police officer responsible for George Floyd's death, and the Israel-Palestine crisis (topic analysis revealed that BLM was blamed for supporting Palestine). E7 (November 23, 2020) relates to the aftermath of the protests in large cities in the United States, especially in Kenosha, Wisconsin, following the acquittal of Kyle Rittenhouse. Finally, E8 (May 28, 2022) likely relates to approximately two years after George Floyd's death and protests.

From our results, offensive discussions were most prominent in May 2020 following George Floyd's death, and the number of offensive tweets declined and stabilized after May 2020. Additionally, real-world events such as protests and police shootings during the study period possibly increased offensive tweets. These results answer RQ1 in part.

3.5.2 Offensive Reply Network Analysis (RQ1)

The offensive reply network statistics is shown in Table 3.8. In the 2020 offensive reply graph, 84.5% of the 873,043 offensive tweets posted by offenders to recipients occurred only once. 84.1% of all recipients are not offenders, 15.9% of all recipients are offenders, and 11.3% of all offenders are recipients. 0.49% of offensive edges were reciprocal, where recipients and offenders responded to each other offensively. 3,816 (0.96%) of offenders engaged in reciprocal offensive discussion (1.0% of offensive tweets) with the recipients, and 9,802 (8.6%) of recipients were repeatedly targeted by offenders who repeatedly replied to them offensively.

In the 2021 offensive reply graph, 89.9% of the 273,028 offensive tweets posted by offenders to recipients occurred between offenders and recipients only once. 86.8% of all recipients are not offenders, 13.2% of all recipients are offenders, and 10.2% of all offenders are recipients. 0.64% of offensive edges were reciprocal, where recipients and offenders

responded offensively to each other. 1,632 (1.1%) of offenders engaged in reciprocal offensive discussion (1.4% of offensive tweets) with the recipients, and 9,802 (8.6%) of recipients were repeatedly targeted by offenders who repeatedly replied to them offensively.

Finally, in the 2022 offensive reply graph, 91.3% of the 143,683 offensive tweets posted by offenders to recipients occurred between offenders and recipients only once. 88.3% of all recipients are not offenders, 11.7% of all recipients are offenders, and 9.1% of all offenders are recipients. 0.9% of offensive edges were reciprocal, where recipients and offenders responded offensively to each other. 1,220 (1.4%) of offenders engaged in reciprocal offensive discussion (2.1% of offensive tweets) with the recipients, and 4,498 (6.7%) of recipients were repeatedly targeted by offenders who repeatedly replied to them offensively.

These results answer the rest of RQ1 and indicate that most offensive tweet recipients were not offenders. The offenders likely targeted them with an offensive tweet due to an innocuous view they held or a tweet they posted about the BLM movement.

Year	# Nodes (recipients, offenders)[5]	# Edges
2020	631,764 (44.5%, 62.6%)	778,308
2021	245,817 (46.3%, 59.9%)	254,402
2022	146,774 (46.0%, 59.4%)	87,240

Table 3.8: Statistics of the offensive reply network.

3.5.3 Examining Emotions Expressed in BLM Tweets (RQ2)

Having identified offensive tweets in BLM movement discussions, we utilized our emotion model to examine the emotions expressed in discussions. This, in turn, is used to answer RQ2, examining emotions and how the emotions expressed differ in offensive and non-offensive tweets.

The results of the temporal analysis marked with change points are shown in Fig. 3.5B for the offensive tweets and in Fig. 3.5C for the non-offensive tweets. The results of the distributions of the five emotions for 2020, 2021, and 2022 are shown in Fig. 3.5D, Fig.

[5]Percentages does not sum to 100% as some recipients are also offenders

3.5E, and Fig. 3.5F, respectively. Through change point analysis, we identified notable social events that might have led to sporadic emotional fluctuations. We used the same change point detection algorithm (PELT) in the temporal analysis of daily tweets to identify periods of high emotional expressions.

Figures 3.5B and 3.5C show the weekly averages in fluctuations of emotions within the offensive and non-offensive tweets. The emotional response within the offensive tweets is dominated by negative emotions (anger, disgust, and fear), with positive emotions (joy and optimism) being the lowest. A similar observation is made in the non-offensive tweets with higher positive emotions (joy and optimism) than the negative emotion (fear). Noticeable changes in emotions are observed for both offensive and non-offensive tweets. For the offensive tweets, the change points correspond to the change points observed in the number of daily tweets as shown in Fig. 3.5A. For the non-offensive tweets (Fig. 3.5C), most change points correspond to the change points in the offensive tweets. The week of September 18, 2020 (E3) coincides with the week of the shooting of James Scurlock in Omaha, Nebraska, during George Floyd's protests, and E7 (February 25, 2022) coincides with the week the three officers involved in the death of George Floyd were found guilty. From Figures 3.5D, 3.5E, and 3.5F, we find that both offensive and non-offensive tweets have higher proportions of anger, disgust, and fear. We also note that in 2020, the proportion of anger and disgust in the offensive tweets was twice that of anger and disgust in the non-offensive tweets. Additionally, [60] analyzed the emotions expressed in tweets collected using Pro and Anti-BLM hashtags and keywords during the 2020 BLM protests. Juxtaposing with the emotion analysis of Pro-BLM tweets of [60] between May 25 and June 30, 2020, we similarly observe that anger and disgust are positively correlated. Compared to our non-offensive and offensive tweets, we similarly observe that anger is the top expressed emotion. Finally, when we filter tweets using the Pro-BLM hashtags (#BlackLivesMatter, #BLM, #GeorgeFloyd, and #JusticeForGeorgeFloyd) shared with [60] between May 25 and June 30, 2020, we find that anger and disgust are higher closely followed by joy and optimism, and then fear contradicting their findings that positivity is higher in tweets with

Pro-BLM hashtags. This contradiction could be due to the difference in hashtags used in data collection.

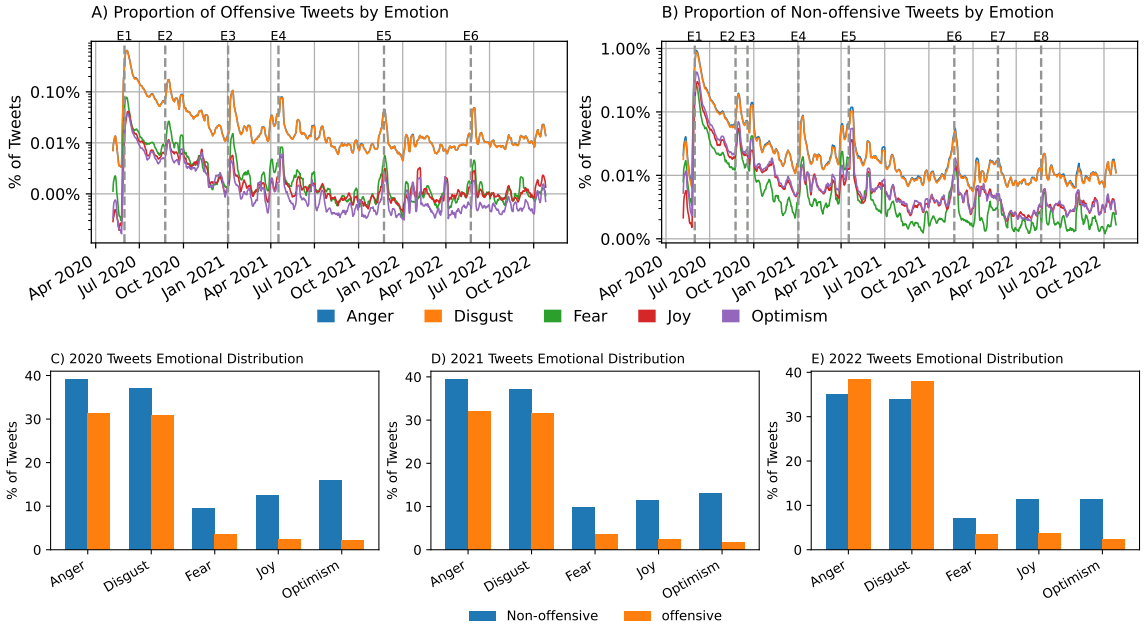


Figure 3.6: Emotion dynamics of offenders. (A) Temporal evolution by emotions for offensive tweets in log-scale. (B) Temporal evolution by emotions for non-offensive tweets in log-scale (C) Emotion distribution of 2020 tweets. (D) Emotion distribution of 2021 tweets. (E) Emotion distribution of 2022 tweets. The temporal evolution of emotions is based on smoothed weekly averages for visual clarity. Anger and disgust correlate with a Pearson correlation score of 0.99 (p-value $\ll 0.0001$). The gray vertical lines signify points with significant changes in emotional distribution.

We further compare emotion dynamics across offenders and recipients in our offensive reply network. We extracted offenders and recipients that do not overlap (i.e., offenders that are not recipients and recipients that are not offenders) and analyzed the emotions in their tweets. The offensive tweets of the offenders and recipients follow similar patterns as the overall offensive tweets where negative emotions (anger, disgust, and fear) dominate, as shown in Figures 3.5B, 3.6A, and 3.7A. The offender's non-offensive tweets, shown in Fig. 3.6B, follow a similar pattern to those of the overall non-offensive tweets. Though anger and disgust dominate, positive emotions (joy and optimism) dominate fear. In contrast, in the recipient's non-offensive tweets, as shown in Fig. 3.7B, fear dominated the positive

emotions throughout the timeline except in 2020. In our robustness checks, we repeat our analysis on offenders with no overlap who had more than 50 replies and recipients with no overlap who received more than 50 offensive tweets, yielding consistent findings.

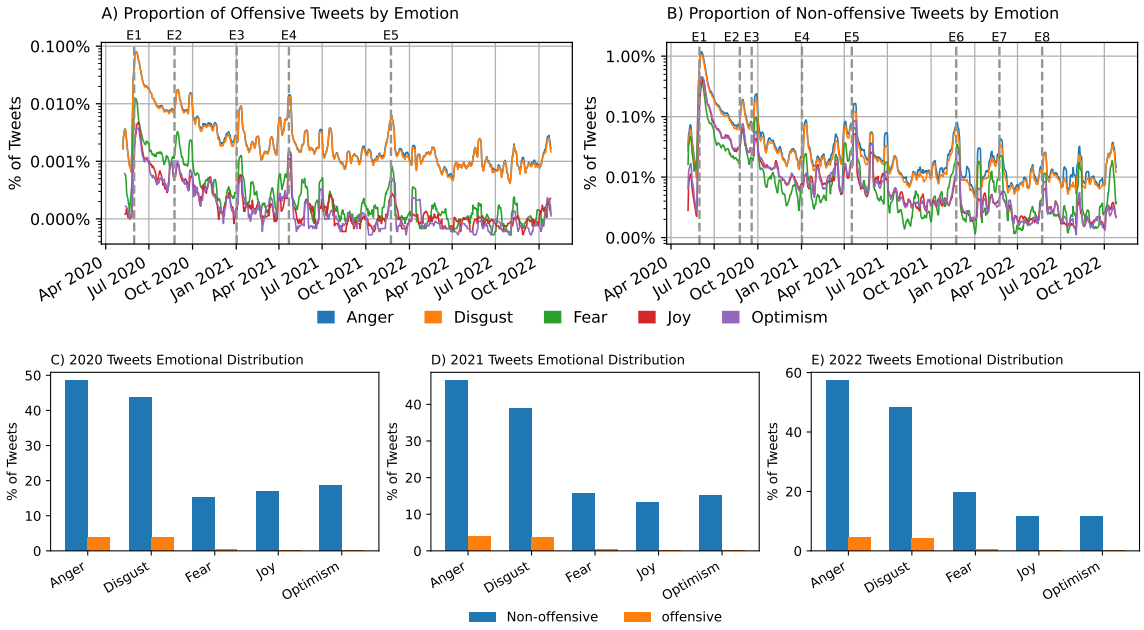


Figure 3.7: Emotion dynamics of recipients. (A) Temporal evolution by emotions for offensive tweets in log-scale. (B) Temporal evolution by emotions for non-offensive tweets in log-scale (C) Emotion distribution of 2020 tweets. (D) Emotion distribution of 2021 tweets. (E) Emotion distribution of 2022 tweets. The temporal evolution of emotions is based on smoothed weekly averages for visual clarity. Anger and disgust correlate with a Pearson correlation score of 0.99 (p-value $\ll 0.0001$). The gray vertical lines signify points with significant changes in emotional distribution.

From our results, we can answer RQ2 that anger and disgust though strongly correlated, were the most predominant emotions expressed in BLM discussions. After E1 (May 31, 2020), there was a reduction in the proportion of emotions in the offensive tweets with few upticks. **While there was a reduction in the proportion of emotions in the non-offensive tweets, non-offensive tweets had more emotional fluctuations throughout the study period. Furthermore, positive emotions (joy and optimism) dominate fear and are more pronounced in the non-offensive tweets of offenders compared to the non-offensive tweets of recipients. Negative emotions**

overpowered the overall sentiment throughout BLM discussions, especially in the offensive tweets.

3.5.4 Analyzing Topics of Discussion (RQ3)

Having identified the emotions expressed in BLM discussions, we utilized topic modeling on the offensive and non-offensive tweets of 2020, 2021, and 2022. This, in turn, is used to answer RQ3, examining the most discussed topics in 2020 and the degree the topics were discussed years (2021 and 2022) after the 2020 BLM protests.

The analysis results of the topics in the 2020 offensive and non-offensive tweets can be found in Table 3.9 without the representative tokens. The topics with the representative tokens are depicted in Tables 3.10 and 3.11. Only the top 9 topics are shown after merging topics and removing topics that the keywords didn't clearly indicate a topic. The topics discovered covered a range of issues, including the death of George Floyd, Breonna Taylor, kneeling during the national anthem, #BlackLivesMatter, and #AllLivesMatter. We describe the top 9 topics in the 2020 offensive tweets.

Offensive		Non-offensive	
Topic #	Topic	Topic #	Topic
0	Floyd	0	Black Lives Matter
1	Kneeling in Sports	2	Kneeling in Sports
2	#BLM	4	Arbery's Shooting
3	Breonna Taylor	5	Police Brutality
7	#DefundThePolice	7	Covid19
8	#BlackLivesMatter /#AllLivesMatter	8	#BlackLivesMatter /#AllLivesMatter
10	Arrest Cops	9	Donation
11	Antifa	10	Petitions
12	#JacobBlake	12	Justice

Table 3.9: The topics discovered by topic modeling in the 2020 offensive and non-offensive tweets without the representative tokens in the topics.

In the offensive tweets, topic "Floyd" primarily discussed George Floyd's death and criticized his character. For example, *"George Floyd acts like a psychopath high on*

Offensive		
Topic #	Topic	Representative Tokens
0	Floyd	floyd, floyds, cop, criminal, murdered, police, cops, officer, officers, justice
1	Kneeling in Sports	nfl, nba, sports, kneeling, anthem, fans, knee, Kaepernick, lebron, athletes
2	#BLM	blm, racist, racists, racism, support, blmprotest, blmterrorists, blmantifaterroriststhugs, thugs, terrorists
3	Breonna Taylor	breonna, taylor, taylors, cops, arrest, killers, murdered, arrested, criminal, bf
7	#DefundThePolice	blacklivesmatter, user, racist, racists, ignorant, people, racism, bitch, hate, alllivesmatter
8	#BlackLivesMatter /#AllLivesMatter	defundthepolice, defund, defundpolice, defunding, abolishthepolice, blacklivesmatter, carenotcops, blm, reform, defundnypd
10	Arrest Cops	breonna, arrest, cops, murdered, taylors, justice, arrested, officers, justiceforbreonnataylor, jail
11	Antifa	antifa, matter, groups, terrorists, cities, democrats, left, violence, rioting, democratic
12	#JacobBlake	jacobblake, Kenosha, kyle, kenoshaprotests, rapist, shooting, kenoshashooting, justiceforjacobblake, kenoshariots, kenoshashooter

Table 3.10: The topics discovered by topic modeling and the most representative tokens in the 2020 offensive tweets.

something, resists arrest continuously, refuses to get into the cop car from 'sudden claustrophobia' despite being in another car moments before and claims he can't breathe before anyone touches him. For this criminal, our cities burn. URL". Tweets in topic "Keenling in Sports" discussed kneeling in general and disapproved of sports teams supporting the BLM movement. For example, *"They mad niggas kneeling during a bullshit football game but these racist mother fucking pigs kneeling on a black mans throat while he's in handcuffs. Someone gotta hang this pig named Derek chauvin and how Asian but buddy NAME. I hope they get what they deserve #GeorgeFloyd".* Topic "#BLM" contains tweets that disapprove of the BLM movement, the BLM protests, and those supporting the movement. For example, *"Black lives matter teaches us that black people are lazy, stupid, and need everything handed to them. Congrats BLM for supporting the KKK. #BlackLivesMatter.*

Non-offensive		
Topic #	Topic	Representative Tokens
0	Black Lives Matter	matter, lives, black, movement, trans, matters, support, mean, rights, racist
2	Kneeling in Sports	nba, nfl, sports, anthem, knee, kneeling, athletes, kaepernick, wnba, jerseys
4	Arbery's Shooting	arbery, ahmaudarbery, mcmichael, shooting, son, jogger, justiceforahmaud, arrested, run, lynching
5	Police Brutality	police, brutality, cops, cop, blacklivesmatter, policebrutality, officers, unarmed, officer, policing
7	Covid19	covid19, coronavirus, covid, pandemic, vaccine, protests, health, quarantine, crisis, blm
8	#BlackLivesMatter /#AllLivesMatter	blacklivesmatter, url, help, alllivesmatter, powerful, read, change, stop, share, thank
9	Donation	donate, donated, fund, donation, donating, donations, charity, charities, fundraiser, ads
10	Petitions	petition, donate, signed, donating, spread, educate, blacklivesmatter, resources, justiceforgeorgefloyd
12	Justice	justiceforgeorgefloyd, justice, justiceforbreonnataylor, justiceforgeorgeflyod, justiceforgeorge, justiceforahmaud, justiceforfloyd, justiceforahmaudarbery, justiceforjacobblake, justiceforcaseygoodson

Table 3.11: The topics discovered by topic modeling and the most representative tokens in the 2020 non-offensive tweets.

You are a garbage movement." In topic "Breonna Taylor", tweets discussed the death of Breonna Taylor, blamed her partner for her death, and called for the prosecution of the officers involved. For example, *"@USER Breonna Taylor was a drug dealing cunt!!!"*. Topic "#BlackLivesMatter/#AllLivesMatter" focused on the debate about black/white/all lives matter. Tweets called out users on racism and accused individuals of being racist and were mainly directed to those with opposing views and those misconstruing the meaning of BLM. For example, *"@USER @USER @USER It is BLACK VS WHITE white people are the only race that are soo racist towards black people yall got mental illnesses and that's why whites can't be trusted #BlackLivesMatter #whitelivesdont matter"*.

Topic "#DefundThePolice" discussed calls to defund the police and police reform.

For example, *"Un. Fucking. Real. I wouldn't want these dumbasses protecting or serving my morning shit let alone our lives. #BlackLivesMatter #DefundThePolice"*. Criticism of the police officers involved in the Death of Breonna Taylor and the justice system was prominent in topic "Arrest Cops". For example, *"They KNOW that they will get a bunch of cop stans if she was suddenly part of a drug ring. They don't want to arrest their cops buddies, they just want media and people off their back. Arrest the cops who murdered Breonna Taylor or burn the whole precinct to the ground"*. Tweets in topic "Antifa" focused on jointly discussing Black Lives Matter and Antifa. Tweets blamed protesters for destroying cities, accused BLM of funding Antifa, and referred to both groups as terrorist organizations. For example, *"@USER Your a fuc-ing criminal you should rot in jail because demoshits have no balls your black lives don't matter assho-es think you can get away with anything. Black lives matter and fuc-ing antifa should be labeled as terrorist groups. If I had my way I would put all in jail"*. Finally, The shooting of Jacob Blake, criticism of his character, the Kenosha protests, and the shooting of protesters by Kyle Rittenhouse was the focus of tweets in topic "#JacobBlake". For example, *"Just saw the video of #JacobBlake getting shot 7 times in the back. Burn the whole city down idgaf. Absolutely wtf was that. I had no doubt the cop was at fault before the video but that was more heinous than I could've imagined. #ACAB anyone who defends this shit can get bent."*

For the 2020 non-offensive tweets, topic "Black Lives Matter" focused on users arguing about Black/White/All lives matter, the meaning of Black Lives Matter, and opposing labels, and argued that other labels are being used to belittle the Black Lives Matter movement. For example, *"@USER @USER @USER @USER @USER You're the one implying they mean something else. No one is saying "Black lives matter more" or "Only black lives matter". That's your projection. Obviously all lives matter, but many black people don't feel like they're being included in that "all". It's really very simple."* The tweets in "Sports" primarily focused on sports organizations and teams supporting the BLM movement and their athletes' gestures. For example, *"I think it's super dope that the NBA put Black Lives Matter on the court"*. In topic "Arbery's Shooting", tweets discussed the killing of Ahmaud

Arbery in Georgia, United States, and the arrest of the individual involved in Ahmaud Arbery's death. For example, *"@USER @USER Need police reform from the top, not vigilante justice and citizens' arrests. That was what the McMichaels' claimed as their reason for killing Ahmaud Arbery. Last thing we need is people like that thinking they have an obligation to enforce their understanding of the law."* The police, policing, police brutality, and calls for police reform were the main focus of topic "Police Brutality". For example, *"just saw some tweets that said not to say 'police brutality' cause it understates the murder, im sorry! let me correct this to say that black people are being murdered here and we cannot pretend it doesnt happen #BlackLivesMatter"*.

Topic "Covid19" discussed the BLM protests amid the coronavirus pandemic, and some tweets blamed the protest for the rise in Covid-19 cases. For example, *"#COVID19, #BLM, and early attempts at election interference have highlighted the online ecosystem that enables the viral spread of hatred, violence, and disinformation. This report finds a path forward for tech companies & govts to moderate online content: URL URL"*. Topic "Donation" primarily focused on soliciting donations to BLM-related charities and discussed using donated funds to bail out arrested protesters. For example, *"Please consider donating to a local bail fund in order to free protestors! I have a list on some in my last RT! #Blacklivesmatter this is my donation to the Louisville bail fund, could anyone match me!? URL"*. Calls to sign petitions to raise awareness and stand against injustice were discussed in Topic "Petitions". For example, *"Hope you're all taking care of yourselves mentally these days since I know everything can be quite overwhelming. However, please continue to use your voice for the #BlackLivesMatter ! Even if you can't go out to protest, rt and sign petitions, keep bringing awareness!"*. Tweets in topic "Justice" generally called for justice for victims of police brutality. For example, *"What the United States has accomplished in the past is far less important than what we should do in the future. #BlackLivesMatter #JusticeForElijah #JusticeForAhmaud #JusticeForGeorgeFloyd #JusticeForElijahMcClain #JusticeforRobertFuller #JusticeForAll #EqualProtectionUnderTheLaw"*.

Tables 3.14 and 3.15 shows the results of 2021 and 2022 offensive and non-offensive

Offensive		
Topic #	Topic	Representative Tokens
0	Ahmaud Arbery	ahmaud, arbery, arberys, murderers, hunted, chased, jogger, defendants, vigilantes, rednecks
1	Breonna Taylor	breonna, taylor, boyfriend, sleeping, bed, taylors, apartment, door, name, killers
2	Rubber Bullets	rubber, bullets, protestors, protesters, protest, storming, protesting, peacefully, protests, capitolbuilding
3	Pregnant Woman	pregnant, belly, woman, gunpoint, robbing, unborn, pistol, robbery, hero, criminal
4	Nancy Pelosi	Nancy, sacrificing, pelosi, sacrifice, thanking, sacrificed, sacrificial, pelosis, schumer, martyr
5	Palestinian Hamas	Hamas, palestine, palestinians, palestinianlivesmatter, palestinian, gaza, palestinianslivesmatter, zionists, israelicrimes, hamasterrorists
6	#DerekChauvin	Derekchauvin, derekchauvintrial, derekchauvinisguilty, justiceforgeorgefloyd, derek, derekchauvintrail, derekchauvinsentencing, derekchauvinverdict, georgefloydisnotontrial, derekchauvins
7	Drug Overdose	Fentanyl, meth, overdose, lethal, overdosed, methamphetamine, overdosing, autopsy, counterfeit, drugs
8	Black Lives Matter	Matter, black, stfu, oxymoron, dumb, thinks, mean, doo, profile, matters

Table 3.12: The topics discovered by topic modeling and the most representative tokens in the 2021 offensive tweets.

tweets topics without the representative tokens. The topics and representative tokens are depicted in Tables 3.12, 3.13, 3.16, and 3.17. We observe that some topics, such as the death of Ahmaud Arbery, Breonna Taylor, BLM, the Kenosha Shooting, Kneeling in Sports, and Racism, continued to be discussed offensively and non-offensively after the 2020 protests. Derek Chauvin is still being discussed due to his ongoing trial and conviction. There were new topics observed, the "Rubber Bullets/Riot/Capitol Protests" topics discussed the January 6 Capitol protests by former president Trump supporters and compared the protesters and BLM protesters were treated differently by the police. The "Drug Overdose/Pregnant Woman" topic criticizes the character of George Floyd, attributing his death to a drug overdose and accusing him of pointing a gun at a pregnant woman. For example, *"@USER*

Non-offensive		
Topic #	Topic	Representative Tokens
0	ahmaud Arbery	Arbery, mcmicheal, jury, arberys, trial, defendants, verdict, ahmaudarberytrial, justiceforahmaud, judge
1	Derek Chauvin	chauvin, derek, neck, chauvins, floyds, verdict, manslaughter, charges, convicted, murdered
3	Breonna Taylor	breonna, breonnataylor, louisville, taylors, sayhername, bed, justice, cops, justiceforbreonnataylor, murdered
4	Kenosha Shooting	Blake, Kenosha, shot, shooting, kyle, blakes, kyleritten-house, armed, officers, filed
5	Black/All Lives Matter	Matter, lives, say, movement, white, matters, racist, support, mean, care
6	Capitol Protests	capitol, protests, protesters, protest, peaceful, protestors, bullets, protesting, riot, rioters
7	Keenling in Sports	Knee, kneeling, nfl, fans, racism, sports, footballers, anthem, kneel, lebron
9	Justice	Justice, verdict, served, hope, floyd, peace, alive, relief, justiceforgeorfloyd, floyds
11	Policing Bill	Policing, bill, senate, passed, congress, vote, filibuster, senators, legislation, chokeholds

Table 3.13: The topics discovered by topic modeling and the most representative tokens in the 2021 non-offensive tweets.

@USER 27 million is not enough ? Do they know George Floyd was a repeat violent offender convict that shortly before his death held a gun against a pregnant black female's stomach during a home invsion?? Are they aware of that? Thank You BLUE Thank You Officer Chauvin" and *"@USER Glad to see they are getting rid of that awful stain of left wing garbage. George Floyd was a MONSTER and I'm glad he isn't with us anymore! I hope other people who rob and assault pregnant women get the same treatment in the future. His entire existence was a pathetic shame."*

The "Nancy Pelosi" topic criticized Representative Nancy Pelosi for her comment that George Floyd sacrificed his life, and the "Barack" topic criticized former president Barack Obama for comparing the Uvalde, Texas school shooting to George Floyd. The introduction of the policing reform bill is discussed in topic "Policing Bill". Topic "Midterm Elections" was primarily political, focusing on the midterm election, Trump, and his lawyer

Offensive			
Topic #	2021	Topic #	2022
0	**Ahmaud Arbery**	0	**Breonna Taylor**
1	**Breonna Taylor**	1	Riots/Protests
2	Rubber Bullets	2	**Ahmaud Arbery**
3	**Pregnant Woman**	3	Barack
4	Nancy Pelosi	4	#CapitolRiot
5	Palestinian Hamas	5	**Drug Overdose**
6	#DerekChauvin	6	**Pregnant Woman**
7	**Drug Overdose**	8	Racism
8	Black Lives Matter	9	Midterm Elections

Table 3.14: The topics discovered by topic modeling in the 2021 and 2022 offensive tweets without the representative tokens in the topics. The highlighted topics are some of the topics in the 2021 offensive tweets that persisted in 2022.

Non-offensive			
Topic #	2021	Topic #	2022
0	Ahmaud Arbery	0	Arbery
1	Derek Chauvin	1	**All Lives Matter**
3	Breonna Taylor	2	George Floyd
4	Kenosha Shooting	3	Arrest Cop
5	**All Lives Matter**	4	Abortion
6	Capitol Protests	5	**#BlackLivesMatter /#AllLivesMatter**
7	**Kneeling in Sports**	6	#NOH8
9	**Justice**	8	**Kneeling in Sports**
11	Policing Bill	9	Drug Overdose

Table 3.15: The topics discovered by topic modeling in the 2021 and 2022 non-offensive tweets without the representative tokens in the topics. The highlighted topics are some of the topics in the 2021 non-offensive tweets that persisted in 2022. After 2020, topics related to Floyd, Breonna, and Riots/Protests are still being discussed.

Giuliani. Tweets that argued about BLM and abortion, abortion and anti-abortion rights, and how supporting BLM and abortion opposes each other were mainly in topic "Abortion". For example, *"@USER I want to speak to the people that were standing for the Black Lives Matter movement and see how they square their conscience if they are fighting for abortion because I would think those two causes would conflict with each other"*. Finally, tweets in

Offensive		
Topic #	Topic	Representative Tokens
0	Breonna Taylor	breonna, warrant, breonnataylor, taylors, justiceforbreonnataylor, raided, ky, justice, charges, sayhername
1	Riots	riots, protests, protest, rioters, rioting, protesters, protesting, riot, insurrection, protestors
2	Ahmaud Arbery	arbery, ahmaudarbery, arberys, prison, sentenced, murderers, mcmichaels, fetterman, sentencing, chased
3	Barack	barack, massacre, children, obamas, shooting, innocent, Hussein, memorialize, comparison, tweets
4	#CapitolRiot	capitolriot, fascism, truthbetold, ukrainewar, humanrights, nojusticenopeace, fascistgop, trumpisacriminal, arresttrumpnow, abortionrights
5	Drug Overdose	fentanyl, overdose, meth, lethal, overdosed, autopsy, drug, fatal, counterfeit, felon
6	Pregnant Woman	pregnant, belly, gun, womans, gunpoint, pistol, robbing, robbery, unborn, felon
7	Innocent Children	children, innocent, teachers, kids, dare, comparison, slaughtered, tweet, families, tragedy
8	Racism	racist, racists, blm, racism, race, antiracist, lisa, bigoted, diversity, whites

Table 3.16: The topics discovered by topic modeling and the most representative tokens in the 2022 offensive tweets.

topic "#NOH8 (No Hate)" used the hashtag and #BLM, among others, to promote human equality and to discuss general issues, especially politics.

Thus, we answer RQ3 - understanding the main discussions in the offensive and non-offensive tweets in the BLM movement, which include discussions on policing and racial injustice. We also confirm that these issues continued to be discussed after the BLM protests in 2020, possibly due to the trials of the individuals involved in the police-related death of victims and observed the discovery of new topics such as Palestinian Hamas, Rubber Bullets/Capitol Protests, Nancy Pelosi, Barack, Policing Bill, Midterm Elections, and Abortion.

Non-offensive		
Topic #	Topic	Representative Tokens
0	Arbery	arbery, mcmichael, arberys, killers, travis, crimes, sentenced, judge, bryan, murderers
1	All Lives Matter	lives, matter, phrase, racist, triggered, mean, response, matters, mattered, says
2	George Floyd	floyd, george, talking, situation, bro, say, lmao, happened, comment, mean
3	Arrest	arrest, cop, resisting, police, officer, unarmed, officers, deserved, compiled, brutality
4	Abortion	abortion, abortions, unborn, parenthood, roe, abortionrights, abortionishealthcare, eugenics, abortionrightsarehumanrights, prochoice
5	#BlackLivesMatter /#AllLivesMatter	Blacklivesmatter, alllivesmatter, racism, racist, whitelivesmatter, supremacy, blacks, stopasianhate, discrimination, twitter
6	#NOH8	noh8, motivation, fbr, resistance, strongertogether2022, waterwave, whiteflag, grassroots, resisters, saturdaythoughts
8	Kneeling in Sports	Nfl, eminem, kaepernick, knee, superbowl, kneeling, anthem, footballers, halftimeshow, helments
9	Drug Overdose	fentanyl, overdose, lethal, meth, autopsy, overdosed, drugs, methamphetamine, fatal, toxicology

Table 3.17: The topics discovered by topic modeling and the most representative tokens in the 2022 non-offensive tweets.

3.6 Implications

Our analysis results contribute to the growing number of works in safety and security in social media, promoting healthy online conversations. Our findings hold substantial implications by offering potential insights for fostering more respectful and constructive discussions on social justice discussions in online spaces. The presence of offensive content in discussions related to BLM, which fights for racial and systemic injustice, further limits the goal and importance of the movement as exposure to such content can increase prejudice towards Blacks or African Americans, the movement or what the movement stands for or increase the lack of trust in authorities, especially the police offline [87, 145].

Our offensive and emotion analysis shows that negative emotions, anger, disgust,

and fear were frequently expressed, particularly anger and disgust were frequently expressed. The authors of offensive tweets likely tweeted out of anger in response to BLM-related discussions as "anger is the emotion of injustice" and a "powerful resource for resisting epistemic injustice" [17]. It does not mean that incivility should be tolerated; our results could guide strategies to moderate online discussions and foster a more healthy, respectful and constructive discussion without leading to tone policing [17, 112] as [51] states "when addressing and identifying forms of epistemic oppression one needs to endeavor not to perpetuate epistemic oppression".

Finally, the primary topics of these discussions, including police brutality and racial injustice, could provide insights and inform policymakers, activists, and community leaders as they address the expressed concerns and grievances. They can be used to gauge public opinions, which can help control the effect of information bias [86] and its impact on readers [222] so that readers' conscious or unconscious attitudes towards the Black community are not exacerbated [87]. Furthermore, we show through analysis of offensive tweets and topics how offensive content can be used to daunt others, possibly to deter them from expressing their opinions, thus preventing an open and productive conversation among users.

3.7 Limitations

We aimed to identify offensive content and the emotions expressed in the BLM movement. We discuss the limitations of this work.

Our offensive tweets classifier is not perfect due to the possibility of the sentiment features confusing the model. In analyzing our model's false negative and false positive predictions, we make the following observations. For false negatives, our model finds it difficult to classify hard to tell implicit offensive content. The following tweet, *"Black Lives Matter means Darkness Lives Matter... URL"* is predicted as non-offensive by the offensive model and predicted to have negative sentiment by the sentiment model. Even though the sentiment model predicted it as having negative sentiment, the offensive model

still misclassified the tweet. A similar observation is made for the tweet *"@USER George Floyd would've already died anyway from a fentanyl/meth overdose long before before Derek Chauvin arrived to tap on his windshield. Floyd was already zombified at that point"*. For false positives, the tweets *"@USER @USER Ahmaud Arbery was out taking a run, and didn't deserve to executed!"* and *"@USER @USER Prosecuting Derek Chauvin also won't bring back George Floyd."* are classified as having a positive sentiment by the sentiment model and the offensive model predicted the tweets as offensive even though both tweets are not offensive but have a negative sentiment. There are also explicit cases predicted as a negative sentiment that our model misclassifies as non-offensive even though the tweet has a negative sentiment and is known to be offensive upon review. For example, *"@USER Are you fucking serious dumbass #BlackLivesMatter"*.

Also, our emotion classifier can misclassify positive tweets as negative tweets or negative tweets as positive tweets, which could have affected our analysis. Another limitation of this work is that we do not consider the demographics (gender, race, and ethnicity) of tweet authors; a better understanding of the tweet content can be achieved by knowing the author of a tweet. We have used the default settings of BERTopic which could have affected the quality of the topics. Finally, our analysis does not consider whether offensive tweets originated from automated accounts.

3.8 Conclusions and Future Work

This research explored the presence of offensive language in BLM discussions in 2020 and years after, the emotions expressed in BLM discussions, and the main topics discussed in the identified offensive and non-offensive tweets. We identified offensive content in BLM-related discussions during the 2020 BLM protests and years after. Results indicate that the number of offensive tweets increased in the weeks following Floyd's death. The number significantly dropped and remained stable afterward. We found that negative emotions (anger, disgust, and fear) were the most expressed in the offensive tweets and were most

expressed in the week following Floyd's death. The topics discussed mainly focused on police brutality and systemic and racial injustice. These topics persisted after the 2020 BLM protests. We also found that most offensive tweets directed to users are unidirectional.

In the future, given the debate that stemmed from the BLM movement being met with opposing labels such as #AllLivesMatter, #WhiteLivesMatter, and #BlueLivesMatter, further research on the communities formed in the reply graph of such discussions and the topics discussed by the communities could lead to an understanding of how different communities discussed the movement offensively. Additionally, a directed network of the replies of the authors of offensive tweets and the receivers of offensive tweets could be studied to understand the types and behaviors of offensive users.

Chapter 4

AAEBERT: Debiasing BERT-based Hate Speech Detection Models via Adversarial Learning

This work has been presented and published in the International Conference on Machine Learning and Applications (ICMLA), 2022.

4.1 Abstract

Hate speech datasets contain bias which machine learning models propagate. When these models classify tweets written in African American English (AAE), they predict AAE tweets as hate/abusive at a higher rate than tweets written in Standard American English (SAE). This paper assesses bias in language models fine-tuned for hate speech detection and the effectiveness of adversarial learning in reducing such bias. We introduce AAEBERT, a pre-trained language model for African American English obtained by re-training BERT-base on AAE tweets. AAEBERT is used to extract the representation of each tweet in the

various hate speech datasets and to classify tweets into two classes - AAE dialect and non-AAE dialect. A three-layer feedforward neural network that takes the representation from AAEBERT and a dialect label as input is used as the adversarial network for debiasing. We evaluate bias in language models fine-tuned for hate speech detection. Then assess the effectiveness of adversarial debiasing in these models by comparing results before and after adversarial debiasing is applied. Analysis reveals that the fine-tuned models are biased towards AAE, and adversarial debiasing is effective in reducing bias.

4.2 Introduction

Social media enables instantaneous access to trending topics and news, provides a medium to keep in touch with friends, and means to connect with people outside our network. Social media platforms have made sharing, viewing, and subscribing to content relatively easy. While this is beneficial, some of the contents are hateful, and offensive [16], [32]. Even worse, when such contents are directed to specific groups, it can lead to social unrest [59].

Due to the rise of social media, researchers and social media platforms use automatic systems based on machine learning and deep neural networks to tackle the problem of hate speech detection. Detection methods based on traditional machine learning require feature engineering and do not generalize well to new datasets. Neural network-based word embeddings can automatically learn word representations reducing the amount of feature engineering needed. Most recently, pre-trained language models like the Bidirectional Encoder Representations from Transformers (BERT) [46] have been used to improve the performance of hate detection systems. Despite these successes, these models have a bias problem.

Hate speech detection models learn from annotated datasets and can assimilate the bias in these datasets. Training traditional machine learning classifiers on these datasets results in classifiers that are racially biased towards African American English (AAE) [42].

These datasets contain texts written in standard English, which differ in terms of syntax, phonology, and lexicon when compared to the AAE variety [26]. BERT [46] based language models fine-tuned on hate speech datasets for hate speech detection also propagate this racial bias [142].

To reduce racial bias, researchers in [234] used adversarial training [73] to demote the racial bias learned by a bidirectional long short term with attention (BLSTMA) model in a two step training procedure. Bias reduction have also been studied in a pre-trained BERT-base model by reweighting hate speech datasets and using the reweighted scores and dataset as model input during fine-tuning [142]. The current approach have focused on recurrent neural network based models [234] and BERT-based pre-trained model [142] without considering other pre-trained models used in hate speech detection.

Pre-training language models with domain-specific corpus have been used to address the problem of models trained on general knowledge corpus [117], [22]. Due to new hateful terms introduced during the pandemic, [123] developed COVID-HateBERT to detect general hate speech and COVID-19 related hate speech. HateBERT [34] was introduced for abusive language detection. AlBERTo [170] and BERTweet [150], are pre-trained language models for Italian and English tweets respectively.

Researchers have argued that the high rate of assigning AAE tweets to negative classes is due to the presence of AAE in the datasets and the high use of words like "n***a" in the AAE tweets [42]. Given the presence of AAE in hate speech datasets, we introduce AAEBERT, a pre-trained language model trained using AAE tweets. We study its effects and the effectiveness of adversarial learning in reducing racial bias in pre-trained models fine-tuned on hate speech datasets. To accomplish the goal of this work:

1. We fine-tune popular pre-trained language models (BERT, BERTweet, and HateBERT) commonly used in hate speech detection. On eight hate speech detection datasets and assess bias in the models obtained from fine-tuning each pre-trained language model on each dataset.

2. We introduce AAEBERT, a pre-trained language model based on BERT-base. AAEBERT is re-trained using 1.2M African-American English tweets and is a language model for African-American English.

3. We develop a three-layer feedforward neural network as the adversary network, trained during the fine-tuning process. Finally, we analyze the effect of adversary debiasing in the models obtained by comparing the models before and after applying adversarial debiasing.

4.3 Related Work

Social media enables communication, instant news, and socialization. Users of social media platforms like Twitter generate an enormous amount of content, some of which are hateful and cannot be efficiently moderated manually. Researchers have explored automatic methods based on machine learning [43] and deep learning techniques [16], [8], [246] to solve this problem.

Fine-tuning transformer-based models such as BERT [46] on downstream tasks have achieved impressive performance. However, they do not perform so well on specialized domains. For example using BERT which was pre-trained on general corpus in biomedical text mining produces undesired results due to the difference in word distribution in both the general corpus and biomedical corpus. To solve this problem, pre-training language models on datasets from specialized domains have been employed. COVID-HateBERT [123] was introduced for detecting Covid-19 related hate speech, ALBERTo-HS [169] for hate speech detection in Italian tweets. BERTweet [150] and ALBERTo [170] are both pre-trained language models for English and Italian tweets. Caselli et. al developed HateBERT [34], a model skewed towards social media and offensive, abusive, and hate related task.

Hate speech detection datasets contain systematic racial bias due to annotation as demonstrated by [42]. Models trained on these datasets automatically inherit the bias. There have been works to mitigate such bias, to mitigate the bias propagated from biased

datasets in trained models, researchers in [142] used different fine-tuning strategies to develop a BERT based hate speech detection model. They mitigate bias in the fine-tuned model by employing a re-weighting mechanism that re-weight the training and validation datasets.

Adversarial training championed by [73] have been used to train multiple networks with one network fooling the other to achieve its objective. Xia et. al [234] used adversarial training to mitigate bias towards AAE texts. They train a classifier that learns to detect hate speech while using an adversarial network to prevent the classifier from learning a representation predictive of AAE attribute. They use a bidirectional long short term memory (BiLSTM) with attention as a feature extractor and multilayer perceptrons (MLPs) as classifer and adversary. Model was trained using a two-phase training procedure.

Our work differs from [142] in the following ways: (1) we do not use a re-weighting mechanism on the training and validation sets but use a simple adversarial network to mitigate bias. (2) we introduce AAEBERT a retrained BERT model skewed towards AAE to extract useful represenation used in bias mitigation. (2) we assess and mitigate bias in other pre-trained language models (BERTweet and HateBERT) fine-tuned for hate speech detection, and extend our analysis to more than three datasets. In addition, (1) we do not apply a two-phase training procedure, rather we employ a simple debiasing procedure performed during model fine-tuning. (2) we use AAEBERT to infer AAE dialect instead of using the model introduced by [26].

4.4 Methodology

This section introduces our solution for debiasing language models fine-tuned on hate speech detection datasets. By exploring the representation of tweets from existing pre-trained language models and the representation of tweets from the AAEBERT model, we introduce a debiasing network with a fusion mechanism capable of reducing bias propagated by pre-trained language models fine-tuned for hate speech detection datasets. The general

description of the architecture is shown in Fig. 4.1. The input is a sequence tweet, the tweet's label, the tweet's AAE dialect label, and a sequence of the tweet's representation from AAEBERT.

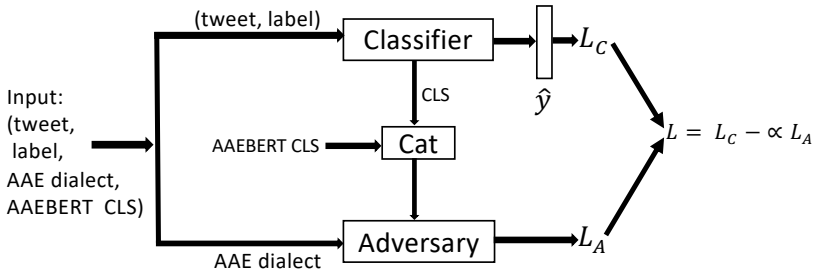


Figure 4.1: Illustration of the proposed debiasing architecture

Specifically, the classifier takes the sequence of a tweet and the tweet's label as input and learns to predict the label. The adversary input is the fusion (concatenation) of a tweet's representation from AAEBERT with the tweet's representation from the classifier and the dialect label (AAE dialect or not) assigned to the tweet by AAEBERT. The adversary learns to predict the dialect label. The network is trained end-to-end and can be divided into three components, the classifier, fusion, and the adversary. We introduce each of these components below.

4.4.1 AAEBERT

We use the huggingface library [232] to retrain BERT$_{base}$ using the race dataset described below. AAEBERT is optimized with Adam [103], trained for 100 epochs on 1 V100 GPU[1] using a batch size of 64 and a learning rate of 5e-5. AAEBERT is trained using the black-aligned corpus from the race dataset containing 1,288,525 tweets. Masked language modeling was used as the training objective. Masked language modeling enables the training of deep bidirectional representation by randomly masking some percentage of the input tokens with a [MASK] token in each input sequence and predicting the masked tokens.

[1]The V100 GPU is a shared resource with a wall time of 72 hours. It took nine days to complete the training.

4.4.2 Classifier

We assess bias by fine-tuning three popular pre-trained language models to detect hate or abusive speech. The models are fine-tuned on datasets annotated for hate speech or abusive language. The following models are considered; First is BERT_{base} [46], pre-trained on BookCorpus and English Wikipedia corpus, and the second is BERTweet [150], pre-trained on a large corpus of tweets. Finally, HateBERT [34] pre-trained on Reddit comments from communities banned for writing hateful or offensive comments. The input to the classifier is a tweet sequence and its corresponding label. The first token in the sequence is a special classification token ([CLS]). The [CLS] token in the final hidden layer is used as the representation of the entire sequence and passed to an output layer for classification. The classifier takes as input a sequence of a hateful or offensive tweet and its label and learns to predict the label well. The classifier is trained using the cross-entropy loss.

4.4.3 Race dataset

We use the African American English (AAE) dataset developed by [26]. Blodgett et al. [26] created the AAE dataset by collecting tweets from the Gardenhose/Decahose Twitter archive and mapped tweet authors to the demography of the location they lived in using the authors' tweet geolocation. The mapping is done by looking up the US Census block group geographic area from which the message originated and using the race information associated with the block group. They defined four covariates - the percentage of non-Hispanic whites, non-Hispanic blacks, Hispanics, and Asians. They utilized a mixed membership demographic-language probabilistic model, which learns a demographic-aligned language model for each demographic category. The model calculates the posterior proportion of language from each category in each tweet. The dataset contains 59.2 million tweets generated by 2.8 million users.

Following [42] and [142], we create a black-aligned corpus by filtering tweets with an average posterior proportion > 0.80 for the non-Hispanic black category and < 0.10 when

the Hispanic and Asian categories are combined—enabling us to obtain tweets written by users who use AAE. Similarly, we create a white-aligned corpus obtained by filtering tweets with an average posterior proportion > 0.80 for the non-Hispanic white category and < 0.10 when the Hispanic and Asian categories are combined. After extraction, we obtained 1,314,176 and 16,077,312 tweets written by non-Hispanic blacks and non-Hispanic whites, respectively. From the black-aligned and white-aligned corpus, we sampled 1000 tweets (999 tweets after preprocessing) used to assess bias in our fine-tuned models. After excluding the sampled tweets and preprocessing, the black-aligned corpus used to train AAEBERT had 1,288,525 tweets.

4.4.4 Fusion

The 768 dimensional CLS representation from the last layer of the classifier is concatenated with the CLS representation from AAEBERT, producing a 1536 dimensional vector representation. This representation serves as the input to the adversary. We experimented with different representations as input to the adversary, using the classifier output logits and classifier [CLS] representation separately as input. The fusion between the classifier [CLS] and AAEBERT [CLS] representation performed the best, and only those results are presented because of space constraints.

4.4.5 Adversary

The adversary is a three-layer feedforward neural network with a Leaky ReLU activation function. The first and second layer contains 256 and 100 neurons, respectively, and the output layer contains two neurons with softmax function. The adversary learns to predict the AAE/Non-AAE dialect of a tweet and uses the cross-entropy loss. It takes as input the fused representations from the classifier and AAEBERT and a label indicating if a tweet is AAE or Non-AAE dialect. Similar to the setup in [221] and [80], the adversary optimizes the equation $L = L_C - \alpha L_A$ as its objective. The variable L_C is the classifier loss, L_A is the adversary loss, and α is a hyperparameter that controls the rate at which

Dataset	Count
Waseem and Hovy [226]	10338
Waseem [224]	5988
Davidson et. al [43]	24773
Golbeck et. al [71]	20305
Founta et al. [62]	45549
OffensEval 2019 [241]	14100
AbusEval [36]	14100
HatEval [20]	11991

Table 4.1: Datasets used in our work

the adversary is maximized, and the classifier minimized. The goal is for the classifier to learn to predict hate speech well and for the adversary to not perform well in predicting AAE dialect from the fused representations.

4.5 Experiments

4.5.0.1 Datasets

We use English Twitter datasets for fine-tuning our models and briefly describe each in this section.

Waseem and Hovy [226] collected 136,052 and labeled 16,914 tweets into three categories - racism, sexism, and neither. After preprocessing, 10,338 tweets are obtained.

Waseem [224] investigated the effect of using datasets annotated for hate speech by experts (feminist and anti-racism experts) and amateurs (recruited from CrowdFlower) on classification models. The dataset contains 5,988 tweets after preprocessing and has 4 classes, racism, sexism, racism and sexism, and neither.

Davidson et. al [43] studied the distinction between hate speech and offensive language by extracting 24,802 tweets labeled into three classes, hate, offensive, and none. After preprocessing, the dataset contained a total of 24,773 tweets.

Golbeck et. al [71] developed a hand-labeled dataset of online harassment containing 35,000 tweets with 20,305 tweets remaining after preprocessing.

Founta et. al. [62] sort out to solve the challenge of having different but related labels (hate, offensive, cyberbullying, and aggressive) in hate speech and developed a dataset

containing 80K tweets. After rehydrating and preprocessing, we obtained 45,549 tweets labeled as abusive, hateful, spam, or normal. In our experiments, we do not consider the spam class.

OffensEval 2019 [241] uses the Offensive Language Identification Dataset (OLID) [240] which is the main dataset used in the SemEval 2019 Task 6 (OffensEval[2]) competition. The dataset contains 14,100 tweets after preprocessing.

AbusEval, [36] created AbuseEval v1.0 dataset, which is the same dataset created by [240] but with the introduction of new labels (implicit and explicit abuse).

The HatEval dataset is the primary dataset used in the SemEval 2019 Task 5, focusing on detecting hate towards women and immigrants on Twitter [20].

4.5.0.2 Data preprocessing

Before training AAEBERT and fine-tuning our models, we preprocessed our datasets by removing duplicate tweets and tweets with two or fewer words. The dataset was normalized by replacing user mentions with @USER, URLs with URL, and emojis with text representation using the Python emoji package. Additional processing of the dataset included removing emojis and hashtags, converting tweets to lowercase, replacing extra blank spaces with a single space, and removing additional empty newlines.

4.5.0.3 Model fine-tuning

The fine-tuning uses the train and test splits provided in the datasets above. For the datasets without train and test splits, we randomly split the entire corpus into two sets; train - 80% and test - 20%. The hate datasets do not have an AAE dialect label for adversarial training. To provide this label, AAEBERT is utilized to classify hateful tweets into two classes - AAE dialect and non-AAE dialect. We fine-tune each of the classifiers described in Section 4.4-A on each dataset for each fine-tuning configuration. Evaluate the performance of the fine-tuned models and access bias without applying adversarial debiasing (as shown

[2]https://competitions.codalab.org/competitions/20011

in Fig. 4.1). Then the process is repeated with adversarial debiasing (as shown in Fig. 4.1). The bias rate is compared as defined in Section 4.5.0.4 with and without adversarial debiasing. To assess bias, we reinitialize our network and we pass the sampled (999 tweets) black-aligned and white-aligned tweets through the classifier (see the top section of Fig. 4.1). We use $\alpha = 1$, experimented with different values of alpha and discuss its effects in Section 4.6.1.

4.5.0.4 Measuring bias

To assess the rate at which fine-tuned models assign black and white aligned tweets to negative classes and the effects of adversarial debiasing in reducing bias, we calculate the percentage of tweets from each racial group assigned to a negative class i. Let 1 represent a tweet that belongs to a negative class c_i and 0 otherwise. Also, let $P(c_i = 1|black) = P(c_i = 1|white)$ denote the null hypothesis that the probability of a tweet belonging to a negative class is independent of the race of the tweet's author. The null hypothesis is rejected in favor of the second hypothesis that black aligned tweets are assigned to a negative class at a higher rate than white aligned tweet or vice versa. The second hypothesis is defined as: $P(c_i = 1|black) > P(c_i = 1|white)$ or $P(c_i = 1|black) < P(c_i = 1|white)$.

We assess the effectiveness of AAEBERT and adversarial debiasing using the black-aligned and white-aligned tweets randomly sampled from the race dataset. Each tweet is passed through each fine-tuned model to predict its probability of belonging to each class. The assessment is repeated with and without applying adversarial debiasing. For each fine-tuned model, a vector of dimension n (the number of tweets in each race dataset) is created containing the probability p_i of belonging to each class i. The proportion of tweets belonging to class i for the black and white groups are given by $\widehat{p_{iblack}} = \frac{1}{n}\sum_{j=1}^{n} p_{ij}$ and $\widehat{p_{iwhite}} = \frac{1}{n}\sum_{j=1}^{n} p_{ij}$ respectively. If the ratio $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ is greater than 1, then black-aligned tweets are assigned to class i at a higher rate than white-aligned tweets. A t-test between $\widehat{p_{iblack}} = \widehat{p_{iwhite}}$ is conducted. P values < 0.001 is indicated with stars (***) and no stars indicate p values > 0.05 in the result tables in Section 4.6.

4.6 Results

This section discusses the results of our experiments. The goal is to study racial bias in fine-tuned language models, the effectiveness of adversarial training, and AAEBERT representation in mitigating racial bias in these models.

We evaluate the performance of models fine-tuned for hate speech detection without adversarial training, assess bias in these models, and evaluate the effectiveness of adversarial training in bias mitigation. Table 4.2 shows the performance evaluation result of the models without adversarial debiasing in terms of macro averaged F1 score, precision, and recall. From Table 4.2, we observe that the models fine-tuned on the Waseem [224] dataset had the least performance compared to models fine-tuned on other datasets in terms of F1 score. The AAEBERT model had the best F1 score. The low performance of the models fine-tuned on the Waseem [226] dataset is as expected given the dataset size. The AAEBERT and HateBERT models performed the best on the Waseem and Hovy [226] dataset. BERTweet obtained the best F1 score in the Davidson et al. [43], Founta et al. [62], and HatEval [20] datasets and tied with HateBERT on the Golbeck et al. [71] dataset. HateBERT outperformed other models in the OffensEval 2019 [241] and AbusEval [36] datasets and is as competitive as AAEBERT on the Waseem and Hovy [224] dataset.

Tables 4.3, 4.4, and 4.5 show the results of racial bias in these models. From the "without adversarial debiasing" column in Table 4.3, we observe that the $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ value of the negative classes of the datasets we fine-tuned on is > 1, indicating that the models are biased as they assigned black-aligned tweets to negative classes at a higher rate than white-aligned tweets. The exception to this is in the racism class of the Waseem and Hovy [226] dataset with a $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ value of 0.971. Indicating the opposite, the model assigned white-aligned tweets to the racism class at a higher rate than black-aligned tweets. When adversarial debiasing is applied as seen in the "with adversarial debiasing" column, bias is reduced to some extent. The highest bias reduction occurred in the racism, sexism, and racism and sexism classes of the Waseem dataset [224] with $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ values of 2.9%, 5.2%, and

Dataset	Model	F1	Precision	Recall
Waseem	BERT	0.403	0.406	0.404
	BERTweet	0.399	0.407	0.402
	HateBERT	0.413	0.415	0.414
	AAEBERT	**0.462**	0.425	0.429
Waseem and Hovy	BERT	0.557	0.566	0.561
	BERTweet	0.551	0.563	0.556
	HateBERT	**0.566**	0.555	0.559
	AAEBERT	**0.566**	0.557	0.561
Davidson et al.	BERT	0.794	0.745	0.763
	BERTweet	**0.803**	0.775	0.787
	HateBERT	0.797	0.745	0.764
	AAEBERT	0.785	0.731	0.749
Golbeck et al.	BERT	0.689	0.618	0.628
	BERTweet	**0.707**	0.628	0.640
	HateBERT	**0.707**	0.626	0.638
	AAEBERT	0.699	0.610	0.618
Founta et al.	BERT	0.761	0.712	0.726
	BERTweet	0.766	0.721	0.733
	HateBERT	**0.776**	0.712	0.727
	AAEBERT	0.775	0.707	0.723
OffensEVal 2019	BERT	0.828	0.810	0.817
	BERTweet	0.813	0.798	0.804
	HateBERT	**0.832**	0.800	0.813
	AAEBERT	0.815	0.777	0.792
AbusEval	BERT	0.571	0.528	0.533
	BERTweet	0.528	0.542	0.533
	HateBERT	**0.688**	0.533	0.537
	AAEBERT	0.600	0.521	0.528
HatEval	BERT	0.682	0.574	0.457
	BERTweet	**0.702**	0.612	0.522
	HateBERT	0.684	0.585	0.480
	AAEBERT	0.687	0.581	0.470

Table 4.2: Evaluation results of fine-tuned models on each hate speech dataset without applying adversarial debiasing

		Without adversarial debiasing					With adversarial debiasing				
Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.074	0.056	2.099	0.0360	1.328	0.120	0.098	4.482	***	1.224
	Implicit	0.035	0.025	3.61	***	1.398	0.044	0.039	4.94	***	1.145
OffensEval	Offensive	0.346	0.204	8.707	***	1.692	0.358	0.259	8.854	***	1.381
HatEval	Hate	0.147	0.089	5.619	***	1.649	0.282	0.196	12.131	***	1.439
Davidson et al.	Hate	0.018	0.014	1.733		1.313	0.083	0.082	1.657		1.016
	Offensive	0.357	0.139	13.423	***	2.572	0.498	0.316	16.618	***	1.576
Founta et al.	Hate	0.070	0.033	5.192	***	2.144	0.051	0.044	11.701	***	1.153
	Abuse	0.257	0.125	8.651	***	2.055	0.235	0.154	9.638	***	1.522
Waseem and Hovy	Racism	0.004	0.004	-0.749		**0.971**	0.004	0.004	2.1	0.0360	1.014
	Sexism	0.209	0.065	11.987	***	3.234	0.155	0.090	10.24	***	1.734
Waseem	Racism	0.017	0.004	15.69	***	4.069	0.013	0.011	14.79	***	1.164
	Sexism	0.111	0.016	14.284	***	6.763	0.068	0.045	14.542	***	1.511
	Racism and Sexism	0.008	0.003	14.365	***	2.989	0.007	0.006	14.726	***	1.211
Golbeck	Harassment	0.093	0.068	4.45	***	1.365	0.226	0.203	11.432	***	1.114

Table 4.3: Racial bias analysis of fine-tuned BERT models. Showing result with and without adversarial debiasing

1.7% respectively. The $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ value of the racism class of Waseem and Hovy [226] became slightly greater than 1 which could indicate that adversarial debiasing is trying to equalize $\widehat{p_{iblack}}$ and $\widehat{p_{iwhite}}$.

Table 4.4 shows the results of fine-tuning BERTweet on different datasets. Similar to the fine-tuned BERT models in Table 4.3, without adversarial debiasing, $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ values of the classes in the datasets are greater than 1 indicating that fine-tuned BERTweet models are biased towards black-aligned tweets. Except for the implicit class of the AbusEval dataset [36]. With adversarial debiasing applied, we observe reduction in $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ values. The $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ are also slightly greater than 1 indicating that adversarial debiasing is effective in reducing bias and in achieving equality between $\widehat{p_{iblack}}$ and $\widehat{p_{iwhite}}$ equal. With the exception of the models fine-tuned on the OffensEval [241] and HatEval [20] datasets though bias is reduced from 1.7 to 1.2 and from 2.022 to 1.238 respectively. After adversarial debiasing, the model obtained from fine-tuning BERTweet on the Davidson et al. [43] dataset obtained $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}} < 1$ in the hate class. Indicating that the model became more biased towards white-aligned tweets than black-aligned tweets. The top 3 reduction in bias occurred in the offensive class of Davidson et al. [43], hate class of Founta et al. [62], and sexism class of Waseem [226] datasets with a $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ value reduction of 1.5%, 1.49%, and 1.41% respectively.

		Without adversarial debiasing					With adversarial debiasing				
Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.083	0.058	2.747	0.006	1.426	0.133	0.127	3.344	***	1.047
	Implicit	0.015	0.016	-0.677		**0.951**	0.050	0.049	3.411	***	1.019
OffensEval	Offensive	0.358	0.208	9.676	***	1.724	0.354	0.295	9.322	***	1.200
HatEval	Hate	0.107	0.053	6.352	***	2.022	0.265	0.214	8.93	***	1.238
Davidson et al.	Hate	0.020	0.015	1.928		1.321	0.059	0.063	-21.651	***	**0.929**
	Offensive	0.476	0.179	17.681	***	2.653	0.744	0.682	22.506	***	1.091
Founta et al.	Hate	0.096	0.038	7.857	***	2.527	0.044	0.043	16.447	***	1.036
	Abuse	0.302	0.148	10.13	***	2.045	0.215	0.199	16.111	***	1.079
Waseem and Hovy	Racism	0.004	0.004	3.581	***	1.071	0.011	0.011	5.958	***	1.028
	Sexism	0.080	0.040	5.34	***	1.98	0.174	0.161	5.25	***	1.078
Waseem	Racism	0.009	0.006	7.389	***	1.538	0.016	0.015	5.827	***	1.043
	Sexism	0.046	0.019	6.483	***	2.492	0.060	0.056	5.148	***	1.074
	Racism and Sexism	0.006	0.004	6.943	***	1.436	0.013	0.013	5.427	***	1.038
Golbeck	Harassment	0.091	0.079	2.227	0.026	1.144	0.238	0.235	3.831	***	1.014

Table 4.4: Racial bias analysis of fine-tuned BERTweet models. Showing result with and without adversarial debiasing

		Without adversarial debiasing					With adversarial debiasing				
Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.079	0.063	1.830		1.254	0.119	0.103	2.809	0.005	1.154
	Implicit	0.027	0.025	0.949		1.078	0.045	0.042	4.498	***	1.068
OffensEval	Offensive	0.346	0.218	8.439	***	1.587	0.370	0.305	7.157	***	1.215
HatEval	Hate	0.155	0.091	6.665	***	1.690	0.305	0.236	10.017	***	1.290
Davidson et al.	Hate	0.016	0.014	1.2		1.162	0.074	0.074	-0.592		**0.995**
	Offensive	0.369	0.139	14.37	***	2.651	0.527	0.339	17.679	***	1.555
Founta et al.	Hate	0.070	0.029	6.171	***	2.468	0.054	0.046	13.054	***	1.163
	Abuse	0.274	0.134	9.229	***	2.051	0.262	0.17	10.483	***	1.542
Waseem and Hovy	Racism	0.004	0.003	5.776	***	1.202	0.004	0.003	7.618	***	1.089
	Sexism	0.145	0.055	9.415	***	2.636	0.147	0.102	8.491	***	1.433
Waseem	Racism	0.019	0.005	11.523	***	3.708	0.012	0.011	11.295	***	1.125
	Sexism	0.052	0.014	9.461	***	3.652	0.058	0.046	10.701	***	1.255
	Racism and Sexism	0.009	0.003	10.498	***	2.533	0.008	0.007	10.701	***	1.124
Golbeck	Harassment	0.084	0.070	2.629	0.009	1.203	0.251	0.249	5.139	***	1.007

Table 4.5: Racial bias analysis of fine-tuned HateBERT models. Showing result with and without adversarial debiasing

Results of fine-tuning the HateBERT model is shown in Table 4.5. In all classes, $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}} > 1$ indicating that fine-tuned HateBERT models are biased. After adversarial debiasing is introduced, bias is reduced in all classes except for the Hate class of the Davidson et al. [43] dataset with a $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ value of 0.995. Indicating that the fine-tuned HateBERT model on the Davidson et al. [43] dataset assigned white-aligned tweets to the hate class at a higher rate than black-aligned tweets. The top three highest reduction of $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ values occurred in the Waseem [226] dataset with 2.5%, 2.3%, and 1.4% reduction in the racism, sexism, and racism and sexism classes respectively.

4.6.1 Effect of α value

The α value determines the rate at which bias is reduced. This section investigates the effects of α on bias reduction and performance when adversarial debiasing is applied. For each of the pre-trained language models, we evaluate the models obtained from fine-tuning when adversarial debiasing is applied using α values of 0.01, 0.05, 0.09, 0.1, 0.3, 0.5, 0.8, and 1. We do not show the results of these experiments due to space. The following are the conclusions from the experiments. First, when $\alpha < 0.5$, performance is as competitive as when the pre-trained models are fine-tuned without adversarial debiasing. There is a reduction in performance when $\alpha \geq 0.5$. Having $\alpha \geq 0.8$ leads to a better reduction in bias. We use $\alpha = 1$ in fine-tuning pre-trained models when adversarial debiasing is applied because it showed a better reduction in bias, and its performance is as competitive as $\alpha = 0.8$. Second, bias reduction is slow and starts improving from $\alpha \geq 0.8$.

4.7 Conclusion

This paper presents an adversarial debiasing network for debiasing BERT-based hate speech detection models. We introduced AAEBERT, a pre-trained language model based on BERT-base for African-American English, and assessed bias in three pre-trained language models used in hate speech detection. We assessed the effect of adversarial de-

biasing in reducing bias by utilizing tweet representations from AAEBERT and fine-tuned pre-trained language models. Bias assessment of fine-tuned models without adversarial debiasing indicates that fine-tuned models are more biased towards AAE than Standard American English (SAE). Analyses of fine-tuning with and without adversarial debiasing, show that adversarial debiasing is effective in reducing bias by achieving $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}} \approx 1$ in some models. Equalization is observed in fine-tuned pre-trained models when α approaches 1.

Chapter 5

Large Language Model Annotation Bias in Hate Speech Detection

This work is in submission at the International AAAI Conference on Web and Social Media (ICWSM), 2025.

5.1 Abstract

Large language models (LLMs) are fast becoming ubiquitous and have shown impressive performance in various natural language processing (NLP) tasks. Annotating data for downstream applications is a resource-intensive task in NLP. Recently, the use of LLMs as a cost-effective data annotator for annotating data used to train other models or as an assistive tool has been explored. This paper examines the risk of using LLMs for data annotation in hate speech detection. We investigate how hate speech detection datasets annotated using GPT-4 can lead to racial bias in online hate detection classifiers. We use GPT-4 to re-annotate seven hate speech detection datasets and trained classifiers on these datasets. Using tweets written in African-American English (AAE) and Standard American English (SAE), we show that classifiers trained on GPT-4 annotated datasets assign tweets written in AAE to negative classes (e.g., hate, offensive, abuse, racism, etc.) at a higher

rate than tweets written in SAE and that the classifiers have a higher false positive rate towards AAE tweets. We explore the effect of using dialect priming in GPT-4 annotation, showing that introducing dialect increases the rate at which AAE tweets are assigned to negative classes.

5.2 Introduction

LLMs have achieved human-level performance in several NLP tasks [65, 198] due to the transformer model architecture [214] and the technological advancement of GPUs and TPUs [69]. The transformer architecture lifted the limitations of recurrent neural networks, such as long short-term memory (LSTM), in modeling dependencies in long sequences and parallelism by relying solely on attention mechanisms. LLMs are pre-trained on large amounts of texts using unsupervised learning via masked language modeling for BERT-based models [46] and next token prediction for autoregressive generative models [125]. The improved performance and sample efficiency of language models on NLP tasks have been attributed to the scaling up in size of these language models [96, 207]. For example, the autoregressive generative model, generative pre-training transformer (GPT) is a model of 117M parameters, succeeded by a 1.5B parameter model GPT-2, followed by GPT-3 with 175B parameters.

LLMs require large pretraining datasets obtained from crawling the internet [172, 173] to learn world knowledge and to prevent overfitting. However, the problem with such datasets is that they contain texts that exhibit biases or stereotypes observed in our society, which has negative implications when models trained on such data are used in real-world applications such as in hate speech or toxicity detection [66]. In hate speech detection on online social media platforms, machine learning models aim to detect hate speech or offensive language towards individuals or groups belonging to a protected category [43, 225]. Research has shown that language models trained on biased datasets propagate these biases [142, 158]. One of the biases propagated by language models fine-tuned for

hate speech detection is racial bias, where the models discriminate against tweets written in African American English (AAE) at a higher rate than tweets written in Standard American English (SAE), meaning that the models fail at protecting the group they were designed to protect [142, 158]. AAE is a dialect of American English with defined syntactic-semantic, phonological, and lexical features [27].

The problem of hate speech detection is non-trivial because of the subjectivity of the task due to individual, cultural, regional, and language differences [43, 195], and the difference in the definition of similar, yet, different phenomena (offensive language, abusive language, aggression, etc.,) [36, 225]These differences limit the generalizability [14, 76] of models due to the lack of a comprehensive dataset labeled into different categories of hate speech since data collection and data annotation of hate speech is costly [223] and have adverse health effects on annotators [216].

With the success of LLMs and how they are becoming ubiquitous, especially the ChatGPT model [192] with 100 million monthly active users in January 2023 [130], an iteration of the InstructGPT model [162] trained using reinforcement learning from human feedback (RLHF), GPT-4 [161], and most recently GPT-4o [160]. The rise in their popularity indicates its potential to be utilized in different applications. Researchers have explored the use of GPT-3 as a low-cost data labeler [223], leveraged ChatGPT for the annotation of implicit hate speech [88], and studied the limitations of using ChatGPT for data annotation [204]. The use of GPT-3 annotated data to train downstream models has been explored by researchers [223]; despite the success of LLMs, they have a high loss in quality compared to state-of-the-art (SOTA) methods in difficult and pragmatic tasks [104] such as aggression and especially for emotion classification task. A significant limitation of using LLMs for annotating datasets used in the training of hate speech detection models is the likely introduction of bias in the models and the propagation of introduced bias by the models, which could lead to the marginalization of already marginalized social groups [25, 28, 42, 142, 158].

In this work, we conduct a systematic analysis focusing on how LLMs, specifically GPT-4, when used for data annotation in hate speech detection, can propagate racial bias

in downstream models trained on the GPT-annotated data. We focus on evaluating hate speech detection classifiers trained on GPT-4-annotated datasets. We show that utilizing GPT-4 for data annotation can introduce racial bias in annotation tasks and downstream models. Our research aims to help maintain civility in conversation on social platforms while highlighting the benefits and, importantly, the risks of using LLMs in annotating data for hate speech detection. We summarize our main contributions below:

- We use GPT-4 to re-annotate seven hate speech detection datasets collected from Twitter using three prompting techniques (general, few-shot learning, and chain-of-thought reasoning).
- We fine-tuned three pre-trained language models often used in the hate speech literature on the re-annotated datasets under each prompting technique and measured racial bias in each model. Specifically, we focus on the racial disparity between text written in AAE and SAE in models trained on GPT-annotated datasets.
- We evaluate racial bias using a corpus of demographically aligned tweets to show how each classifier performs on AAE and SAE tweets and AUC-based metrics to calculate the false positive rates of each classifier on the test sets conditioned on dialect.

Extensive evaluation of 63 (seven datasets, three models, and three prompting techniques) classifiers shows evidence of racial bias across all the classifiers and prompting techniques, with AAE tweets assigned to negative classes at a higher rate than SAE tweets and the models having more false positives for AAE tweets than SAE tweets. Compared to models trained on human-annotated data, models trained on GPT-4 annotated data can increase the rate of classifying AAE tweets to negative classes. We expect that if GPT-4 is used for data annotation and subsequently to train downstream models deployed in the field, the models will discriminate against those who write in AAE, most who are African-American known to experience racial discrimination in a wide range of applications such as housing [134] and criminal justice [178] which feeds into the ideology and stereotypes about African-Americans [23, 67].

5.3 Related Work

5.3.1 Hate Speech Detection

The problem of hate speech in online social platforms as a critical threat [111] has been tackled by researchers using traditional machine learning approaches [61, 191], deep neural network methods [16, 129], transformer-based approaches such as the use of BERT in [79, 126], COVID-Twitter-BERT in [124], and the retraining of BERT_base on COVID-19 related hateful tweets and on posts from banned Reddit subcommunities to produce COVID-HateBERT [123] and HateBERT [35] respectively. Most recently, the capabilities of LLMs have been explored in hate speech detection [79, 82, 104, 122, 217].

5.3.2 Racial Bias and Toxicity in Language Models

Past work has studied racial bias in machine learning [42] and deep learning models [142, 158, 234]. Using a regularized logistic regression model trained on hate speech dataset, [42] assessed racial disparity in tweets written in AAE and SAE using demographically aligned Twitter corpus [26]. For deep learning, [234] used an adversarial network to demote the dependence of a bidirectional LSTM encoder and multi-layer perceptron (MLPs) based hate speech classifier on protected attributes in a two-phased setting to reduce the high rate of false positives for AAE tweets. With the arrival of BERT-based language models, [142] assessed racial bias using the demographic data set by [26] in a BERT [46] model fine-tuned on hate speech datasets using a reweighting mechanism as regularization during fine-tuning. In [158], the authors assessed and reduced racial bias in various BERT-based models. They re-trained BERT [46] on a subset of the demographic data set [26] to obtain AAEBERT. Using the representation of tweets from AAEBERT and the BERT-based models being fine-tuned in an adversarial setting, they reduced racial bias towards AAE tweets. More recently, [85] demonstrated dialect prejudice in LLMs using matched guise probing. They showed the potential harm that could result in using LLMs to make decisions about people based on their language. Results indicate that speakers of AAE are

more likely to be assigned less attractive jobs, be convicted of crimes, and be sentenced to death by LLMs due to the features of AAE than speakers of SAE. They also found that methods for mitigating covert and overt stereotypes in LLMs, especially LLMs trained using human feedback, exacerbate but do not alleviate them. The authors in [66] introduced the RealToxicityPrompt data set of sentence prefixes paired with their toxicity score from Perspective API[1]. With the data set, they showed that pre-trained autoregressive language models can be prompted to generate toxic text even with non-toxic prompts and explored effective detoxification methods, concluding that toxic generation persists even in the best method (domain-adaptive pretraining on non-toxic corpus), hindering the safe deployment of language models. Using the same dataset, [189] focusing on the generation of biased text by GPT-2 and T5 [173] demonstrated that language models are aware of their biases and the toxicity of the text they generated. They developed a decoding algorithm that uses the description of an undesired outcome to reduce the probability of language models generating biased text.

5.3.3 Large Language Model Annotation

Researchers have evaluated the performance of LLMs as annotators for various NLP tasks. From using GPT-3 with different annotation strategies for annotating multiple NLP tasks (from sentiment analysis to named entity recognition) and fine-tuning BERT_base model on the GPT-annotated data [49] to using ChatGPT as an assistive tool during annotation [139], and to using ChatGPT as the sole annotator [68, 88]. The study by [207] suggests using GPT-3 for downstream tasks. However, the study by [223] indicates that downstream models such as PEGASUS_large and RoBERTa_large fine-tuned with GPT-3 annotated datasets produce good performance suggesting that using GPT-3 directly for downstream tasks may not produce the best performance. The potential of ChatGPT for data annotation tasks was demonstrated in [88], where the effectiveness of ChatGPT in classifying implicit hate speech and the quality of ChatGPT's natural language expla-

[1] https://github.com/conversationai/perspectiveapi

nations (NLEs) of implicit hate speech was investigated. ChatGPT was used to classify instances in the LatentHatred dataset, and human raters were employed to measure the quality of ChatGPT's classifications and explanations. Results indicate that ChatGPT, with 80% agreement with humans, is effective in implicit hate speech classification. For the 20% disagreement, ChatGPT's results more likely align with human perception. The NLEs generated by ChatGPT reinforce human perception. They are clearer than human-written NLEs, which can be detrimental when using ChatGPT for data annotation because laypeople can be easily misled when ChatGPT is wrong. Whether LLMs such as ChatGPT can replace humans in data annotation tasks was studied in [204]. The researchers argued that LLMs can inherit the bias in their training datasets, which can reduce data annotation quality. They further argued that LLMs trained on general knowledge datasets may not have the domain-specific knowledge to perform specialized annotation, which can result in inaccurate annotation.

In contrast to our work, all these ideas focus on evaluating models trained on human-annotated hate speech data sets for racial bias or evaluating socially undesirable attributes or biases in text generated by generative models. The works on data annotation explored the effectiveness of ChatGPT in classifying implicit hate, the possibility of ChatGPT-like LLMs replacing human annotators, and the implications of using ChatGPT-like LLMs for data annotation. While they discussed the advantages and disadvantages of using ChatGPT, the empirical evidence of its negative implications is lacking in terms of racial bias. Our work differs from these works by qualitatively showing evidence of racial bias in using LLMs such as GPT-4 to annotate data used in downstream models, specifically in hate speech detection.

5.4 Methodology

This section details our methodology for assessing bias in downstream models trained on LLM-annotated hate speech detection datasets.

5.4.1 Data

We utilize two types of datasets in our study, as shown in Fig 5.1. The race dataset described below is used to extract datasets used in training an AAE language model and a dialect classifier. The hate speech datasets described below are used for training hate speech classifiers.

5.4.1.1 Race Dataset

Following [42, 142, 158], we use a race dataset introduced by [26] to measure racial bias in classifiers trained on LLM annotated datasets and to train a dialect classifier. Blodgett *et al.* [26] collected tweets mapped according to the location the tweet author lived in using the geo-location published by the tweet author. They matched each tweet to the US Census block group they were sent in. Using the race and ethnicity information from the US Census block group, each user is mapped as non-Hispanic whites, non-Hispanic blacks, Hispanics, and Asians. They trained a mixed-method probabilistic model that learns demographically-aligned language models for each demographic. Each model calculates the posterior probability of using a language in a tweet. Their linguistic analysis reveals that AAE with a high posterior probability of non-Hispanic black language contain phonological, syntactic, and lexical variations that differ from SAE. The dataset contained 59.2 million tweets. We follow the authors in [142, 158] and extract a black-and-white corpus containing tweets with a posterior probability > 0.8. We obtained a black-and-white corpus of 1.28M and 19M tweets, respectively. From the black corpus, we randomly sample two sets of tweets. The first set contains 1.22M tweets used to train an AAE language model [158] as described in Section 5.4.2; from the remaining set of tweets, we randomly sample the second set containing 1K tweets which we call *black-aligned* tweets used to fine-tune a dialect classifier described in Section 5.4.2 and for racial bias assessment. We randomly sample 1K tweets from the white corpus, which we call *white-aligned* tweets also used to fine-tune a dialect classifier and for racial bias assessment. Fig. 5.1 shows the data extraction process of the white-aligned and *black-aligned* tweets.

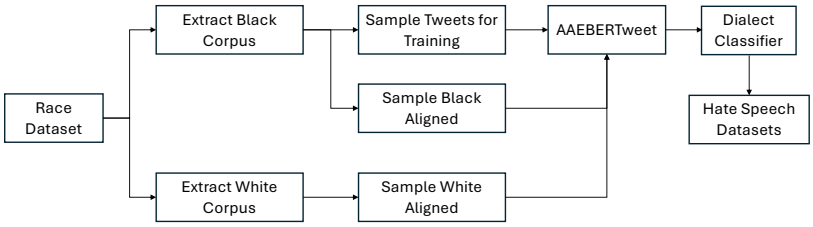


Figure 5.1: An overview of the white-aligned and black-aligned datasets creation from the race dataset.

5.4.1.2 Hate Speech Datasets

To understand the racial bias in hate speech detection models trained on LLM annotated data, we focus our analyses on seven corpora of tweets [20, 36, 43, 62, 71, 224, 239] widely used in hate speech detection [35, 97, 118, 123, 163, 212] and racial bias assessment [42, 142, 158, 186]. The statistics of these datasets are shown in Table 5.1[2]. We utilize random samples of these datasets described below in our experiments due to the cost of annotating large samples with OpenAI's GPT-4 using different prompting strategies for annotation discussed in Section 5.4.3. Most of the datasets were annotated by crowd-sourcing.

Waseem [224] sampled and annotated 6,900 tweets from the [226] dataset using amateur annotators (from the crowd-sourcing platform, CrowdFlower) and expert annotators (feminist and anti-racism activists). They annotated tweets into four classes: racism, sexism, both (racism and sexism), and neither, of the 62 racism, 530 sexism, 24 racism and sexism, and 5,372 neither tweets. We randomly sampled 500 tweets from neither class and retained all tweets in the other classes.

Davidson [43] collected 25K tweets using a set of hate speech lexicons from HateBase (a database of hate speech terms) and labeled the tweets into three categories, hate speech, offensive (but not hate speech), or neither using CrowdFlower workers. We randomly sampled 500 tweets from each class.

Golbeck [71] introduced a dataset containing 35K (20K after duplicates are re-

[2]After tweet rehydration, the total count of tweets in some datasets does not sum to the count originally published because some tweets have been removed by X (Twitter).

moved) tweets using a set of phrases related to Islamophobia, anti-semitism, homophobia, anti-black racism, and sexism. Two annotators labeled the tweets into overlapping categories: threats, hate speech, potentially offensive, and harassment. A third annotator broke the tie if the two annotators disagreed on a label. After rounds of sample tweet coding, they settled on two categories (harassment and non-harassment) in the final corpus, for which we randomly sampled 500 tweets from each category.

Founta [62] studied the correlation between different forms of online abuse, e.g., hate speech, offensive language, abusive language, cyberbullying behavior, aggressive behavior, spam, and normal. Using a boosted sampling technique, they developed a dataset of 80K tweets to study the correlation. In various exploratory rounds, they used CrowdFlower workers to label the tweets into some categories mentioned. Finally, they decided upon four labels: abusive, hateful, normal, and spam. Following [158], we don't utilize the spam label in our analysis. We randomly sampled 500 tweets from the abusive, hateful, and normal classes.

OffensEval [239] is a competition where participants participate in the SemEval [241] tasks focused on the identification of offensive language, type of offensive language, and targets of offensive language. The SemEval [241] tasks use the Offensive Language Identification Dataset (OLID) [239] containing 14K tweets obtained using 50% keywords from politics and 50% keywords from non-politics. Two annotators from the crowd-sourcing platform, Figure Eight, annotated the dataset using a hierarchical annotation scheme. In layer A (identification of offensive language) of the scheme, tweets are labeled into two categories: offensive or not offensive. If there were ties, a third annotator broke the tie. The original dataset was split into training and testing sets. We retain the 240 offensive and 620 non-offensive tweets in the testing set and sample 500 tweets from the offensive and non-offensive tweets in the training set.

AbusEval [36] re-annotated the OffensEval/OLID dataset [239] for explicitness and added an extra layer to the labels. Using a new annotation guideline that distinguishes between abusive and offensive language, three annotators re-annotated the OffensEval dataset

Dataset	Count	Count after sampling
Waseem	5,988	1,116
Davidson	24,773	1,500
Founta	45,549	1,500
Golbeck	20,305	1,000
OffensEval	14,100	1,860
AbusEval	14,100	2,360
HatEval	11,991	2,000

Table 5.1: Statistics of the datasets.

into three classes: explicit abuse, implicit abuse, and not abusive. The dataset is the same size as the OffensEval/OLID dataset. Like OffensEval, the original dataset was split into training and testing sets. We retained 106 explicit abuse, 72 implicit abuse, and 682 non-abusive tweets in the test set and randomly sampled 500 tweets from each class in the training set.

HatEval [20] is a multilingual (English and Spanish) dataset that is also one of the tasks in SemEval [241]. In this task, hate speech against women or immigrants is to be detected in a tweet. The dataset composed of 13K tweets was collected using keywords and hashtags derogatory towards the targets and annotated by three annotators using the crowd-sourcing platform, *Figure Eight* [3]. The HatEval dataset was initially split into train, validation, and test sets. From the training and testing sets, we randomly sample 500 English tweets from the hate and non-hate classes. We do not use the validation set in our experiments as done in [158].

5.4.2 AAE Language Model and Dialect Classifier

The authors in [158] introduced AAEBERT, an African-American English language model. AAEBERT was obtained by retraining BERT [46] on a subset of tweets written by non-Hispanic Blacks from the race dataset [26]. While BERT [46] have shown improved

[3] www.figure-eight.com

performance on several NLP tasks, BERTweet [150], pre-trained on English Twitter data have shown better performance on several Tweet NLP tasks. Due to the performance of BERTweet [150] on English tweets and the fact that our language of interest (AAE) is a variation of English, BERTweet will more likely capture the nuances of AAE than BERT [46]. In this work, we reproduce AAEBERT [158] by retraining BERTweet [150] on the 1.22M black corpus extracted from the race dataset as discussed in Section 5.4.1.1 and call this retrained model ***AAEBERTweet***. We follow the implementation details of AAEBERT as described in [158] using masked language modeling as the training objective, a maximum sequence length of 100, batch size of 64, and training for 100 epochs on one V100 GPU.

We train a dialect classifier to infer the dialect of each tweet in the hate speech datasets described in Section 5.4.1.2 because a race label is required to assess racial bias using the AUC metrics described in Section 5.4.5.2. The authors in [158] directly used the AAEBERT language model with sigmoid activation function to classify a tweet as AAE with a threshold > 0.5 and as SAE otherwise. Contrary to this, we fine-tune AAEBERTweet on the 1K black-aligned and 1K white-aligned tweets to obtain a dialect model used to infer dialects as shown in Fig 5.1. The fine-tuned model was used to classify the tweets in each hate speech dataset discussed in Section 5.4.1.2 as AAE and SAE. The model trained with a learning rate of $1e-5$, batch size of 32, and for 20 epochs achieved 0.848, 0.847, and 0.847 precision, recall, and F1 scores, respectively. To show that our dialect classifier obtained by fine-tuning AAEBERTweet performs better than AAEBERT [158], we retrain BERT to obtain AAEBERT as described in [158], then fine-tuned AAEBERT on the 1K black-aligned and 1K white-aligned tweets. The resulting dialect model achieved 0.800, 0.800, and 0.800 precision, recall, and F1 scores, a less-performing dialect model compared to our dialect model obtained from fine-tuning AAEBERTweet. The per-class performance of our dialect model is shown in Table 5.2. We preprocess each tweet in the black-and-white-aligned tweets by replacing hyperlinks with HTTPURL, removing the # sign in hashtags, replacing handles with @USER, replacing extra white space with single space, replacing

numbers with NUMBER, removing punctuations, and ensuring each tweet contained more than three words.

We used the USERLEVELRACE dataset [171] to validate our dialect classifier. The USERLEVELRACE dataset is a Twitter dataset of users who self-reported their race/ethnicity through a survey. The dataset contains 5.4M tweets from 4,132 users, of which 374 are African American (AA) and 3,184 are White. Due to X's (Twitter) terms of service, the authors could not release the actual tweets to us upon request. Instead, they provided us with each user's ID, age, gender, and race. We randomly sample 14 AA and 14 White users. For each of the users, we collect the user's most recent tweets from the user's timeline using Twarc[4], a Python library for collecting Twitter JSON data via the Twitter API[5]

The data collection resulted in a dataset of 14,794 tweets, of which AA users wrote 5,103 tweets and White users wrote 9,691 tweets. We sampled 5K tweets from tweets written by AA users and 5K by white users. The dialect classifier was used to predict the dialect of the sampled tweets and achieved 0.734, 0.693, and 0.679 precision, recall, and F1 scores, respectively. The per-class performance of the dialect model on the USERLEVELRACE dataset is shown in Table 5.3. Upon analyzing the tweets qualitatively, we note that most AA users did not tweet in AAE, which most likely explains the low recall and F1 scores. We chose to sample 14 users from each group because, under the new X (Twitter) rules, data collection from the platform has become expensive as it is no longer free for academic research. Also, only five requests can be made per 15 minutes due to rate limitations.

[4]https://twarc-project.readthedocs.io/en/latest/

[5]Twitter's timeline API provides 3200 most recent tweets. Users had varying numbers of tweets; AA users tweeted 364.5 tweets on average, and White users tweeted 692.2. we could not retrieve 3200 tweets for all users, possibly due to the recent changes in Twitter API or users not having enough posts.

Target	F1	Precision	Recall
AAE	0.846	0.852	0.839
SAE	0.849	0.843	0.856

Table 5.2: Performance of the dialect classification model for the AAE and non-AAE (SAE) classes. Evaluation metrics are macro averages.

Target	F1	Precision	Recall
AAE	0.611	0.832	0.483
SAE	0.746	0.636	0.903

Table 5.3: Performance of the dialect classification model on the dataset of users who self reported their race/ethnicity (AA and White). Evaluation metrics are macro averages.

5.4.3 Prompt Annotation

We employ various prompting strategies for data annotation to determine bias in classifiers trained on LLM-annotated datasets. We used the GPT-4 model from the official OpenAI API endpoints to run the various prompts for annotating each dataset. We used a variation of prompting strategy utilized in hate speech detection and annotation [79, 88, 122]. Each prompting strategy is described below with examples in Table 5.4.

5.4.3.1 General Prompt Annotation

The general (Gen) prompt technique allows the adaptation of LLMs for the specific task of hate speech annotation. Given a tweet x, the input to GPT-4 in this annotation strategy is formatted as: *Given the tweet in triple quotes: """x""". Do you think the tweet is [classes]? Only answer with one of the following: [classes]. Do not provide an explanation for your answer.* Where [classes] represent the original classes or categories in each human-annotated dataset, for example, in the Davidson dataset, [classes] = hate or offensive or normal. GPT-4 will then output y, for example, either "hate", "offensive" or "normal" representing the annotation for x for the Davidson dataset.

5.4.3.2 Few-shot Prompt Annotation

Few-shot (FS) learning has improved LLMs' performance in many NLP tasks [207]. In this annotation setting, examples, also known as few-shot demonstrations with answers or solutions, are included in the prompt given to an LLM, essentially demonstrating the task to the LLM; the LLM learns in context via prompting. We explore whether racial bias persists in this annotation setting as it is likely that data annotation can be performed in the real world using an LLM with a few labeled examples to improve annotation and to avoid over-exposure to hateful content. We randomly sampled two exemplars from each class in each dataset, which were used as part of the prompt to assess racial bias in downstream models trained on datasets annotated with few-shot demonstrations. One of the exemplars is detailed in Table 5.4.

5.4.3.3 Chain-of-Thought Prompt Annotation

Finally, we explored Chain-of-Thought (CoT) prompting annotation, a series of intermediate natural language reasoning steps that lead to the final answer [227]. The Chain-of-Thought prompt has been shown to enhance the ability of LLMs to solve complex reasoning tasks in NLP [227], and it consists of triples: [x, chain of thought, y]. We modify the few-shot exemplars in the few-shot prompt annotation setting for CoT annotation as shown in Table 5.4. The intermediate natural language reasoning steps are designed by elaborating the definitions of hate speech as defined by the authors of each dataset. Each example used in the few-shot annotation setting is augmented with an answer with comprehensive reasoning to explain why the example belongs to a particular class or category. As in few-shot prompt annotation, CoT reasoning is likely to be used in the real world for data annotation, where a few examples are provided together with a reasoned explanation of why a text is hateful or not hateful. We simulate that scenario in this setting

We tested different variations of these prompts and settled for the stated prompts because they worked and exhibited good performance across all the seven datasets analyzed and across different numbers of class labels. As discussed in Section 5.4.1, we sampled from

the hate speech datasets because of the cost associated with annotating using FS and CoT prompting techniques via Open AI as charges are token-based[6].

5.4.4 Hate Speech Classifiers

For each GPT-4 annotated dataset, we train a classifier on the training dataset to predict the class of each tweet in the test dataset. For the datasets not initially split into train and test sets by the original authors, we randomly split those datasets (Waseem, Davidson, Founta, and Golbeck) into train and test sets using the 80:20 splits. We use the same set of pre-trained models used in [158], BERT [46] (bert-base-uncased on HuggingFace), BERTweet [150] (vinai/betweet-base on HuggingFace), and HateBERT [35]. We fine-tuned the pre-trained models on GPT4-annotated datasets to obtain twenty-one hate speech classifiers (seven datasets on three pre-trained models). We fine-tuned the pre-trained models on human-annotated datasets for comparison purposes and obtained twenty-one hate speech classifiers. Each classifier was trained using a learning rate of $1e-5$, a batch size of 32, a maximum sequence length of 100, 5 epochs, an Adam optimizer, and cross-entropy loss. We used the same pre-processing steps used in Section 5.4.2 to pre-process each tweet, except filtering tweets that do not have at least four words.

5.4.5 Bias Metrics

We use the evaluation metrics described below to evaluate bias in classifiers trained on GPT-4-annotated datasets annotated using different prompting strategies.

5.4.5.1 Hypothesis-based Metric

The hypothesis-based evaluation metric [42, 142, 158] assesses whether there exists a difference in the probability of a tweet being predicted as a particular class is due to the tweet author's race (of which we use the dialect of the tweet as a proxy for race). The

[6]https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4-gpt-4-turbo-and-gpt-4o

evaluation is based on estimating the proportion of tweets in each dataset that each classifier classifies as belonging to each class using the sampled black-aligned and white-aligned tweets discussed in Section 5.4.1.1. We define a null hypothesis (H_N) that there is no racial bias if the probability of a tweet belonging to a negative class is independent of the author's race. We test $H_N : P(c_i = 1|black) = P(c_i = 1|white)$ for each negative class c_i, where $c_i = 1$ represents membership in the class and $c_i = 0$ represents otherwise. We reject the null hypothesis H_N in favor of the alternative H_A that black-aligned tweets are classified to negative classes c_i at a higher rate than white-aligned tweets if $P(c_i = 1|black) > P(c_i = 1|white)$ and the difference is statistically significant. If $P(c_i = 1|black) < P(c_i = 1|white)$, then white-aligned tweets are assigned to negative classes at a higher rate.

We create a vector per class for each racial group (black and white) in which each element is the probability p_i of a tweet belonging to a negative class i as predicted by a classifier. We obtain vectors of dimension $n = 1000$ (the number of tweets in the black-aligned and white-aligned datasets). For each group, we calculate the proportion of tweets assigned to negative class i as $\widehat{p_{iblack}} = \frac{1}{n}\sum_{j=1}^{n} p_{ij}$ for black-aligned and $\widehat{p_{iwhite}} = \frac{1}{n}\sum_{j=1}^{n} p_{ij}$ for white-aligned. We test $\widehat{p_{iblack}} = \widehat{p_{iwhite}}$ for significance using t-test. If the magnitude of the difference $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}} > 1$, then black-aligned tweets are assigned to a negative class at a higher rate than white-aligned.

5.4.5.2 AUC-based Metric

Machine learning classifiers can exhibit unintended bias as the systemic differences in performance for different demographic groups [29]. The AUC-based metrics introduced by the Google Conversational AI Team [29] have been used to measure identity-based (such as "gay", "muslim", etc.) unintended bias in machine learning classifiers for hate speech detection [135, 211]. We use AUC-based metrics described below to assess racial bias towards tweets written in AAE and SAE by classifiers trained on GPT-4-annotated datasets. We focus on the ability of the classifiers to reduce false positive rates on non-hateful tweets inferred to be written in AAE known empirically to introduce model bias. The AUC metrics include

Subgroup AUC, Background Positive Subgroup Negative (BPSN), and Generalized Mean of Bias AUCs. For these metrics, we convert datasets with multi-class GPT-4-annotated labels into binary labels, re-trained our classifiers on the binary labels (hate and non-hate) and evaluate the reduction of unintended bias towards a group by the classifiers. Evaluation is restricted to the test set of each dataset and not on the black-aligned and white-aligned datasets to understand how the classifiers perform in hate speech detection and bias reduction.

Subgroup AUC: We restrict the test set to hateful and non-hateful tweets written in AAE and SAE. The ROC-AUC score is calculated for each group (AAE and SAE), resulting in the Subgroup AUC for a group. This metric measures the model's ability to separate hateful and not hateful tweets in the context of a specific group. A higher score indicates that the model is doing an excellent job of separating hateful and non-hateful posts particular to the racial group.

BPSN (Background Positive, Subgroup Negative AUC): We restrict the test set to non-hateful tweets written in AAE and hateful tweets not in AAE. The BPSN AUC is obtained by calculating the ROC-AUC score of this set. The false positive rate of each classifier in the context of each specific group is measured by this metric. A model is less likely to confuse non-hateful tweets written by a group with hateful tweets not written by the group if the BPSN score is high, meaning the model can reduce bias towards a specific group. We consider BPSN a stronger metric than Subgroup AUC because it aligns with the focus of this paper, which is the false positive rate towards certain groups.

Generalized Mean of Bias AUCs: As part of their Kaggle competition[7], the Google Conversational AI Team introduced this metric which combines the per-group Bias AUCs into an overall measure as $M_p(m_s) = (\frac{1}{n}\sum_{s=1}^{N} m_s^p)^{\frac{1}{p}}$ where, M_p is the p^{th} power-mean function, m_s is the bias metric calculated for subgroup s, N is the number of groups which is 2, and p is set to -5 as done in the competition. We report the following metrics for our datasets:

[7]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview/evaluation

- **GMB-Subgroup-AUC**: GMB AUC with Subgroup AUC as the bias metric
- **GMB-BPSN-AUC**: GMB AUC with BPSN AUC as the bias metric

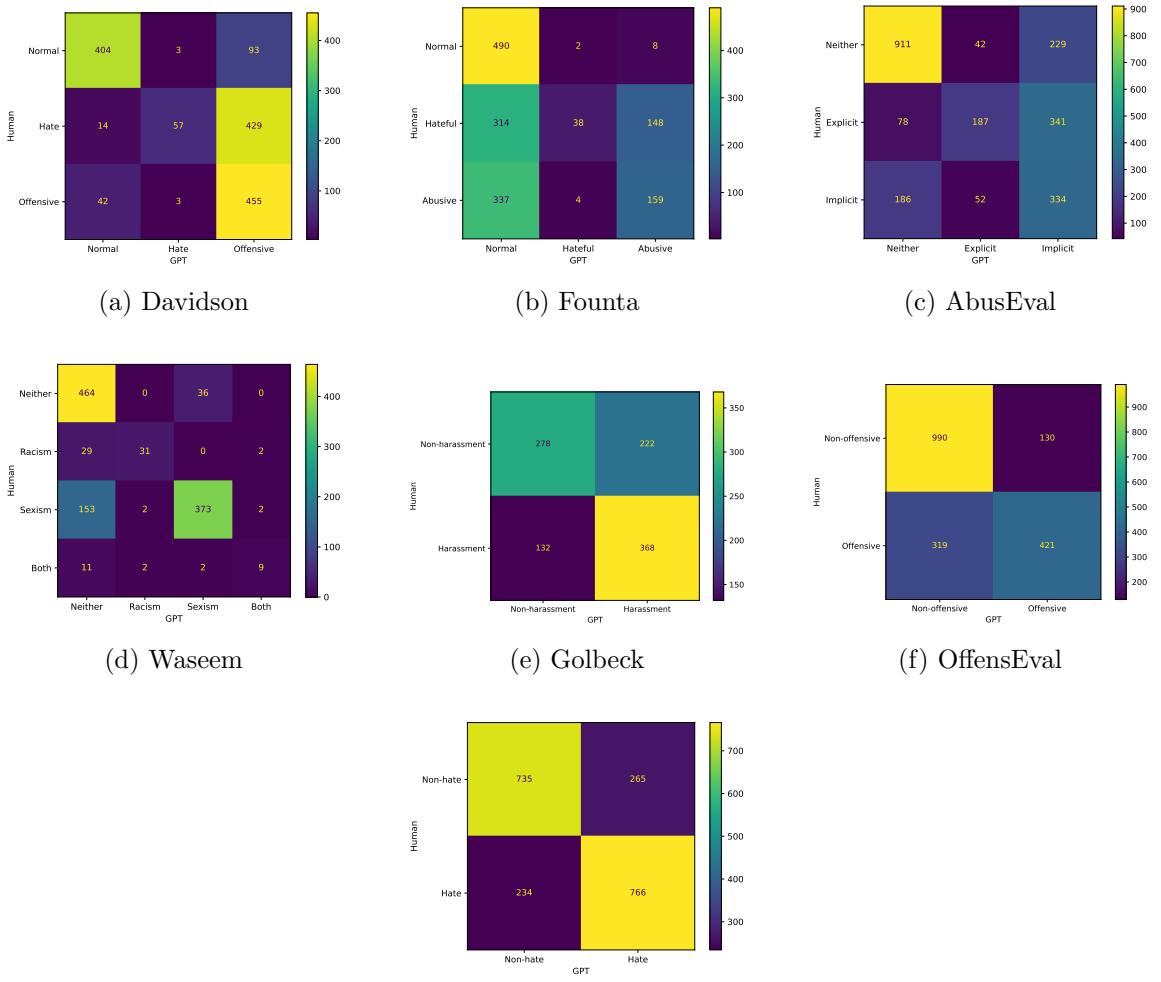
5.5 Results

In this section, we discuss the results of our study examining racial bias in using LLMs for data annotation in hate speech detection.

5.5.0.1 Performance

The overall multi-class and binary classification performance for various prompting strategies is summarized in Tables 5.5 and 5.6, respectively. From the multi-class classification results in Table 5.5, we observe that for general prompt annotation, BERTweet is competitive, for FS and CoT, HateBERT and BERTweet outperform almost all models, respectively. Overall, the use of FS learning prompt annotation increases performance. For the binary classification in Table 5.6, BERTweet outperforms other models in five and six datasets across the prompt annotation strategies, respectively. Similar to multi-class classification, FS prompt annotation increases performance consistently across models and datasets except in the BERTweet model fine-tuned on the AbusEval dataset and in the BERT and BERTweet models fine-tuned on the HatEval datasets. When compared to the model performance on human-annotated datasets as shown in Tables 5.7 and 5.8 in the Appendix for multi-class and binary label classification, models fine-tuned on GPT-4 annotated datasets generally outperforms models fine-tuned on human-annotated datasets in binary classification. In multi-class classification, models fine-tuned on human-annotated datasets generally perform better than models fine-tuned on datasets annotated using general prompt annotation.

We compare how well GPT-4 annotation using general prompt strategy matches human annotation and show the confusion matrix in Fig 5.2. For the Davidson dataset (Fig 5.2a), GPT-4 tends to label most of the tweets annotated as hate by humans as



(a) Davidson (b) Founta (c) AbusEval

(d) Waseem (e) Golbeck (f) OffensEval

(g) HatEval

Figure 5.2: Confusion matrix of human annotation and GPT-4 general prompt annotation on full (training and testing) datasets.

offensive in 29% of the tweets. GPT-4 labels a sizable amount of tweets annotated as hateful and abusive by humans as normal in 21% and 22% of the tweets, respectively, for the Founta dataset (Fig 5.2b). 14% of the tweets annotated as explicit by humans were labeled as implicit by GPT-4 in the AbusEval dataset. We observe that GPT-4 does a good job in correctly annotating Sexist tweets in 33% of the tweets. In the binary datasets, GPT-4 disagrees with humans by re-labeling harassment tweets to non-harassment in 22% of the tweets, offensive tweets to non-offensive in 17% of the tweets, and hateful tweets to non-hateful in 12% of the tweets for the Golbeck, OffensEval, and HatEval datasets, respectively. The overall performance of GPT-4 general prompt annotation and human-annotation are shown in Table 5.9, and the performance when each dataset is conditioned on each dialect (AAE and SAE) is shown in Table 5.10. From Table 5.9, in all of the metrics, the level of agreement is above 50% except in the recall, F1 and accuracy metrics of the Founta dataset. From Table 5.10, in general, the rate at which GPT-4 general prompt annotation aligns with human-annotation is slightly better if tweets are written in SAE in terms of precision, F1, and accuracy for most datasets.

The FS and CoT prompt annotation overall performance results are shown in Tables 5.11 and 5.12, respectively, and the performance results when the datasets are conditioned on dialect are shown in Tables 5.13 and 5.14. When the performance of FS and CoT annotation is compared to the general prompt annotation results in Tables 5.9 and 5.10, we observe that the level of agreement between GPT-4 annotation and human annotation increases across almost all datasets with or without conditioning on dialect and across nearly all metrics. While FS and CoT prompt annotation with GPT-4 increases the level of agreement between human annotation, it could further introduce bias in the annotation because humans provide exemplars and exemplars with detailed explanations or reasoning in both strategies, which is a means for human prejudice or bias to manifest.

5.5.0.2 Bias - Hypothesis based

Table 5.15 shows the results of the BERT model fine-tuned on datasets annotated by GPT-4 using general prompt annotation. From Table 5.15, we observe racial disparities in the performance of most of the classifiers. There are statistically significant differences ($p \ll 0.05$) in most classifiers except in two instances. In all the statistically significant instances, we observe that black-aligned tweets are assigned to negative classes at a higher rate than white-aligned tweets except for one instance, in the implicit class of the AbusEval classifier where black-aligned tweets are assigned to a negative class at a lower rate than white-aligned tweets. For the Davidson and Waseem classifiers, we observe no significant difference in the rates at which tweets are classified as hate and racism, respectively, with the rates remaining low. Black-aligned tweets are classified frequently as offensive at 1.4 times, hate at 1.7 times, and abuse at 1.6 times compared to white-aligned tweets in the OffensEval, HatEval, Davidson, and Founta classifiers. Similar observation on racial disparities is made in the BERT model fine-tuned on GPT-4 annotated datasets using FS and CoT prompt annotation as shown in Table 5.16 of the Appendix. Comparing the human annotation results in Table 5.15 with the general, FS, and CoT annotation results in the Tables 5.15 and 5.16 (See Appendix), we see that there are statistically significant ($p \ll 0.05$) instances, OffensEval, HatEval, Founta, Golbeck, offensive class of Davidson using general annotation, and in the sexism class of Waseem using CoT annotation, where GPT-4 annotation increases the rate at which black-aligned tweets are classified as negative classes more than white-aligned tweets. There are statistically significant ($p \ll 0.05$) instances, explicit class of AbusEval, sexism class of Waseem using general and FS annotation, and offensive class of Davidson using FS and CoT annotation, where GPT-4 annotation reduces the rate of assigning black-aligned tweets to negative classes. There are statistically significant ($p \ll 0.05$) instances, such as in the hate class of Davidson using FS and CoT annotation and the racism and sexism class of Waseem, where there is a change in direction.

The results of the BERTweet model fine-tuned on datasets annotated using general prompt annotation are shown in Table 5.17, and the results of using FS and CoT prompt

annotation are shown in Table 5.18 of the Appendix. From Table 5.17, we see results similar to the BERT model; racial disparities are observed in the performance of most of the classifiers, although they are lesser in magnitude. Most of the classifiers have statistically significant differences ($p < 0.05$). In all the statistically significant instances, we observe that black-aligned tweets are assigned to negative classes at a higher rate than white-aligned tweets, except for one instance, in the hate class of the Davidson classifier, where black-aligned tweets are assigned to a negative class at a lower rate than white-aligned tweets. The largest disparity is observed in the Founta classifier, where 22% of black-aligned tweets are classified as abuse compared to 14% of white-aligned tweets. Comparing the human annotation results with the general, FS, and CoT annotation results as shown in Tables 5.17 and 5.18 (See Appendix), there are increases in the rate at which black-aligned tweets are classified as negative classes more than white-aligned tweets in some statistically significant ($p < 0.05$) instances, OffensEval, HatEval, Golbeck, abuse class of Founta, hate class of Founta using FS and CoT annotation, implicit class of AbusEval, and racism and sexism class of Waseem. There are instances ($p < 0.05$), offensive class of Davidson, hate class of Davison using FS and CoT annotation, hate class of Founta using general annotation, explicit class of AbusEval using general annotation, sexism class of Waseem, and racism class of Waseem using general annotation, where there are reductions in the rate of assigning black-aligned tweets to negative classes using prompt annotations. A significant ($p < 0.05$) change in direction is observed in the hate class of the Davidson classifier fine-tuned on the Davidson dataset annotated using general prompting annotation.

The results of the HateBERT classifier obtained by fine-tuning datasets annotated using general prompt annotation as shown in Table 5.19 and FS and CoT prompt annotation as shown in Table 5.20 of the Appendix are consistent with the BERT and BERTweet results. Racial disparities persist in most of the classifiers, and there are statistically significant differences ($p \ll 0.05$) in most of the classifiers except, in one instance, in the sexism class of the Waseem classifier. In all the statistically significant instances, black-aligned tweets are assigned to negative classes at a higher rate than white-aligned tweets, except

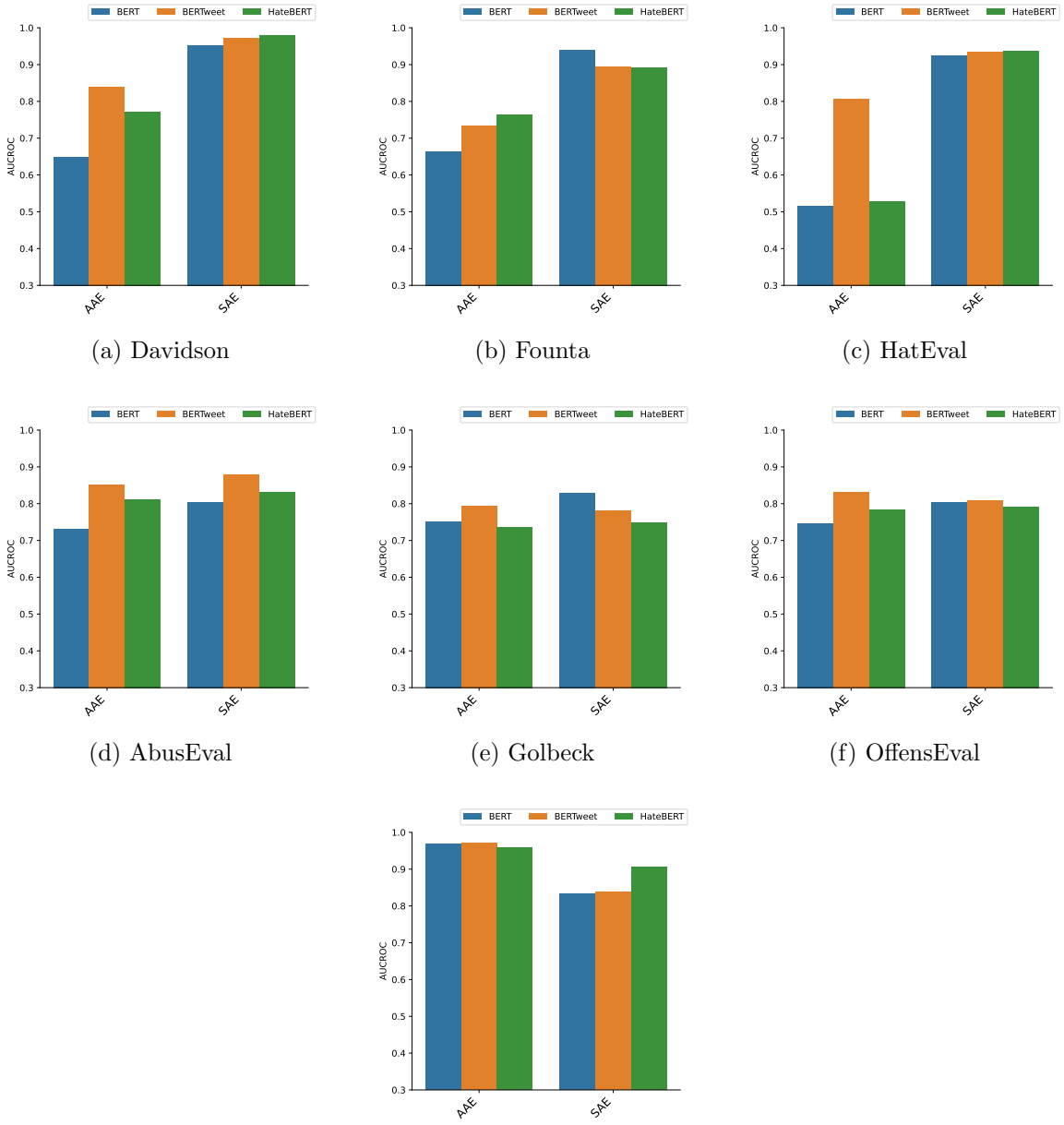
in the Davidson classifier, where black-aligned tweets are less likely to be assigned to the hate class. However, they are more likely to be assigned to the offensive class. When compared to the classifiers fine-tuned on human annotated datasets focusing on significant ($p \ll 0.05$) instances, the rate of classifying black-aligned tweets to negative classes increases in AbusEval, OffensEval, HatEval, Founta, and Golbeck classifiers and the racism class of Waseem and the sexism class of Waseem using CoT annotation. There are a few decreases in the offensive class of Davidson, the hate class of Davidson using CoT, the sexism class of Wasee using FS, and a change in direction in the hate class of Davidson using general prompt annotation.

These results demonstrate that using an LLM, such as GPT-4, for data annotation can introduce racial bias in the annotated dataset. If the biased dataset is used to fine-tune a model for downstream tasks, as we have seen in hate speech detection, the downstream models will propagate the racial bias introduced by the LLM. We have shown that tweets written in AAE are disproportionately assigned to negative classes at a higher rate than tweets written in SAE, and while some classifiers trained on LLM annotated datasets can decrease this rate, most increase it.

5.5.0.3 Bias - Subgroup & BPSN AUCs

From Table 5.6, For GMB-Subgroup-AUC, BERTweet consistently outperforms other models in six, five, and six datasets for general, FS, and CoT prompt annotations, respectively, indicating that it is most successful at accurately classifying tweets written in AAE and SAE. Additionally, BERTweet also outperforms other models on five datasets for both general and FS annotation and on all datasets annotation using CoT reasoning annotation in Generalized Mean Bias with BPSN AUC, suggesting that BERTweet is less likely to confuse non-offensive tweets written in AAE with offensive tweets not written in AAE, i.e., BERTweet significantly reduces false positive rate.

The dialect-wise BPSN AUC metric results for the datasets annotated using general prompt annotation are reported in Fig 5.3. In Fig 5.3, we observe that tweets written in

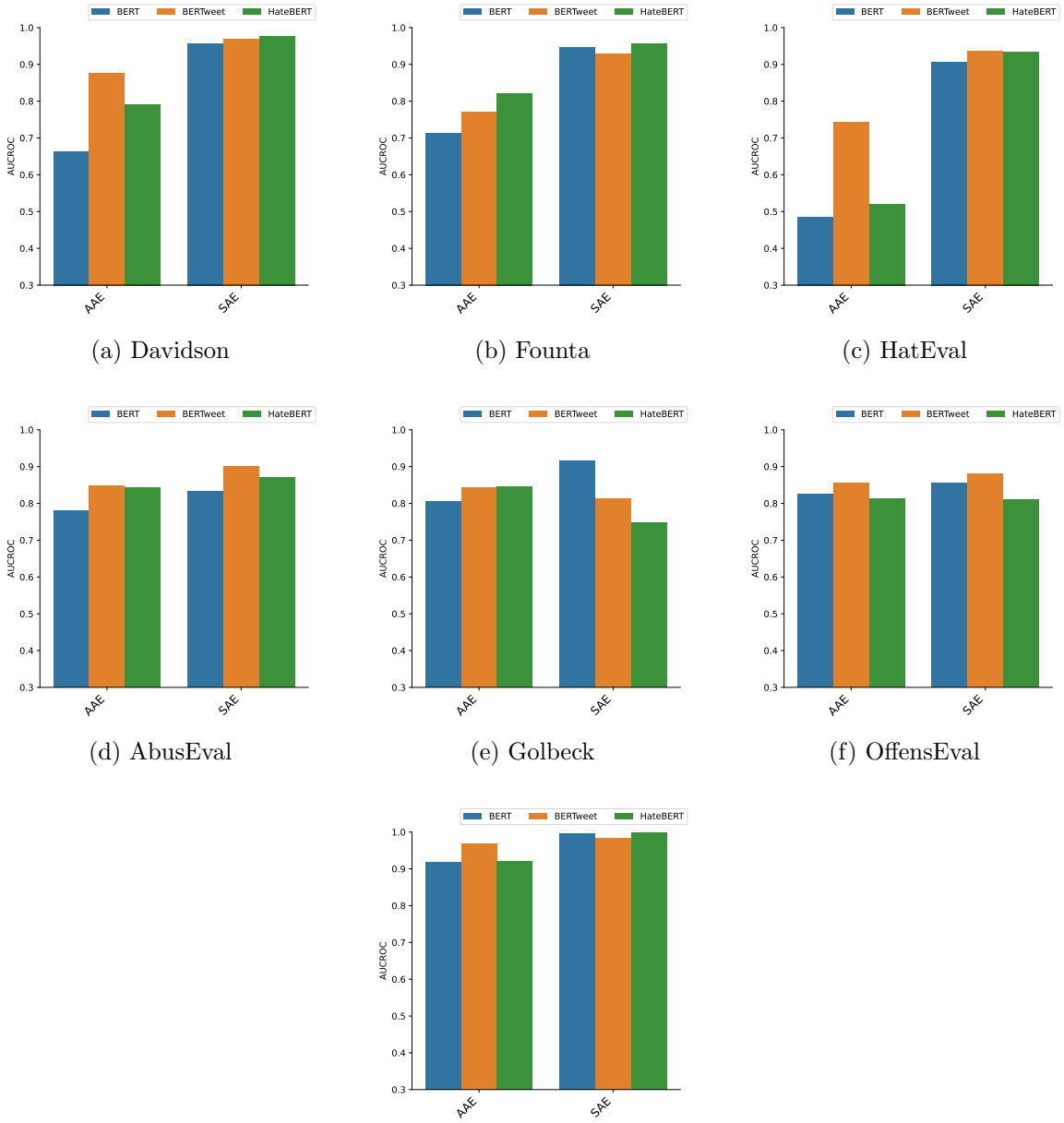


(a) Davidson (b) Founta (c) HatEval

(d) AbusEval (e) Golbeck (f) OffensEval

(g) Waseem

Figure 5.3: Dialect-wise results for BPSN AUC on general prompt annotated test datasets.



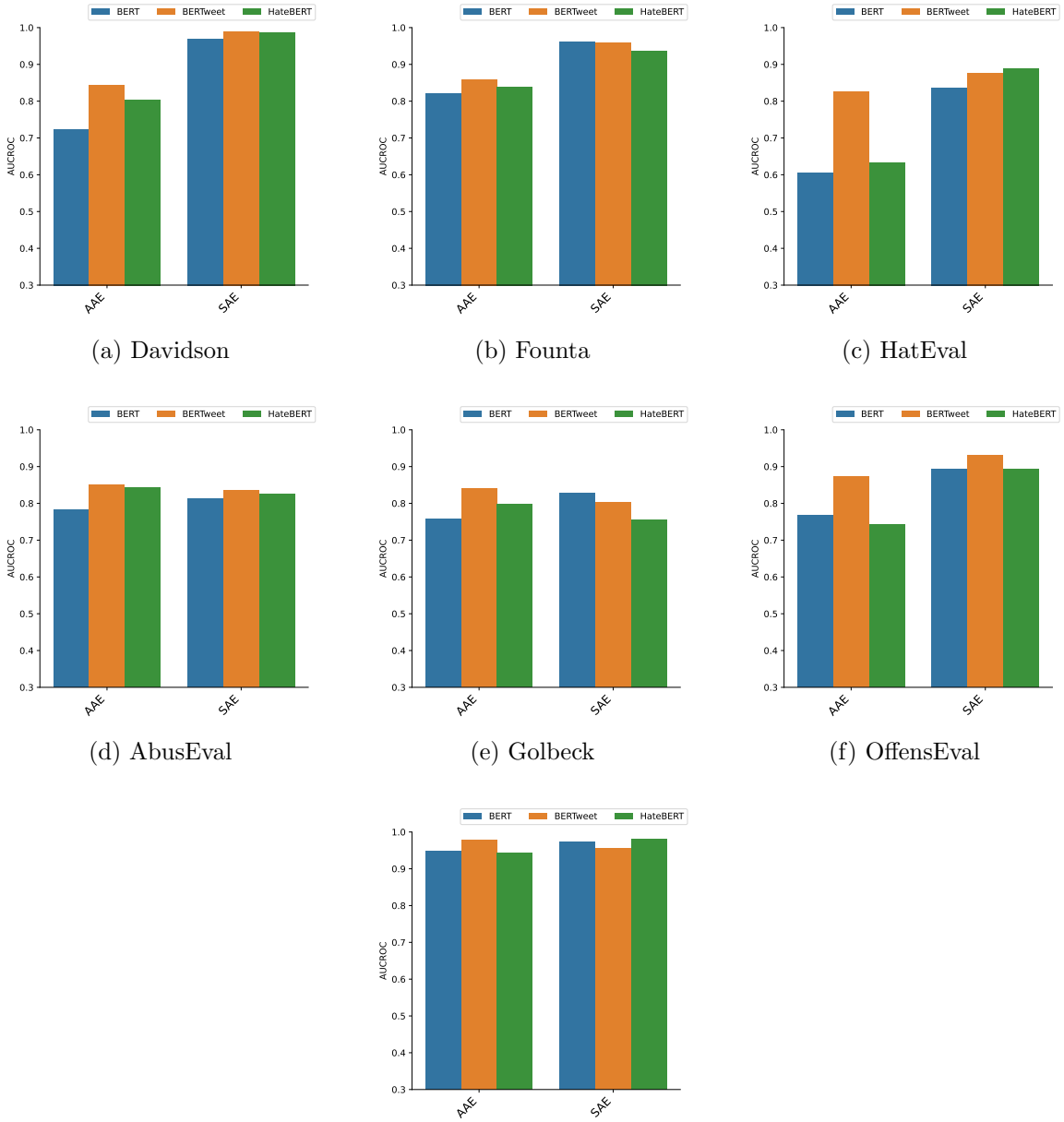
(a) Davidson (b) Founta (c) HatEval

(d) AbusEval (e) Golbeck (f) OffensEval

(g) Waseem

Figure 5.4: Dialect-wise results for BPSN AUC using few-shot annotation.

AAE seem to be biased toward having more false positives due to the low BPSN AUC scores than tweets written in SAE for all models for the Davidson (Fig 5.3a), Founta (Fig 5.3b), HatEval (Fig 5.3c), and AbusEval (Fig 5.3d) datasets. The BERTweet model is slightly biased toward more false positives for tweets written in SAE (1%) when compared to the



(a) Davidson (b) Founta (c) HatEval

(d) AbusEval (e) Golbeck (f) OffensEval

(g) Waseem

Figure 5.5: Dialect-wise results for BPSN AUC using chain-of-thought annotation.

AAE tweets (2%) for the Golbeck (Fig 5.3e) and OffensEval (Fig 5.3f) datasets. All models are biased toward having more false positives for the SAE tweets than AAE in the Waseem (Fig 5.3g) dataset. The BPSN-AUC metric results for FS prompt annotation are consistent with the general prompt annotation except for the change in direction in the Waseem dataset

and the increased difference in BPSN scores of the BERTweet and HateBERT models across the two groups as shown in Fig 5.4 of the Appendix indicating more false positives towards SAE tweets. Fig 5.5 in the Appendix show the BPSN AUC metric results for the CoT prompt annotation which is consistent with general and FS prompt annotation except in the Waseem (Fig 5.5g) where there is no large difference in the BPSN scores of the AAE and SAE tweets and in AbusEval (Fig 5.5d) where the BERTweet and HateBERT models are slightly more biased towards SAE tweets than AAE tweets.

Overall, the BERTweet model achieves an increased BPSN AUC performance across all datasets and annotation strategies for AAE tweets, suggesting its ability to reduce false positive rate as also shown in the GMB-BPSN column in Table 5.6. AAE tweets seem to be biased toward having more false positives in most models and datasets. FS annotation increases dialect-wise (figures are not shown due to space) Subgroup AUC score in almost all models and datasets indicating its effects in correctly separating offensive and non-offensive AAE and SAE tweets as also seen under GMB-Subgroup-AUC summary in Table 5.6 when compared to general annotation. A similar observation is made for CoT to some extent. FS annotation improves the ability to reduce false positives, as seen in the increment in BPSN AUC score across models, datasets, and groups (AAE and SAE) compared to general annotation. We see similar observations for CoT annotation to an extent.

5.5.0.4 Dialect Priming

We explore the effect of dialect information in LLM annotation, [186] used dialect priming in human annotation task where annotators are instructed to consider the dialect of a tweet as a proxy for race when deciding a label for a tweet. They infer dialect using the statistical model developed by [26]. Motivated by this idea, we included dialect in each of the annotation strategies (general, FS, and CoT), used GPT-4 to re-annotate our sampled datasets as discussed in Section 5.4.1 and inferred dialect using our dialect model discussed in Section 5.4.2. For the FS examples used in the FS and CoT annotation settings, we randomly sample balanced exemplars by dialect. From each class in each dataset, we

randomly sampled one exemplar from tweets inferred to be AAE and tweets inferred to be SAE, making it two exemplars per class. We slightly modify the prompts we used in the setting without dialect priming as discussed in Section 5.4.3 as shown in Table **??**. In this experiment, we only consider the Founta, Davidson, HatEval, and OffensEval datasets because they consistently showed racial bias in models fine-tuned on them using GPT-4 as seen in the hypothesis 5.5.0.2 and AUC 5.5.0.3 results sections. We only show the hypothesis-based results for all models; the BERT model results for the general prompt annotation are shown in Table 5.21, and the FS and CoT prompt annotation results are shown in Table 5.22. The BERTweet and HateBERT models results are shown in Tables 5.23 and 5.25 for general prompt annotation, respectively. While the FS and CoT annotation results are shown in Tables 5.24 and 5.26, respectively. Results in Tables 5.21, 5.22, 5.23, 5.24, 5.25, and 5.26 indicate that racial disparity persists and that FS and CoT annotation with dialect priming does not help in bias mitigation. Instead, they increase model bias towards AAE tweets.

5.6 Broader Perspectives

Our work has implications for data annotation using LLMs. As data annotation is a labor-intensive and time-consuming process for various NLP tasks, especially in online hate speech annotation where annotators can be exposed to hateful content that can negatively affect them [216], researchers have explored the use of LLMs such as GPT-3 to reduce the cost of annotation and have shown that downstream models trained on LLM-annotated data can achieve good performance [223]. As we have demonstrated empirically in this study, if LLM-annotated datasets are biased, they may cause unfair treatment of certain groups, such as African Americans who write in African American English. If the biased data is used to train hate speech detection models used in online social platforms such as Twitter (X), tweets written in AAE might be flagged as hateful more frequently than tweets written in SAE, which could lead to the marginalization of users who use AAE in voicing their

opinions or struggles for example during social movements [206, 209] through the removal of such content by the platform.

5.7 Limitations

Our study has several limitations. First, the subsets of the race dataset [26] used in the training of AAEBERTweet and the fine-tuning of AAEBERTweet to obtain the dialect classifier do not strictly contain tweets written in AAE by African-Americans and tweets written in SAE by white Americans. Also, not all African Americans use AAE, and not all AAE users are African-American, although its use suggests closeness to the African-American community [26]. Therefore, the dialect classifier can predict a tweet to be AAE even though it is SAE and vice versa. Second, our analyses are limited to the three language models fine-tuned for hate speech detection. Third, the three classifiers are trained on subsets of the seven hate speech detection datasets considered, and the bias assessed in these classifiers could be lower bound estimates. Fourth, we have only focused on one LLM, GPT-4. Fifth, due to the number of datasets studied, we settled for a prompt that worked across all datasets, which might not have been the best for each dataset for each of the prompt annotation strategies considered. Finally, the FS exemplars used in the FS and CoT annotations used the human ground-truth label. We didn't change the label. The human ground-truth label could be wrong, which might have affected the LLM's annotation; we did not make any corrections to simulate a real-world annotation task where a team or an organization utilizing FS or CoT annotation has labeled a few examples.

5.8 Conclusions and Future Work

In this paper, we have shown that large language models, even though they are capable of annotating data for online hate detection, can introduce racial bias to the data annotation process, which can lead to unfair treatment of already marginalized groups. We used GPT-4 to re-annotate seven hate speech detection datasets using general, few-shot

learning, and chain-of-thought reasoning prompt annotation strategies. We then used these GPT-4-annotated datasets to fine-tune three classifiers (BERT, BERTweet and HateBERT) and understand the racial bias in LLM annotation and its propagation in the downstream models trained on them. As our evaluation is limited to only GPT-4 and racial bias, it is important to extend our analysis to other LLMs and other biases and explore mitigation approaches in the future.

General prompt	Given the tweet in triple quotes: """@USER fuck you faggot""". Do you think the tweet is hate or offensive or normal? Only answer with one of the following: hate or offensive or normal. Do not provide an explanation for your answer.
Few-shot learning promt	Question: is the tweet in triple quotes """Johnny Rebel - If I could be a nigger for a day: URL""" hate or offensive or normal? Only respond with hate or offensive or normal. Answer: hate. ... Question: is the tweet in triple quotes """@USER fuck you faggot""" hate or offensive or normal? Only respond with hate or offensive or normal. Answer:
Chain-of-Thought prompt	Question: Is the tweet in triple quotes """@USER @USER. LOVE ISLAM. Even the trees will say "Come kill the Jews " (and other kufr). AMEN. URL""" harassment or not harassment? Answer with only one of the following classes: harassment or not harassment. The answer should be in the last words of your final sentence after your explanation. Answer: This tweet is harassment because it calls for violence towards a group of people - the jews and other kufr, who are people who don't believe in God, making them feel unsafe. So, the answer is harassment. ... Question: Is the tweet in triple quotes """RT @USER: @USER @USER @USER @USER @USER The Jews have caused all the world's problems since time started""" harassment or not harassment? Answer with only one of the following classes: harassment or not harassment. The answer should be in the last words of your final sentence after your explanation. Answer:

Table 5.4: Annotation prompt samples from the Davidson and Golbeck datasets for the three prompting strategies.

		F1 Score		
Dataset	Model	Gen	FS	CoT
Waseem	BERT	0.418	0.429	0.426
	BERTweet	0.426	0.428	**0.429**
	HateBERT	**0.428**	**0.433**	0.412
Davidson	BERT	0.558	0.698	0.726
	BERTweet	**0.563**	0.753	**0.767**
	HateBERT	**0.563**	**0.763**	0.762
Founta	BERT	**0.490**	0.564	0.550
	BERTweet	**0.490**	0.605	0.539
	HateBERT	0.488	**0.624**	**0.743**
AbusEval	BERT	0.407	0.491	0.518
	BERTweet	**0.504**	**0.576**	**0.647**
	HateBERT	0.471	0.529	0.577

Table 5.5: Classifier performance after fine-tuning on each GPT-annotated dataset with multi-class labels for each prompting strategy. Evaluation metrics are macro averages.

		F1 Score			GMB-Sub			GMB-BPSN		
Dataset	Model	Gen	FS	CoT	Gen	FS	CoT	Gen	FS	CoT
Waseem	BERT	0.855	0.888	0.866	0.916	0.957	0.960	0.888	0.953	0.960
	BERTweet	0.876	**0.901**	**0.873**	**0.930**	**0.974**	**0.912**	0.892	**0.977**	**0.967**
	HateBERT	**0.880**	0.892	0.867	0.909	0.948	0.823	**0.932**	0.956	0.962
Davidson	BERT	0.823	0.841	0.849	0.808	0.856	0.868	0.724	0.739	0.798
	BERTweet	**0.857**	**0.900**	**0.900**	**0.926**	**0.937**	0.935	**0.892**	**0.916**	**0.900**
	HateBERT	0.840	0.855	0.864	0.922	0.924	**0.940**	0.840	0.857	0.868
Founta	BERT	0.686	0.736	0.857	0.830	0.859	0.908	0.737	0.786	0.875
	BERTweet	**0.756**	**0.811**	**0.865**	0.823	0.872	**0.914**	0.790	0.828	**0.901**
	HateBERT	0.743	0.772	0.834	**0.841**	**0.901**	0.910	**0.813**	**0.874**	0.879
Golbeck	BERT	**0.692**	**0.737**	**0.654**	0.798	**0.832**	0.780	0.783	**0.852**	0.789
	BERTweet	0.576	0.648	0.643	**0.802**	0.822	**0.811**	**0.788**	0.827	**0.820**
	HateBERT	0.633	0.686	0.601	0.758	0.770	0.771	0.741	0.789	0.772
OffenEval	BERT	0.605	0.642	0.613	0.775	0.838	0.806	0.771	0.841	0.818
	BERTweet	**0.732**	**0.736**	**0.662**	**0.815**	**0.874**	**0.912**	**0.819**	**0.868**	**0.900**
	HateBERT	0.625	0.630	0.578	0.787	0.815	0.823	0.787	0.812	0.800
AbusEval	BERT	0.669	0.700	0.695	0.770	0.807	0.799	0.762	0.806	0.797
	BERTweet	**0.799**	**0.780**	**0.783**	**0.866**	**0.876**	**0.841**	**0.865**	**0.873**	**0.844**
	HateBERT	0.729	0.729	0.742	0.823	0.856	0.836	0.820	0.857	0.836
HatEval	BERT	0.698	0.674	0.639	0.758	0.730	0.736	0.587	0.554	0.671
	BERTweet	**0.773**	**0.733**	**0.766**	**0.880**	**0.853**	**0.855**	**0.856**	**0.809**	**0.849**
	HateBERT	0.715	0.684	0.644	0.787	0.783	0.773	0.600	0.592	0.703

Table 5.6: Classifier performance after fine-tuning on each GPT-annotated dataset with binary labels for each prompting strategy. Evaluation metrics are macro averages.

Dataset	Model	F1
Waseem	BERT	0.409
	BERTweet	**0.423**
	HateBERT	0.416
Davidson	BERT	0.693
	BERTweet	0.771
	HateBERT	**0.783**
Founta	BERT	**0.738**
	BERTweet	0.713
	HateBERT	0.716
AbusEval	BERT	0.457
	BERTweet	**0.570**
	HateBERT	0.518

Table 5.7: Classifier performance after fine-tuning on each human-annotated dataset with multi-class labels for each prompting strategy. Evaluation metrics are macro averages.

Dataset	Model	F1
Waseem	BERT	0.838
	BERTweet	**0.852**
	HateBERT	**0.852**
Davidson	BERT	0.838
	BERTweet	**0.888**
	HateBERT	0.885
Founta	BERT	0.859
	BERTweet	**0.865**
	HateBERT	0.861
Golbeck	BERT	0.527
	BERTweet	0.575
	HateBERT	**0.601**
OffenEval	BERT	0.636
	BERTweet	**0.715**
	HateBERT	0.657
AbusEval	BERT	0.635
	BERTweet	**0.748**
	HateBERT	0.657
HatEval	BERT	0.503
	BERTweet	**0.527**
	HateBERT	0.449

Table 5.8: Classifier performance after fine-tuning on each human-annotated dataset with multi-class labels for each prompting strategy. Evaluation metrics are macro averages.

Dataset	Precision	Recall	F1	Accuracy
Davidson	0.75	0.61	0.55	0.61
Founta	0.60	0.46	0.38	0.46
AbusEval	0.60	0.55	0.55	0.61
Waseem	0.80	0.63	0.68	0.79
Golbeck	0.65	0.65	0.64	0.65
OffensEval	0.76	0.73	0.73	0.76
HatEval	0.75	0.75	0.75	0.75

Table 5.9: Performance of human vs GPT-4 general prompt annotation. Evaluation metrics are macro averages.

	AAE				SAE			
Dataset	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Davidson	0.70	0.55	0.50	0.62	0.73	0.63	0.54	0.61
Founta	0.55	0.48	0.34	0.35	0.60	0.45	0.38	0.48
AbusEval	0.50	0.50	0.50	0.65	0.61	0.56	0.55	0.60
Waseem	0.78	0.68	0.71	0.79	0.80	0.61	0.67	0.79
Golbeck	0.68	0.68	0.68	0.68	0.65	0.64	0.64	0.64
OffensEval	0.82	0.81	0.81	0.82	0.75	0.72	0.72	0.75
HatEval	0.74	0.72	0.73	0.75	0.75	0.75	0.75	0.75

Table 5.10: Performance of human vs GPT-4 general prompt annotation with datasets conditioned on dialect. Evaluation metrics are macro averages.

Dataset	Precision	Recall	F1	Accuracy
Davidson	0.78	0.75	0.75	0.75
Founta	0.64	0.51	0.46	0.51
AbusEval	0.58	0.57	0.57	0.64
Waseem	0.80	0.75	0.78	0.84
Golbeck	0.67	0.67	0.67	0.67
OffensEval	0.78	0.75	0.76	0.78
HatEval	0.76	0.76	0.76	0.76

Table 5.11: Performance of human vs GPT-4 few-shot learning prompt annotation. Evaluation metrics are macro averages.

	AAE				SAE			
Dataset	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Davidson	0.71	0.67	0.68	0.71	0.77	0.77	0.75	0.78
Founta	0.59	0.61	0.52	0.54	0.65	0.48	0.43	0.50
AbusEval	0.50	0.52	0.50	0.68	0.59	0.57	0.58	0.64
Waseem	0.83	0.88	0.82	0.87	0.80	0.73	0.76	0.84
Golbeck	0.71	0.71	0.71	0.71	0.67	0.67	0.67	0.67
OffensEval	0.86	0.86	0.86	0.86	0.77	0.74	0.74	0.77
HatEval	0.75	0.74	0.75	0.76	0.76	0.76	0.76	0.76

Table 5.12: Performance of human vs GPT-4 few-shot learning prompt annotation with datasets conditioned on dialect. Evaluation metrics are macro averages.

Dataset	Precision	Recall	F1	Accuracy
Davidson	0.79	0.76	0.76	0.76
Founta	0.69	0.60	0.56	0.60
AbusEval	0.62	0.61	0.61	0.66
Waseem	0.81	0.74	0.77	0.84
Golbeck	0.67	0.66	0.66	0.66
OffensEval	0.79	0.71	0.72	0.76
HatEval	0.79	0.78	0.78	0.78

Table 5.13: Performance of human vs GPT-4 chain-of-thought prompt annotation. Evaluation metrics are macro averages.

	AAE				SAE			
Dataset	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Davidson	0.77	0.71	0.73	0.74	0.78	0.77	0.75	0.77
Founta	0.69	0.74	0.67	0.68	0.68	0.56	0.51	0.58
AbusEval	0.54	0.55	0.53	0.70	0.62	0.61	0.61	0.66
Waseem	0.86	0.88	0.85	0.87	0.81	0.71	0.75	0.84
Golbeck	0.69	0.69	0.69	0.69	0.66	0.66	0.66	0.66
OffensEval	0.88	0.87	0.87	0.88	0.78	0.69	0.69	0.74
HatEval	0.81	0.80	0.80	0.81	0.78	0.78	0.77	0.77

Table 5.14: Performance of human vs GPT-4 chain-of-thought prompt annotation with datasets conditioned on dialect. Evaluation metrics are macro averages.

		Human annotated					GPT-4 annotated (general prompt)				
Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.334	0.292	8	***	1.144	0.206	0.184	8.546	***	1.116
	Implicit	0.234	0.276	-7.751	***	0.848	0.285	0.327	-5.97	***	0.872
OffensEval	Offensive	0.488	0.408	8.123	***	1.195	0.386	0.275	12.007	***	1.406
HatEval	Hate	0.497	0.379	12.541	***	1.311	0.444	0.264	17.17	***	1.678
Davidson	Hate	0.345	0.331	3.173	0.002	1.042	0.050	0.051	-1.143		0.985
	Offensive	0.397	0.241	19.505	***	1.644	0.497	0.291	15.226	***	1.708
Founta	Hate	0.372	0.367	0.781		1.013	0.031	0.027	4.429	***	1.149
	Abuse	0.366	0.273	9.72	***	1.342	0.163	0.099	9.488	***	1.657
Waseem	Racism	0.099	0.104	-9.205	***	0.954	0.048	0.048	-0.905		0.992
	Sexism	0.269	0.232	5.99	***	1.158	0.124	0.112	3.432	***	1.106
	R & S	0.045	0.046	-2.483	0.013	0.981	0.025	0.025	2.476	0.013	1.036
Golbeck	Harassment	0.501	0.461	8.578	***	1.088	0.584	0.499	11.141	***	1.170

Table 5.15: Racial bias analysis of fine-tuned BERT model on human annotated datasets (left) and GPT-4 (right) annotated datasets using general prompt annotation.

Dataset	class	Few-shot Prompt annotation					CoT Prompt Annotation				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.323	0.307	3.735	***	1.053	0.325	0.310	3.806	***	1.049
	Implicit	0.157	0.164	-1.876		0.958	0.166	0.184	-4.607	***	0.903
OffensEval	Offensive	0.387	0.283	10.649	***	1.368	0.348	0.262	9.71	***	1.327
HatEval	Hate	0.439	0.277	16.322	***	1.588	0.489	0.320	16.233	***	1.529
Davidson	Hate	0.193	0.223	-10.378	***	0.865	0.189	0.207	-6.559	***	0.913
	Offensive	0.512	0.314	20.523	***	1.632	0.522	0.329	19.096	***	1.589
Founta	Hate	0.100	0.047	14.582	***	2.121	0.101	0.072	12.428	***	1.409
	Abuse	0.150	0.108	8.618	***	1.394	0.239	0.129	12.207	***	1.854
Waseem	Racism	0.078	0.081	-7.814	***	0.965	0.061	0.062	-2.165	0.031	0.983
	Sexism	0.185	0.166	5.879	***	1.119	0.170	0.138	7.861	***	1.235
	R & S	0.036	0.035	2.987	0.003	1.020	0.033	0.031	5.072	***	1.074
Golbeck	Harassment	0.552	0.456	12.274	***	1.211	0.613	0.522	13.114	***	1.175

Table 5.16: Racial bias analysis of fine-tuned BERT model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right).

Dataset	class	Human annotated					GPT-4 annotated (general prompt)				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.214	0.188	5.946	***	1.136	0.165	0.146	9.272	***	1.127
	Implicit	0.257	0.269	-5.267	***	0.957	0.231	0.221	3.032	0.002	1.044
OffensEval	Offensive	0.415	0.393	6.075	***	1.058	0.294	0.257	9.879	***	1.144
HatEval	Hate	0.426	0.331	18.85	***	1.286	0.319	0.238	13.578	***	1.343
Davidson	Hate	0.246	0.210	7.955	***	1.169	0.064	0.065	-2.179	0.029	0.988
	Offensive	0.366	0.237	22.687	***	1.544	0.568	0.427	14.721	***	1.332
Founta	Hate	0.315	0.234	16.865	***	1.343	0.049	0.038	11.38	***	1.300
	Abuse	0.459	0.441	2.146	0.032	1.042	0.217	0.141	11.993	***	1.543
Waseem	Racism	0.071	0.069	10.46	***	1.036	0.051	0.049	8.747	***	1.031
	Sexism	0.237	0.212	6.087	***	1.118	0.137	0.131	2.91	0.004	1.045
	R & S	0.052	0.051	4.854	***	1.017	0.042	0.041	6.746	***	1.025
Golbeck	Harassment	0.474	0.471	2.11	0.035	1.005	0.574	0.555	7.816	***	1.034

Table 5.17: Racial bias analysis of fine-tuned BERTweet model on human annotated datasets (left) and GPT-4 (right) annotated datasets using general prompt annotation.

Dataset	class	Few-shot Prompt annotation					CoT Prompt Annotation				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.190	0.150	6.512	***	1.264	0.171	0.134	6.04	***	1.272
	Implicit	0.169	0.161	4.551	***	1.047	0.171	0.169	1.562		1.015
OffensEval	Offensive	0.316	0.270	8.715	***	1.171	0.289	0.232	12.054	***	1.246
HatEval	Hate	0.331	0.254	13.473	***	1.303	0.330	0.246	17.084	***	1.345
Davidson	Hate	0.141	0.135	2.213	0.027	1.049	0.144	0.126	6.261	***	1.147
	Offensive	0.475	0.374	19.107	***	1.269	0.483	0.377	16.802	***	1.281
Founta	Hate	0.174	0.065	24.99	***	2.693	0.169	0.076	28.812	***	2.233
	Abuse	0.173	0.096	20.796	***	1.815	0.325	0.216	15.339	***	1.505
Waseem	Racism	0.063	0.061	10.357	***	1.036	0.060	0.058	11.948	***	1.036
	Sexism	0.181	0.168	4.858	***	1.078	0.177	0.164	5.638	***	1.081
	R & S	0.050	0.048	9.919	***	1.036	0.047	0.046	9.344	***	1.028
Golbeck	Harassment	0.546	0.533	5.177	***	1.025	0.571	0.555	7.212	***	1.028

Table 5.18: Racial bias analysis of fine-tuned BERTweet model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right).

Dataset	class	Human annotated					GPT-4 annotated (general prompt)				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.242	0.217	4.988	***	1.115	0.169	0.147	7.207	***	1.149
	Implicit	0.197	0.190	1.992	0.047	1.032	0.236	0.215	4.42	***	1.096
OffensEval	Offensive	0.449	0.408	7.439	***	1.100	0.344	0.281	10.661	***	1.226
HatEval	Hate	0.524	0.435	14.574	***	1.203	0.413	0.311	15.015	***	1.329
Davidson	Hate	0.228	0.180	9.036	***	1.267	0.051	0.060	-9.956	***	0.853
	Offensive	0.344	0.191	20.355	***	1.796	0.535	0.323	19.051	***	1.657
Founta	Hate	0.321	0.281	8.587	***	1.143	0.041	0.036	5.574	***	1.144
	Abuse	0.389	0.313	10.512	***	1.241	0.179	0.121	9.659	***	1.478
Waseem	Racism	0.085	0.077	7.717	***	1.107	0.058	0.052	8.613	***	1.125
	Sexism	0.265	0.236	4.858	***	1.123	0.109	0.111	-0.811		0.979
	R & S	0.051	0.052	-0.752		0.982	0.033	0.035	-2.206	0.028	0.938
Golbeck	Harassment	0.466	0.453	5.04	***	1.029	0.573	0.542	7.906	***	1.059

Table 5.19: Racial bias analysis of fine-tuned HateBERT model on human annotated datasets (left) and GPT-4 (right) annotated datasets using general prompt annotation.

Dataset	class	Few-shot Prompt annotation					CoT Prompt Annotation				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.217	0.179	6.527	***	1.211	0.207	0.174	5.816	***	1.192
	Implicit	0.144	0.133	4.988	***	1.083	0.153	0.146	2.815	0.005	1.045
OffensEval	Offensive	0.357	0.294	10.195	***	1.212	0.349	0.267	13.016	***	1.310
HatEval	Hate	0.419	0.323	14.512	***	1.298	0.433	0.350	14.201	***	1.239
Davidson	Hate	0.122	0.124	-0.53		0.985	0.115	0.104	2.933	0.003	1.102
	Offensive	0.478	0.283	22.179	***	1.685	0.492	0.305	20.111	***	1.615
Founta	Hate	0.112	0.055	13.134	***	2.039	0.130	0.064	14.379	***	2.029
	Abuse	0.141	0.096	10.735	***	1.475	0.276	0.192	11.634	***	1.434
Waseem	Racism	0.078	0.068	9.471	***	1.143	0.074	0.063	11.93	***	1.169
	Sexism	0.170	0.152	4.151	***	1.115	0.153	0.136	5.226	***	1.125
	R & S	0.044	0.045	-0.054		0.999	0.040	0.041	-0.152		0.996
Golbeck	Harassment	0.551	0.515	8.364	***	1.070	0.576	0.541	8.741	***	1.064

Table 5.20: Racial bias analysis of fine-tuned HateBERT model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right).

Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
OffensEval	Offensive	0.375	0.260	11.794	***	1.444
HatEval	Hate	0.374	0.260	13.459	***	1.440
Davidson	Hate	0.070	0.075	-5.332	***	0.933
	Offensive	0.412	0.297	10.607	***	1.385
Founta	Hate	0.026	0.025	1.251		1.047
	Abuse	0.114	0.090	4.321	***	1.257

Table 5.21: Racial bias analysis of fine-tuned BERT model on GPT-4 annotated datasets using general prompt annotation with dialect priming.

Dataset	class	Few-shot Prompt annotation					CoT Prompt Annotation				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
OffensEval	Offensive	0.334	0.214	12.773	***	1.564	0.304	0.228	9.044	***	1.333
HatEval	Hate	0.388	0.212	17.361	***	1.831	0.472	0.286	17.099	***	1.652
Davidson	Hate	0.157	0.153	2.116	0.034	1.026	0.101	0.089	8.251	***	1.138
	Offensive	0.394	0.265	15.033	***	1.488	0.477	0.306	15.207	***	1.559
Founta	Hate	0.056	0.045	5.383	***	1.243	0.069	0.057	10.221	***	1.209
	Abuse	0.130	0.095	6.206	***	1.365	0.319	0.170	12.794	***	1.880

Table 5.22: Racial bias analysis of fine-tuned BERT model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right) with dialect priming.

Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
OffensEval	Offensive	0.262	0.214	10.97	***	1.228
HatEval	Hate	0.306	0.241	11.526	***	1.270
Davidson	Hate	0.071	0.070	2.008	0.045	1.016
	Offensive	0.439	0.366	9.091	***	1.199
Founta	Hate	0.040	0.035	6.396	***	1.147
	Abuse	0.146	0.117	6.642	***	1.242
Golbeck	Harassment	0.552	0.537	6.475	***	1.028

Table 5.23: Racial bias analysis of fine-tuned BERTweet model on GPT-4 annotated datasets using general prompt annotation with dialect priming.

		Few-shot Prompt annotation					CoT Prompt Annotation				
Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
OffensEval	Offensive	0.265	0.211	13.841	***	1.251	0.260	0.219	11.544	***	1.188
HatEval	Hate	0.310	0.226	13.133	***	1.372	0.332	0.245	14.859	***	1.357
Davidson	Hate	0.118	0.102	6.662	***	1.150	0.108	0.100	9.263	***	1.087
	Offensive	0.339	0.285	10.795	***	1.189	0.494	0.394	14.732	***	1.254
Founta	Hate	0.070	0.054	8.604	***	1.307	0.094	0.076	16.866	***	1.233
	Abuse	0.134	0.100	8.813	***	1.343	0.539	0.391	13.284	***	1.377
Golbeck	Harassment	0.320	0.295	7.492	***	1.085	0.494	0.478	5.365	***	1.034

Table 5.24: Racial bias analysis of fine-tuned BERTweet model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right) with dialect priming.

Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
OffensEval	Offensive	0.292	0.254	7.433	***	1.150
HatEval	Hate	0.377	0.313	10.59	***	1.206
Davidson	Hate	0.056	0.064	-8.499	***	0.869
	Offensive	0.393	0.282	12.13	***	1.392
Founta	Hate	0.033	0.034	-0.781		0.978
	Abuse	0.123	0.111	2.789	0.005	1.110
Golbeck	Harassment	0.585	0.557	7.124	***	1.050

Table 5.25: Racial bias analysis of fine-tuned HateBERT model on GPT-4 annotated datasets using general prompt annotation with dialect priming.

		Few-shot Prompt annotation					CoT Prompt Annotation				
Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
OffensEval	Offensive	0.310	0.242	14.634	***	1.280	0.320	0.250	14.005	***	1.277
HatEval	Hate	0.359	0.256	14.249	***	1.403	0.440	0.337	15.697	***	1.305
Davidson	Hate	0.094	0.089	2.069	0.039	1.057	0.082	0.081	1.018		1.021
	Offensive	0.348	0.235	15.85	***	1.483	0.490	0.312	19.373	***	1.573
Founta	Hate	0.061	0.054	3.498	***	1.126	0.078	0.065	10.596	***	1.208
	Abuse	0.139	0.105	7.016	***	1.321	0.416	0.281	13.571	***	1.481
Golbeck	Harassment	0.391	0.342	9.746	***	1.142	0.501	0.466	7.833	***	1.075

Table 5.26: Racial bias analysis of fine-tuned HateBERT model on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right) with dialect priming.

Chapter 6

AI-Cybersecurity Education Through Designing AI-based Cyberharassment Detection Lab

This work, accepted at the Frontiers in Education Conference (FIE) 2023, has been published as an Archive preprint.

6.1 Abstract

Cyberharassment is a critical, socially relevant cybersecurity problem because of the adverse effects it can have on targeted groups or individuals. While progress has been made in understanding cyber-harassment, its detection, attacks on artificial intelligence (AI) based cyberharassment systems, and the social problems in cyberharassment detectors, little has been done in designing experiential learning educational materials that engage students in this emerging social cybersecurity in the era of AI. Experiential learning opportunities are usually provided through capstone projects and engineering design courses in STEM programs such as computer science. While capstone projects are an excellent example of experiential learning, given the interdisciplinary nature of this emerging social cybersecurity

problem, it can be challenging to use them to engage non-computing students without prior knowledge of AI. Because of this, we were motivated to develop a hands-on lab platform that provided experiential learning experiences to non-computing students with little or no background knowledge in AI and discussed the lessons learned in developing this lab. In this lab used by social science students at North Carolina A&T State University across two semesters (spring and fall) in 2022, students are given a detailed lab manual and are to complete a set of well-detailed tasks. Through this process, students learn AI concepts and the application of AI for cyberharassment detection. Using pre- and post-surveys, we asked students to rate their knowledge or skills in AI and their understanding of the concepts learned. The results revealed that the students moderately understood the concepts of AI and cyberharassment. Comparing the learning experiences of students in the Spring and Fall semesters using the post-surveys, in the Spring semester, while students understood the purpose of detecting cyberharassment, their knowledge of how AI works, state-of-the-art (SOTA) cyber-harassment detectors and automated cyberharassment detection did not improve. After refining the lab using the student feedback from the Spring semester, students' knowledge of AI and automated cyberharassment detection significantly improved. We learn that supplementing the hands-on lab with a theoretical background lecture improves understanding of the lab contents, indicating that supplementing experiential learning with theoretical constructs improves the student learning experience. These findings confirm that the developed lab is viable for teaching AI-driven socially relevant cybersecurity to non-computing students and can be used by other institutions.

6.2 Introduction

With the rise of social media, cyberharassment (*e.g.*, cyberbullying and cyberhate) has become more prevalent in daily interactions [205]. It often involves *inappropriate online behavior* and *deliberate cyber threats* against individuals (such as teenagers [202]), or specific social groups on the grounds of characteristics such as race, sexual orientation, gen-

der, or religious affiliation [220]. Cyberharassment is identified as a critical socially-relevant cybersecurity problem [2, 6], since it can have significant negative impacts on the safety and emotional well-being of targeted groups, especially teens and minority communities. The Cyberbullying Research Center's research reported that 37% of middle and high school students have been cyberbullied during their lifetime [164], and this number is expected to further increase as teens continue to have an increased online presence. Cyberharassment can even result in catastrophic consequences of increased suicide among the affected teens who are unable to appropriately get away from the harassment [54]. The shift from traditional text-based cyberharassment to *multimodal* (*i.e.*, both texts and images) [4] cyberharassment poses a challenge to effective cyberharassment detection.

Artificial Intelligence (AI)/Machine Learning (ML) has immense potential to solve this critical problem. Automatic detection methods of both text-based and image-based cyberbullying using AI techniques have emerged [248]. Internet companies such as Facebook and Google have also deployed AI algorithms to detect toxic content on social media [1, 3]. Meanwhile, adversaries may exploit vulnerabilities of AI-based classifiers to evade existing cyberharassment detectors [74, 121, 245]. There exist *social problems*, such as fairness and ethics, in AI models for cyberharassment detection. For example, some particular demographic groups are unfairly treated by AI-based detectors [158, 187]. Concerns have been raised that the vulnerabilities of AI models as well as the robustness against attacks are biased towards underrepresented groups [146]. As such, an unfair AI-based cyberharassment detection system may perpetuate and aggravate existing prejudices and inequalities in society.

As cyberharassment grows online, particularly on social media, there is a need to equip computing students with the AI skills and knowledge required to design and develop AI-based systems to detect and remove cyberharassment. As the field of cyberharassment is interdisciplinary, to develop better detection systems, non-computing students, especially social science students, need to have a general understanding of AI and how it is being used in detecting cyberharassment. A major concern with University training is how siloed it can

be and how challenging it is for young adults to truly explore their options. Interventions such as the one outlined here would enable effective collaboration between Computer science and social science students and researchers to create better cyberharassment detection systems. It is, therefore, imperative to teach AI-based cyberharassment in universities to both computing and non-computing students. This intervention also benefits social science students by providing insights into how they can address social problems using computer science, a discipline they may have otherwise had no exposure. This practical application of their social scientific passions may influence the degree minors or graduate programs students pursue moving forward as more researchers are needed in this intersectional area since this is a growing area of concern.

To teach and engage students in learning cybersecurity and AI-related topics such as data science, instructors have adopted a wide range of pedagogical methods such as flipped classroom [58, 155], project-based learning [154, 215], gamification [137], among others. Experiential learning has been regarded as one of the best ways to train future engineers by engineering educators [184]. Towards this end, experiential learning could be used in teaching AI socially relevant cybersecurity to non-computing students.

Experiential learning, simply put, is learning from experience or learning by doing. More formally, experiential learning is a type of active learning where students learn through experience [105, 106]. Experiential learning is active rather than passive. Instructors have recognized how instrumental experiential learning is in providing students with valuable hands-on experience in an AI/ML-related field such as data science [9, 13, 18, 180]. In our study, we teach AI-based socially relevant cybersecurity for cyberharassment detection for two semesters using a hands-on experiential lab. Before the introduction of the lab, a questionnaire was used to rate the AI skills and knowledge of the students. After the end of the lab, another questionnaire was used to ask the students to rate their skills and knowledge of AI and AI-based cyberharassment detection covered in the lab.

Our analysis and statistical results (using sample t-test) showed that experiential learning engages students in learning AI-based cyberharassment, and it is viable for teaching

AI skills to non-computing students. Also, our findings show that having a theoretical lecture before the experiential lab improves understanding of the lab contents. These findings confirm that the developed lab is viable for teaching AI-driven socially relevant cybersecurity to non-computing students and can be used in other institutions.

6.3 Related Work

Instructors have mainly employed active learning paradigms such as experiential learning to teach AI/ML-related courses, mostly in engineering and Computer Science. The shortcoming of standard pedagogical methods in data science in online courses and data science specializations are detailed in [193]. Noting that experiential learning mitigates the shortcomings as it focuses on problems to be solved instead of on specific methods being used. Using the experiential learning style theory introduced in [106], they developed a framework to create experiences in a deep learning course. Finally, they review dataset repositories used in data science and propose requirements for an experiential learning platform to offer experiences.

Understanding the importance of capstone projects in data science courses, [9] observed that more attention needs to be paid to how these projects or curriculum are structured. In their work, they develop an interdisciplinary, client-sponsored capstone program in data science and machine learning. In the program, students from different undergraduate and graduate degree programs engage in experiential learning by completing a large-scale data science or machine learning capstone project toward the program's end— the projects were framed to be challenging and encompass all aspects of data science. The curriculum was split into modules focusing on the data science pipeline, ethics, and communication. It was developed using evidence-based approaches from capstone and design programs in engineering, practicums, and other project-based courses.

In [18], the challenges in using capstone projects as experiential learning opportunities in data science courses due to resource constraints and data legalities involved in

students working with clients on clients' real-world data sets are emphasized. A novel client-facing consulting data science course that provides experiential learning to undergraduate and graduate students is developed to tackle this issue.

In an ethics and data analytics course where engineering students developed solutions for smart grid, smart health, and smart mobility, [132] explored how professional responsibility is understood by engineering students working on a solution to a real-world problem proposed by a client. The authors acknowledge that as technology such as AI/ML advances, it presents complex challenges that require an interdisciplinary approach. In line with this, they introduce students to real-life problems presented as socio-technical challenges embedded in a learning context using Challenge Based Learning (CBL), an experiential learning method.

The most recent works more closely related to our work are the works of [182] and [81]. AI education is a challenging task because it is not well studied, making it one of the challenges in engineering education [182]. With this knowledge, through a series of directives, acts, and laws, the United States of America government highlighted the importance of the Department of Defence to embrace AI at speed and scale [182]. In [182], the Department of Defence (DoD) and the United States Air Force (USAF) partnered with MIT to design and develop educational research activities that will provide AI training for DoD and USAF personnel with diverse professional and educational backgrounds from high school to graduate degrees, and to the general public. The educational research activities explored various ways to teach AI education, online and in-person, to deliver experiential learning experiences to learners of diverse backgrounds. In the program's first iteration, they developed and studied the learning journeys of three different types of USAF employees (leaders, developers, and users). The findings will be used in future iterations of the program.

The authors in [215] design a data science course for non-computer science students. The course goal was to develop a new method for designing a data science course suitable for teaching students with different backgrounds in data science. The goal of the new method was to ensure that students gained skills on how to set up, manage, and conduct data science

projects. The data science course was taught in Business Analytics and Data Science and the Digital Humanities master's programs. Students were introduced to KNIME, an open-source analytics platform for creating data science workflows without coding. Students were also encouraged to use Python and R programming languages if they were already familiar with them. The results reveal that the course is well designed and structured for students with different backgrounds, that the students gained skills to carry out a data science project, and that students liked the analytics platform used in the project work.

The popularity of AI/ML has led to the proliferation of research studies on designing better data science education materials, as shown in some of the works reviewed. These works mainly focused on designing data science courses for engineering and computing students, with a few focusing on non-computing students. Most importantly, these works do not focus on experiential, hands-on labs that will provide students with the experience to bridge the gap between theoretical knowledge and practice. We fill this gap by developing experiential, hands-on labs for non-computing students and discuss the lessons learned in developing these hands-on labs.

6.4 Design & Development of AI Socially Relevant Cyber-harassment Lab

6.4.1 Lab Structure

The experiential learning laboratory consists of two labs where students become familiar with the basics of AI and the AI/ML pipeline for applying ML to a problem.

6.4.1.1 Objectives

We designed the experiential learning laboratories with specific learning objectives. Essential for guiding student learning, the learning objectives guided and helped develop our laboratories. This ensured that our study covered the fundamental elements of AI and different dimensions of AI-driven social cybersecurity. It also demonstrated the interplay

between AI and cybersecurity and how AI is used for cyberharassment detection. Specifically, our learning objective is to develop hands-on experiential labs that will increase general awareness of socially relevant cybersecurity and AI, which is suitable for teaching AI socially relevant cybersecurity to non-computing students.

6.4.1.2 Lab

Our lab adopts a phased design approach. Initially, AI and cybersecurity experts in our team designed the preliminary lab and implemented and integrated the cyberharassment detection code in the lab. The lab was designed considering the student profiles, AI and cyberharassment learning objectives, and desired skills. Subsequently, social scientists in our team engaged in dialogues with the AI and cybersecurity scholars to enhance the lab's reach and inclusivity. After the concerns of the social scientists have been addressed in the next iteration of the lab, the researchers collaboratively designed the lecture sessions and lab assignments. The collaborative and interdisciplinary approach ensured that the lab was accessible to a broad range of learners.

Our lab, *Cyberbullying Detection Using AI*, is designed to guide students through a series of learning objectives and the AI development process. The learning objectives include understanding AI, understanding the concept and severity of cyberharassment, the importance of using AI in addressing this widespread online social issue, and introduction to the development and the use of AI systems for cyberharassment detection. For the AI development process, the AI experts ensured that the design followed the AI development pipeline (data collection, verification and preprocessing, feature extraction, AI system training, and testing, and use of trained AI systems on a task) to provide students with the fundamentals of the processes followed to develop an AI system. In parallel, the social scientists offered profound insights into the peculiarities of cyberharassment, highlighting its significance and the pressing need for AI intervention. All researchers from diverse academic backgrounds involved in this work cross-verified the lab content to ensure the lab is easily accessible to non-computing students. This process ensures that those new to AI and

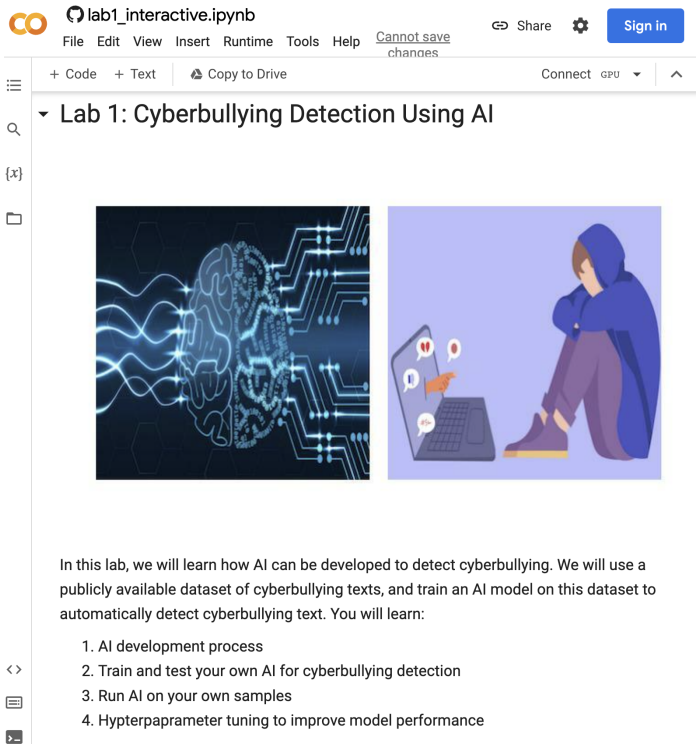


Figure 6.1: A screenshot illustrating the lab interface

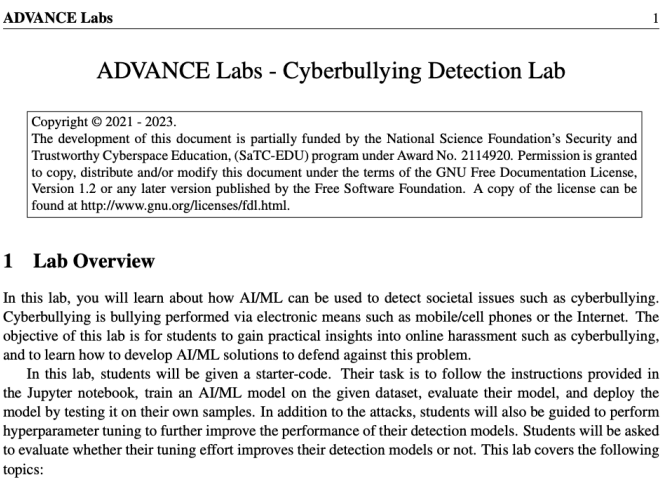
cyberharassment can easily grasp and engage with our lab materials. Figure 6.2 shows a comprehensive instruction manual prepared for the lab to facilitate independent learning and ensure students accomplish the learning objectives outside of the classroom.

6.4.1.3 Lab Delivery

To facilitate a comprehensive understanding of our lab, we implement a three-fold approach that includes a lecture phase, an experiential hands-on experience phase, and a phase dedicated to independent work.

Before the students can work on the hands-on lab, a background lecture designed and developed by the team of researchers is given to the students. In the lecture, the students get acquainted with the nature and nuances of cyberharassment, AI, the AI development process, and the need for utilizing AI for cyberharassment detection.

After introducing the students to the fundamentals needed to complete the lab



ADVANCE Labs 1

ADVANCE Labs - Cyberbullying Detection Lab

The development of this document is partially funded by the National Science Foundation's Security and Trustworthy Cyberspace Education, (SaTC-EDU) program under Award No. 2114920.

1 Lab Overview

In this lab, you will learn about how AI/ML can be used to detect societal issues such as cyberbullying. Cyberbullying is bullying performed via electronic means such as mobile/cell phones or the Internet. The objective of this lab is for students to gain practical insights into online harassment such as cyberbullying, and to learn how to develop AI/ML solutions to defend against this problem.

In this lab, students will be given a starter-code. Their task is to follow the instructions provided in the Jupyter notebook, train an AI/ML model on the given dataset, evaluate their model, and deploy the model by testing it on their own samples. In addition to the attacks, students will also be guided to perform hyperparameter tuning to further improve the performance of their detection models. Students will be asked to evaluate whether their tuning effort improves their detection models or not. This lab covers the following topics:

Figure 6.2: A screenshot illustrating Lab 1 instruction manual

through the lecture, students are introduced to the hands-on lab and guided through the hands-on experience platform depicted in Figure 6.1. The lab is developed on the Google Colab platform. The hands-on experience allows students to apply theory in practice. It facilitates a deeper understanding of how AI solutions are developed and, most importantly, how AI can be utilized to mitigate cyberharassment. We have chosen to utilize the Google Colab platform for several compelling reasons. First and foremost, Google Colab provides an interactive environment that integrates text and code cells. This not only enables us to write and execute Python code for deploying AI models and detecting cyberharassment content, but it also allows us to provide clear, step-by-step explanations alongside the code. Moreover, these text cells can be utilized to embed visual aids and explanations such as Figure 6.3, facilitating a more comprehensive understanding of the concepts and processes involved. Secondly, Google Colab offers access to free GPU resources. This is a significant advantage for our participants as the computational power of GPUs can greatly expedite the execution of AI models, ensuring that experiments are completed within a reasonable time frame. Furthermore, Google Colab's cloud-based nature eliminates the need for complex setups on personal machines, lowering the entry barrier for participants. This easy access, combined with the platform's robust functionality, makes Google Colab an ideal tool for our hands-on experiments in AI education.

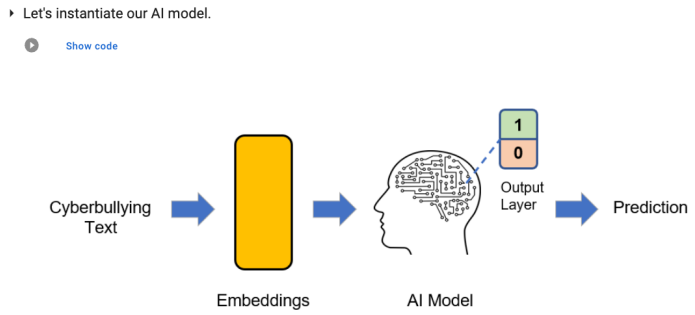


Figure 6.3: The visual aid for model explanation in Lab 1

Figure 6.4: One example of a lab activity

In our labs, we have meticulously curated post-lab assignments that not only enhance the engagement factor but also deepen students' comprehension of the material. For example, as depicted in Figure 6.4, students are tasked with using arbitrary input statements to test their AI's capability to recognize cyberharassment content. In the Google Colab platform, the correct execution of a cell could depend on the execution of the previous cell, and students are made aware of this, which is also in the lab manual. In the example activity shown in Figure 6.4, students must complete all prior steps to access the developed AI model for cyberharassment detection.

Additionally, we have formulated post-lesson discussion questions. Figure 6.5 depicts an example question. We aim to stimulate students' critical thinking by encouraging them to consider other real-world problems that could be alleviated through AI.

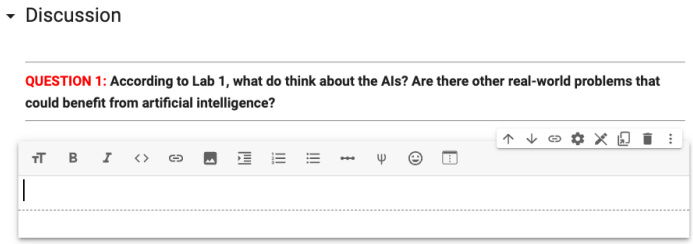


Figure 6.5: One example of Lab 1's discussion questions

	Spring 2022	Fall 2022
Demographic	Percent	Percent
Male	23.81%	14.29%
Female	71.43%	85.71%
Freshman	23.81%	4.76%
Sophomore	33.33%	66.67%
Junior	38.10%	9.52%
Senior	4.76%	14.29%
Graduate	0.00%	4.76%

Table 6.1: Demographics of students who participated in the Social Statistics 1 course in the Spring and Fall 2022 semesters.

6.4.2 Course Structure

The Sociology program at North Carolina A&T State University includes courses on Social Statistics, parts 1 and 2. Students from Social Statistics 1 were included in this intervention. Students in Social Statistics 1 learn how to interpret and describe data. Students are exposed to topics such as distributions; descriptive statistics (e.g., measures of central tendency and dispersion); statistical null-hypothesis testing, and independent and dependent samples t-tests. They also learn the basic operation of the statistical computing software program SPSS. Throughout the course, students learn the value statistical analysis offers to our attempts to address real-world social problems. This course runs 15 weeks and includes lecture and lab time for analysis. Students attend two 75-minute classes each week. Lectures involve definitions and the demonstration of practice problems. Lab time involves hands-on computing using SPSS. In Table 6.1, we report some basic demographics of the students in the Social Statistics 1 class in the Spring and Fall semesters.

6.5 Methods

6.5.1 Survey Design

To evaluate the impact of our lab on students, we designed two surveys (pre-survey and post-survey) to collect quantitative and qualitative data from the students. Students rate their knowledge about AI and machine learning methods for cyberharassment detection. For both surveys, students rated their knowledge using a 5-point scale, with 1 representing "Proficient or strongly agree" and 5 representing "None or strongly disagree". In the post-survey, students also rate the lab from different perspectives. The post-survey also includes three open questions to understand what helped the students the most in understanding the concepts taught, the difficulties faced in using the lab, and any suggestions for improving the learning experience of the lab.

6.5.2 Data Collection

Before our data collection, we obtained our institution's Institution Review Board (IRB) approval. All the students were notified that the survey would remain confidential and only the research team could view the data. Students were also informed that their participation was optional. Before the class, students filled out the pre-survey to evaluate their knowledge level, and after the lab, students filled the post-survey. Our surveys are conducted using the Qualtrics platform [5].

6.5.3 Data Analysis

To analyze the collected data, we first compared the average knowledge score between pre-survey and post-survey. Then we used the sample t-test to determine the existence of statistical significance between the observed differences. We use a significance level of 0.05. If the p-value of the t-test is less than 0.05, we conclude a significant difference exists. Scipy, an open-source Python library for scientific computing, was used to analyze the collected data.

1 = Proficient, 5 = None	Pre	Post
Automated cyberharassment detection.	3.95	3.56
State-of-The-Art cyberharassment Detectors.	4.33	3.67
How Machine Learning Works.	4.10	3.89
The lab engaged me in learning the topic of AI Driven Socially-Relevant Cybersecurity.	-	2.78
I enjoyed the learning experience of this lab(s)	-	2.89
I think the learning experience with the lab(s) is effective	-	2.78
I am satisfied with the level of effort the lab requires for learning this topic	-	2.78
After using the lab(s), I have better understanding about the concepts learned	-	3.44

Table 6.2: Spring 2022 semester survey results.

6.6 Results and Discussion

We analyze the data for the Spring 2022 and Fall 2022 semesters and compare the learning experiences of both semesters because the student feedback from the Fall 2022 semester, as discussed in the lesson learned section 6.6.10, provided data and insights that were used to inform and refine the lab in the Spring 2022 semester. Before introducing the labs to the students, they were invited to complete a pre-questionnaire, and after completing the labs, they were invited to complete a post-questionnaire. The pre and post-questionnaire contained eleven questions in total. The first three questions were asked in both the pre and post-questionnaire. The remaining eight questions were only asked in the post-questionnaire in addition to the first three questions. Of the eight questions, the last three were open-ended questions. In the initial data analysis, we focus on the first three questions in the pre-and-post questionnaires administered to the students. In the final data analysis, we

1 = Proficient, 5 = None	Pre	Post
Automated Cyberharassment Detection.	4.38	3.31
State-of-The-Art cyberharassment detectors.	4.71	3.44
How Machine Learning Works.	4.43	3.13
The lab engaged me in learning the topic of AI Driven Socially-Relevant Cybersecurity.	-	2.56
I enjoyed the learning experience of this lab(s).	-	2.44
I think the learning experience with the lab(s) is effective.	-	2.13
I am satisfied with the level of effort the lab requires for learning this topic.	-	2.56
After using the lab(s), I have better understanding about the concepts learned.	-	2.31

Table 6.3: Fall 2022 semester survey results.

focus on only the last eight questions in the post-questionnaire that the students completed after completing the labs to gauge their understanding and perception of the lab. The data analysis focused on survey participation, knowledge in automated cyberharassment detection, evaluation of ML classifiers, current state-of-the-art cyberharassment systems, how ML works, general learning experience and engagement of the labs, and student qualitative feedback. Tables 6.2 and 6.3 show the Spring 2022 and Fall 2022 survey results.

6.6.1 Survey Participation

The survey was presented to students in the Spring 2022 and Fall 2022 semesters. Of the 21 students enrolled in the course in Spring 2022, 21 (100%) students completed the pre-survey, and 9 (43.9%) students completed the post-survey. In the Fall 2022 semester, 21 students were enrolled in the class, of which all students completed the pre-survey, and 16 (80%) students completed the surveys—showing that the Fall 2022 session of the course had more student participation than when the lab was first introduced in Spring 2022.

6.6.2 Cyberharassment Detection

During the pre-survey of Spring 2022, 19.05% of the students rated their knowledge or skills in automated cyberharassment detection in the *Proficient* and *Good* category, about 14.29% of the students rated themselves as having moderate knowledge or skill, and more than half (66.67%) of the students rated their knowledge or skills in the *A little* and *None* category. After the post-survey, 33.33% of the students rated their knowledge or skills in the *Proficient* and *Good* category, 0% rated themselves as moderate, and 66.66% rated their knowledge in the *A little* and *None* category.

In the pre-survey Fall 2022, 4.76% of the students rated their knowledge in the *Proficient* and *Good* category, 19.05% as having moderate skill or knowledge, and 76.19% rated their knowledge or skills in the *A little* and *None* category. On the other hand, in the post-survey of Fall 2022, 25% of the students rated their knowledge in the *Proficient* and *Good* categories, 25% rated themselves as moderate, and 50% rated their knowledge in the *A little* and *None* category.

Most students had little knowledge of automated cyberharassment detection before enrolling in the class and participating in the lab. After completing the lab and comparing the means of the pre-survey and post-survey of Spring 2022 as shown in Table 6.2, there is no significant difference ($p > 0.05$) in the knowledge of automated cyberharassment detection. For the Fall 2022 semester, there is a significant difference ($p < 0.05$) in the knowledge of automated cyberharassment detection as indicated in Table 6.3. Indicating that the improvements made after the Spring 2020 semester helped improve students' knowledge or skill in the Fall 2022 semester.

6.6.3 State of the Art Cyberharassment Systems

We determined the knowledge or skills of students in the state-of-the-art cyberharassment systems before and after the lab. During the pre-survey of Spring 2022, 4.76% of the students rated their knowledge or skills in state-of-the-art systems in the *Proficient* and *Good* categories, about 19.05% of the students rated themselves as having moderate

knowledge or skill, and more than half (76.19%) of the students rated their knowledge or skills in the *A little* and *None* category. After the post-survey, 33.33% of the students rated their knowledge or skills in the *Proficient* and *Good* category, 0% rated themselves as moderate, and 66.66% rated their knowledge in the *A little* and *None* category.

In the pre-survey Fall 2022, 4.76% of the students rated their knowledge in the *Proficient* and *Good* categories, 4.76% as having moderate skill or knowledge, and 90.47% rated their knowledge or skills in the *A little* and *None* category. On the other hand, in the post-survey of Fall 2022, 18.75% of the students rated their knowledge in the *Proficient* and *Good* categories, 25% rated themselves as moderate, and 56.25% rated their knowledge in the *A little* and *None* category.

Many students had little knowledge of the state-of-the-art cyberharassment system before the lab. After completing the lab and comparing the means of the pre-survey and post-survey of Spring 2022, there is no significant difference ($p > 0.05$) in the knowledge of state-of-the-art systems. For the Fall 2022 semester, there is a significant difference ($p < 0.05$) in the knowledge of state-of-the-art systems. Similar to the knowledge of cyberharassment detection, the improvements made after the Spring 2020 semester helped in improving students' knowledge or skill in the Fall 2022 semester.

6.6.4 How Machine Learning Works

We also determined if the students knew how machine learning worked before the lab and if they learned how machine learning worked after the lab. From the pre-survey of Spring 2022, results show that 9.52% of the students rated their knowledge in how ML worked in the *Proficient* and *Good* categories, about 28.57% of the students rated themselves as having moderate knowledge, and more than half (61.9%) of the students rated their knowledge or skills in the *A little* and *None* category. After the post-survey, 22.22% of the students rated their knowledge in the *Proficient* and *Good* category, 0% rated themselves as moderate, and 77.77% rated their knowledge in the *A little* and *None* category.

In the pre-survey Fall 2022, 4.76% of the students rated their knowledge in the

Proficient and *Good* categories, 14.29% as having moderate skill or knowledge, and 80.96% rated their knowledge or skills in the *A little* and *None* category. On the other hand, in the post-survey of Fall 2022, 26.66% of the students rated their knowledge in the *Proficient* and *Good* categories, 33.33% rated themselves as moderate, and 40% rated their knowledge in the *A little* and *None* category.

In both semesters, more students had little knowledge of how ML worked before the lab. After completing the lab and comparing the means of the pre-survey and post-survey of Spring 2022, there is no significant difference ($p > 0.05$) in the knowledge of how ML worked. For the Fall 2022 semester, there is a significant difference ($p < 0.05$) in the knowledge of how ML worked. The lab refinements made after the Spring 2020 semester helped improve students' knowledge or skill in the Fall 2022 semester.

6.6.5 Lab Engagement

The students responded positively to how the lab engaged them in learning about AI-driven socially relevant cybersecurity. In the Spring of 2022, 44.44% of the students rated how engaging the lab was in the *Strongly agree* and *Somewhat agree* categories, 22.22% rated engagement as moderate, and 33.33% rated engagement in the *Somewhat disagree* and *Strongly disagree* category. In the Fall of 2022, more than half (62.5%) of the students rated how engaging the lab was in the *Strongly agree* and *Somewhat agree* categories, 12.50% rated engagement as moderate, and 25% rated engagement in the *Somewhat disagree* and *Strongly disagree* category.

The mean response by students in the Spring and Fall semesters was 2.78 and 2.56, respectively. The mean values are between the *Somewhat agree* and *Neither agree nor disagree* with the mean value of Fall closer to *Somewhat agree* than the Spring mean, indicating that the students' lab engagement was positive and not very negative.

6.6.6 Learning Experience

The student's response to the overall learning experience and the effectiveness of the experience was somewhat positive. In the Spring of 2022, 33.33% of the students rated the learning experience in the *Strongly agree* and *Somewhat agree* categories, 44.44% rated the learning experience as moderate, and 22.22% rated the learning experience in the *Somewhat disagree* and *Strongly disagree* category. For the effectiveness of the learning experience, 44.44% of the students rated the effectiveness in the *Strongly agree* and *Somewhat agree* category, 33.33% rated effectiveness as moderate, and 22.22% rated learning experience in the *Somewhat disagree* and *Strongly disagree* category.

In the Fall of 2022, more than half (68.75)% of the students rated the learning experience in the *Strongly agree* and *Somewhat agree* categories, 12.50% rated learning experience as moderate, and 18.75% rated learning experience in the *Somewhat disagree* and *Strongly disagree* category. For the effectiveness of the learning experience, 43.75% of the students rated the effectiveness in the *Strongly agree* and *Somewhat agree* category, 37.50% rated effectiveness as moderate, and 18.75% rated learning experience in the *Somewhat disagree* and *Strongly disagree* category.

For the overall learning experience, the mean response by students in the Spring and Fall semesters was 2.89 and 2.44, respectively. The mean value of the Fall semester is closer to *Somewhat agree*, indicating that the students had a more positive learning experience in the Fall semester. For the effectiveness of the learning experience, the mean response from students in the Spring and Fall semesters was 2.78 and 2.13, respectively. Similar to the learning experience, the mean value of the Fall semester is closer to *Somewhat agree*, indicating that in the Fall semester, the effectiveness of the learning experience was more positive for the students.

6.6.7 Lab Difficulty

We gauged the difficulty of the lab for the students, and the response indicated that the students neither agreed nor disagreed that the labs were demanding. In the Spring

of 2022, 44.44% of the students rated the level of effort required to learn the topic in the *Strongly agree* and *Somewhat agree* categories, 22.22% rated engagement as moderate, and 33.33% rated learning effort in the *Somewhat disagree* and *Strongly disagree* category. In the Fall of 2022, more than half (62.5%) of the students rated how engaging the lab was in the *Strongly agree* and *Somewhat agree* categories, 25% rated learning effort as moderate, and 12.50% rated learning effort in the *Somewhat disagree* and *Strongly disagree* category.

The mean response by students in the Spring and Fall semesters was 2.78 and 2.56, respectively. The mean values are between the *Somewhat agree* and *Neither agree nor disagree* with the mean value of Fall closer to *Somewhat agree* than the Spring mean, indicating that the level of effort required to learn the topic by the students was better in the Fall than in Spring even though the students neither agree nor disagree.

6.6.8 Understanding of Concepts

To understand whether the students better understood the concepts learned, we determined students' grasp of the concepts introduced in the lab. In the Spring of 2022, 22.22% of the students rated understanding of concepts in the *Strongly agree* and *Somewhat agree* categories, 33.33% rated understanding of concepts as moderate, and 44.44% rated understanding of concepts in the *Somewhat disagree* and *Strongly disagree* category. In the Fall of 2022, 43.75% of the students rated understanding of concepts in the *Strongly agree* and *Somewhat agree* categories, 25% rated understanding of concepts as moderate, and 31.25% rated learning effort in the *Somewhat disagree* and *Strongly disagree* category.

The mean response by students in the Spring and Fall semesters was 3.44 and 2.31, respectively. The mean value for the Spring semester is between *Neither agree nor disagree* and *Somewhat disagree*. For the Fall semester, the mean value is between *Somewhat agree* and *Neither agree nor disagree*. These values indicate that in the Spring semester, students understanding of concepts learned was closer to negative (Somewhat disagree). However, in the Fall semester, students better understood the concepts learned.

6.6.9 Qualitative Feedback

Our last three survey questions were open-ended questions about what has been the most helpful for learning, what has caused the most difficulty using the lab, and how the lab can be improved. We used these as the qualitative data source, providing insights into students' perceptions and preferences. In general, the students understood the purpose of the lab and cyberharassment, found the terminology confusing, and wanted the lab to be more fun and engaging. In the Spring 2022 semester, in response to "What has been most helpful for your learning in using the lab so far". Notable student responses were: *"I understand the purpose of cyberbullying and its purpose and how it is designed to be successful"* and *"The guest speakers coming in to help"*. For the Fall 2022 semester, the notable student responses were: *"The most helpful part for my learning has been the hands-on activity, being able to ask questions while going through the work and having a guest speaker gave some new insight."*, *"I learned a lot about cyber bullying that I did not know about and the different forms it can come in."*, *"The lab was helpful with detecting cyber harassment."*, *"It was nice knowing the set up and watching the steps be performed"*, and *"The videos and zoom call"*.

For the question "In terms of your learning, what has caused you the most difficulty in using the lab so far". The notable student responses in the Spring 2022 semester were: *"I had a hard time understanding how to actually complete on my own"*, *"The terminology"*, *"I could not stay focus and lacked engagement"*, and *"Being online."* Notable responses in the Fall semester were: *"It was difficult that when there was a troubleshooting error that I could not walk through it with someone like I could in person."*, *"The most difficult part is not usage of the lab itself; it is remembering certain aspects of what to do and what to look for when in the lab."*, *"The most difficulty experienced in the labs is facing errors and technical difficulties."*, *"I think the directions were hard to follow because the instructor seemed like he assumed we knew something about this content."*, and *"Fully understanding what was being done in the lab was the most difficult."*

Finally, for the question "What suggestion(s) can you make that would enhance your

learning experience with the lab". Notable student responses in the Spring 2022 semester were: *"Make the lab more engaging/fun"*, *"Break down steps on how to actually complete the activity"*, *"I would say, trying using a different online platform for this lab, to make everything a little bit easier for students to understand."*, *"Being in person"*, and *"Better terminology"*. In the Fall 2022 semester, notable responses were: *"The instructor was helpful it's just hard to learn over the computer such a difficult thing to do."*, *"I cannot think of anything at the moment. I really enjoyed learning about this lab and how it worked."*, *"An in person option"*, *"provide a tutorial video."*, *"I would say slowing down the directions."*, and *"Teach more about how to face technical difficulties."*

6.6.10 Lesson Learned

The authors learned the following lessons in implementing a cloud-based laboratory experience. We outline tips for developing a cloud-based laboratory for teaching AI socially relevant cybersecurity.

6.6.10.1 Code Dilution

The lab is implemented on Google Colab, Google's cloud-based jupyter notebook platform. The initial implementation of the lab on Google Colab presented the students with the raw code implementation of cyberharassment detection using PyTorch, an open-source Python framework for developing ML, especially deep learning systems. From the Spring 2022 feedback from the students, we observed that the students were not positive towards the code implementation since they have very little programming experience. This frustrated students and slowed interest and learning. With this knowledge, in preparation for the Fall 2022 semester, we improved the learning experience by re-implementing the lab with the code hidden to enable the students to think about social issues and focus on understanding how AI can be used to approach social issues such as cyberharassment.

6.6.10.2 Lab Instructions

Developing the lab instructions in the lab manual is crucial to enhancing the student experience. If the lab instructions are not very detailed, with step-by-step instructions on how to complete the lab activities, the students struggle with understanding and completing the lab independently. From the Spring 2022 student feedback, students complained that the instruction manual needed more detail and felt the instructors assumed they were familiar with AI and AI terminology. In the Fall 2022 semester, we improved the lab manual by toning down the terminology and providing more step-by-step descriptions so the students could complete the labs independently and understand the purpose and rationale behind each step.

6.6.10.3 Pre-lab Lecture

Before allowing the students to complete the lab independently, we prepared lecture slides about the lab and introduced them to the problem and the activities they would be completing. The students found this pre-lab lecture particularly helpful.

Other best practices include recording the pre-lab lecture so that the students can refer back to it when working on the lab activities, anticipating possible technical difficulties, and including steps to solve the issues in the lab manual. Having an in-person option where the students can complete the labs during class could help with engagement and technical issues they might encounter.

6.7 Limitations

Our work has limitations. First, our study is focused on one public institution in the United States, limiting our findings' generalizability. Second, the generalizability of our findings is also limited by the focus of our study on one non-computing program - Social Statistics, excluding other non-computing programs. Third, the number of participants in our study further limits our conclusions. Finally, the current lab only focuses on the general

concepts of AI and how it can be utilized to address social issues such as cyberharassment. It does not cover other areas of AI, such as how adversarial attacks can fool cyberharassment models, multi-modal cyberharassment detection, and the issue of fairness and bias in cyberharassment models.

6.8 Conclusion and Future Work

We have developed an AI socially relevant cybersecurity lab for cyberharassment detection for non-computing students. We introduced a cyberharassment detection lab's development, implementation, and assessment. The development process has been guided by the learning objective of introducing a hands-on experiential lab that will increase general awareness of socially relevant cybersecurity and AI and is suitable for teaching AI socially relevant cybersecurity to non-computing students. Our lab offers meaningful experiential learning opportunities that allow students to work on real-world social issues such as cyberharassment. After incorporating student feedback in the redesign of the lab used in the Fall semester, the knowledge or skills of most students in automated cyberharassment detection and how ML works improved significantly compared to the Spring semester. Also, students found the detection of cyberharassment helpful and understood the purpose of using AI for social issues. We plan on continuing to refine the lab and use the knowledge gained in developing four labs currently under development that cover multi-modal (text and image) cyberharassment detection, adversarial attacks on cyberharassment systems, bias mitigation in cyberharassment systems, and the use of generative AI models such as ChatGPT for cyberharassment detection. Additionally, we plan on developing these labs for computer science and engineering students in the future.

6.9 Acknowledgment

The work was supported by National Science Foundation (NSF) under the Grant No. 2239605, 2228616 and 2114920.

Chapter 7

Conclusion and Future Work

7.1 Conclusions

Online social platforms have become part of our daily routines as they make it easy to make new connections, assess current news and trends, disseminate information reaching a broader audience range, and keep up with politics and our network. The benefits are obvious, but the dangers are also apparent, as the content disseminated on these platforms can be offensive. The total elimination of offensive content is difficult because platforms must balance content moderation and freedom of expression. The moderation of offensive content in itself is a difficult task because of its subjectivity. Despite this, prominent platforms have invested in technical solutions, such as employing machine learning, precisely deep learning solutions, to moderate offensive content. Researchers have introduced various machine learning and deep learning solutions to tackle the problem of offensive language on online social media platforms. In this dissertation, we contribute to this line of research that promotes healthy online conversations by analyzing offensive language on social media, specifically during global events such as Black Lives Matter. We further investigated the bias, specifically racial bias, in offensive language detection models and explored the use of experiential learning hands-on labs to engage non-computing students in AI-centered social cybersecurity education.

In Chapter 3, we conducted a large-scale analysis of offensive content and emotional dynamics in Black Lives Matter-related tweets following the protests in 2020 and the years after. We collected more than 20 million tweets, trained an offensive language detection model that uses sentiment features to improve detection, trained an emotion classification model that uses deep attention of the sentiment features from a sentiment model to classify the emotions expressed in tweets effectively, used topic modeling to evaluate primary topics discussed offensively, and employed network analysis on the network of authors who replied offensively to other users to understand the nature of offensive discourse. Our analysis showed that negative emotions such as anger and disgust were the most expressed emotions throughout the study period and could have led to the outburst of offensive tweets. Topics such as police brutality and systemic injustice were the center of discussions. The network analysis reveals that most offender-recipient conversations are unidirectional from the offender to the recipient, likely because of a supportive tweet by the recipient and the offender disagreeing with the recipient's stance. Our analysis helps promote healthy online conversation, helps policymakers create policies that address the issues being discussed and could help reduce the effect of information bias toward the black community.

In Chapter 4, we evaluate the level of bias in offensive language detection models. Our study focuses on transformer-based models such as BERT, BERTweet, and HateBERT, commonly used in the literature for offensive language detection. We assessed the bias towards African-American English (AAE) in these models fine-tuned on eight publicly available offensive language detection datasets. The analysis revealed that the evaluated models classify tweets written in AAE into negative (hate, abuse, racism, etc.) classes more than tweets written in Standard American English (SAE). To mitigate bias in such models, we introduced AAEBERT, a language model that is AAE-aligned, and used the representation of tweets from AAEBERT and the models being trained. We achieved a reduction in bias towards AAE through adversarial learning.

In Chapter 5, we extend the work done in Chapter 4 by exploring the implications of using LLMs to annotate offensive language detection datasets used in fine-tuning down-

stream models for offensive language detection. We used three prompting techniques to annotate seven offensive language datasets and fine-tuned three models on these datasets.

Finally, in Chapter 6, we tackle the issue of bridging the gap between artificial intelligence (AI) education in computing/engineering and non-computing disciplines in the era of socially relevant cybersecurity. In this work, we developed an experiential learning hands-on lab and learning materials to engage students in non-computing AI cybersecurity.

7.2 Future Recommendations

This dissertation recommends the inclusion of more diverse languages in the training of LLMs. For data annotation used in the training of downstream models, diverse annotators, such as annotators familiar with the African-American Vernacular, should be recruited for the annotation task. While limited resources can hinder the application of these recommendations, it opens up an avenue for more research to create new ways to de-bias models algorithmically. While using an adversarial network works in debiasing deep learning offensive content detection models, the models suffer from a slight reduction in performance. The robustness of debiased offensive content detection models has yet to be explored in the literature. We recommend a robustness assessment of models debiased using various debiasing techniques against adversarial examples.

Bibliography

[1] AI advances to better detect hate speech. https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/.

[2] Center for Informed Democracy & Social-Cybersecurity. https://www.cmu.edu/ideas-social-cybersecurity/.

[3] Google's Hate Speech Detection A.I. Has a Racial Bias Problem. https://fortune.com/2019/08/16/google-jigsaw-perspective-racial-bias/.

[4] Hateful Memes Challenge and Data Set. https://ai.facebook.com/tools/hatefulmemes/.

[5] Qualitrics. https://www.qualtrics.com/.

[6] Social-Cybersecurity. http://www.casos.cs.cmu.edu/projects/projects/social_cyber_security.php.

[7] Shivang Agarwal and C Ravindranath Chowdary. Combating hate speech using an adaptive ensemble learning model with a case study on covid-19. *Expert Systems with Applications*, 185:115632, 2021.

[8] Sweta Agrawal and Amit Awekar. Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer, 2018.

[9] Genevera I Allen. Experiential learning in data science: Developing an interdisciplinary, client-sponsored capstone program. pages 516–522, 2021.

[10] Raghad Alshalan, Hend Al-Khalifa, Duaa Alsaeed, Heyam Al-Baity, and Shahad Alshalan. Detection of hate speech in covid-19–related tweets in the arab region: Deep learning and topic modeling approach. *Journal of Medical Internet Research*, 22(12):e22609, 2020.

[11] Jisun An, Haewoon Kwak, Claire Seungeun Lee, Bogang Jun, and Yong-Yeol Ahn. Predicting anti-asian hateful users on twitter during covid-19. *arXiv preprint arXiv:2109.07296*, 2021.

[12] Monica Anderson, Michael Barthel, Andrew Perrin, and Emily A Vogels. # blacklivesmatter surges on twitter after george floyd's death. 2020.

[13] Paul Anderson, James Bowring, Renée McCauley, George Pothering, and Christopher Starr. An undergraduate degree in data science: curriculum and a decade of implementation experience. pages 145–150, 2014.

[14] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54, 2019.

[15] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, and Idowu Ademola Osinuga. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, 38:100311, 2020.

[16] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760, 2017.

[17] Alison Bailey. On anger, silence, and epistemic injustice. *Royal Institute of Philosophy Supplements*, 84:93–115, 2018.

[18] Arko Barman, Su Chen, Andersen Chang, and Genevera Allen. Experiential learning in data science through a novel client-facing consulting course. pages 1–9, 2022.

[19] Sérgio Barreto, Ricardo Moura, Jonnathan Carvalho, Aline Paes, and Alexandre Plastino. Sentiment analysis in tweets: an assessment study from classical to modern word representation models. *Data Mining and Knowledge Discovery*, 37(1):318–380, 2023.

[20] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63, 2019.

[21] Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*, 2018.

[22] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[23] Hilary B Bergsieker, Lisa M Leslie, Vanessa S Constantine, and Susan T Fiske. Stereotyping by omission: eliminate the negative, accentuate the positive. *Journal of personality and social psychology*, 102(6):1214, 2012.

[24] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[25] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

[26] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*, 2016.

[27] Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*, 2017.

[28] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

[29] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[30] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017.

[31] Simon Buckingham Shum and Ruth Deakin Crick. Learning analytics for 21st century competencies. *Journal of Learning Analytics*, 3(2):6–21, 2016.

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

[33] Francisco J Cantú-Ortiz, Nathalíe Galeano Sánchez, Leonardo Garrido, Hugo Terashima-Marin, and Ramón F Brena. An artificial intelligence educational strategy for the digital transformation. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 14:1195–1209, 2020.

[34] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*, 2020.

[35] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, 2021.

[36] Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th language resources and evaluation conference*, pages 6193–6202, 2020.

[37] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22, 2009.

[38] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Measuring# gamergate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th international conference on world wide web companion*, pages 1285–1290, 2017.

[39] cjadams, Sorensen Jeffrey, Elliott Julia, Dixon Lucas, McDonald Mark, nithum, and Cukierski Will. Toxic comment classification challenge. `https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge`, 2017.

[40] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska De Jong. Improving cyberbullying detection with user context. In *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35*, pages 693–696. Springer, 2013.

[41] Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1014–1023, 2023.

[42] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*, 2019.

[43] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

[44] Munmun De Choudhury, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber. Social media participation in an activist movement for racial equality. In *Proceedings of the international aaai conference on web and social media*, volume 10, pages 92–101, 2016.

[45] Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[48] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 11–17, 2011.

[49] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. Is gpt-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, 2023.

[50] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30, 2015.

[51] Kristie Dotson. A cautionary tale: On limiting epistemic oppression. *Frontiers: A Journal of Women Studies*, 33(1):24–47, 2012.

[52] Kiela Douwe, Firooz Hamed, and Mohan Aravind. Hateful memes challenge and dataset for research on harmful multimodal content. `https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/`, 2020.

[53] Bruce Drake. The darkest side of online harassment: Menacing behavior. 2015.

[54] Daniel Ducharme. *Machine Learning for the Automated Identification of Cyberbullying and Cyberharassment*. PhD thesis, University of Rhode Island, 2017.

[55] Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986.

[56] Johannes C Eichstaedt, Garrick T Sherman, Salvatore Giorgi, Steven O Roberts, Megan E Reynolds, Lyle H Ungar, and Sharath Chandra Guntuku. The emotional and mental health impact of the murder of george floyd on the us population. *Proceedings of the National Academy of Sciences*, 118(39):e2109139118, 2021.

[57] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[58] Sunet Eybers and Mariè Hattingh. Teaching data science to post graduate students: A preliminary study using a" flip" class room approach. *International Association for Development of the Information Society*, 2016.

[59] Md Fahim and Swapna S Gokhale. Detecting offensive content on twitter during proud boys riots. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1582–1587. IEEE, 2021.

[60] Anjalie Field, Chan Young Park, Antonio Theophilo, Jamelle Watson-Daniels, and Yulia Tsvetkov. An analysis of emotions and the prominence of positivity in# blacklivesmatter tweets. *Proceedings of the National Academy of Sciences*, 119(35):e2205767119, 2022.

[61] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[62] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[63] Deen Freelon, Charlton D McIlwain, and Meredith Clark. Beyond the hashtags:# ferguson,# blacklivesmatter, and the online struggle for offline justice. *Center for Media & Social Impact, American University, Forthcoming*, 2016.

[64] Ryan J Gallagher, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. Divergent discourse between protests and counter-protests:# blacklivesmatter and# alllivesmatter. *PloS one*, 13(4):e0195644, 2018.

[65] Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H Andrew Schwartz. Empirical evaluation of pre-trained transformers for human-level nlp: The role of sample size and dimensionality. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4515. NIH Public Access, 2021.

[66] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

[67] Negin Ghavami and Letitia Anne Peplau. An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly*, 37(1):113–127, 2013.

[68] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

[69] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE, 2020.

[70] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[71] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233, 2017.

[72] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478, 2020.

[73] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[74] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2015.

[75] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[76] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12, 2018.

[77] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[78] Cristela Guerra. Where did 'me too' come from? activist tarana burke, long before hashtags. `https://www.bostonglobe.com/lifestyle/2017/10/17/alyssa-milano-credits-activist-tarana-burke-with-founding-metoo-movement-years-ago/o2Jv29v6ljObkKPTPB9KGP/story.html`, 2017.

[79] Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573. IEEE, 2023.

[80] Xudong Han, Timothy Baldwin, and Trevor Cohn. Diverse adversaries for mitigating bias in training. *arXiv preprint arXiv:2101.10001*, 2021.

[81] Hashim A Hashim, Catherine Tatarniuk, and Brad Harasymchuk. First year engineering design: Course design, projects, challenges, and outcomes. pages 1–9, 2022.

[82] Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 61–61. IEEE Computer Society, 2023.

[83] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. Searching for safety online: Managing" trolling" in a feminist forum. *The information society*, 18(5):371–384, 2002.

[84] Sameer Hinduja and Justin W Patchin. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3):206–221, 2010.

[85] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Dialect prejudice predicts ai decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742*, 2024.

[86] J Brian Houston, Glenn J Hansen, and Gwendelyn S Nisbett. Influence of user comments on perceptions of media bias and third-person effect in online news. *Electronic News*, 5(2):79–92, 2011.

[87] Mark Hsueh, Kumar Yogeeswaran, and Sanna Malinen. "leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human communication research*, 41(4):557–576, 2015.

[88] Fan Huang, Haewoon Kwak, and Jisun An. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297, 2023.

[89] Jelani Ince, Fabio Rojas, and Clayton A Davis. The social media response to black lives matter: How twitter users interact with black lives matter through hashtag use. *Ethnic and racial studies*, 40(11):1814–1830, 2017.

[90] Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. Fermi at semeval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 70–74, 2019.

[91] Internet Live Stats. Twitter Usage Statistics. `https://www.internetlivestats.com/twitter-statistics/`.

[92] Johannes Jakubik, Michael Vössing, Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. Online emotions during the storming of the us capitol: evidence from the social media network parler. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 423–434, 2023.

[93] Keenan Jones, Jason RC Nurse, and Shujun Li. Out of the shadows: Analyzing anonymous' twitter resurgence during the 2020 black lives matter protests. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 417–428, 2022.

[94] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. A just and comprehensive strategy for using nlp to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, 2019.

[95] Martin Kandlhofer, Gerald Steinbauer, Sabine Hirschmugl-Gaisch, and Petra Huber. Artificial intelligence and computer science in education: From kindergarten to university. In *2016 IEEE frontiers in education conference (FIE)*, pages 1–9. IEEE, 2016.

[96] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[97] Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. Mind your language: Abuse and offense detection for code-switched languages. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9951–9952, 2019.

[98] George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. Technology solutions to combat online harassment. In *Proceedings of the first workshop on abusive language online*, pages 73–77, 2017.

[99] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

[100] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[101] Bumsoo Kim, Eric Cooks, and Seong-Kyu Kim. Exploring incivility and moral foundations toward asians in english-speaking tweets in hate crime-reporting cities during the covid-19 pandemic. *Internet Research*, 2021.

[102] Jihie Kim and Jaebong Yoo. Role of sentiment in message propagation: Reply vs. retweet behavior in political communication. In *2012 international conference on social informatics*, pages 131–136. IEEE, 2012.

[103] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[104] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861, 2023.

[105] Alice Y Kolb and David A Kolb. Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of management learning & education*, 4(2):193–212, 2005.

[106] David A Kolb. *Experiential learning: Experience as the source of learning and development*. FT press, 2014.

[107] Deepak Kumar, Jeff Hancock, Kurt Thomas, and Zakir Durumeric. Understanding the behaviors of toxic accounts on reddit. In *Proceedings of the ACM Web Conference 2023*, pages 2797–2807, 2023.

[108] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318, 2021.

[109] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 1–11, 2018.

[110] Sumit Kumar and Raj Ratn Pranesh. Tweetblm: A hate speech dataset and analysis of black lives matter-related microblogs on twitter. *arXiv preprint arXiv:2108.12521*, 2021.

[111] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart and Gianluca Stringhini. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267. IEEE, 2021.

[112] Joseph Kwarteng, Serena Coppolino Perfumi, Tracie Farrell, and Miriam Fernandez. Misogynoir: public online response towards self-reported misogynoir. In *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 228–235, 2021.

[113] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*, 2013.

[114] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[115] Anti-Defamation League. Online hate and harassment. the american experience 2021. *Center for Technology and Society: New York, NY, USA*, pages 10–23, 2021.

[116] Claire Seungeun Lee and Ahnlee Jang. Questing for justice on twitter: Topic modeling of# stopasianhate discourses in the wake of atlanta shooting. *Crime & Delinquency*, 69(13-14):2874–2900, 2023.

[117] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[118] Younghun Lee, Seunghyun Yoon, and Kyomin Jung. Comparative studies of detecting abusive language on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 101–106, 2018.

[119] Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, 2021.

[120] Rebecca Leung and Robert Williams. # metoo and intersectionality: An examination of the# metoo movement through the r. kelly scandal. *Journal of Communication Inquiry*, 43(4):349–371, 2019.

[121] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. 2019.

[122] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36, 2024.

[123] Mingqi Li, Song Liao, Ebuka Okpala, Max Tong, Matthew Costello, Long Cheng, Hongxin Hu, and Feng Luo. Covid-hatebert: a pre-trained language model for covid-19 related hate speech detection. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 233–238. IEEE, 2021.

[124] Song Liao, Ebuka Okpala, Long Cheng, Mingqi Li, Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Matthew Costello. Analysis of covid-19 offensive tweets and their targets. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4473–4484, 2023.

[125] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018.

[126] Ping Liu, Wen Li, and Liang Zou. Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *SemEval@ NAACL-HLT*, pages 87–91, 2019.

[127] Rosemary Luckin, Mutlu Cukurova, Carmel Kent, and Benedict du Boulay. Empowering educators to be ai-ready. *Computers and Education: Artificial Intelligence*, 3:100076, 2022.

[128] Bernhard Lutz, Marc TP Adam, Stefan Feuerriegel, Nicolas Pröllochs, and Dirk Neumann. Affective information processing of fake news: Evidence from neurois. *European Journal of Information Systems*, pages 1–20, 2023.

[129] Krishanu Maity, AS Poornash, Shaubhik Bhattacharya, Salisa Phosit, Sawarod Kongsamlit, Sriparna Saha, and Kitsuchart Pasupa. Hatethaisent: Sentiment-aided hate speech detection in thai language. *IEEE Transactions on Computational Social Systems*, 2024.

[130] Mariella Moon. ChatGPT reportedly reached 100 million users in January. `https://www.engadget.com/chatgpt-100-million-users-january-130619073.html`, 2023.

[131] Lina Markauskaite, Rebecca Marrone, Oleksandra Poquet, Simon Knight, Roberto Martinez-Maldonado, Sarah Howard, Jo Tondeur, Maarten De Laat, Simon Buckingham Shum, Dragan Gašević, et al. Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with ai? *Computers and Education: Artificial Intelligence*, 3:100056, 2022.

[132] Diana Adela Martin and Gunter Bombaerts. Enacting socio-technical responsibility through challenge based learning in an ethics and data analytics course. pages 1–7, 2022.

[133] Adrienne Massanari. # gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New media & society*, 19(3):329–346, 2017.

[134] Douglas S Massey and Garvey Lundy. Use of black english and racial discrimination in urban housing markets: New methods and findings. *Urban affairs review*, 36(4):452–469, 2001.

[135] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875, 2021.

[136] Hollister Matissa and World Economic Forum. Artificial intelligence. curation: Desautels faculty of management, mcgill university. `https://intelligence.weforum.org/topics/a1Gb0000000pTDREA2`, 2024.

[137] Richard Matovu, Joshua C Nwokeji, Terry Holmes, and Tajmilur Rahman. Teaching and learning cybersecurity awareness with gamification in smaller universities and colleges. pages 1–9, 2022.

[138] Long Cheng Matthew Costello, Hongxin Hu Feng Luo, and Nishant Vishwamitra Song Liao. Covid-19: A pandemic of anti-asian cyberhate. *Journal of Hate Studies*, 17(1):108–118, 2021.

[139] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*, 2023.

[140] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.

[141] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 85–94, 2017.

[142] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861, 2020.

[143] Aaron Mueller, Zach Wood-Doughty, Silvio Amir, Mark Dredze, and Alicia Lynn Nobles. Demographic representation and collective storytelling in the me too twitter hashtag activism movement. *Proceedings of the ACM on human-computer interaction*, 5(CSCW1):1–28, 2021.

[144] Marcia Mundt, Karen Ross, and Charla M Burnett. Scaling social movements through social media: The case of black lives matter. *Social Media+ Society*, 4(4):2056305118807911, 2018.

[145] Diana C Mutz and Byron Reeves. The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review*, 99(1):1–15, 2005.

[146] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. *CoRR*, abs/2006.12621, 2020.

[147] NASEM. Data science for undergraduates: Opportunities and options (p. 138). `https://doi.org/10.17226/25104`, 2018.

[148] Christof Naumzik and Stefan Feuerriegel. Detecting false rumors from retweet dynamics on social media. In *Proceedings of the ACM web conference 2022*, pages 2798–2809, 2022.

[149] Huy Nghiem and Fred Morstatter. " stop asian hate!": Refining detection of anti-asian hate speech during the covid-19 pandemic. *arXiv preprint arXiv:2112.02265*, 2021.

[150] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*, 2020.

[151] Thu T Nguyen, Shaniece Criss, Eli K Michaels, Rebekah I Cross, Jackson S Michaels, Pallavi Dwivedi, Dina Huang, Erica Hsu, Krishay Mukhija, Leah H Nguyen, et al. Progress and push-back: How the killings of ahmaud arbery, breonna taylor, and george floyd impacted public discourse on race and racism on twitter. *SSM-population health*, 15:100922, 2021.

[152] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

[153] NSCAI. Final report. `https://www.nscai.gov/2021-final-report/`, 2021.

[154] Joshua C Nwokeji and Psem Stephen T Frezza. Cross-course project-based learning in requirements engineering: An eight-year retrospective. pages 1–9, 2017.

[155] Joshua Chibuike Nwokeji, Richard Stachel, and Terry Holmes. Effect of instructional methods on student performance in flipped classroom. pages 1–9, 2019.

[156] Adewale Obadimu, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal. Identifying toxicity within youtube video comment. In *Social, Cultural, and Behavioral Modeling: 12th International Conference, SBP-BRiMS 2019, Washington, DC, USA, July 9–12, 2019, Proceedings 12*, pages 214–223. Springer, 2019.

[157] Abby Ohlheiser. The woman behind 'me too' knew the power of the phrase when she created it — 10 years ago. https://www.washingtonpost.com/news/the-intersect/wp/2017/10/19/the-woman-behind-me-too-knew-the-power-of-the-phrase-when-she-created-it-10-years-ago 2017.

[158] Ebuka Okpala, Long Cheng, Nicodemus Mbwambo, and Feng Luo. Aaebert: Debiasing bert-based hate speech detection models via adversarial learning. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1606–1612. IEEE, 2022.

[159] Angela Onwuachi-Willig. What about# ustoo?: The invisibility of race in the# metoo movement. *Yale LJF*, 128:105, 2018.

[160] OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/, 2024.

[161] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5), 2023.

[162] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[163] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, 2018.

[164] Justin W. Patchin. Cyberbullying Statistics. https://cyberbullying.org/2019-cyberbullying-data, 2019.

[165] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993*, 2017.

[166] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*, 2020.

[167] Hao Peng, Ceren Budak, and Daniel M Romero. Event-driven analysis of crowd dynamics in the black lives matter online social movement. In *The World Wide Web Conference*, pages 3137–3143, 2019.

[168] Jacqueline Peng, Jun Shen Fung, Muhammad Murtaza, Afnan Rahman, Pallav Walia, David Obande, and Anish R Verma. A sentiment analysis of the black lives matter movement using twitter. *STEM Fellowship Journal*, (0):1–11, 2022.

[169] Marco Polignano, Pierpaolo Basile, Marco De Gemmis, and Giovanni Semeraro. Hate speech detection through alberto italian language understanding model. In *NL4AI@ AI* IA*, pages 1–13, 2019.

[170] Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, and Valerio Basile. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR, 2019.

[171] Daniel Preoţiuc-Pietro and Lyle Ungar. User-level race and ethnicity predictors from twitter text. In *Proceedings of the 27th international conference on computational linguistics*, pages 1534–1545, 2018.

[172] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[173] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[174] Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. Joint modelling of emotion and abusive language detection. *arXiv preprint arXiv:2005.14028*, 2020.

[175] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):1–10, 2018.

[176] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[177] Kandace Redd. Unity against hate rally starts in sacramento. `https://www.abc10.com/article/news/community/race-and-culture/national-unity-against-hate-rally-confronts-anti-asian-hate/103-baf3f170-539a-4dc7-bcdc-03d1496b16c0`, 2021.

[178] John R Rickford and Sharese King. Language and linguistics on trial: Hearing rachel jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, pages 948–988, 2016.

[179] Claire E Robertson, Nicolas Pröllochs, Kaoru Schwarzenegger, Philip Pärnamets, Jay J Van Bavel, and Stefan Feuerriegel. Negativity drives online news consumption. *Nature Human Behaviour*, 7(5):812–822, 2023.

[180] Stephanie Rosenthal and Tingting Chung. A data science major: Building skills and confidence. pages 178–184, 2020.

[181] Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv preprint arXiv:1910.02334*, 2019.

[182] Andres F Salazar-Gomez, Aikaterini Bagiati, Nicholas Minicucci, Kathleen D Kennedy, Xiaoxue Du, and Cynthia Breazeal. Designing and implementing an ai education program for learners with diverse background at scale. pages 1–8, 2022.

[183] Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10:1–34, 2020.

[184] Lakshminarayanan Samavedham and Kiruthika Ragupathi. Facilitating 21st century skills in engineering students. *The Journal of Engineering Education*, 26(1):38–49, 2012.

[185] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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

[187] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. pages 1668–1678, July 2019.

[188] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*, 2021.

[189] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

[190] Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. "go eat a bat, chang!": An early look on the emergence of sinophobic behavior on web communities in the face of covid-19, 2020.

[191] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10, 2017.

[192] John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, et al. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*, 2:4, 2022.

[193] Emilio Serrano, Martin Molina, Daniel Manrique, and Luis Baumela. Experiential learning in data science: From the dataset repository to the platform of experiences. pages 122–130, 2017.

[194] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 687–690, 2016.

[195] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*, 2023.

[196] Daniel Sousa, Luís Sarmento, and Eduarda Mendes Rodrigues. Characterization of the twitter@ replies network: are user ties social or topical? In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 63–70, 2010.

[197] Jane Southworth, Kati Migliaccio, Joe Glover, David Reed, Christopher McCarty, Joel Brendemuhl, Aaron Thomas, et al. Developing a model for ai across the curriculum: Transforming the higher education landscape via innovation in ai literacy. *Computers and Education: Artificial Intelligence*, 4:100127, 2023.

[198] Aarohi Srivastava, Denis Kleyjo, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolatingthe capabilities of language models. *Transactions on Machine Learning Research*, (5), 2023.

[199] Alex T St. Louis, Penny Thompson, Tracey N Sulak, Marty L Harvill, and Michael E Moore. Infusing 21st century skill development into the undergraduate curriculum: the formation of the ibears network. *Journal of Microbiology & Biology Education*, 22(2):10–1128, 2021.

[200] Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4):217–248, 2013.

[201] Jiahong Su, Yuchun Zhong, and Davy Tsz Kit Ng. A meta-review of literature on educational approaches for teaching ai at the k-12 levels in the asia-pacific region. *Computers and Education: Artificial Intelligence*, 3:100065, 2022.

[202] Dorothy Espelage; Susan Swearer. Research on school bullying and victimization: What have we learned and where do we go from here? *School Psychology Review*, page 365–383, 2013.

[203] Keeanga-Yamahtta Taylor. *From# BlackLivesMatter to black liberation*. Haymarket Books, 2016.

[204] Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*, 2023.

[205] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, and G. Stringhini. Sok: Hate, harassment, and the changing landscape of online abuse. pages 473–493, 2021.

[206] Neil Thompson. Social movements, social justice and social work. *British journal of social work*, 32(6):711–722, 2002.

[207] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[208] Xin Tong, Yixuan Li, Jiayi Li, Rongqi Bei, and Luyao Zhang. What are people talking about in# backlivesmatter and# stopasianhate? exploring and categorizing twitter topics emerging in online social movements through the latent dirichlet allocation model. *arXiv preprint arXiv:2205.14725*, 2022.

[209] Tom R Tyler and Heather J Smith. Social justice and social movements. 1995.

[210] Joshua Uyheng and Kathleen M Carley. Characterizing network dynamics of online hate communities around the covid-19 pandemic. *Applied Network Science*, 6(1):1–21, 2021.

[211] Ameya Vaidya, Feng Mai, and Yue Ning. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693, 2020.

[212] Betty Van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, 2018.

[213] Dunya Van Troost, Jacquelien Van Stekelenburg, and Bert Klandermans. Emotions of protest. *Emotions in politics: The affect dimension in political tension*, pages 186–203, 2013.

[214] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[215] Yllka Velaj, Dominik Dolezal, Roland Ambros, Claudia Plant, and Renate Motschnig. Designing a data science course for non-computer science students: Practical considerations and findings. pages 1–9, 2022.

[216] Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. Challenges and frontiers in abusive content detection. Association for Computational Linguistics, 2019.

[217] Nishant Vishwamitra, Keyan Guo, Farhan Tajwar Romit, Isabelle Ondracek, Long Cheng, Ziming Zhao, and Hongxin Hu. Moderating new waves of online hate with chain-of-thought reasoning in large language models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 178–178. IEEE Computer Society, 2024.

[218] Nishant Vishwamitra, Ruijia Roger Hu, Feng Luo, Long Cheng, Matthew Costello, and Yin Yang. On analyzing covid-19-related hate speech using bert attention. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 669–676. IEEE, 2020.

[219] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1231–1245, 2017.

[220] Sebastian Wachs, Michelle F. Wright, and Alexander T. Vazsonyi. Understanding the overlap between cyberbullying and cyberhate perpetration: Moderating effects of toxic online disinhibition. *Criminal Behaviour and Mental Health*, 29(3):179–188, 2019.

[221] Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

[222] Joseph B Walther and Jeong-woo Jang. Communication processes in participatory websites. *Journal of Computer-Mediated Communication*, 18(1):2–15, 2012.

[223] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? gpt-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, 2021.

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

[225] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*, 2017.

[226] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[227] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[228] Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*, 2021.

[229] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. The role and limits of principles in ai ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 195–200, 2019.

[230] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. Overview of the germeval 2018 shared task on the identification of offensive language. 2018.

[231] Jamillah Bowman Williams. Maximizing# metoo: Intersectionality & the movement. *BCL Rev.*, 62:1797, 2021.

[232] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[233] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

[234] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. Demoting racial bias in hate speech detection. *arXiv preprint arXiv:2005.12246*, 2020.

[235] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW2):1–23, 2020.

[236] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666, 2012.

[237] Kimmy Yam. Anti-asian hate crimes increased by nearly 150 `https://www.nbcnews.com/news/asian-america/anti-asian-hate-crimes-increased-nearly-150-2020-mostly-n-n1260264`, 2021.

[238] Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the third workshop on abusive language online*, pages 11–18, 2019.

[239] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*, 2019.

[240] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*, 2019.

[241] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*, 2019.

[242] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014, 2018.

[243] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, 2018.

[244] Ping Zhang. The affective response model: A theoretical framework of affective concepts and their relationships in the ict context. *MIS quarterly*, pages 247–274, 2013.

[245] Wei Emma Zhang, Quan Z. Sheng, Ahoud Abdulrahmn F. Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3):1–41, 2020.

[246] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer, 2018.

[247] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. Content-driven detection of cyberbullying on the instagram social network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3952–3958, 2016.

[248] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. Content-driven detection of cyberbullying on the instagram social network. page 3952–3958, 2016.

[249] Assem Zhunis, Gabriel Lima, Hyeonho Song, Jiyoung Han, and Meeyoung Cha. Emotion bubbles: Emotional composition of online discourse before and after the covid-19 outbreak. In *Proceedings of the ACM Web Conference 2022*, pages 2603–2613, 2022.

[250] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis, 2020.