Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

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

Ebuka Okpala, Long Cheng, Kehinde Elelu

College of Engineering, Computing and Applied Sciences
Clemson University
{eokpala, lcheng2, kelelu}@clemson.edu

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 attention 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.

1 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 (Taylor 2016) 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 (Freelon, McIlwain, and Clark 2016). 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 (Anderson et al. 2020). 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 (Nguyen et al. 2021).

While the movement has drawn researchers to study its different aspects (Ince, Rojas, and Davis 2017; Peng, Budak, and Romero 2019; Tong et al. 2022) 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 (Eichstaedt et al. 2021). The findings relate to the findings that offensive language has adverse health effects that can lead to suicide (Hinduja and Patchin 2010).

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 (Van Troost, Van Stekelenburg, and Klandermans 2013; Jasper 2018). 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 (Van Troost, Van Stekelenburg, and Klandermans 2013). People can experience emotions without being directly confronted by the triggering situation (Van Troost, Van Stekelenburg, and Klandermans 2013). 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?

2 Related Work

Emotion is one of the most complex affective concepts and is defined as reactions attributed to stimuli (response to situational events) (Zhang 2013). Emotions play a significant role in online behavior, as found in the dynamics of retweets (Kim and Yoo 2012; Stieglitz and Dang-Xuan 2013) and online consumption (Robertson et al. 2023). Social media users primarily rely on affective rather than cognitive information processing, as suggested by psychology research (Lutz et al. 2023), making emotions essential drivers of online behavior (Brady et al. 2017; Naumzik and Feuerriegel 2022; Robertson et al. 2023; Jakubik et al. 2023). 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 (Kumar and Pranesh 2021) 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 (Fortuna and Nunes 2018), and deep learning techniques (Badjatiya et al. 2017), 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. (Liu, Li, and Zou 2019) obtained first place by fine-tuning the BERT (Devlin et al. 2018) model. Hate speech and offensive language (Nghiem and Morstatter 2021; An et al. 2021; Uyheng and Carley 2021) were intensively studied during the COVID-19 pandemic.

Researchers have also considered sentiment features by including these as features in supervised machine learning and deep learning approaches (Schmidt and Wiegand 2017). For instance, a multitask framework is developed in (Rajamanickam et al. 2020) 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 (Ince, Rojas, and Davis 2017), #AllLivesMatter (Gallagher et al. 2018), the scaling of the movement (Mundt, Ross, and Burnett 2018), the resurgence of Anonymous during BLM protests (Jones, Nurse, and Li 2022), the common and different topics in BLM and Stop Asian Hate movements (Tong et al. 2022), the social media engagement in the movement over time and the relationship between online engagement and offline activities (De Choudhury et al. 2016), and the analysis of sentiment and emotions during the movement (Peng et al. 2022; Field et al. 2022).

3 Methodology

In this section, we detail our methodology for collecting a dataset of 21 million tweets and our classification techniques.

3.1 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 containing the hashtags and keywords listed in the Appendix A.2. 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 (Kumar and Pranesh 2021). 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.

3.2 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 (Schmidt and Wiegand 2017). 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 (Wulczyn, Thain, and Dixon 2017) in YouTube (Obadimu et al. 2019), to understand behaviors of toxic account on Reddit (Kumar et al.

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

[2]https://perspectiveapi.com/

2023), and to filter potentially offensive tweets in COVID-19 dataset (Liao et al. 2023). Following (Liao et al. 2023), 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. 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 3.3 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. See Appendix A.3 for details on why we chose 0.7 as the threshold and how we handle potential bias of Perspective API.

In the offensive language detection literature, researchers have adopted different definitions of offensive language (Fortuna and Nunes 2018). The comprehensive definition in (Caselli et al. 2020b) 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 in (Caselli et al. 2020b) 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) (Zampieri et al. 2019) 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. The detailed annotation process is given in Section A.4 of the Appendix. Our annotated dataset consists of 2,465 tweets after removing duplicates and tweets of less than four words.

3.3 Sentiment Classification

We fine-tuned the pre-trained language model[3] BERTweet (Nguyen, Vu, and Nguyen 2020) (vinai/betweet-base on HuggingFace) on the Twitter Sentiment140 dataset (Go, Bhayani, and Huang 2009). 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 to reduce the ambiguity in the model, i.e., reduce the possibility of the model confusing neutral classes with negative and between negative and neutral classes. The fine-tuned BERTweet model achieved an F1 score of 0.89 and was used to classify the potentially offensive tweets identified by Perspective and all the 21M tweets in our dataset. Our model outperforms the fine-tuned BERT (Devlin et al. 2018) model (0.87) used in (Peng et al. 2022) and is competitive when compared to the adapted BERTweet + SVM model (0.905) used in (Barreto et al. 2023) in terms of F1 scores. The per class performance of the sentiment model is shown in Table 1. Fine-tuning details of the model is provided in Section A.5 of the Appendix.

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

Table 1: Performance of the sentiment model for the negative and positive classes. Evaluation metrics are macro averages.

3.4 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.

We implement our offensive model by fine-tuning BERTweet (Nguyen, Vu, and Nguyen 2020) (vinai/betweet-base on HuggingFace). During fine-tuning, the latent representations of the input text from our fine-tuned sentiment model (frozen) and the offensive model being fine-tuned are fused to obtain a joint representation used in detection. We denote the representation (special classification token CLS) extracted from the penultimate layer of the sentiment model by vector $\mathbf{S}$ (with dimension 768), similarly, denote the representation from the offensive model as $\mathbf{O}$ (with dimension 768). The concatenation ($\mathbf{S} \oplus \mathbf{O}$) of the two vectors 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. 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 (Schmidt and Wiegand 2017). Our model architecture is shown in Fig 1. 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 an AUROC

[3]Code available at https://github.com/burunkus/blm-research

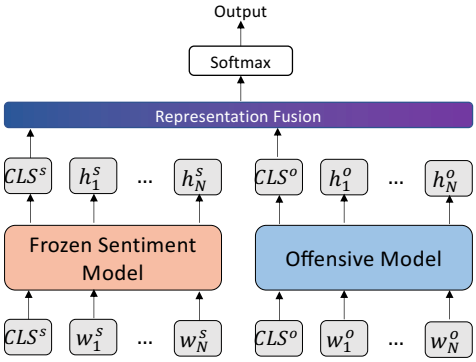


Figure 1: An overview of the proposed offensive language classification model with fused representation of the input text from the sentiment and offensive models.

Target	F1	Precision	Recall	AUROC
Non-offensive	0.787	0.798	0.776	0.864
Offensive	0.841	0.832	0.850	

Table 2: Performance of the offensive detection model for the offensive and non-offensive classes. Evaluation metrics are macro averages.

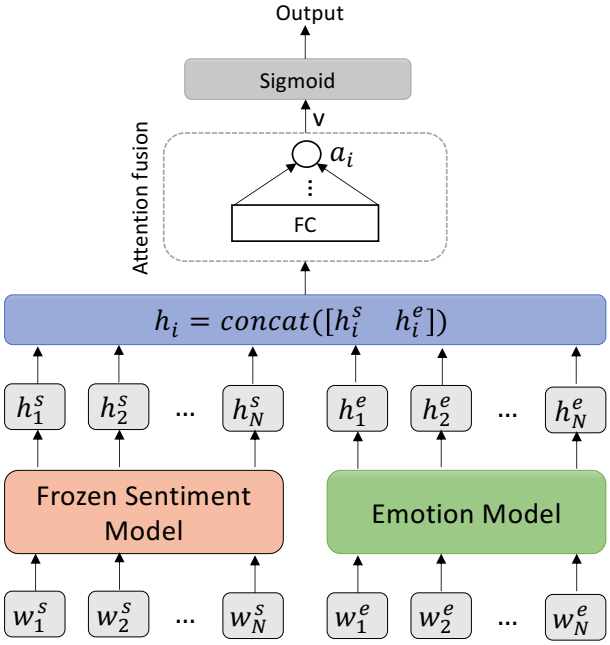


Figure 2: An overview of the proposed Emotion classification model with deep attention fusion. FC indicates a fully connected layer.

score of 0.864, a macro F1, macro precision, and macro recall of 0.814, 0.815, and 0.813, respectively. After training, the fine-tuned model is used to classify each of the 21M tweets in our dataset. Table 2 summarizes the classifier's performance.

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. See Appendix Section A.6 for comparison details with state-of-the-art models and significance test.

3.5 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 (Ekman 1992), and most emotion analysis studies focus on these emotions. In (Ekman 1992), the author further argues that these emotions can differ in antecedent events, behavioral response, physiology, etc. To classify emotions, following (Zhunis et al. 2022), we used the Semeval-2018 Twitter dataset (Mohammad et al. 2018) 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 preprocessing as in our sentiment analysis.

The emotion classification model is similar to the bidirectional LSTM (BiLSTM) architecture with a deep self-attention mechanism introduced by (Baziotis et al. 2018) and used in (Zhunis et al. 2022) to study emotion during COVID-19 pandemic. In our model, we use BERTweet and used our fine-tuned sentiment model from Section 3.3 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 emotion model architecture is shown in Fig 2. The network consists of the fine-tuned sentiment model with weights frozen, the pre-trained language model (BERTweet) being fine-tuned, 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 (Pavlopoulos, Malakasiotis, and Androutsopoulos 2017), (Baziotis et al. 2018) to amplify the influence of each word:

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

$$v = \sum_{i=1}^{N} a_i h_i \quad (2)$$

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 (Zhunis et al. 2022)) and a micro F1 score of 68.7%. Following (Zhunis et al. 2022), we focus our analysis on emotions (anger, disgust, fear, joy, and optimism) with F1-scores above 0.7. Section A.7 of the Appendix contains the per label performance details of our model.

3.6 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 (Grootendorst 2022), 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. See Section A.8 of the Appendix for details about BERTopic, model settings, and topic validation.

3.7 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 (Kumar et al. 2023). 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.

4 Results and Discussion

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

4.1 Offensive and Non-offensive Content (RQ1)

In answer to RQ1, we examined the presence and increase of offensive content in BLM tweets. Table 3 summarizes the statistics of offensive and non-offensive tweets in our dataset. From Table 3, 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: 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). We used the Pruned Exact Linear Time (PELT) (Killick, Fearnhead, and Eckley 2012) 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 and the possible events and dates are shown in Table 4. From Fig. 3A, we find a notable uptick on May 31, 2020 (E1). The noteworthy increase in 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. 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.

4.2 Offensive Reply Network Analysis (RQ1)

The offensive reply network statistics is shown in Table 5. In the 2020 offensive reply graph, 84.5% of the 873,043 offensive tweets posted by offenders to recipients occurred only

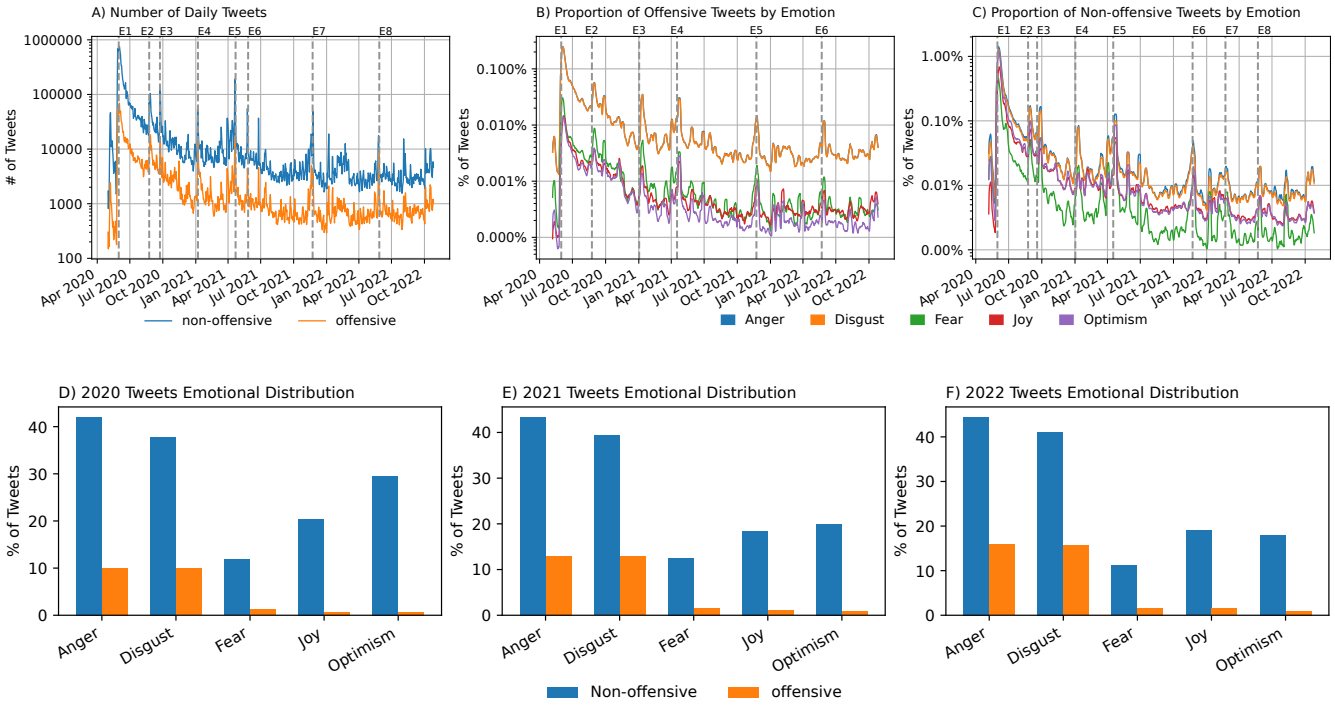


Figure 3: (A) Daily tweet count in log scale. (B) Temporal evolution by emotions for offensive tweets in log-scale. (C) Temporal evolution by emotions for non-offensive tweets in log-scale (D) Emotion distribution of 2020 tweets. (E) Emotion distribution of 2021 tweets. (F) Emotion distribution of 2022 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.

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. Similar observation is made in the reply network of 2021 and 2022. 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.

4.3 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. 3B for the offensive tweets and in Fig. 3C for the non-offensive tweets. The results of the distributions of the five emotions for 2020, 2021, and 2022 are shown in Fig. 3D, Fig. 3E, and Fig. 3F, 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 3B and 3C 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. 3A. For the non-offensive tweets (Fig. 3C), 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

Index	Change Point	Real-world Event
E1	May 31, 2020	Protests over George Floyd's death
E2	August 24, 2020	Aftermath of Jacob Blake's shooting
E3	September 23, 2020	Protests over Breonna Taylor's death and officer indictment
E4	January 7, 2021	Day after US Capitol attack by pro-Trump protesters
E5	April 22, 2021	Aftermath of the shooting of Daunte Wright and killing of Ma'Khia Bryant
E6	May 27, 2021	One year after George Floyd's death, the trial of Derek Chauvin, and the Palestine support by BLM movement
E7	November 23, 2020	Aftermath of the protests in large cities in the US, especially in Kenosha and Kyle Rittenhouse's acquittal
E8	May 28, 2022	Two years after George Floyd's death and protests

Table 4: Change points of offensive and non-offensive tweets and the possible corresponding real-world event.

E7 (February 25, 2022) coincides with the week the three officers involved in the death of George Floyd were found guilty. From Figures[4] 3D, 3E, and 3F, 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, (Field et al. 2022) 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 (Field et al. 2022) 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 (Field et al. 2022) 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.

[4]Emotions can sum up to more than 100% because each tweet can have multiple emotions

Year	# Nodes (recipients, offenders)	# 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 5: Statistics of the offensive reply network. Percentages does not sum to 100% as some recipients are also offenders.

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 3B, 6A, and 7A in the Appendix. The offender's non-offensive tweets, shown in Fig. 6B in the Appendix, 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. 7B in the Appendix, 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.

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**.

4.4 Analyzing Topics of Discussion (RQ3)

We used topic modeling on the 2020, 2021, and 2022 offensive and non-offensive tweets. This, in turn, is used to answer RQ3, examining the most discussed topics in 2020 and the degree to which 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 6 without the representative tokens. 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, #Black-

LivesMatter, and #AllLivesMatter. We describe the top three topics in the 2020 offensive and non-offensive tweets. The 2021 and 2022 offensive and non-offensive tweets topics are discussed in Section A.1 of the Appendix.

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 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 "Kneeling 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. You are a garbage movement."*. Tweets in topics 3, 7, 8, 10, 11, and 12 blamed Breonna Taylor's partner for her death, called for police reform and defunding, called out users for being racist or having opposing views about the BLM movement and debated black/white/all lives matter, criticized the police officers in the death of Breonna Taylor and the justice system, blamed protesters for destroying cities and accusing BLM of funding Antifa, and the shooting of Jacob Blake and criticism of his character respectively.

Focusing on 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 "Kneeling 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."*

Tweets in topic 5 discussed the police, policing, and police brutality and called for police reform. Tweets in topic 7 discussed the BLM protests amid the coronavirus pandemic, and some tweets blamed the protest for the rise in COVID-19 cases. Topic 9 primarily focused on soliciting donations to BLM-related charities and discussed using donated funds to bail out arrested protesters. In topic 10, tweets called for the signing of petitions to raise awareness and stand against injustice. Tweets in topic 12 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"*. Finally, tweets in topic 13 primarily focused on discussing the Blake Lives Matter protests in general. For example, *"Hey so it's August and #BlackLivesMatter is still something you should be following on. We can't let history erase the MONTHS of protests. We must show a rainbow of skin colors marching when we put pictures in the history books."*

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. These topics are discussed in detail in Appendix A.1.

5 Broader Perspectives

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 (Hsueh, Yogeeswaran, and Malinen 2015).

Our offensive and emotion analysis shows that negative emotions, 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" (Bailey 2018). 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 (Bailey 2018) as (Dotson 2012) 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

Offensive		Non-offensive	
#	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	9	Donation
10	Arrest Cops	10	Petitions
11	Antifa	12	Justice
12	#JacobBlake	13	Protest

Table 6: The topics discovered by topic modeling in the 2020 offensive and non-offensive tweets without the representative tokens in the topics.

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 (Houston, Hansen, and Nisbett 2011) and its impact on readers (Walther and Jang 2012) so that readers' conscious or unconscious attitudes towards the Black community are not exacerbated (Hsueh, Yogeeswaran, and Malinen 2015). 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.

6 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. For false positives, the tweet*"@USER @USER Ahmaud Arbery was out taking a run, and didn't deserve to executed!"*.is classified as having a positive sentiment and the offensive model predicted the tweets as offensive even though both tweet is 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.

7 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.

Ethical Statement

We reflect on the ethical and privacy implications of our work due to its sensitive nature. 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). Before our analyses (in

2023), we performed a non-compliance (e.g., deleted tweets or from suspended accounts) check of tweet IDs to ensure compliance with Twitter rules. All non-compliant tweets are excluded from our analyses. The example tweet quotations shown have been modified to protect the identity of the original author. Some content is offensive and sensitive; readers should read content cautiously. We follow the guidelines in (Vidgen et al. 2019) to avoid vicarious trauma[5]. Our institution's instutional review board (IRB) approved this study.

Acknowledgments

This work is partially supported by National Science Foundation (NSF) under the Grant No. 2239605, 2228616, 2228617 and 2114920.

References

An, J.; Kwak, H.; Lee, C. S.; Jun, B.; and Ahn, Y.-Y. 2021. Predicting anti-Asian hateful users on Twitter during COVID-19. *arXiv preprint arXiv:2109.07296*.

Anderson, M.; Barthel, M.; Perrin, A.; and Vogels, E. A. 2020. # BlackLivesMatter surges on Twitter after George Floyd's death.

Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, 759–760.

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

Barreto, S.; Moura, R.; Carvalho, J.; Paes, A.; and Plastino, A. 2023. Sentiment analysis in tweets: an assessment study from classical to modern word representation models. *Data Mining and Knowledge Discovery*, 37(1): 318–380.

Baziotis, C.; Athanasiou, N.; Chronopoulou, A.; Kolovou, A.; Paraskevopoulos, G.; Ellinas, N.; Narayanan, S.; and Potamianos, A. 2018. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Brady, W. J.; Wills, J. A.; Jost, J. T.; Tucker, J. A.; and Van Bavel, J. J. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28): 7313–7318.

Caselli, T.; Basile, V.; Mitrović, J.; and Granitzer, M. 2020a. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Caselli, T.; Basile, V.; Mitrović, J.; Kartoziya, I.; and Granitzer, M. 2020b. 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*, 6193–6202.

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

Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J.; and Blei, D. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

De Choudhury, M.; Jhaver, S.; Sugar, B.; and Weber, I. 2016. Social media participation in an activist movement for racial equality. In *Proceedings of the international aaai conference on web and social media*, volume 10, 92–101.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dinakar, K.; Reichart, R.; and Lieberman, H. 2011. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 11–17.

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

Efron, B.; and Tibshirani, R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, 54–75.

Eichstaedt, J. C.; Sherman, G. T.; Giorgi, S.; Roberts, S. O.; Reynolds, M. E.; Ungar, L. H.; and Guntuku, S. C. 2021. 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.

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

Field, A.; Park, C. Y.; Theophilo, A.; Watson-Daniels, J.; and Tsvetkov, Y. 2022. An analysis of emotions and the prominence of positivity in# BlackLivesMatter tweets. *Proceedings of the National Academy of Sciences*, 119(35): e2205767119.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2025-04-01.

Fortuna, P.; and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4): 1–30.

Freelon, D.; McIlwain, C. D.; and Clark, M. 2016. Beyond the hashtags:# Ferguson,# Blacklivesmatter, and the online struggle for offline justice. *Center for Media & Social Impact, American University, Forthcoming*.

Gallagher, R. J.; Reagan, A. J.; Danforth, C. M.; and Dodds, P. S. 2018. Divergent discourse between protests and counter-protests:# BlackLivesMatter and# AllLivesMatter. *PloS one*, 13(4): e0195644.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12): 2009.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Hinduja, S.; and Patchin, J. W. 2010. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3): 206–221.

Houston, J. B.; Hansen, G. J.; and Nisbett, G. S. 2011. Influence of user comments on perceptions of media bias and third-person effect in online news. *Electronic News*, 5(2): 79–92.

Hsueh, M.; Yogeeswaran, K.; and Malinen, S. 2015. "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human communication research*, 41(4): 557–576.

Ince, J.; Rojas, F.; and Davis, C. A. 2017. 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.

Jakubik, J.; Vössing, M.; Pröllochs, N.; Bär, D.; and Feuerriegel, S. 2023. 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, 423–434.

Jasper, J. M. 2018. *The emotions of protest*. University of Chicago Press.

Jones, K.; Nurse, J. R.; and Li, S. 2022. 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, 417–428.

Killick, R.; Fearnhead, P.; and Eckley, I. A. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500): 1590–1598.

Kim, J.; and Yoo, J. 2012. Role of sentiment in message propagation: Reply vs. retweet behavior in political communication. In *2012 international conference on social informatics*, 131–136. IEEE.

Kumar, D.; Hancock, J.; Thomas, K.; and Durumeric, Z. 2023. Understanding the behaviors of toxic accounts on reddit. In *Proceedings of the ACM Web Conference 2023*, 2797–2807.

Kumar, S.; and Pranesh, R. R. 2021. Tweetblm: A hate speech dataset and analysis of black lives matter-related microblogs on twitter. *arXiv preprint arXiv:2108.12521*.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Leonardelli, E.; Menini, S.; Aprosio, A. P.; Guerini, M.; and Tonelli, S. 2021. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10528–10539.

Liao, S.; Okpala, E.; Cheng, L.; Li, M.; Vishwamitra, N.; Hu, H.; Luo, F.; and Costello, M. 2023. Analysis of COVID-19 Offensive Tweets and Their Targets. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4473–4484.

Liu, P.; Li, W.; and Zou, L. 2019. NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers. In *SemEval@ NAACL-HLT*, 87–91.

Lutz, B.; Adam, M. T.; Feuerriegel, S.; Pröllochs, N.; and Neumann, D. 2023. Affective information processing of fake news: Evidence from NeuroIS. *European Journal of Information Systems*, 1–20.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, 1–17.

Mundt, M.; Ross, K.; and Burnett, C. M. 2018. Scaling social movements through social media: The case of Black Lives Matter. *Social Media+ Society*, 4(4): 2056305118807911.

Naumzik, C.; and Feuerriegel, S. 2022. Detecting false rumors from retweet dynamics on social media. In *Proceedings of the ACM web conference 2022*, 2798–2809.

Nghiem, H.; and Morstatter, F. 2021. " Stop Asian Hate!": Refining Detection of Anti-Asian Hate Speech During the COVID-19 Pandemic. *arXiv preprint arXiv:2112.02265*.

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200*.

Nguyen, T. T.; Criss, S.; Michaels, E. K.; Cross, R. I.; Michaels, J. S.; Dwivedi, P.; Huang, D.; Hsu, E.; Mukhija, K.; Nguyen, L. H.; et al. 2021. 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.

Obadimu, A.; Mead, E.; Hussain, M. N.; and Agarwal, N. 2019. 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*, 214–223. Springer.

Okpala, E.; Cheng, L.; Mbwambo, N.; and Luo, F. 2022. AAEBERT: Debiasing BERT-based Hate Speech Detection Models via Adversarial Learning. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1606–1612. IEEE.

Pavlopoulos, J.; Malakasiotis, P.; and Androutsopoulos, I. 2017. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993*.

Peng, H.; Budak, C.; and Romero, D. M. 2019. Event-driven analysis of crowd dynamics in the Black Lives Matter online social movement. In *The World Wide Web Conference*, 3137–3143.

Peng, J.; Fung, J. S.; Murtaza, M.; Rahman, A.; Walia, P.; Obande, D.; and Verma, A. R. 2022. A sentiment analysis of the Black Lives Matter movement using Twitter. *STEM Fellowship Journal*, (0): 1–11.

Rajamanickam, S.; Mishra, P.; Yannakoudakis, H.; and Shutova, E. 2020. Joint modelling of emotion and abusive language detection. *arXiv preprint arXiv:2005.14028*.

Rajkomar, A.; Oren, E.; Chen, K.; Dai, A. M.; Hajaj, N.; Hardt, M.; Liu, P. J.; Liu, X.; Marcus, J.; Sun, M.; et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1): 1–10.

Raschka, S. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Robertson, C. E.; Pröllochs, N.; Schwarzenegger, K.; Pärnamets, P.; Van Bavel, J. J.; and Feuerriegel, S. 2023. Negativity drives online news consumption. *Nature Human Behaviour*, 7(5): 812–822.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1668–1678.

Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Schmidt, A.; and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, 1–10.

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

Taylor, K.-Y. 2016. *From# BlackLivesMatter to black liberation*. Haymarket Books.

Tong, X.; Li, Y.; Li, J.; Bei, R.; and Zhang, L. 2022. 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*.

Uyheng, J.; and Carley, K. M. 2021. Characterizing network dynamics of online hate communities around the COVID-19 pandemic. *Applied Network Science*, 6(1): 1–21.

Van Troost, D.; Van Stekelenburg, J.; and Klandermans, B. 2013. Emotions of protest. *Emotions in politics: The affect dimension in political tension*, 186–203.

Vidgen, B.; Harris, A.; Nguyen, D.; Tromble, R.; Hale, S.; and Margetts, H. 2019. Challenges and frontiers in abusive content detection. Association for Computational Linguistics.

Walther, J. B.; and Jang, J.-w. 2012. Communication processes in participatory websites. *Journal of Computer-Mediated Communication*, 18(1): 2–15.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, 1391–1399.

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Zhang, P. 2013. The affective response model: A theoretical framework of affective concepts and their relationships in the ICT context. *MIS quarterly*, 247–274.

Zhunis, A.; Lima, G.; Song, H.; Han, J.; and Cha, M. 2022. Emotion bubbles: Emotional composition of online discourse before and after the COVID-19 outbreak. In *Proceedings of the ACM Web Conference 2022*, 2603–2613.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? No
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? No
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, in the limitation section

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? No

(c) Did you include any new assets in the supplemental material or as a URL? No

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No because the Sentiment140 dataset is public.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

A Appendix

A.1 2021 and 2022 Offensive and Non-offensive Topics Discussion

Table 7 shows the results of 2021 and 2022 without the representative tokens. We observe that some topics, such as the death of Breonna Taylor, Ahmaud Arbery, George Floyd, kneeling in Sports, and BLM 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. For example, in the "#DerekChauvin" topic, a user wrote *"I wish #DerekChauvin gets raped everyday for the rest of his life in jail by #thewhatsappblackdude until he starts to like (it) black people #JusticeForGeorgeFloyd"*. In the non-offensive topic "Derek Chauvin", a user wrote *"The conviction of Derek Chauvin yesterday over the death of George Floyd last May is not only the time to celebrate, but also to reflect on America's racial injustice issues, police brutality/ethics, & how far we've come since last summer"*. 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 @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"*.

There were new topics observed, the "Palestinian Hamas" topic, tweets mainly called the world's attention to the conflict between Palestine and Israel. Discussions expressed disappointment in the level of support the conflict is getting compared to the BLM movement in 2020 and also criticized the BLM movement for supporting Hamas. For example, *"It would take IQ of Monkeys for some1 to Support a #Terrorists Organization like #Hamas ! #fact #truth #EraseHamas #BlackLivesMatter = ignorant clowns supporting terrorism against #Israel #IsraelUnderAttack Kill the #PalestinianTerrorists ! #CNN #FoxNews"* and *"@Blklivesmatter Well at least the truth has come out. We now know (BLM) Black Lives Matter is a terrorist organization. I am a proud American infidel! MOLONLABE! I hope Israel kills all of them"*..

The "Rubber Bullets/Riot/Capitol Protests" topics discussed the January 6 Capitol protests by former President Trump supporters and compared how the protesters and BLM protesters were treated differently by the police. For example, in the "Rubber Bullets" topic in 2021, a user wrote *"so Black Lives Matter protesters get teargas pepper spray and rubber bullets but when these blue lives matter Trumpie fuckers want to literally attack police and Government property nothing happens right? practice what you preach you racist trumpie bitches"*. In the 2021 non-offensive topic "Capitol Protests", a user wrote *"Merrick Garland needs to step up and really do the job. The people who stormed the Capitol have been given a slap on the wrist, while Black Lives Matter people have been treated differently"*.

The "Nancy Pelosi" topic criticized Representative Nancy Pelosi for her comment that George Floyd sacrificed his life. For example, *"Pelosi Hammered Over Comments Thanking*

2021		2022	
Offensive	Non-offensive	Offensive	Non-offensive
0. **Ahmaud Arbery**	0. Ahmaud Arbery	0. **Breonna Taylor**	0. Arbery
1. **Breonna Taylor**	1. Derek Chauvin	1. Riots/Protests	1. All Lives Matter
2. Rubber Bullets	3. Breonna Taylor	2. **Ahmaud Arbery**	2. George Floyd
3. **Pregnant Woman**	4. Kenosha Shooting	3. Barack	3. Arrest Cop
4. Nancy Pelosi	5. All Lives Matter	4. #CapitolRiot	4. Abortion
5. Palestinian Hamas	6. Capitol Protests	5. **Drug Overdose**	5. #BlackLivesMatter/#AllLivesMatter
6. #DerekChauvin	7. Kneeling in Sports	6. **Pregnant Woman**	6. #NOH8
7. **Drug Overdose**	9. Justice	8. Racism	8. Kneeling in Sports
8. Black Lives Matter	11. Policing Bill	9. Midterm Elections	9. Drug Overdose

Table 7: The topics discovered by topic modeling in the 2021 and 2022 offensive and non-offensive tweets without the representative tokens in the topics. The highlighted topics are some of the topics in the offensive tweets that persisted in both 2021 and 2022. After 2020, topics related to Floyd, Breonna, and Riots/Protests are still being discussed.

George Floyd 'For Sacrificing Your Life For Justice' URL Wouldn't you just love to take that scarf around her neck and choke her to death?". The "Barack" topic criticized former president Barack Obama for comparing the Uvalde, Texas school shooting to George Floyd. For example, *"Barack Obama Adds Fuel To Flames Inserting George Floyd Into TRAGIC Uvalde Massacre https://t.co/0eUYPfSuNj Words from a hate America communist Muslim thug"*.

The introduction of the policing reform bill is discussed in topic "Policing Bill". For example, *"Today the House reintroduced The George Floyd Justice in Policing Act, including important policing reform measures. Thanks @USER & @USER for taking this step to reform policing in America. This bill is an important contribution to much-needed structural change"*. Topic "Midterm Elections" was primarily political, focusing on the midterm election, Trump, and his lawyer 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 topic "#NOH8 (No Hate)" used the hashtag and #BLM, among others, to promote human equality and to discuss general issues, especially politics.

A.2 List of Search Terms for Data Collection

The list of hashtags and keywords used in retrieving data from Twitter using the Twarc API is given below:

#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.

A.3 Perspective API Threshold & Bias

Perspective assigns a toxicity probability score between 0 and 1 to a text, with higher values indicating high perceived toxicity. The Perspective API suggests using a 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 (Okpala et al. 2022). In particular, Perspective API is biased towards African American English (AAE) (Sap et al. 2019). Due to the potential bias of the Perspective API (Sap et al. 2021, 2019), relabeling tweets helps mitigate bias. While we do not provide dialect or race priming to our data annotators (Sap et al. 2019), 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 AAE that some lexical markers of AAE are reclaimed offensive slurs that are used safely and are not particularly offensive (Sap et al. 2021).

A.4 Data Annotation Decision Tree

The data annotators 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 (Landis and Koch 1977). 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 agree-

ment (Landis and Koch 1977). The modified guideline is formulated as a decision tree as shown in Fig 4. 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 = 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 (Dinakar, Reichart, and Lieberman 2011), achieved a moderate agreement strength (Landis and Koch 1977), 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 (Leonardelli et al. 2021).

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

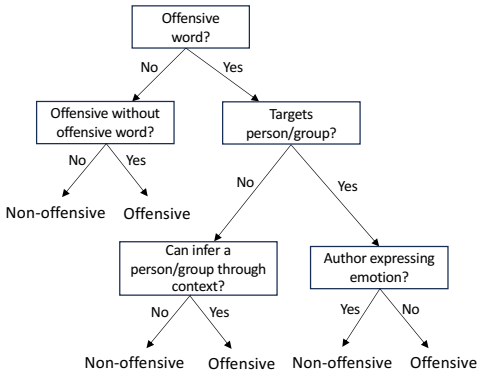


Figure 4: Annotation decision tree used in data labeling. A tweet is labeled offensive if it contains an offensive word, targets a person or group, and the tweet's author is not simply expressing emotion. By emotion, we mean when we don't have enough context to label a tweet as offensive, such as when a user uses an offensive word as commonly used on social media (e.g., @USER I miss you bitch!!!). If a tweet contains an offensive word but does not target a person or group, and the person or group can be inferred from context, then the tweet is labeled as offensive. Finally, a tweet is offensive when it is offensive without having any offensive words.

A.5 Sentiment Model Details

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 size of 256.

A.6 Offensive Model Details

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, 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 (Efron and Tibshirani 1986; Rajkomar et al. 2018; Raschka 2018). 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. We observed that there is significant difference (p-value < 0.05) in AUROC scores using paired t-test.

We compare our results to the fine-tuned BERT, BERTweet, and HateBERT (Caselli et al. 2020a) models 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, fine-tuning of HateBERT uses the fine-tuning hyperparameter specified in the original work (Caselli et al. 2020a). Our model achieved an F1 score of .814 and outperformed other models - BERT (.679), BERTweet (.791), and HateBERT (.710) on the BLM dataset in terms of macro F1 score.

A.7 Emotion Model Details

The per class performance details of our model is depicted in Table 8.

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

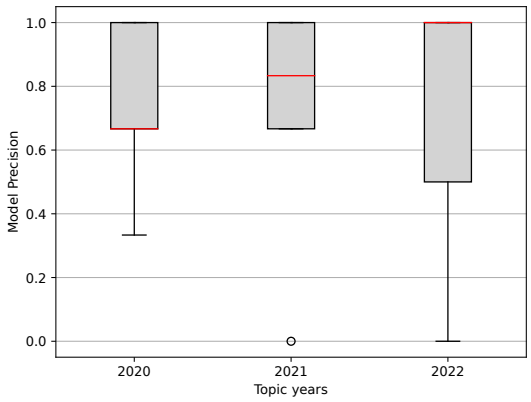
A.8 Topic Modeling Details

Unlike Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), 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. 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[6]. 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 9 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.

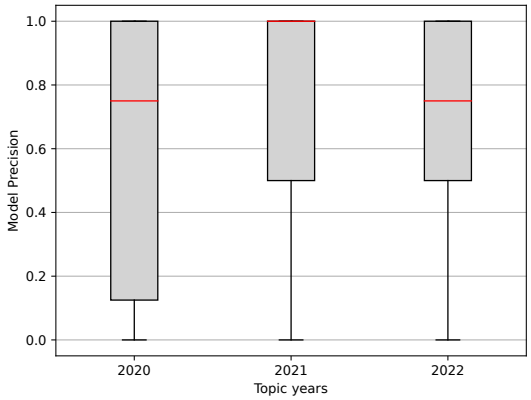
We use the default configurations for each of the main steps of BERTopic (Grootendorst 2022) for topic modeling. To validate our topic model, especially how well the inferred topics correspond with human concepts, we utilized word intrusion (Chang et al. 2009). Word intrusion measures the semantic cohesiveness 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 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 5. From Fig 5, we observe that in each year, the level of agreement is good, indicating that the inferred topics are semantically meaningful.

A.9 Offenders and Recipients Emotion Dynamics

[6]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 5: The model precision of the topic models. Higher is better. The red line within the boxplot represents the median.

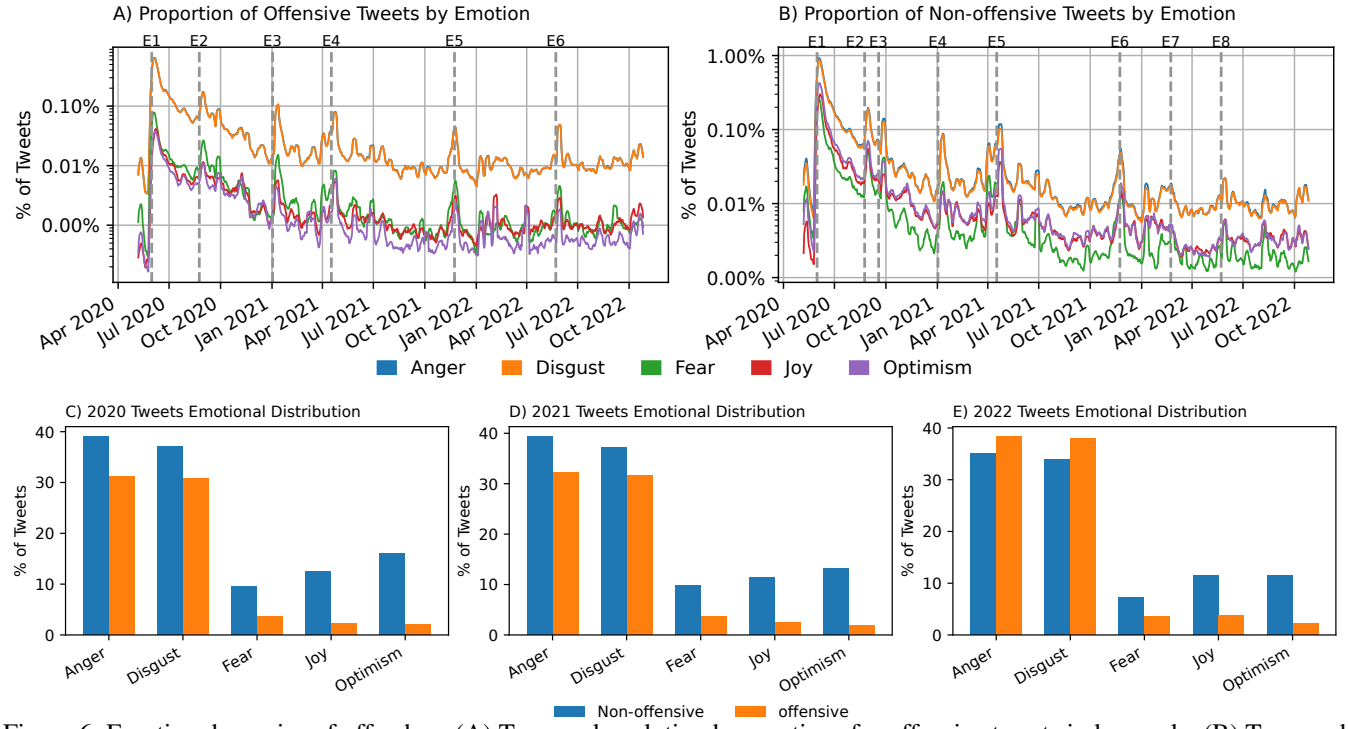


Figure 6: Emotion dynamics of offenders. (A) Temporal evolution by emotions for offensive tweets in log-scale. (B) Temporal evolution by emotions for non-offensive tweets in log-scale (C) Emotion distribution of 2020 tweets. (D) Emotion distribution of 2021 tweets. (E) Emotion distribution of 2022 tweets. 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.

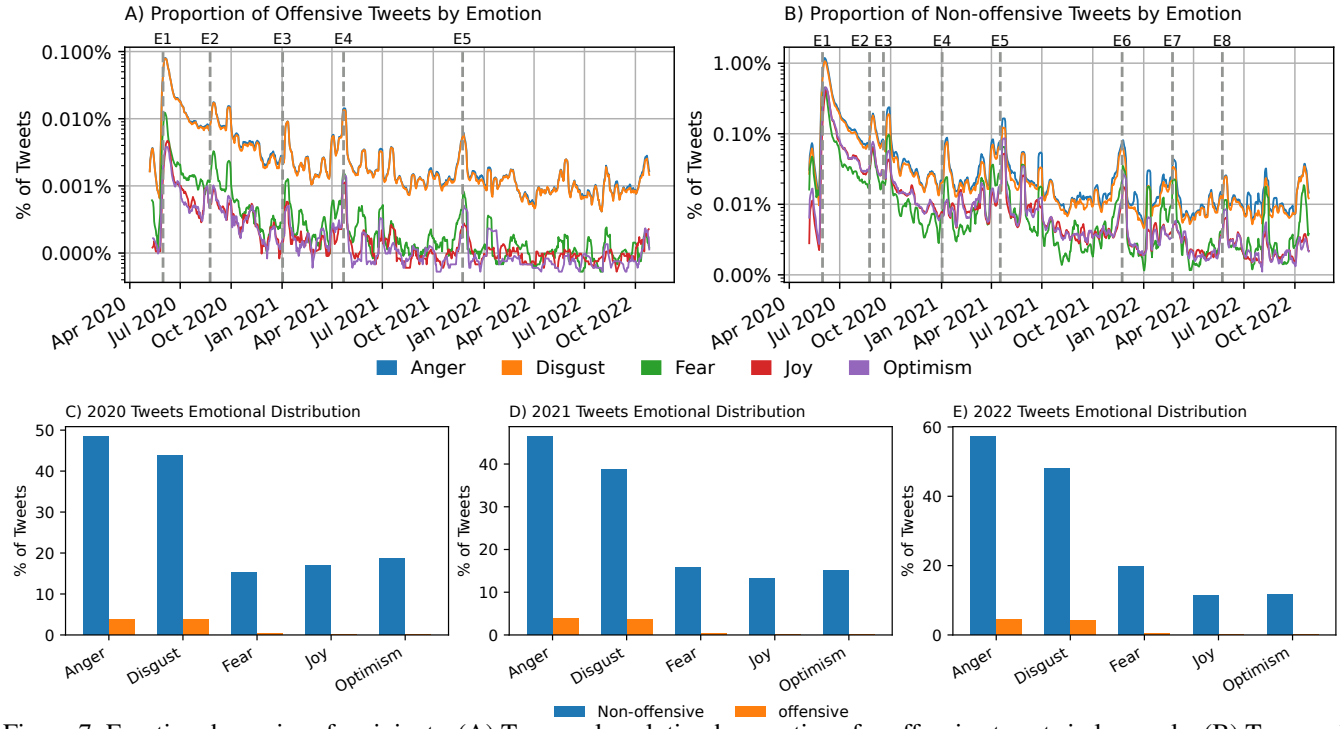


Figure 7: Emotion dynamics of recipients. (A) Temporal evolution by emotions for offensive tweets in log-scale. (B) Temporal evolution by emotions for non-offensive tweets in log-scale (C) Emotion distribution of 2020 tweets. (D) Emotion distribution of 2021 tweets. (E) Emotion distribution of 2022 tweets. 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.