

Hate speech and hate crimes: a data-driven study of evolving discourse around marginalized groups

Malvina Bozhidarova^{† ‡‡}, Jonathn Chang^{‡ ‡‡}, Aaishah Ale-rasool^{* ‡‡}, Yuxiang Liu^{‡ ‡‡}, Chongyao Ma^{§ ‡‡}, Andrea L. Bertozzi[‡], P. Jeffrey Brantingham[¶], Junyuan Lin^{||}, Sanjukta Krishnagopal^{‡ **}

^{*}*Department of Mathematics, University of Texas at Dallas, Richardson, TX, USA*

[†]*Department of Mathematics, University of Nottingham, Nottingham, UK*

[‡]*Department of Mathematics, University of California, Los Angeles, Los Angeles, CA, USA*

[§]*Department of Mathematics, Colby College, Waterville, ME, USA*

[¶]*Department of Anthropology, University of California, Los Angeles, Los Angeles, CA, USA*

^{||}*Department of Mathematics, Loyola Marymount University, Los Angeles, CA, USA*

^{**}*Berkeley Artificial Intelligence Laboratory, University of California, Berkeley, Berkeley, CA, USA*

^{††} Equal contribution ^{‡‡} Equal contribution. ^{**}Corresponding author: sanjukta@math.ucla.edu

Abstract—This study explores the dynamic relationship between online discourse, as observed in tweets, and physical hate crimes, focusing on marginalized groups. Leveraging natural language processing techniques, including keyword extraction and topic modeling, we analyze the evolution of online discourse after events affecting these groups. Examining sentiment and polarizing tweets, we establish correlations with hate crimes in Black and LGBTQ+ communities. Using a knowledge graph, we connect tweets, users, topics, and hate crimes, enabling network analyses. Our findings reveal divergent patterns in the evolution of user communities for Black and LGBTQ+ groups, with notable differences in sentiment among influential users. This analysis sheds light on distinctive online discourse patterns and emphasizes the need to monitor hate speech to prevent hate crimes, especially following significant events impacting marginalized communities.

Index Terms—hate speech, knowledge graph, topic modeling, sentiment analysis, dynamic network analysis

I. INTRODUCTION

In the wake of escalating racial tensions and recent surges in xenophobia, hate crime directed at marginalized communities has re-emerged as a significant problem in many settings [1]. There is widespread concern that social media plays a significant role in promoting hateful behavior [2]. While hate speech is a general challenge with wide relevance to society, its effects are particularly severe following major social disruptions such as the murder of George Floyd or the outbreak of the COVID-19 pandemic [3].

Prior research has examined the complex interplay between online discourse and offline actions [4,5]. Müller and Schwarz

This material is based upon work supported by the National Science Foundation grants DMS-2027277 and DMS-2318817 and AFOSR-MURI grant FA9550-22-1-0380. MB was supported by the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 777826 (NoMADS). Any opinions, findings, and conclusions expressed in this material are those of the authors.

[6] compared occurrences of Muslim-related comments on Twitter to the frequency of anti-Muslim hate crimes. They found that tweets about Muslims by prominent online figures triggered increases in xenophobic tweets by others, cable news mentions of Muslims, and hate crimes against Muslims in the following days. Similar results on social media enabling the spread of extreme viewpoints that migrate to the physical world have been echoed in [2,7]. Conversely, Lupu et al. [8] found that offline events, such as protests and elections, appear to trigger a rise in online hate speech. Lastly, Awan and Zempi [9] report on qualitative interviews suggesting that individuals targeted online are also frequently targeted offline.

The present work seeks to understand the broader contextual relationships between online hate speech and offline hate crime. We focus on a collection of tweets from California and examine their connections with hate crimes recorded from a publicly available dataset [10]. The Twitter dataset is processed using keyword filtering and topic modeling to identify tweets related to Black, Asian, Hispanic, Jewish, and LGBTQ+ groups. We refer to tweets sorted by topic as “topic groups” since a tweet’s membership in a group is dependent on the content of the tweet, not on the characteristics of the individual who posted the tweet. As hate speech is not defined consistently, we use sentiment analysis to study the behavior of tweets with negative sentiment. We present statistics for negative sentiment tweets aimed at Black, Asian, LGBTQ+, Hispanic, and Jewish communities, but restrict ourselves to Black and LGBTQ+ groups for detailed analysis.

First, we perform time series analysis to identify significant temporal correlations between tweet topic groups and hate crimes for Black and LGBTQ+ groups and find that online and offline activity both peak around major societal events pertaining to these groups.

Following this, we draw on network analyses to study the

social behavior among the Black and LGBTQ+ topic groups. Network models provide a robust framework to study time-evolving relationships that underlie various complex phenomena (e.g. social isolation and unemployment [11], friendship and violence among classmates [12]) They can also provide concise graphical descriptions of data that lend easily to analysis of the most central nodes [13].

In our research, we use a knowledge graph [14] to structure tweets and their multitude of attributes, leveraging the graphical structure to uncover connections between users and their discussions. We employ dynamic community detection [15] to compare the temporal evolution of communities of users associated with the Black and LGBTQ+ topic groups. In both topic groups, the emergence of new communities co-occurs with large-scale real-world events. In addition, the community of users in the Black topic group displays patterns of emergence and dissolution, while the community of users in the LGBTQ+ topic group tend to split over time but displays less dynamic activity. We also construct a user network, a complete graph connecting active users in each topic group weighted by the similarity of their tweets. Finally, we define ‘influence’ as a way to measure user importance by combining network-based centrality metrics on the user network with follower count. We find that the most influential users in the Black topic group user network tweet more negatively and have more followers than the most central users in the LGBTQ+ topic group user network, highlighting important differences in their online behavior.

This paper is structured as follows. In Section II, we introduce the raw data and filter it by keywords related to five underrepresented groups. In Section III, we perform sentiment analysis and topic modeling on the filtered tweets and assign topics to these five targeted groups. In Section IV, we conduct time series analysis on the Black and LGBTQ+ topic groups, comparing the frequencies of low-sentiment tweets and hate crimes related to these topic groups. In Section V, we construct the knowledge graph and perform dynamic community detection to compare the evolution of communities on Black and LGBTQ+ topic group graphs. In Section VI, we construct a weighted user network and compute centrality to compare the behavior of the most influential users among the Black and LGBTQ+ topic groups. In Section VII, we summarize our findings and outline possible future directions.

II. DATASETS

A. Twitter Dataset

To model online discourse, we collect a Twitter (now known as X) dataset consisting of more than 28 million tweets dating from March 11, 2020, to June 17, 2021. The tweets were collected using the Twitter API and were selected using the bounding box attribute to include those with user attributes or tweet locations tied to the greater Los Angeles area [16]. Users may be associated with Los Angeles but active in another location. Thus, most tweets are concentrated in Los Angeles, but other locations in California are also observed in the dataset. We focus on Los Angeles due to its status

Group	Keyword (Phrases)
Black	‘blm’, ‘george floyd’, ‘blackout tuesday’, ‘say her name’
Asian	‘amplify asian voices’, ‘chinese virus’, ‘stop asian hate’
LGBTQ+	‘homophobia’, ‘trans lives matter’, ‘drag queen’
Hispanic	‘illegal aliens’, ‘border wall’, ‘hispanic lives matter’
Jewish	‘jewish pride’, ‘anti semitism’, ‘nazism’, ‘scapegoat’

TABLE I: Keyword Examples

as a large and diverse metropolitan city in the US, but our ensuing methodologies can be extended to other areas given appropriate data.

Each tweet in the dataset is represented as an object containing a unique ID number, the date of its creation, raw text information, and separate fields for the tweet’s hashtags and mentions. The tweet object also includes specific information about the user who posted it: the user’s number of followers at the time of posting the tweet, their unique ID number, and their screen name at the time of posting the tweet. In addition, 81% of the tweets have the “lang” attribute set to English, but this increases to 88% after filtering by keywords. We do not filter tweets by language because we are interested in representing the cultural diversity of online discussion in California.

We process the data by filtering the tweets by their textual information through keyword extraction. The keywords are selected to match the most polarizing tweets, so they are typically derogatory or empowering towards specific groups, including slang or racial slurs. We source the keywords from academic journals and online glossaries that have accumulated lists of words relevant to our groups of interest [17,18]. Examples of these keywords are given in Table I, and the full list is given in a GitHub repository (VIII). We omit words that do not specifically target the five groups as well as generic keywords that are typically non-polarizing, such as “racism”, “justice”, and “equality”. We augment the keyword list by including plurals, removing spaces, and replacing spaces with hyphens. The pre-augmented list contains 162 items, while the augmented list contains 694. We also clean the tweets by making all characters lowercase and removing punctuation, emojis, and non-alphanumeric characters. After filtering the tweets by keeping only the tweets that contain keywords in our augmented list, the processed dataset contains 74639 tweets.

B. Hate Crimes Data

The hate crime dataset is taken from the FBI Crime Data Explorer, a digital repository under the FBI’s Uniform Crime Reporting (UCR) program [10]. This data contains approximately 220k incidents from 1991 to 2021, including reports from all U.S. counties that contribute to the FBI.

The FBI defines hate crimes as offenses “motivated, in whole or in part, by an offender’s bias against a race, gender, gender identity, religion, disability, sexual orientation, or ethnicity, and committed against people, property, or society” [19]. Due to the substantial evidence required to classify an incident as a hate crime and the lack of reporting from certain counties, the FBI’s dataset has been criticized for underestimation [6]. Still, it remains the most comprehensive, publicly

available source of information on hate crimes. Each recorded incident includes details such as the date, location type (e.g., household, street, park), offense type (e.g., vandalism, assault, intimidation), and the race and number of offenders if known. The FBI also identified 35 motivating biases (e.g., anti-Black, anti-Asian) and classified each incident according to established evidence of bias.

In this work, we consider incidents that occurred in the state of California between March 11, 2020 and July 17, 2021, totaling 1192 hate crimes. We exclude crime data before March 11, 2020 and after July 17, 2021, in order to align the timestamps of hate crime and Twitter datasets. Within this timeframe, most hate crime incidents are racially motivated, followed by sexuality-based offenses. The primary offense types are property vandalism and simple assaults, followed by intimidation. Figure 1 shows that the five most common bias types for hate crime events are anti-Black, anti-Gay (Male), anti-Hispanic, anti-Jewish, and anti-Asian.

III. NATURAL LANGUAGE PROCESSING

In this section, we discuss sentiment analysis and topic modeling, which are Natural Language Processing (NLP) methods used to extract information from the Twitter dataset. The NLP pipeline is outlined in Figure 2.

A. Sentiment Analysis

We use the Valence Aware Dictionary for sEntiment Reasoning (VADER) [20] model for sentiment analysis. VADER is a state-of-the-art lexicon and rule-based sentiment analysis tool to extract social media sentiments. We select VADER for its efficiency in computation time as well as its interpretability and consistency in performance without the need of finetuning [21]. Here, sentiment analysis is used to gauge public opinion, allowing us to identify, and analyze, the most polarizing tweets and users, providing insight into the most significant (and potentially harmful) trends within these online communities.

B. Topic Modeling

We leverage BERTopic, a widely used semantic-based topic modeling technique, to identify the most relevant topics in the filtered dataset. BERTopic first generates document embedding using the BERT model, then clusters these embeddings, and finally generates topic representations using a class-based TF-IDF procedure [22]. This process categorizes the keyword-filtered tweets into a concise list of interpretable topics relevant to our groups of interest. BERTopic also serves as a noise filter by placing tweets with spam-like or noisy text, analogous to that generated by bots, into distinct topics of their own.

The topic model gives 270 topics, with the most prominent topic having 1540 tweets. We manually categorize topics into the following topic groups: 141 topics with 25467 tweets are related to a Black topic group, 16 topics with 3088 tweets to an LGBTQ+ topic group, 15 topics with 1846 tweets to an Asian topic group, 21 topics with 6585 tweets to a Jewish topic group, and 5 topics with 1000 tweets to a Hispanic topic group. The remaining topics are discarded. Topics that appear

in more than one topic group are counted in each of the topic groups that they appear in, allowing us to identify connections between topic groups for the knowledge graph construction (see below). Around 38% of the tweets, including a significant number of relevant tweets, are classified into an ‘irrelevant’ or ‘miscellaneous’ topic by BERTopic. To reduce the number of outliers, we also extract, from the clustering layer of BERTopic, the probability distribution over the topics for each tweet. Then, we move those tweets, originally classified into the miscellaneous topic, into one of the relevant topics if its probability for that topic is ≥ 0.01 . This allows us to classify each tweet into several topics, which we refer to as “multi-topic modeling”. We use the hyperparameters `min_topic_size` set to 30 to minimize the number of extraneous or repeated topics, as well as `top_n_words` and `n_gram_range` set to 15 and (1,3), respectively, to generate adequately sized topic names.

Figure 3 displays the weekly frequency of tweets in each topic as a fraction of the total tweets for two representative topics in each of our five topic groups. We see that tweet spikes in each topic group coincide with real-world events related to those topic groups. For example, the increase in the proportion of tweets with topics `1_lives_matter` and `4_chauvin_derek` corresponds to the Murder of George Floyd. Similarly, when Pride Month began in June 2020, the proportion of tweets with topics `18_homophobia` and `28_gay_lgbtq` increased. This provides evidence that the topic groups identified by BERTopic are meaningful.

IV. TEMPORAL DATA ANALYSIS

In this section, we examine the temporal correlation between online discourse and hate crimes targeting specific groups, and their relationship to real-world events. We focus on the Black and LGBTQ+ topic groups due to the availability of significantly larger amounts of data compared to other groups. We consider data from March 2020 to December 2020 and exclude data after December 31, 2020 due to an officially recognized discrepancy in collected crime statistics [23] outside of this period that may yield misinformed correlations between discourse and crimes. We extract tweets related to the LGBTQ+ and Black topic groups and study the temporal correlation between various quantities characterizing online activity (e.g., daily tweet count, average sentiment) and daily occurrences of hate crimes. We investigate trends in this correlation in the aftermath of major events such as the murder of George Floyd and Pride Month. To minimize noise in the daily data, we employ a 30-day rolling window for our time series analysis.

Figures 4a and 4b illustrate the number of hate crimes and the total number of tweets in the Black and LGBTQ+ topic groups. In both figures, we see that the number of hate crimes and tweets increased after May 2020. In Figure 4a, there is a peak in anti-Black hate crimes following a peak in the Black topic group shortly after the murder of George Floyd, which received widespread public attention. Interestingly, while the frequency of tweets decreased rapidly over the subsequent

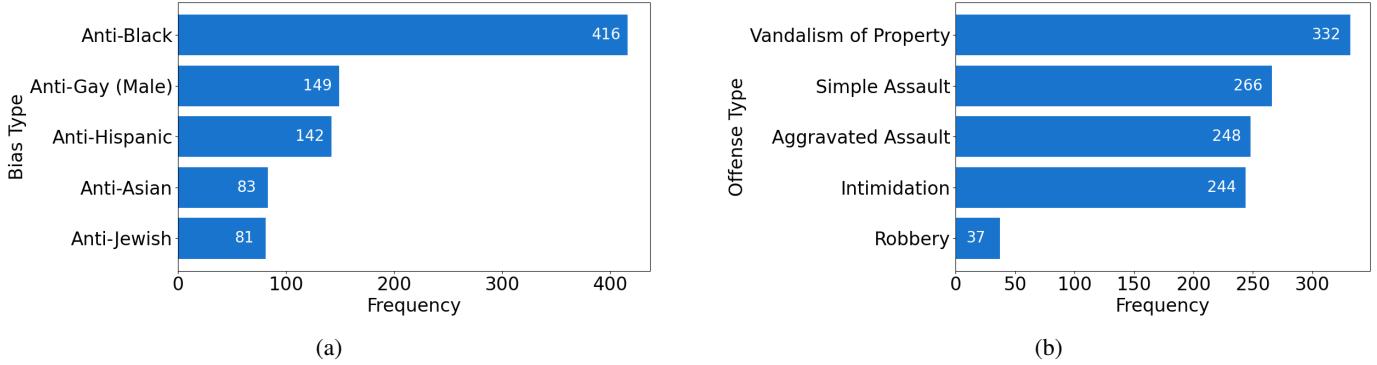


Fig. 1: Five most common motivating biases (left) and offense types (right) of hate crimes in California from 2020-03-11 to 2021-06-17.

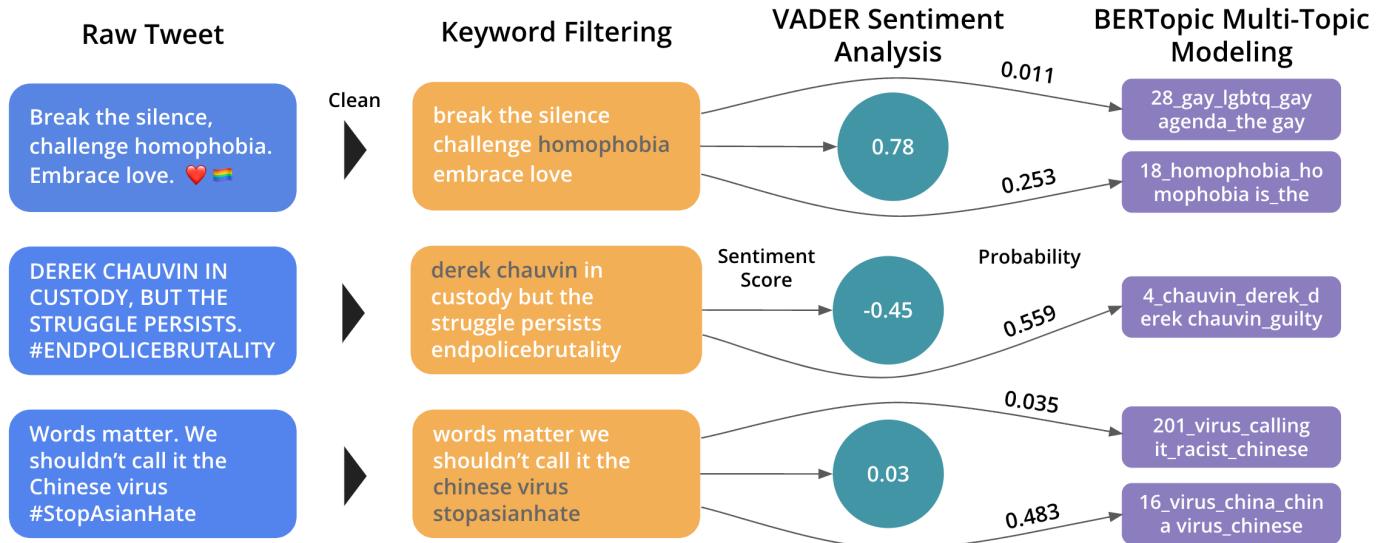


Fig. 2: Natural Language Processing (NLP) Pipeline.

month, the number of hate crimes stayed relatively high for up to six months. A similar peak in the LGBTQ+ topic group and hate crimes can be seen in Figure 4b around Pride Month; while the number of tweets decreased following this, analogous to the Black topic group, anti-LGBTQ+ hate crimes are more consistent throughout 2020.

Next, we explore the potential correlation between online sentiment and hate crimes. Figures 4c and 4d show the average number of hate crimes compared to the average sentiment of tweets in the Black and LGBTQ+ topic groups, respectively. Sentiments take values between -1 and 1, and positive values are associated with positive sentiments whereas negative values are associated with negative sentiment. The absolute value of sentiment indicates how strong the sentiment is. Here we threshold the strength of a tweet to give a binary classifier; we consider a tweet to be ‘strong’ if it has a sentiment with an absolute value greater than 0.5. The results illustrate that the peak in the average sentiment for each group aligns with a corresponding major event—the murder of George Floyd and Pride Month, respectively. However, this peak is around 0

(neutral average sentiment) for the Black topic group, whereas it’s 0.05 (positive average sentiment) for the LGBTQ+ groups, indicating the difference in the nature of the public events (death versus Pride week), and their effect on public sentiment. Additionally, the range of sentiment is larger (captured on the y-axis on the right) for LGBTQ+ groups.

Additionally, peak sentiment is aligned with peak hate crimes (particularly evident in the black topic group), which may initially seem counterintuitive as one might expect that higher mean sentiment in online discussions would correspond to fewer hate crimes. However, this can be explained by the unfolding of events after the murder of George Floyd. Immediately after, there were widespread protests and demonstrations against racial injustice and police brutality. These events galvanized a global movement advocating for racial equality and social justice and resulted in many people expressing positive sentiments online. We want to emphasize that we do not make any claims of causality here.

Finally, we investigate a correlation between hate crimes and the strength of tweets in topic groups. Figures 4e and

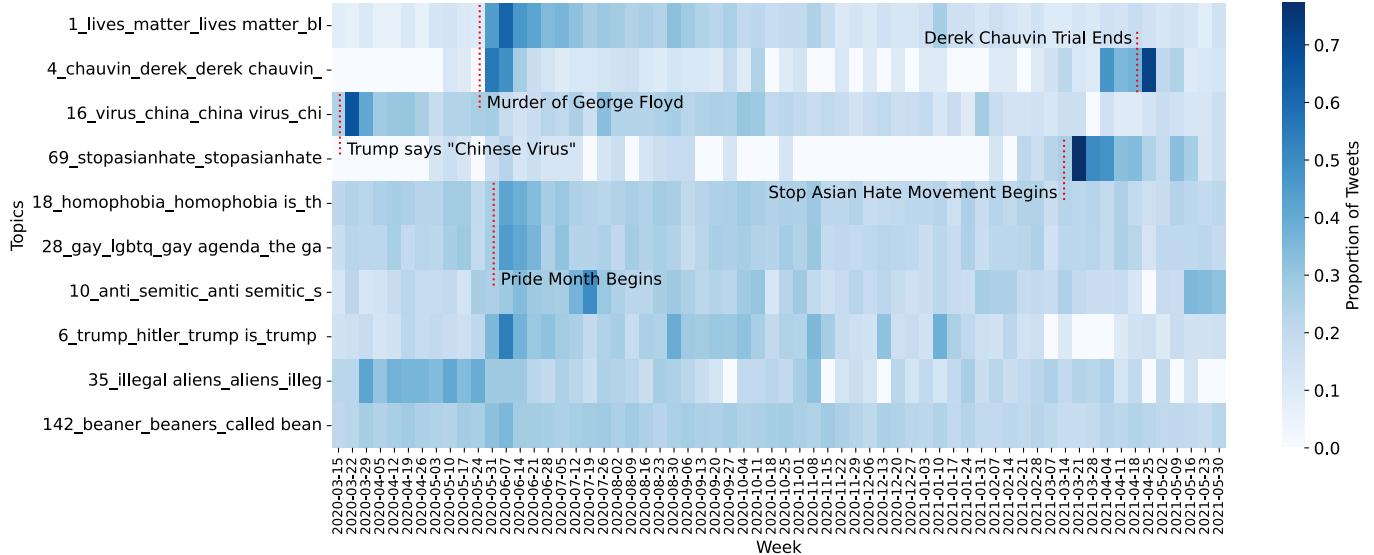


Fig. 3: Proportion of tweets in topics from 2020-03 to 2021-06 weekly, annotated with relevant events. Two representative topics per group are selected randomly from each of the five topic groups: (top to bottom) Black, Asian, LGBTQ+, Jewish, and Hispanic. The shade of blue indicates the frequency of tweets from a specific topic (truncated) that occurred in a given week, relative to the total number of tweets in that topic. Real-world events relevant to a specific group are denoted by red dotted lines in the rows relevant to those groups.

4f illustrate the frequency of hate crimes targeting Black and LGBTQ+ topic groups alongside the proportion of tweets exhibiting a strong sentiment. We observe an interesting pattern where the rise in anti-Black hate crimes in June 2020 follows an increase in the proportion of strong sentiment tweets in May 2020. By contrast, an increase in anti-LGBTQ+ hate crimes corresponds to a lower proportion of strong tweets within that topic group. We also observe that the proportion of tweets that have strong sentiment is on average slightly higher in LGBTQ+ communities than black communities (their range is captured on the y-axis on the right). Moreover, the highest proportion of strong sentiment LGBTQ+ topic tweets aligns with the lowest proportion of strong Black topic tweets, suggesting that users may shift their focus between these topic groups. The results suggest that there is not one simple relationship between the strength of sentiments in online content and hate crime, but rather something that varies by group and context.

Next, we quantify the relationship between the temporal evolution of the above quantities. Table II presents the Spearman correlation coefficients r_s [24] and their associated p -values for each pair of time series in Figure 4. The r_s values show the strength and direction of monotonic relationships between each pair, while the p -values show the significance of the correlation. Table II shows that the correlation between the number of group-topic tweets and anti-group hate crimes differs for the Black and LGBTQ+ groups. While this correlation is not significantly different than 0 ($p = 0.773$) for the Black group, it is moderate and negative for the LGBTQ+ group. This suggests that changes in the number of Black-topic tweets do not appear to correlate with the occurrence of anti-Black

hate crimes. In contrast, the negative correlation between the number of LGBTQ+ topic tweets and anti-LGBTQ+ hate crimes may be indicative of increased awareness, advocacy, education, and community support within online discussions, potentially contributing to a reduced occurrence of such hate crimes. We also see that the average sentiment of tweets in both the Black and LGBTQ+ topic groups exhibits a negative correlation with the number of hate crimes towards the specific group, implying that sentiment of online discourse can affect public perspective and negative sentiment can lead to social environments that promote hate crimes.

Finally, we find that the correlation between the proportion of strong sentiment tweets and the number of hate crimes towards the targeted group display opposing patterns for the Black group (negative correlation) and the LGBTQ+ group (positive correlation). We postulate that the strong sentiment tweets in the Black group may lead to an increased level of awareness, concern, or activism, which may contribute to lower hate crime rates. Inversely, strong sentiment tweets in the LGBTQ+ group may lead to increased hate. Although speculative, these hypotheses could inform targeted strategies to detect and prevent the escalation of strong sentiment tweets into physical acts of hate.

V. KNOWLEDGE GRAPH

A knowledge graph (KG) is a network consisting of nodes (real-world entities such as objects, events, and people) and edges (the relationships encoded between them) [14]. KGs provide a robust methodology to organize, structure, and analyze large quantities of data with complex relationships between different types of objects such as text data from

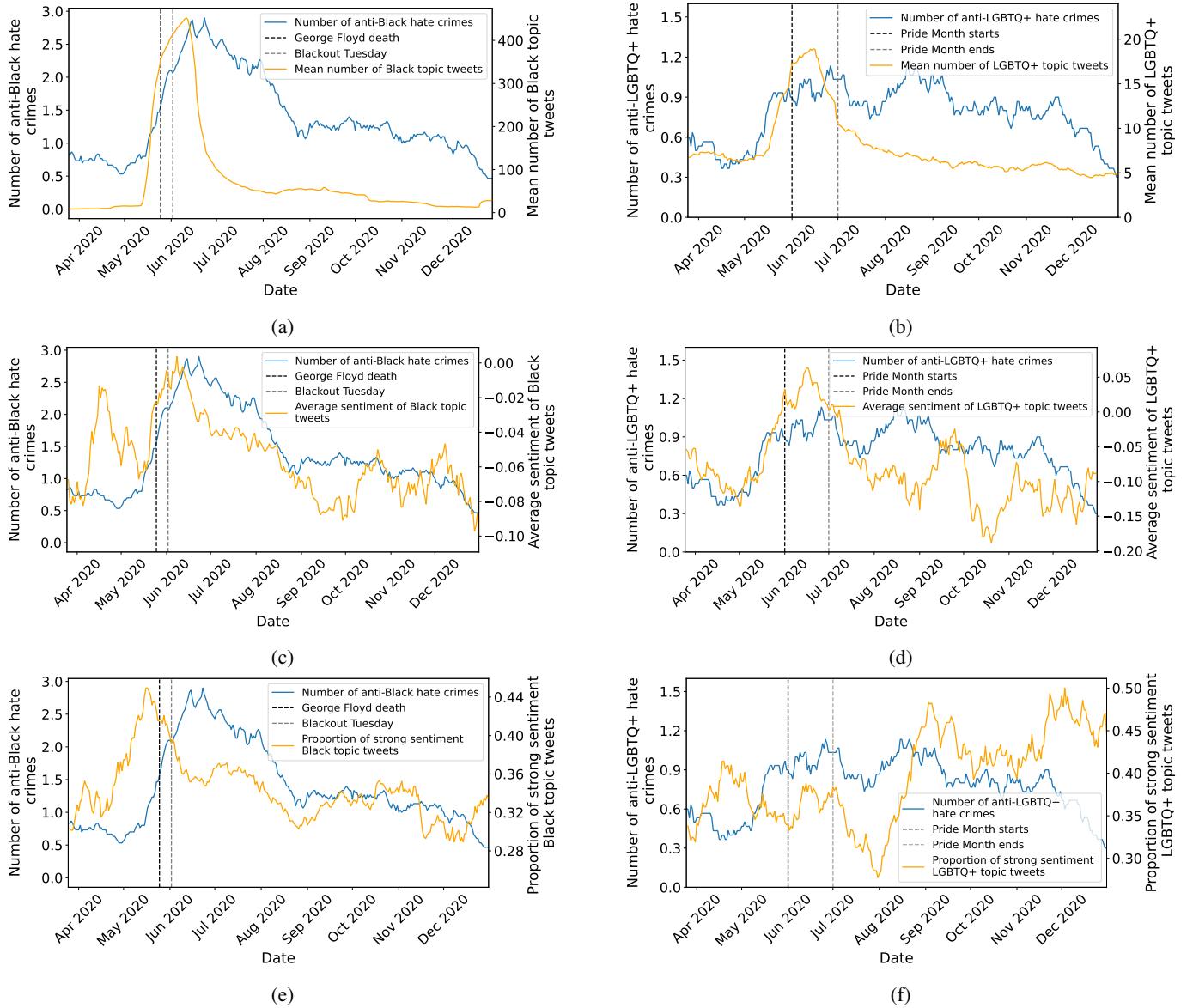


Fig. 4: The number of (left) anti-Black and (right) anti-LGBTQ+ hate crimes versus (a,b) the number, (c,d) the average sentiment, and (e,f) the proportion of “strong” sentiments, of all tweets in the respective topic groups. In each plot, the left y-axis corresponds to the blue curve, while the right y-axis corresponds to the orange curve.

various online platforms, and have been used successfully in a variety of data-driven research problems [25]. A KG is particularly useful to identify complex patterns among topic groups and hate crimes that are not apparent from our time series analysis, by extracting subgraphs (e.g. extracting clusters of users and analyzing the sentiments of the most prevalent topics). We construct a KG using the semantic, sentiment, and topic information extracted from the tweets.

A. KG Construction

Our KG is a weighted and directed graph $G = (V, R, \Delta)$ where V is the set of all nodes (consisting of seven different entity types), R is the set of nine edge (relation) types, and $\Delta \subset V \times V \times R \times \mathbb{R}$ is the set of weighted triples. Each

weighted triple $\alpha = (h, t, r, w) \in \Delta$ consists of a head and tail node $h, t \in V$, a relation type $r \in R$, and a weight $w \in [0, 1]$ denoting the strength of the relation. The entity and relation types are summarized in our ontology (Figure 5). Five of these—tweeted, replied_to, mentioned, has_hashtag, tweeted_on—are directly extracted from the Twitter dataset. For the in_topic relation, we weight the edge using the probabilities calculated during topic modeling in Section III.

We filter for tweets with at least a 1% likelihood of being in at least one of the 191 topics relevant to our five groups of interest, which leaves us with 57,142 tweets. The mean number of in_topic relations is about four per tweet. The user and tweet entities, along with their relations, form a

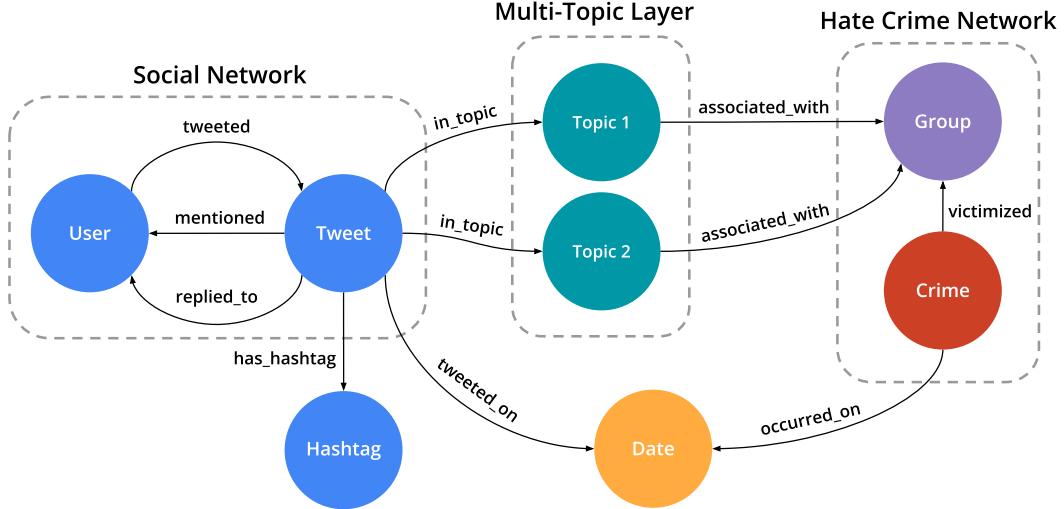


Fig. 5: Ontology describing the entity and relation types in the KG. Blue entities are extracted from the Twitter dataset, teal entities form the Multi-Topic Layer connecting each tweet to its relevant topics, and the red and purple group and crime entities establish a network of real-world victims. Tweets and crimes are also connected by the date that they occurred.

Correlation with hate crimes of the specific group	r_s	p-value
Number of Black topic tweets	0.017	0.773
Number of LGBTQ+ topic tweets	-0.545	5.60e-24
Average sentiment of Black topic tweets	-0.387	7.69e-12
Average sentiment of LGBTQ+ topic tweets	-0.601	4.70e-30
Proportion of strong sentiment Black topic tweets	-0.352	6.19e-10
Proportion of strong sentiment LGBTQ+ topic tweets	0.463	6.96e-17

TABLE II: The Spearman correlation coefficients and corresponding p -values between the number of anti-Black or anti-LGBTQ+ hate crimes and the number of tweets (top rows), the average sentiment of tweets (mid rows), and the proportion of strong sentiment tweets (bottom rows) in the Black or LGBTQ+ topic group respectively.

Entity	Count	Relation	Count
Tweet	57,142	tweeted	57,142
User	22,065	mentioned	604
Hashtags	10,948	replied_to	2,533
Crime	738	in_topic	214,821
Date	464	has_hashtag	55,127
Topic	191	tweeted_on	57,142
Group	5	occurred_on	738
Total	91,553	victimized	738
		associated_with	198
		Total	389,043

TABLE III: KG Statistics

“Twitter Social Network”—a subgraph encoding the complex interactions between users of different levels of influence. Incorporating the topic entities into the subgraph (which are generated by tweet text) allows us to perform analysis among tweets related to a certain topic group and examine the discourse over various subject matters in a group. Tweets and hate crimes are implicitly connected through the date entity. This enables temporal analysis of the KG through dynamic community detection.

We also introduce a group entity, which allows us to connect tweets about marginalized groups with the hate crimes targeting those groups. This is enabled by the multi-topic modeling, allowing us to categorize tweets into our desired topic group communities. The entity and relation counts for the completed KG are displayed in Table III.

B. Dynamic Community Detection

To better understand the dynamics of the social network, we examine the evolution of networks involved in the Black and LGBTQ+ topic groups, focusing on the users discussing topics relevant to each topic group over time. Easy extraction of relevant quantities is enabled by the knowledge graph.

We extract user-tweet, tweet-topic, and relevant user-user triples to form respective sub-graphs for the Black and LGBTQ+ topic groups. We then partition undirected subgraphs into monthly snapshots and apply Louvain community detection [26] to each snapshot. We omit the edge weight for tweet-topic (“in_topic”) relations during community detection to allow better clustering of communities. We record the users within each Louvain community from monthly snapshots. We then use the Jaccard index $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ to track common users in the communities over time [27]. Thus, users within a community tweet on similar topics. We connect two user communities A at time t and B at time $t + 1$ with an edge

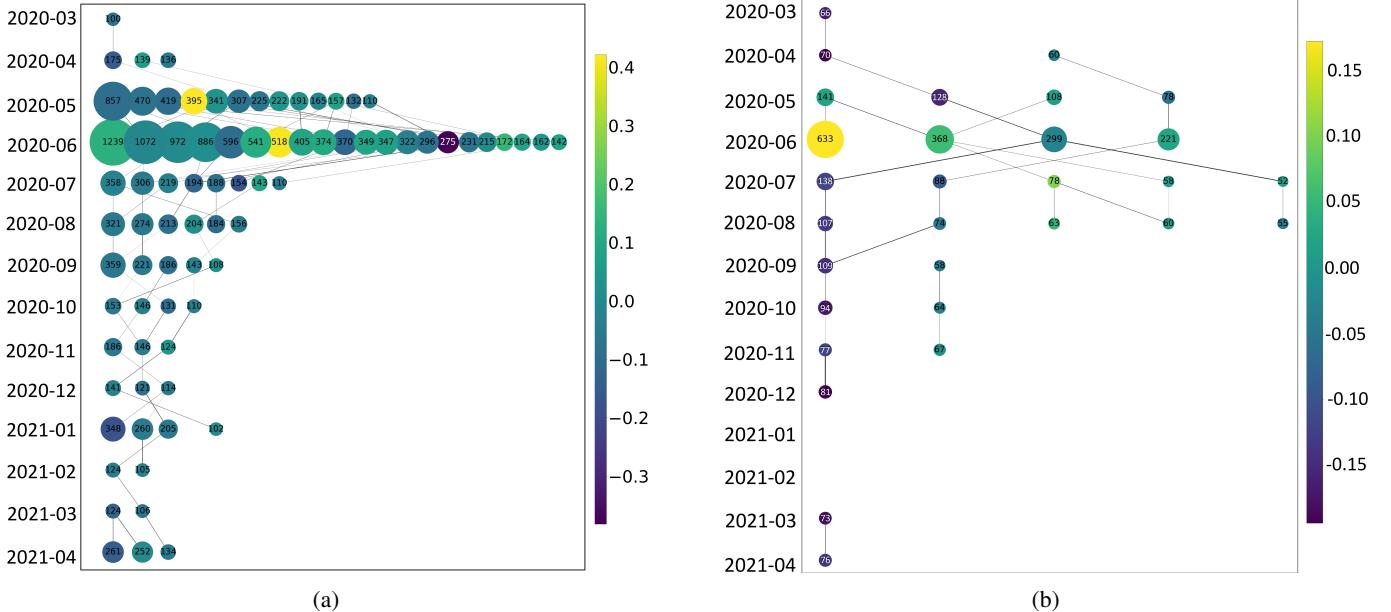


Fig. 6: The evolution of users communities in the a) Black-topic group network and b) LGBTQ+ topic group network ($\gamma = 0.01$). The node size indicates the size of a community (also shown in numbers), the color shows the mean sentiment of the users in the community, and the edge width indicates the number of common users in the connected communities.

if $J(A, B) \geq \gamma$ with $\gamma = 0.01$. To reduce the noise, we only consider communities with 100 or more users for the Black topic group and the communities with 50 or more users for the LGBTQ+ topic group.

Figure 6 illustrates the dynamics of these user communities over time. Notably, neither of these user communities remains constant over the period from March 2020 to June 2021. Instead, communities appear to emerge and dissolve over time. During May and June 2020, many new, big communities appeared in association with the Black topic group. But these new groups persist for one or two months only. The surge in new communities coincides with George Floyd’s murder, capturing a relationship between this public event and the change in online discourse patterns. Moreover, most of these new communities are connected to tweets with positive sentiments, indicating an expression of solidarity with and support for the Black community.

Similarly, the largest communities in the LGBTQ+ topic group appeared in June 2020, corresponding to Pride Month. However, here, not as many new communities emerged and instead communities tended to split into sub-communities over time. This may indicate that discussions within the LGBTQ+ community became more diverse and multifaceted during this period, leading to splitting into sub-topics or sub-interests. Most of these communities expressed positive sentiment during Pride Month, as one may expect, with the largest community having the highest positive sentiment. However, throughout the rest of the time period, the average sentiment in the largest LGBTQ+ topic group communities is negative.

Overall, the differing patterns of community evolution between Black and LGBTQ+ topic groups highlight each net-

work’s unique characteristics and dynamics. Moreover, in both networks, the change in sentiment and communities over time is closely connected to real-life events, showing that online discussions and sentiment are sensitive to external factors and societal occurrences.

VI. USER NETWORK

To analyze the most central users among the Black and LGBTQ+ topic groups, we build Black or LGBTQ+ “user networks” (similarity networks with users as nodes). These networks examine the similarities between the tweets of active (frequently-posting) users through edges with weight given by similarity in their tweeted topics, using the topic probabilities generated by BERTopic.

A. User Network Construction

First, we filter the dataset to include 1195 users who tweeted at least ten times in the overall time period. Then, between every pair of users A, B we assign the weight

$$w_{A,B} = \sum_{(t_1, t_2) \in T_A \times T_B} \sum_{\tau \in L} t_1^{(\tau)} \cdot t_2^{(\tau)}, \quad (1)$$

where T_A and T_B are the collections of tweets posted by users A and B , respectively, L is a set of selected topics, and $t_i^{(\tau)}$ refers to the probability that the tweet t_i is in topic τ assigned by BERTopic. This formulation defines a “similarity” between users. Edges with high weights connect users who have similar topic distributions. This produces a complete graph where the weighted degree of a user indicates their overall similarity to other users in the network in terms of (1). Users who tweet on the most common topics tend to have larger weighted degrees

as their tweets are more likely to be on the same topics as other users. Summing over every pair of tweets between users also gives weight to users who tweet more frequently. We create two different networks by varying L : a user network using only Black topics and a user network using only LGBTQ+ topics. As such, the top users are different in each network.

B. Exploring Influential Users

Next, we compute the eigenvector centrality of each user, which measures the importance of the user through the importance of its neighbors. Because edge weights (and consequently centrality) do not account for a user's follower count which can be an important quantity, we introduce influence I :

$$I = c \cdot \ln(f + 1),$$

where c is the user's eigenvector centrality and f their follower count. We take the natural logarithm of the follower count because the distribution of the follower counts is highly skewed. Note that the follower count attribute may include followers (users) that are not in this network or in the dataset, nevertheless it provides a natural estimate of user influence.

We analyze and compare the five most influential users in the Black and LGBTQ+ user networks. Several of their statistics are shown in Table IV. We omit the ID and screen names of the users. We see that the most influential users in the Black user network have more followers and tweet more frequently compared to users in the LGBTQ+ user network. Also, the average sentiments of the top five users in the Black user network are all negative, while the average sentiments of the top five users in the LGBTQ+ user network are all positive.

Next, for each network, we calculate the mean and standard error of user influences, grouped by the average sentiment of their tweets (Figure 7). To limit noise, we omit data points that average less than ten users. For the Black user network, we see that the distribution of influence means is skewed to the right with the peak occurring at -0.25 , indicating that users who tweet with sentiments between -0.2 and -0.3 are the most influential. The LGBTQ+ user network shows nearly the opposite result with the most influential users tweeting with positive sentiments. The difference in peaks of influence suggests that the most popular users in the Black user network gain traction through posting negatively, or that popularity prompts negativity, as opposed to the trend of positivity among influential users in the LGBTQ+ user network. Furthermore, mean influence has greater standard error among users in the LGBTQ+ user network, perhaps indicating that average sentiment among the LGBTQ+ topic group is more heterogeneous. Interestingly, the mean of the average sentiments of the five most influential users for each network is less strong than each network's influence peak averaged over all active users, implying that the most influential users are not as polarizing as their respective networks overall.

VII. CONCLUSION AND FUTURE WORK

In this work, we conduct a multi-faceted quantitative analysis of online sentiment and hate crimes. Our time series analysis reveals notable temporal correlations between negative

User (Black)	Centrality	Followers	# Tweets	Sentiment
User 1	0.298	885	417	-0.223
User 2	0.330	163	405	-0.191
User 3	0.224	1254	371	-0.080
User 4	0.115	17869	104	-0.235
User 5	0.112	10411	142	-0.018
User (LGBTQ+)	Centrality	Followers	# Tweets	Sentiment
User 1	0.244	6747	69	0.018
User 2	0.289	606	40	0.264
User 3	0.206	1424	52	0.127
User 4	0.217	592	28	0.616
User 5	0.199	433	32	0.336

TABLE IV: User statistics for five most influential users in Black and LGBTQ+ user networks.

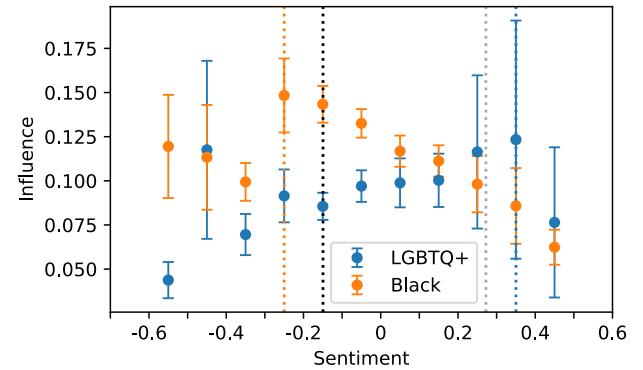


Fig. 7: Influence versus sentiment comparison for the 1195 users in the Black and LGBTQ+ user networks. Each data point marks the mean (and standard error) of influence of users whose tweets have sentiments inside the interval on the x-axis, omitting data points that average less than ten users. Blue and orange dotted lines denote the sentiment interval corresponding to the largest influence in the Black and LGBTQ+ user networks, respectively. Black and grey dotted lines denote the average sentiments of the five most influential users in the Black and LGBTQ+ user networks respectively.

sentiment tweets and physical violence against the Black and LGBTQ+ communities and identifies how they are related to large-scale events. The intensity of online sentiment appears to be related to hate crimes, but causality is difficult to infer.

Our topic model-based knowledge graph (KG) provides a framework for studying the link between online social networks and hate crime. Dynamic community detection reflects differences in the discourse on the KG around different topic groups and reveals that the Black topic group communities exhibit a pattern of emergence and dissolution, reflecting the dynamic appearance and disappearance of new topics. On the other hand, the LGBTQ+ topic group communities tend to split, which suggests that discussions within the LGBTQ+ topic group are prone to split into sub-topics or sub-interests.

Finally, we introduce a measure of ‘influence’ by combining network centrality and follower counts of users. Through

user network analyses, we find that the most influential users in the Black topic group have higher engagement (tweet frequency and follower count) than the most influential users in the LGBTQ+ topic group. They also post significantly more negatively than users in the LGBTQ+ topic group. Our analysis suggests that the nature of tweets driving engagement is different in different groups—negative sentiments gain traction in the Black topic group, whereas positive sentiments gain traction in the LGBTQ+ topic group. One can leverage this finding to proactively monitor online sentiment trends, informing protective measures for vulnerable groups.

Our research is a robust, large-scale, data-driven analysis, and provides a preliminary analysis on trends in online discourse and hate crimes against marginalized communities. This framework can be extended in many possible subsequent directions, including a more comprehensive analysis, subject to availability of appropriate data, on other areas of the country and other groups, and particularly how separate online communities might affect each others' sentiment over time. Another direction of future work could explore lagged correlations (i.e., where correlations between a pair of time-series where one is time-delayed compared to the other) in order to study time-delayed correlations between online discussions on hate crimes. Lastly, one could enrich the nodes and edges of the knowledge graph with additional features, enabling more detailed analyses.

VIII. CODE AND DATA AVAILABILITY

A complete online repository of this project's codes is available from <https://github.com/jimliu01/kg-twitter-hate-crimes>. For confidentiality, the Twitter dataset is not published. However, tweets are publicly available via Twitter (now known as X) and we list all the processing steps in the paper.

REFERENCES

- [1] M. Kaplan, "The rise in Anti-Asian racism and xenophobia in the time of COVID-19," *Journal of Hate Studies*, 2023. [Online]. Available: <https://dx.doi.org/10.33972/jhs.221>
- [2] K. Müller and C. Schwarz, "Fanning the Flames of Hate: Social Media and Hate Crime," *Journal of the European Economic Association*, vol. 19, no. 4, pp. 2131–2167, 10 2020. [Online]. Available: <https://doi.org/10.1093/jeea/jvaa045>
- [3] P. Ramsden, "How the pandemic changed social media and George Floyd's death created a collective conscience," June 2020. [Online]. Available: <http://bit.ly/3L22pHP>
- [4] J. D. Gallacher, M. W. Heerdink, and M. Hewstone, "Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters," *Social Media + Society*, vol. 7, no. 1, 2021. [Online]. Available: <https://doi.org/10.1177/2056305120984445>
- [5] E. W. Bernroider, G. Harindranath, and S. Kamel, "From connective actions in social movements to offline collective actions: An individual level perspective," *Information Technology & People*, vol. 35, no. 8, p. 205–230, 2022.
- [6] K. Müller and C. Schwarz, "From hashtag to hate crime: Twitter and antiminority sentiment," *American Economic Journal: Applied Economics*, vol. 15, no. 3, pp. 270–312, July 2023. [Online]. Available: <https://www.aeaweb.org/articles?id=10.1257/app.20210211>
- [7] M. L. Williams, P. Burnap, A. Javed, H. Liu, and S. Ozalp, "Hate in the machine: Anti-Black and Anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime," *The British Journal of Criminology*, vol. 60, no. 1, pp. 93–117, 07 2019. [Online]. Available: <https://doi.org/10.1093/bjc/azz049>
- [8] Y. Lupu, R. Sear, N. Velásquez, R. Leahy, N. J. Restrepo, B. Goldberg, and N. F. Johnson, "Offline events and online hate," *PLoS one*, vol. 18, no. 1, pp. 1–14, 2023.
- [9] I. Awan and I. Zempi, "'I Will Blow Your Face OFF'—VIRTUAL and Physical World Anti-muslim Hate Crime," *The British Journal of Criminology*, vol. 57, no. 2, pp. 362–380, 12 2015. [Online]. Available: <https://doi.org/10.1093/bjc/azv122>
- [10] Federal Bureau of Investigation, "Crime data explorer," 2023, accessed: 2023-08-07. [Online]. Available: <https://cde.ucr.cjis.gov>
- [11] J. J. Rözer, B. Hofstra, M. E. Brashears, and B. Volker, "Does unemployment lead to isolation? the consequences of unemployment for social networks," *Social Networks*, vol. 63, pp. 100–111, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S037887332030040X>
- [12] M. Wittek, C. Kroneberg, and K. Lämmermann, "Who is fighting with whom? how ethnic origin shapes friendship, dislike, and physical violence relations in German secondary schools," *Social Networks*, vol. 60, pp. 34–47, 2020, social Network Research on Negative Ties and Signed Graphs. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0378873318300212>
- [13] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks*, vol. 32, no. 3, pp. 245–251, 2010. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0378873310000183>
- [14] "What is a knowledge graph?" <https://www.ibm.com/topics/knowledge-graph>, accessed: 2023-06-20.
- [15] C. Adams, M. Bozhidarova, J. Chen, A. Gao, Z. Liu, J. Hunter Prinski, J. Lin, R. Sonthalia, A. L. Bertozzi, and P. Jeffrey Brantingham, "Knowledge graphs of the QAnon Twitter network," in *2022 IEEE International Conference on Big Data (Big Data)*, 2022, pp. 2903–2912.
- [16] Twitter, "Twitter API data dictionary tweet location metadata," 2023, accessed: 2023-09-09. [Online]. Available: <https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/geo>
- [17] T. T. Nguyen, S. Criss, P. Dwivedi, D. Huang, J. Keralis, E. Hsu, L. Phan, L. H. Nguyen, I. Yardi, M. M. Glymour, A. M. Allen, D. H. Chae, G. C. Gee, and Q. C. Nguyen, "Exploring u.s. shifts in Anti-Asian sentiment with the emergence of COVID-19," *International Journal of Environmental Research and Public Health*, vol. 17, no. 19, 2020. [Online]. Available: <https://www.mdpi.com/1660-4601/17/19/7032>
- [18] PFLAG, "LGBTQ+ glossary," 2022, accessed: 2023-08-07. [Online]. Available: <https://pflag.org/glossary/>
- [19] "FBI hate crime statistics." [Online]. Available: <https://www.fbi.gov/how-we-can-help-you/more-fbi-services-and-information/ucr/hate-crime>
- [20] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>
- [21] R. Jain, A. Kumar, A. Nayyar, K. Dewan, R. Garg, S. Raman, and S. Ganguly, "Explaining sentiment analysis results on social media texts through visualization," *Multimedia Tools and Applications*, vol. 82, no. 15, pp. 22613–22629, 2023. [Online]. Available: <https://doi.org/10.1007/s11042-023-14432-y>
- [22] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.
- [23] Department of Justice, "Updated 2021 hate crime statistics," 2023, accessed: 2023-08-07. [Online]. Available: <https://www.justice.gov/crs/highlights/2021-hate-crime-statistics>
- [24] C. Spearman, "Demonstration of formulae for true measurement of correlation," *The American journal of psychology*, pp. 161–169, 1907.
- [25] X. Zou, "A survey on application of knowledge graph," *Journal of Physics: Conference Series*, vol. 1487, no. 1, p. 012016, mar 2020. [Online]. Available: <https://dx.doi.org/10.1088/1742-6596/1487/1/012016>
- [26] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics Theory and Experiment*, vol. 2008, 04 2008.
- [27] D. Ogwok and E. M. Ehlers, "Jaccard index in ensemble image segmentation: An approach," in *Proceedings of the 2022 5th International Conference on Computational Intelligence and Intelligent Systems. CIIS 2022: 2022 The 5th International Conference on Computational Intelligence and Intelligent Systems, Quzhou China, 2022*. [Online]. Available: <https://doi.org/10.1145/3581792.3581794>