Analyzing *#BlackLivesMatter* related Social Media Posts published from Racially diverse Geographic Regions in the United States

1st Doron Reid
Electrical Engineering & Computer Science
Howard University
Washington DC, United States
doron.reid@bison.howard.edu

2nd Esau Hutcherson
Computer Science and Engineering
Texas A& M University
College Station, Texas, United States
esau.hutcherson@tamu.edu

3th Natasha Tonge
Department of Psychology
George Mason University
Fairfax, VA, United States
ntonge@gmu.edu

4rd Paria Rezaei
Electrical Engineering & Computer Science
Howard University
Washington DC, United States
Paria.Rezaei@bison.howard.edu

5th Salvatore Giorgi
Computer and Information Science
University of Pennsylvania
Philadelphia, United States
sal.giorgi@gmail.com

6th Sharath Guntuku
Computer and Information Science
University of Pennsylvania
Philadelphia, United States
sharathg@seas.upenn.edu

7th Anietie Andy
Electrical Engineering & Computer Science
Howard University
Washington, DC, United States
anietie.andy@howard.edu

2025 IEEE International Conference on Big Data (BigData) | 979-8-3315-9447-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIGDATA66926.2025.11402492

***Abstract*—The Black Lives Matter (BLM) movement is focused on protesting police brutality against Black people and communities – mostly in the United States (US). The racial demographic composition of US counties varies. On social media platforms, individuals publish posts about BLM. In this work, we hypothesize that the topics of discussions expressed in BLM related social media posts published from US counties with a higher percentage of African American residents varies from those expressed in BLM related social meida posts published from US counties with fewer African American residents. Using a large language model (LLM), we extract the topics expressed in millions of BLM related social media posts published over an 8-year time period from different US counties. With this dataset, we conduct analyses to determine if there are differences in the topics expressed in social media posts published from US counties with 20% or more African American residents compared to those published from US counties with less than 20% African American residents. Our findings show that the topics on *"police violence"* and *"social issues"* are more associated with the posts published from US counties with 20% or more African Americans, while the topics on *"law enforcement"* and *"politics"* are more associated with the posts published from US counties with less than 20% African Americans. These findings can inform the policies and interactions of relevant organizations and law enforcement agencies – as it relates to BLM, with members of these communities. We discuss the findings from this work and their implications in the discussions section.**

***Index Terms*—Black lives matter, LLM, county.**

Identify applicable funding agency here. If none, delete this.

I. Introduction

The US has more than 3,000 counties, some of which have varying racial demographic compositions. Given these demographic differences, we hypothesize that there are differences in some of the topics discussed around BLM by people who reside in US counties with different racial demographics, specifically US counties with a population of more African Americans compared to US counties with fewer African Americans.

On social media platforms such as X (formerly Twitter), some individuals publish posts about BLM in which they express their views and opinions about the BLM movement [1]. Prior work [1], [4]–[7] analyzing social media posts related to BLM did not study if/how the topics expressed around BLM varied across geographical regions. Prior work [1] analyzed social media data related to COVID-19 and determined that the topics of discussion around COVID-19 vaccines varied across geographical regions in the US; for example, it was found that social media posts published from (a) evangelical hubs tended to focus on topics around operation warp speed and thanking God and (b) Hispanic centers focused on topics related to concerns around food and water.

In this work, we analyze millions of BLM related social media posts published over a period of 8 years on X (formerly Twitter). For each of these posts, we obtain meta information related to the post such as (a) the user id of the individual who published the post, (b) the US county from which the

user indicated to have published the post from, and (c) the timestamp in which the post was published. Using a large language model (LLM), we extract the topics expressed in each of the social media posts in our dataset. With this dataset, we conduct analysis to determine if there are differences in the topics expressed in BLM related social media posts published from US counties with more African American residents compared to US counties with fewer African American residents.

II. Related Works

There has been discourse on the efficacy of Black Lives Matter as a movement for its intended aims and how social media intersects with BLM-related initiatives. Detractors of the BLM movement, especially around the minimal violent protests that arose, point to the fact that violent protests hurt the movement and served to undermine its sociopolitical aims. However, it has been shown that while minimal violence in protests does not reduce prejudice, it does increase support from groups that are not generally favorable towards BLM [6]. Additionally, it has been shown that BLM protests have increased public discourse around racial inequalities for sustained periods of time [7]. Past work has been conducted on the topic of social media utilization of the Black Lives Matter movement. Various works have shown that social media activity spikes around real-world politicized events [1]. BLM supporters on Instagram leveraged the visual focus of the platform to spark Black Lives Matter protests in 2020 [3]. Furthermore, research has shown that user engagement with Black Lives Matter was more concentrated in certain regions, such as Washington D.C. or Minnesota, than in other regions in 2020 [3]. This shows that there are geographic differences in how users engage with Black Lives Matter from a social media standpoint. Other work has showcased the ability, given the advancement of natural language processing, to classify social media posts as distinct emotional categories [4]. Emotions such as anger and disgust were shown to be classified accurately using BERT as a base model to be fine-tuned [4]. Advancements in emotion classification in textual data allow us to gain new insights into how users interact with certain politicized topics. A comprehensive study analyzed how adolescents of different racial and ethnic groups interacted with media types. Particularly, black youth were more likely to turn to view traditional and social media in comparison to other youth groups in regard to BLM [5]. Surveys showcase that social media is an important political vehicle for Black people in the U.S and should be considered when analyzing African American sentiment of various political events. [1].

These prior works serve as an underpinning in our research goal. Our work differ from these prior works in that we are analyzing the BLM related social media posts to determine if and how the topics expressed in posts published from US counties with more African American residents vary from the topics and subtopics expressed in posts published from US counties with less African American residents.

III. Dataset

In this work, we use the dataset from prior work [1] which consists of 63,884,799 X posts published between January 2013 and December 2021 that mentioned the following keywords, *BlackLivesMatter*, *AllLivesMatter*, and *BlueLivesMatter*. Each of these X posts has a large number of metadata associated with it such as user handles, their bios i.e. free text description of a users bio, and places – a physical location a user associates with a given X post. As it relates to the specific metadata, places, which is associated with the X posts, users can manually tag and associate a location with their post. From the dataset in [1], we identified 10,820,854 X posts in which the places meta data was populated with a location; we then identified the US county associated with the location stated in the places metadata associated with each of these X posts. Prior work analyzing social media data have focused on analyzing data by users who publish a lot of posts because this gives better insights [8]–[11]; for example, in [11], data from users with 100 or more social media posts were used for their experiments. Similarly, in this work, we focus on analyzing X posts by users who published 100 or more posts – all of these posts contained the keyword *BlackLivesMatter* or a variation of it e.g. *BLM*.

The focus of this work is to determine if/how the topics of discussions around BLM varies across US counties with different racial demographic compositions, specifically, US counties with a higher percentage of African Americans compared to US counties with fewer African Americans. Using data from the US County Health Ranking [2], we identified the racial composition of each of the US counties in which users in our dataset published posts from. Prior work [9], compared the predictors of COVID-19 deaths and cases between US counties with $>=$13% Black residents compared to US counties with $<$ 13% Black residents. In this work, to determine the optimal percentage of African-American residents in US counties to use for our analyses, we did the following: using a sample of our dataset, we varied the percentage of African Americans in US counties in our dataset starting from 50% down to 15% and observed that US counties with $>=$ 20% African Americans gave the optimal number of US counties with African Americans and the number of posts belonging to each group i.e. US counties with $>=$ 20% or more African Americans vs US counties with $<$ 20% African Americans. Table I shows the number of posts used for the analyses in this work.

IV. Methodology

For each of the posts in our dataset i.e. Table I, using the OpenAI Python API, we used a LLM, specifically GPT4 to extract the topic expressed in each of the posts. We gave GPT4 the following prompt: *"Given the Twitter message*

[1]Brooke Auxier, "Social Media Continue to Be Important Political Outlets for Black Americans," Pew Research Center, December 11, 2020, https://www.pewresearch.org/short-reads/2020/12/11/social-media-continue-to-be-important-political-outlets-for-black-americans/

[2]https://www.countyhealthrankings.org/reports/2022-county-health-rankings-national-findings-report

County Percentage	Posts	Counties
>= 20% African Americans	1,223,932	208
< 20% African Americans	2,045,187	845

TABLE I

Number of posts collected from US Counties with a population of >= 20% African Americans and number of posts collected from US Counties with a population of < 20% African Americans

below, please identify and update the list of Topics found in this tweet. Each of these elements will help in the further segmentation and clustering of the message. Topic is defined as Primary subject of the tweet". We did not tune any of the LLM parameters. We observed that the extracted topics were not very specific in some cases. In this work, we are interested in specific details of the topic, hence, using GPT4, we extracted the subtopics expressed in each of the X posts in our dataset, where subtopic is defined as "the more specific areas of the main topic". Specifically, similar to what we did for extracting the topics, we prompted GPT4 to extract the subtopics expressed in each post: *"Given the Twitter message below, please identify and update the list of Subtopics found in this tweet. Each of these elements will help in the further segmentation and clustering of the message. Subtopic is defined as the more specific areas of the main topic"*. We did not fine tune any of the LLM parameters. Table II shows examples of X posts and the GPT4 extracted subtopics associated with these posts.

To determine that the subtopics identified by GPT4 were accurate, we randomly selected 3,000 X posts from our dataset with their corresponding GPT4 predicted subtopics. Two PhD students familiar with the BLM movement independently reviewed each of the 3,000 X posts and the GPT4 predicted subtopics to determine the accuracy of the predictions. Each Ph.D. student made note/record of if each predicted GPT4 subtopic was accurate or not. Using Cohens Kappa, the inter-annotator reliability was calculated for the ratings of the two PhD students; inter-annotator reliability for subtopics was 0.7273.

V. Analysis:

In this section, we conduct an analysis to determine what subtopics are most associated with X posts published from US counties with 20% or more African Americans compared to counties with less than 20% African Americans. From our dataset, we examined the subtopics representing more than 1% of posts in our dataset and found the following subtopics: *social issues* (14.82%), *black lives matter movement* (3.97%), *politics* (3.90%), *law enforcement* (3.72%), *protest* (3.61%), *social justice* (2.83%), *black lives matter* (2.76%), *unknown* (2.42%), *police brutality* (1.92%), *racism* (1.58%), *racial injustice* (1.52%), *black lives matter protest* (1.37%), *police violence* (1.09%), and *political commentary* (1.01%). Notably, these 14 subtopics had some topical overlap. As a result, we combined topically similar subtopics to arrive at the top 5 subtopics for further analysis by racial demographic. These top 5 subtopics are: *Black Lives Matter Movement*, *Law Enforcement*, *Police Violence*, *Politics*, and *Social Issues*.

We conducted a chi square test. The results of the chi square test showed there is a significant difference in the number of times subtopics were mentioned and whether the associated US county had 20% or more African American residents or less than 20% African American residents ($x^2(4) = 10427, p < .001$) . We conducted posthoc chi-square tests with 10 comparisons of the residuals to investigate which subtopics differed by racial demographic makeup. We used bonferroni correction to adjust alpha to $p < .005$ because of the multiple tests. Posthoc tests with bonferroni correction revealed that each cell contributed significantly to the chi-square test result; however, standardized residuals were largest for the subtopics: *law enforcement*, *social issues*, and *politics* subtopics. The subtopics *BLM*, *police violence*, and *social issues* were significantly more associated with US counties with 20% or more African Americans than would be expected by chance; however, the subtopics on *law enforcement* and *politics* were significantly more associated with US counties with less than 20% African Americans (all the p-values $< .001$). Residuals were largest for the subtopics *law enforcement* and *politics* indicating that discrepancies between demographics were the largest for these two subtopics.

VI. Discussion

The findings from this work indicate that there are statistically significant differences in the subtopics expressed in BLM related social media posts published from US counties with 20% or more African American residents compared to those published from US counties with less than 20% African American residents. In this section, we discuss the findings of the analyses done in this work.

We conducted analyses to determine the specific subtopics that are most associated with posts published in US counties with 20% or more African American residents compared to those published from US counties with less than 20% African American residents. We find that the subtopcis on *BLM*, *police violence*, and *social issues* were significantly more associated with US counties with 20% or more African Americans than would be expected by chance and the subtopics on *law enforcement* and *politics* were significantly more associated with US counties with less than 20% African American residents. These findings can provide insights to relevant organizations and law enforcement agencies on what the concerns are in these US counties, as it relates to BLM, thereby informing their policies and approach to interacting with these communities and members of these communities and resolving these issues.

VII. Limitation

This work has some limitations. Below we list some of these limitations.

- One of the limitations of this work is that X users could manually tag a certain location when they publish their posts, in spite of the actual location of the user. Hence, in

Social media post	Subtopic
”Police arrest two in shooting of Minnesota Black Lives Matter protesters”	Arrest related to Black Lives Matter protesters shooting
”#blacklivesmatter is NOT a hate group! My life matters and if you don’t agree with #blacklivesmatter then you don’t think my life does. Join the movement because America has a serious problem ! #stephonclark #michaelbrown #TrayvonMartin”	Misconception of the Movement, Police Brutality
”Masked militants ransacking the Justice Center in downtown Portland. Many came prepared with chemicals to start fires and weapons to break windows. #antifa #BlackLivesMatter”	Ransacking of the Justice Center in Portland
”Author Addresses Social Justice and Black Trauma in New Book https://t.co/35FqHMOn91 blackwomen #blackwomenlead #BlackGirlMagic #blackgirlsrock #SocialJustice #trauma #blm #BlackLivesMatter #blacktwitter #BlackExcellence #blackauthor #blackauthors #PoliceBrutality #black”	Social Justice, Black Trauma”
”@Amecarnelian @PlayersTribune @Layshiac @WNBA Rioting and violence is to be met with violence everytime i could give a shit less about BlackLivesMatter and its racist movement. If you think you can take on the rest of america and police go for it, it’ll end badly haha”	Rioting and Violence
”We need mental health workers on the police force right now. We need mental health services available to everyone right now. nnAnd we need to end systemic racism, right now.nnNone of this will happen right now but we need to work towards it every single day.nnBlackLivesMatter”	Mental Health Services, Police Force, Systemic Racism

TABLE II
EXAMPLE OF BLM RELATED SOCIAL MEDIA POSTS AND THE SUBTOPICS ASSOCIATED WITH THESE POSTS

some cases, the location data may not be a true indication of where the X post was published from.

- The dataset used in the analyses in this work is from users who publish 100 or more posts, hence, the findings from this work may not apply if we analyze posts by users with fewer published posts.

VIII. CONCLUSION

In conclusion, in this work, we conduct analysis to show that there are differences in the subtopics expressed in X posts published from US counties with 20% or more African American residents compared to to posts published from US counties with less than 20% African American residents.

IX. ACKNOWLEDGMENT

This material is based on work supported by the Department of Defense/Department of Air Force, Air Force Office of Scientific Research under the Department of Defense/Department of Air Force, Air Force Office of Scientific Research Contract No. FA9550-23-D-0001. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of Defense/Department of Air Force, Air Force Office of Scientific Research.

REFERENCES

[1] Giorgi, Salvatore and Guntuku, Sharath Chandra and Himelein-Wachowiak, McKenzie and Kwarteng, Amy and Hwang, Sy and Rahman, Muhammad and Curtis, Brenda, “Twitter corpus of the# blacklivesmatter movement and counter protests: 2013 to 2021,” Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 1228–1235, 2022.

[2] Tong, Xin and Li, Yixuan and Li, Jiayi and Bei, Rongqi and Zhang, Luyao, What are People Talking about in# backlivesmatter and# stopasianhate? Exploring and Categorizing Twitter Topics Emerged in Online Social Movements through the Latent Dirichlet Allocation Model, Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp.723–738.

[3] Chang, Ho-Chun Herbert and Richardson, Allissa and Ferrara, Emilio, “# justiceforgeorgefloyd: how instagram facilitated the 2020 black lives matter protests,” PLoS one, vol. 17, 2022, pp. e0277864.

[4] Field, Anjalie and Park, Chan Young and Theophilo, Antonio and Watson-Daniels, Jamelle and Tsvetkov, Yulia, “An analysis of emotions and the prominence of positivity in# BlackLivesMatter tweets,” Proceedings of the National Academy of Sciences, 2022.

[5] Arielle Baskin-Sommers and Cortney Simmons and May Conley and Shou-An Chang and Suzanne Estrada and Meghan Collins and William Pelham and Emil Beckford and Haley Mitchell-Adams and Nia Berrian and Susan F. Tapert and Dylan G. Gee and B. J. Casey, “Adolescent civic engagement: Lessons from Black Lives Matter,” Proceedings of the National Academy of Sciences, vol. 118 pp. e2109860118 2021.

[6] Shuman, Eric and Hasan-Aslih, Siwar and van Zomeren, Martijn and Saguy, Tamar and Halperin, Eran, “Protest movements involving limited violence can sometimes be effective: Evidence from the 2020 BlackLivesMatter protests,” Proceedings of the National Academy of Sciences, vol. 119, pp. e2118990119, 2022 .

[7] Dunivin, Zackary Okun and Yan, Harry Yaojun and Ince, Jelani and Rojas, Fabio, ”Black Lives Matter protests shift public discourse”, Proceedings of the National Academy of Sciences, vol. 119 pp. e2117320119 2022 .

[8] Guntuku, Sharath Chandra and Buttenheim, Alison M and Sherman, Garrick and Merchant, Raina M, ” Twitter discourse reveals geographical and temporal variation in concerns about COVID-19 vaccines in the United States”, Vaccine, vol. 39 pp. 4034–4038 2021 .

[9] Millett, Gregorio A and Jones, Austin T and Benkeser, David and Baral, Stefan and Mercer, Laina and Beyrer, Chris and Honermann, Brian and Lankiewicz, Elise and Mena, Leandro and Crowley, Jeffrey S and others, ” Assessing differential impacts of COVID-19 on black communities”, Annals of epidemiology, vol. 47 pp. 37–44 2020 .

[10] Andy, Anietie U., Sharath C. Guntuku, Srinath Adusumalli, David A. Asch, Peter W. Groeneveld, Lyle H. Ungar, and Raina M. Merchant, ” Predicting cardiovascular risk using social media data: performance evaluation of machine-learning models.,JMIR cardio 5, no. 1 (2021): e24473.

[11] Yates, Andrew, Arman Cohan, and Nazli Goharian, ” Depression and Self-Harm Risk Assessment in Online Forums”, EMNLP 2017.