#BlackLivesMatter Movement and Consequences of Racism: A Data and Sentiment Analysis on Tweets in the USA

by

Amir Hossein Zolfaghari

A thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science (MSc) in Computational Sciences

The Faculty of Graduate Studies

Laurentian University

Sudbury, Ontario, Canada

© Amir Hossein Zolfaghari, 2021

THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE
Laurentian Université/Université Laurentienne
Faculty of Graduate Studies/Faculté des études supérieures

Title of Thesis Titre de la thèse	#BlackLivesMatter Movement and Consequences of Racism: A Data and Sentiment Analysis on Tweets in the USA
Name of Candidate Nom du candidat	Zolfaghari, Amir Hossein
Degree Diplôme	Master of Science
Department/Program Département/Programme	Computational Sciences
Date of Defence Date de la soutenance	March 17, 2021

APPROVED/APPROUVÉ

Thesis Examiners/Examinateurs de thèse:

Dr. Kalprum Passi
(Supervisor/Directeur(trice) de thèse)

Dr. Ratvinder Grewal
(Committee member/Membre du comité)

Dr. Parveen Nangia
(Committee member/Membre du comité)

Approved for the Faculty of Graduate Studies
Approuvé pour la Faculté des études supérieures
Tammy Eger, PhD
Vice-President, Research
Vice-rectrice à la recherche

Dr. Shervin Assari
(External Examiner/Examinateur externe)

ACCESSIBILITY CLAUSE AND PERMISSION TO USE

I, **Amir Hossein Zolfaghari**, hereby grant to Laurentian University and/or its agents the non-exclusive license to archive and make accessible my thesis, dissertation, or project report in whole or in part in all forms of media, now or for the duration of my copyright ownership. I retain all other ownership rights to the copyright of the thesis, dissertation or project report. I also reserve the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that this copy is being made available in this form by the authority of the copyright owner solely for the purpose of private study and research and may not be copied or reproduced except as permitted by the copyright laws without written authority from the copyright owner.

Abstract

Introduction: A movement arose in the middle of a challenging pandemic time. In a year that everybody keeps their six feet distance and mask on, many came to the streets or started publishing social media contents asking for Black rights. It was after an injustice killing of a black man - George Floyd - by a police officer that #BlackLivesMatter trended again as the top conversation in the world. Hence, it became our question that how racism - specifically on social media - is associated with blacks' mattering lives.

Methods: We carried out an ecological retrospective study on Twitter data for the year 2020, which had location tags inside the USA. We created inclusion criteria to shape our dataset based on that and categorizing tweets into separate groups. Our groups were (1) "BLM" for those supporting the "BlackLivesMatter" movement; (2) "Anti-BLM" containing tweets in opposition to the first group; (3) "Ambiguous" who had both previous group contents; and (4) the "Racists" comprising those who included offensive n-words in their tweets. We employed some statistical data by utilizing previous research for the "Life Expectancy", "Poverty Rates," "Educational Attainment," and "Race Compositions" factors of the black and white population in the USA by the states. We employed additional techniques to identify genders and classify records in reference to their states. Moreover, we applied the sentiment analysis using Python. We calculated the final rates considering each group's statistics compared to the sum of all tweets published in each state. The analysis of the final rates in correlation with employed tables was done by IBM SPSS Statistics 26.

Results: We found 43,830,301 tweets with location data inside the USA in this time frame, and 306,925 of them applied for our study. A noticeable initial observation was the sharp increase of the #BlackLivesMatter after George Floyd's demise on May 25, 2020, while this hashtag has a history back to 2013. There is a positive correlation between the rates of offensive-content tweets and the life expectancy of Black males. The same tweets showed an association that wherever racism is higher, more are suffering poverty. This is rather surprising that the BlackLivesMatter movement supporters were mostly among those with the bachelor or advance degree educational attainments. By contrast, if a state had lower rates of high school degrees, more racists tweets exist there. The rates of aggressive tweets are higher in areas with more black populations and are weaker in states having white people's domination. Regarding the sentiment analysis, the majority of tweets are written in objective forms, and it had a slight increase after the mentioned event. The polarities were also mostly in a neutral way. The most negative sense belonged to the BLM supporters, with the rate of 46% before and 33% after the event.

Conclusion: This project was undertaken to evaluate the relationship between rates of cyber-racism and anti-racism posts to some real-world indicators. We considered our inclusion criteria in reference to the cruel killing of a black man – George Floyd - by police to investigate the published tweets classified by supporters and opponents of this story, in addition to those using offensive language towards black people. This study showed a strong correlation between these concepts while contents on the world wide web could impute the day-to-day life conversation. Hence, it shows how a drop in racist behaviors can lead to a world with higher life expectancy, wealth, and education.

Reduction of color discrimination on social media and particularly toward blacks could help to have a healthier community. Contrarily, the rise of these bigoted contents results in disastrous consequences on these racialized populations.

Keywords: BlackLivesMatter, Racism, Twitter, Tweet, Sentiment, Analysis, Life Expectancy, Poverty, Educational Attainment, Composition Rates, USA, Location, States, n-words.

Acknowledgements

I would like to show my sincere gratitude to my supervisor, **Professor Kalpdrum Passi**, who is a perfect inspiration for me. I never forget his kind support in the hard days of doing my thesis, especially the early pandemic days when I had challenging hard times of my life. Thank you very much, Dr. Passi, for being patient, a good listener, and a great advisor to me.

I would like to pay my special regards to **Dr. Serge Demers**, who helped me with his kind decisions and a great understanding of the issues that a new international student face.

I am indebted to **Dr. David Lesbarrères** the same way while I found him an honorable man.

I wish to thank all my colleagues whose guides were a milestone in completing this thesis, especially Mahsa Roozrokh, and Jaydeep Motisariya.

I also acknowledge the support of Mohammad Safavi, Emmanuel Ngoma, Armaan Sandhu Singh, Zahra Banaiyan, and other kind friends in Canada.

Finally, yet importantly, I would like to whole-heartedly thank my loving **Father** and **Mother**, my dear **sister**, and my caring **brother** for their endless love, continuous encouragement, and exciting talks.

Table of Contents

Thesis Defense Committee II
Abstract III
Acknowledgements VI
List of Figures X
List of Tables XI
Chapter 1 ..... Introduction 1
1.1 Background: Hashtag activism 1
1.1.1 History: The Case of #BlackLivesMatter 2
1.2 Why Study on Racism 3
1.2.1 Human Nature 3
1.2.2 What are the consequences of racism? 3
1.2.3 Types of Racism 4
1.2.3.1 Within Individual Racism (Internalized) 4
1.2.3.2 Between Individual Racism (Discrimination) 4
1.2.3.3 Institutional Racism 4
1.2.3.4 Structural Racism 5
1.3 Why Twitter 5
1.3.1 Racial content on Twitter 5
1.3.2 Twitter and Health 5
1.4 Objectives 7
Chapter 2 ..... Literature Review 8
2.1 Overview 8
2.2 The Role of Racism 8
2. 3 Discussion: The Role of Twitter? 13
2.4 Racism on Social Media 17
2.5 Why study “Black lives matter” in academia 19
2.6 Twitter mining for health data 20
2.7 Why Genes do not Interfere? 24
2.8 Racism Consequences 26
2.9 Sentiment Analysis 28

Chapter 3 ..... Data Extraction & Preprocessing 30
3.1 Overview 30
3.2 About Twitter 30
3.3 Twitter Data Characteristics 31
3.4 Dataset and Variables 33
3.5 Inclusion Criteria 37
3.6 About Python 41
3.7 Data Preprocessing 43
3.8 Data Preprocessing Methodology 43
3.9 Processing Time 46
3.10 Study Limitations 46
Chapter 4 ..... Analysis of Data, Sentiment, and Statistics 48
4.1 Overview 48
4.2 Study Design 48
4.2.1 Ecological Study 48
4.2.2 Statistical Analysis by Correlation Coefficients 50
4.3 Natural Language Processing Toolkit (NLTK) 52
4.3.1 Word Tokenization 52
4.3.2 Stop Word Removal 53
4.3.3 Word Lemmatization and Stemming 55
4.4 Sentiment Analysis 56
4.5 Statistical Analysis 57
4.5.1 Spearman's correlation 58
4.5.2 Life Expectancy 60
4.5.3 Poverty Rate 61
4.5.4 Educational Attainment 62
4.5.5 Race Composition 62
4.5.6 Emojis 62
Chapter 5 ..... Results and Discussion 65
5.1 Overview 65

5.2 Tweets Review 65
5.3 States' Statistics 70
5.4 Principal Analysis 79
5.4.1 Correlation with Life Expectancy 80
5.4.2 Correlation with Poverty Rate 83
5.4.3 Correlation with Educational Attainment 86
5.4.4 Correlation with Race Composition 87
5.5 Analysis Based on the Sentiment Outcomes 88
5.5.1 Sentiment Details for Each Group 88
5.5.2 Sentiment Details Regarding the Event 96
5.5.3 Sentiment Compared to Tables 100
5.5.4 Sentiment Polarity Tweet Examples 103
5.6 Supplementary Results 104
5.6.1 Word Cloud 104
5.6.2 Emojis 106
5.7 Discussion: Summary of results 107
5.8 Discussion: Why we studied "Black Lives Matter" with Respect to George Floyd 109
5.9 Discussion: Why black males are more impacted? 110
5.10 Discussion: The Imputation of Social Media and Real-World 111
5.11 A talk to policy makers 112
Chapter 6 ..... Conclusions and Future Work 114
6.1 Conclusions 114
6.2 Future Works 116
References 118
Appendix 1 Tweet Examples 137
Appendix 2 Tweet statistics table 139
Appendix 3 Life expectation table 141
Appendix 4 Word Clouds 143
Appendix 5 Emojis 145

List of Figures

Figure 3.1 The characteristics of a sample Twitter data. 33
Figure 3.2 An Example of one record inside a file in a JSON viewer software. 37
Figure 3.3 The code snippet for defining dictionary variable types of words to use as the study inclusion criteria. 38
Figure 3.4 The code snippet related to the function for grouping the tweets based on the study inclusion criteria. 40
Figure 3.5 The top popular programming languages in reliance on their percentage of users that searched for instructions in Google. 42
Figure 3.6 Code snippets for the dictionaries used as stopwords 43
Figure 4.1 Various types of mathematical correlation regarding the degree level two variable are associated. 51
Figure 4.2 An instance of a monotonic function. 59
Figure 5.1. Increase in the number of tweets supporting Black Lives Matter movement in 2020 67
Figure 5.2. Comparing the trends of each group in 2020 69
Figure 5.3. Comparing the rates of BLM tweets in each state on the map of the USA. 73
Figure 5.4. Comparing the rates of Anti-BLM tweets in each state on the map of the USA. 74
Figure 5.5. The rates of Racist tweets in each state on the map of the USA. 75
Figure 5.6 Comparing sentiment polarity percentages for each group. 90
Figure 5.7 Sentiment subjectivity percentages for each group. 91
Figure 5.8 Scatter plot of sentiment polarity and subjectivity scores for BLM tweets. 92
Figure 5.9 Scatter plot of sentiment polarity and subjectivity scores for Anti-BLM tweets. 93
Figure 5.10 Scatter plot of sentiment polarity and subjectivity scores for ambiguous tweets. 94
Figure 5.11 Scatter plot of sentiment polarity and subjectivity scores for racist tweets. 95
Figure 5.12 Scatter plot of sentiment polarity and subjectivity scores comparing male and female for BLM tweets. 96
Figure 5.13 Scatter plot of sentiment polarity and subjectivity scores comparing male and female for racist tweets. 96
Figure 5.14 Comparing the sentiment polarity percentages before and after the event for each group. 98
Figure 5.15 Comparing sentiment subjectivity percentages for BLM group regarding the before and after the event. 99
Figure 5.16 Comparing sentiment subjectivity percentages for Racist group regarding the before and after the event. 100
Figure 5.17 Word Cloud of BLM group tweets. 105

List of Tables

Table 3.1 Comparison and the pros and cons of Twitter data providers. 32
Table 3.2 Variable list of a tweet inside a JSON file. 35
Table 3.3 Selected variable list of a tweet and their descriptions. 36
Table 3.4 List of new variables created in this study for each tweet in parallel to their descriptions and how that variable was formulated. 45
Table 4.1 An example of a tokenized tweet. 53
Table 4.2 An example of word lemmatization and stemming. 56
Table 5.1 Number of tweets in each group detailed by month 66
Table 5.2 Number of tweets in each group detailed by gender 67
Table 5.3. Number of tweets in each category regarding the area. 71
Table 5.4. Number of tweets in two main groups regarding the area and the event of George Floyd' demise. 76
Table 5.5. Number of tweets that have gender data in three main groups regarding the area. 78
Table 5.6 Spearman's rho statistical analysis for percentage of racist tweets compared to life expectations of "White people minus Black" regarding gender. 81
Table 5.7 Spearman's rho statistical analysis for rate of BLM and Anti-BLM tweets regarding gender, compare to "life expectations," "Poverty Rates," "Educational Attainments," and "Race Composition" statistics. 82
Table 5.8 Spearman's rho statistical analysis for percentage of Racist tweets compare to "Poverty Rates," "Educational Attainments," and "Percentage of Race Composition" in each state. 84
Table 5.9 Spearman's rho statistical analysis for the percentage of racist tweets compared to life expectations tables regarding polarity score. 101
Table 5.10 Spearman's rho statistical analysis for the percentage of Racist tweets compare to "Poverty Rates," "Educational Attainments," and "Percentage of Race Composition," considering the polarity scores of sentiment analysis. 102
Table 5.11 Examples of tweet categorization in Sentiment Polarity process. 103
Table 0.1 Sum of all tweets that had locations regarding the month published. 139
Table 0.2 Life expectation table of white and black people in addition to the genderLife expectation table of white and black people in addition to the genderTable 0.3 141

Chapter 1

Introduction

In what circumstances "Social Movements" are going to arise, that is a research question making scientists busy for a long time. The rapid development of social media networks has led scholars to inquire about this question from another perspective and ask about social media's impact on social movements. Social movements were able to reconstruct societies on different levels, and now specialists are looking at it as a medium that forms collective action and agendas of social movements. Its fascinating feature is its ubiquity, low-cost, and availability for all citizens to join and let the world hear their voices. Hence, platforms like Twitter, Facebook, YouTube, and numerous online blogs have arguably given a voice to individuals who otherwise could not declare their opinions [1].

1.1 Background: Hashtag activism

A hashtag is a keyword used to categorize the contents published on social media and is prefaced by the symbol, #. It is also a linking term for all the contents in that category on microblogging services such as Twitter. These hashtags help to archive content on Twitter and make it easy to see how trends enhance [2].

This symbol now embraces an outstanding character in popular culture and is commonly linked to most social media activities. It is used as a common phenomenon and for various purposes, while Zappavigna [3] defines it as a labeling system for the subjects discussed by a micro-post to construct the metadata and combine it with the original text. Also, numerous advertising

campaigns used hashtags as their core commercial message, while there are many examples like "Dutch Reformed Church" in 2014, #OscarPistorius as a worldwide search phrase on Google in 2014 or #Bahrianused as a trend of Arab spring in the same year [4].

We can trace it back to 2007 when Chris Messina suggested it, regarding the hashtag history. Initially, he was the first person who recommended that this symbol should also be used on Twitter. Also, some believed that the first hashtag on Twitter, which was used for this concept, was #barcamp - for organizing sort of technology conferences. Later, the idea was refined and it was shown how this approach could appear on Twitter. At that time, the term 'channels' was used and further, Stowe Boyd introduced it as “hashtags” [5].

1.1.1 History: The Case of #BlackLivesMatter

The story gets back to an evening of February 2012, when Trayvon Martin [6] decided to tour around, and a neighbor who was suspiciously watching him called the police. His only crime was walking black, and despite suggested by 911 to stay away, the neighbor confronted him, which ended shooting this innocent teenage black dead. Soon, the point that he was unarmed put him in the center of the spotlight and conversation. However, the story turned once in the summer of 2013, that killer was found unguilty for Martin's death - a failure in the justice system to consider the avoidable loss of a black fellow. That made Alicia Garza, Patrisse Cullors, and Opal Tometi construct #BlackLivesMatter and popularize it on social media - a call to fight for Black people [7].

The same story recurred again and again, such as the killing of Michael Brown [8] by a police officer in Fergusen, Missouri, in August 2014, or a more recent one George Floyd [9] who was

killed on May 25, 2020. In contrast, it shows how some part of society needs a change toward its beliefs about the black community.

Now, #BlackLivesMatter is the flag of those demanding a shift in the world against racism. It is a figure in social movements for many local branches fighting against racial injustice.

1.2 Why Study on Racism

1.2.1 Human Nature

As humans, we need to express our opinions, feelings of pride and pain, and intellectual powers. That is how we join a community of people to have a sense of value and influence. In this day and age, this demand emerges in a microblogging service that allows users to declare their views with short messages that roughly resemble their ideas.

1.2.2 What are the consequences of racism?

There is an increasing number of researches that show inequalities leading to health problems. The situation is the same even in the well-developed states of Europe, and the resolution of socioeconomic health disparities is a significant cause of public health disappointments. [10].

Researchers identified a term for this purpose: "Social Gradient [11]" which defines a condition when individuals face less socioeconomic situations and worsen their health than those who are more advantaged. Eventually, it shortens their lives and poses them with severe health risks. It is currently among the main robust research outcomes in public health, as there are manifestations confirming that when there are health inequalities enduring during life and across societies, it leads to health problems [12].

1.2.3 Types of Racism

A variety of topics influence proper care, and health results from racial disparities, and a system is required to address these inequalities in the social environment [13]. Here we are going to describe different types of racism [14].

1.2.3.1 Within Individual Racism (Internalized)

Internalized or also known as individual racism, involves personal demonstrations of racism that reside inside a person. According to Lawrence et al. [14], some instances are prejudice, xenophobia, internalized injustice and privilege, and feelings regarding race led by the dominant society.

1.2.3.2 Between Individual Racism (Discrimination)

Discrimination is when internalized racism is taken into action or as stated by the "Canadian Human Rights Commission" [15] it is a practice or a resolution that maltreats an individual or a society for purposes such as race.

1.2.3.3 Institutional Racism

If the racism gets into practice inside an organization or by using a specific inter-agency power like police, it is called Institutional Racism. Randall [16] defines it as any activity, based on race, intentional or unintentional, that influences someone or society according to skin color or race is racism triggered by a person or the institution itself. The killing of George Floyd [17] by a policeman placing legs on his neck for close to nine minutes was an event some people in the USA considered it as an example of this category. However, in this study, we believe that the assessment of this should be judged by authorities.

1.2.3.4 Structural Racism

Structural racism though, is the normalization of all these definitions above [14]. It is a collective of discrimination efforts based on race in various systems such as education, job, incomes, healthcare, and criminal justice that distribute resources unequally and foster adverse health consequences [18].

1.3 Why Twitter

Twitter provides a free, fast, and federal messaging service that allows worldwide to share their 280 characters note in parallel with other 335 million active users [19]. This is the path that population uses to express themselves that sometimes can show the undercover beliefs.

1.3.1 Racial content on Twitter

Criss et al. [20] studied the effects of racial content on Twitter with their participants and proved the adverse effects of it on mental, emotional, and physical health. They stated that even some Twitter users with racial interests tend to cooperate to create an "echo chamber" aiming to reinforce racial content. A striking result is the feelings of disappointment, rage, and sorrow in other users, which compromise their psychological health, comprising depressive indications, tensions, and nervousness. This view is also supported by Ferrara et al. [21] that points out content published on social media has emotional influences on the audiences. They observed a group of random Twitter users and found that emotional contagion aroused both negative and positive sentiments on other users.

1.3.2 Twitter and Health

Many studies [22–27] have shown that social networking sites are powerful platforms to research on health either as a medium for health care interventions, users' behavior studies, or health content analysis. Twitter is the favorite platform of at least 22% of Americans [28]. Regarding American adults, 61% of them are using the virtual world as a source of health information [29],[30] whereas Twitter is the most common used social media in this specific topic [31].

Joseph-Shehu et al. [26] explored the effects of the information and communication technologies tools such as social networks, websites, text messages, and cell-phone apps on health-promoting lifestyle behavior and reported its positive impact. They determined the essential impression of these technologies on preventing and managing diseases, healthy BMI, and its rule to promote physical and psychic health.

Scientists found Twitter beneficial in medicine because it enables medical professionals to reach a broad audience, such as physicians, trainees, or patients [32,33]. Anyone with an account can access broad terms of information, reply, ask questions, add to favourite, or even retweet them to share them with friends. These features can address two advantages, direct medical professionals to the content with the most reaction and guide users to reliable sources. Still more remarkable, we have "hashtags" indicating the tweet conversation's topic category, which encourages other users to continue discussing that disease or question type. The potential of Twitter to share and advance biomedical research also made several medical experts to employ it for its features that allow them to promote their research, new treatments, or clinical difficulties that require further study, panel discussion, or investigations [31].

1.4 Objectives

The main objective of this thesis is finding the relation between racial content published on Twitter and some real-life factors linked to the life of black people. The applicable objective of this study is to research the relationship between the tweets related to racism and their association with Blacks who are highly exposed and impacted by this content.

The goals can be written specifically in this way:

1. To extract tweets with racial content published publicly on Twitter in the year 2020
2. To identify the category and some specifications of these tweets
3. To show the correlation between the state-level percentage of racial tweets published in this term and the life expectancy tables of Black people posted by USA national health centers.
4. To show the correlation between the percentage of tweets in each state and three tables of "Poverty Rate," "Educational Attainment," and "Race Composition".

The primary research questions are:

1. What is the rate of racist tweets on Twitter?
2. What is the relation between racism tweets and the mentioned real-world factors?
3. What are the associations between tweet rates and these factors?

Chapter 2

Literature Review

2.1 Overview

An outline of similar studies related to the primary topic of this thesis will be presented. Studies that analyzed racism's role in human health and how it impacted communities will be covered first. Next, the scholars who employed Twitter data in their papers will be discussed. Finally, studies that include sentiment analysis on tweets will be discussed.

2.2 The Role of Racism

In the U.S., Black people experience racism, and this racial discrimination can have many side effects on gestation and childbirth. Although this discrimination is much less than in the past, it cannot be said that it has completely disappeared. This discrimination has led to black families being isolated and often living further away from society. This isolation can have many effects on pregnant women. They have less access to parks, supermarkets, and other public areas, and a lack of healthy communication with the society around them leads to feelings of humiliation and loneliness. The results of Chae et al. (2018) show that this discrimination in pregnant women leads to preterm birth, and their babies are more likely to be underweight at birth [34].

Williams and Mohammed, [13] in 2009, attempted to see the evidence and research gaps about discrimination and health. They found that racism can affect people's health negatively facing discrimination. They found that discrimination and health have an opposite association after they

reviewed the available articles on racism and health indexed in PubMed between 2005 and 2007. This pattern is currently appearing in a broader range of contexts and for a broader group of results. Finding the effects of perceived discrimination on health will need more work since discrimination relates to racism's health-relevant aspects. Discrimination should be measured extensively and precisely; also, its stressful impacts need to be assessed, and the mechanisms that link between discrimination and health need to be identified.

Racial health disparities are widespread; as referring to the Kung et al. study (2008) [35], Black people in the US are suffering more death rates among principal causes of death compared to the whites. Considering the fifteen foremost reasons that take the human life, such as cardiovascular conditions, cancer, stroke, diabetes, renal failure, hypertension, liver cirrhosis, and homicide, the rates are disappointingly high. African Americans and American Indians have higher death rates by age as compared to whites. . Likewise, in 2001, Levine et al. [36] found that almost 100,000 black people die prematurely each year because of racial health disparities. Although life expectancy for blacks and whites has increased, the gap between the races in life expectancy is high, about 5.1 years in 2005, according to the National Center for Health Statistics, 2007. Similarly, infant death rate has declined over time for blacks and whites, but the racial gap increased since 1950 [37]. Williams believes that the results of researches about health issues show that the disparities are even getting worse. As reported by NCHS [37], statistics show that a clear example is heart disease and cancer, which are the two significant causes of death in the United States. Evidence indicates that black and white people had different death rates in these circumstances in 1950; however, now, African Americans have even more death rates than whites.

Hicken and colleagues (2018) [38] performed a similar series of experiments in 2017 on cultural and structural racism resulting in health inequalities and attain with some suggestions. Regarding findings, they have observed that stress is an active response to anxiety on challenging racism and discrimination and causes dysfunction in the body. They also argued that structural racism addresses how society was developed to value and promote white well-being and somehow differs from another term - cultural racism - that preserves structural racism within specific interconnected cultural processes. They combined the conceptual and experimental evidence, which indicates structural racism as the realization of cultural racism. They also applied some methods involving race theory, as they believed these items are significant and often ignored in racial health inequalities speech.

LaVeist et al. (2011) [39] demonstrated that disparities in health due to race and ethnicity could give rise to direct and indirect health care costs, while productivity loss is an example of an indirect one. The authors carry out three analyses to prove their ideas: direct medical costs; indirect ones; and costs of premature death. To evaluate the second, they employed the Medical Expenditure Panel Survey (2002–2006) data to show that by reducing the health disparities among racial or enthnic minorities would result in cost savings. They assessed these indirect costs to evaluate productivity loss, challenging racial inequalities. Moreover, they assessed premature death costs by deriving data from the National Vital Statistics Reports (2003–2006) for their analysis. If health disparities for minorities were removed, direct medical care costs would reduce by about $230 billion, and indirect costs related to illness and premature death would reduce by more than $1 trillion between 2003 and 2006. This enormous number of premature deaths imposes a substantial loss of human resources and productivity that might otherwise have provided to civilization's advancement.

Experiencing racism can lead to illness in individuals. People who are discriminated against and excluded from society are more prone to cardiovascular reactions. In general, people who encounter racism are more stressed, and this constant stress can lead to diseases such as heart problems, cancer, and stroke and increase their mortality rate. Repeated experiences of racism may result in a heightened pro-inflammatory, which in turn leads to cardiovascular disease and other diseases of the immune system. Research has also shown that factors related to racism can accelerate the aging process in individuals [40].

The living environment and the degree of separation from society can impact pregnant mothers and newborns. Among the most important of these effects are preterm delivery and low birth weight. The isolation experienced by blacks has many adverse effects on pregnant women and newborn infants. This isolation and living in unconventional environments cause many problems during pregnancy and childbirth. In an experiment conducted between 1998 and 2002, the results showed that children born under these conditions had a worse general condition in comparison to average. Interestingly, white mothers living in isolation also experience some of these crises, but not as much as black mothers. In addition to social isolation, factors such as discrimination and humiliation, and lack of self-esteem also seem influential [41].

Atopic dermatitis (AD) is a skin disease that is common among children and is more common in African American children than in whites. The point in this distinction is that the factors that cause this skin condition are often prevalent among African American children. Factors such as low family incomes, having less educated caregivers, and exposure to secondhand smoke can be very effective in determining the severity of the disease. These factors are much more common among African American families than whites, making the disease more common for them. Also, African American families live more in isolation and away from society, and this isolation and

distance from society are among the fundamental reasons for developing atopic dermatitis. Overall, socioeconomic status, race, and physical environment effectively determine the severity of the disease, and these factors are far more significant among African American families than whites [42].

Racial discrimination is an important and influential social factor that has a vast impact on children and adolescents' health. The social environment of children and adolescents has a significant impact on their growth and formation and helps manage their problems and prevent them. According to the World Health Organization, public health factors are defined as the forms of the environment in which people are born, grow, live, and work in. These elements are directly related to financial, political, and cultural conditions. Racism leads to chronic stress, and this chronic stress affects all of these factors (political, social, and economic). As a result, it can, directly and indirectly, affect children and adolescents' growth and development and shape it abnormally [43].

Greevy et al. (2004) [44] had a linguistics-based plan to detect racism content on the web and categorize them into predefined categories. This method was a way to replace manual labeling of the offensive material with keywords. Their goal was to enhance productivity and save time by using new techniques. They developed an automated system to classify racist texts and evaluate them using a support vector machine (SVM). Regarding their analysis, they used various approaches. They believed that identifying texts as racism is challenging, so they employed three types of interpretations for this term: bag-of-words, bigrams, and part-of-speech tags. The output precision and recall of the performed shows that these classifiers can be well-effective for practice, yet it is a computationally expensive solution. Although stopping racist content on the web by automated algorithms can lead to a clean and utopian environment for all, it could also

open some doors for censorship and stopping free speech by dictators in specific countries. Literature can sometimes be tricky even by punctuation, so it is not easy to group text by algorithms disregarding this attention. They could also give some examples of the content categorized to enable readers to see how these various solutions were conducted.

2. 3 Discussion: The Role of Twitter?

Twitter can somehow be a sample representative of the social conversation outside the virtual world. It is not a body of evidence that can be imputed to the all of the community, yet it can reflect people's thoughts on specific subjects, such as racism. Considering this, we can infer that social media talks, the studies on it, and the outcome statistics can all be exemplary instance of what is happening in the real world.

Twitter is one of the most widely used media, which is an outstanding tool for sharing ideas and exchanging opinions due to its structure. But unfortunately, many conversations and exchanges are not in the right direction, and many racists share their hateful ideas in this space.

Chaudhry [45] expanded this discussion on his paper entitled "Not So Black and White" and addressed that users widely tweeted the hashtag of Ferguson after the murder of "Mike Brown" [46] in Ferguson area, and along with this hashtag, we witnessed many racist and hate speech. At that time, many used the hashtag "blacklivesdontmatter," and no one seemed to blame the killer. Although most of them condemned the murder, they did not pay much attention to it due to the fact that the victim was black. Many racists were afraid to see a white man being punished because of a black man, and it was unacceptable to them, even if that white man was a murderer.

In 2015, Chaudhry [47] examined three separate projects that analyzed racist speech on Twitter and discussed their findings, in addition to why Twitter is an essential source of data gathering for studying racism. He believes this novel technology of social media opens up new horizons for obtaining these types of studies. Scientists can mutate Twitter from a social media app into an immeasurable platform for research studies. In contrast, some others could also transform this playground into an environment to distribute the hateful speech that they cannot usually express out of the internet world. Hence, Twitter is a tool that empowers us to publish wisdom by discovering new findings employing such datasets and directing the readers through what is happening in the shadows.

In this regard, the first study belongs to Chaudhry (2015) [48], who followed the frequency of racism content on Twitter to compare them with the population rates of Indigenous people across Canada and reported that users tend to publish racist content like the way they are talking about it in the real world.

He used API methods to collect the rates of specific racist words published on Twitter regarding some places in Canada for three consecutive months in 2013. These places were Vancouver, Toronto, Montreal, and Winnipeg. The reason for choosing these places was that they are considered civilized cities in Canada and even experienced the most reports of racial harassment in 2010.

He aimed to observe how Canadian racists behave on Twitter and what words they use. He accessed Twitter by searching for keywords or geographical location, using location tags to access tweets for geo-research. This will limit the study to solely about one percent of all tweets that use location tags, yet it is the only available solution. Merely a tiny percentage of users allow their location to be identified, even though location tags seem to be increasing. He also

used another method and searched a series of racist words on Twitter to see how many times they were used. These racist words included: native, white trash, nigger, paki, and chink, since they were common racist words frequently heard among racist groups. The word ‘native”, which is a nonracial term (used generally in English), was unexpectedly used negatively on Twitter.

He reported the findings of how these racist tweets were used as 50% “real-time” response, negative stereotypes by 28%, 12% for the informal use of the slur, and the minority of the rest for reacting to racism and abuse objective. By real-time response, he means that the users tweeted about something that happened to them at the current moment.

Another experiment was conducted in 2014 [49] to explore the racist words used on Twitter and how often they were repeated. In this experiment, the API method was used, and racist words were assessed for nine consecutive days in November. The list of words selected for the search was extracted from Wikipedia. These words were: “white boy,” “paki”, “whitey,” “pikey,” “nigga”, “spic”, “crow,” “squinty,” and “wigga.” They also double-checked to make sure the words are correctly considered racist content and not something like part of a person’s username.

Moreover, they used a manual method for software validation and in-depth analysis, plus this that the non-English tweets were eliminated and not reviewed. Their total tweet numbers were 126,975, averaging 14,100 tweets per day. Their findings showed that racist users used “white boy” as the most word with 48.9%; subsequently “paki” by 11.7%, following “whitey” by 7.9%, and afterward the rest of the words for less than 4 percent each.

Hate Map Geography [50] is another project that used the geographical codes of tweets to review racism. According to the researchers, there was so much controversy about bullying and hate

speech on social media that it prompted them to look at how social media altered its obligation and became a tool for spreading hate and how certain words can be used to denigrate a group.

In total, researchers found more than 150,000 tweets with geographical codes, including hateful words, while the number carries both racist and homophobic tweets. This research showed that the most used words were: “nigger,” “gook,’ “chink,” “wetback”, and “spick.” The findings manifest that some specific places had more negative tweets, and some words were commonly used in particular places such as “wetback”, around Texas. The exciting method they followed in their research was to show the categorized collected information since they ordered it to show the use of words on Twitter by region. They could also finalize their information to create a heat map out of these hate words.

Twitter seems to have been very efficient and useful for researchers, physicians, and scientists. Nature Medicine [51] had featured a story of a practitioner who assumes that through Twitter, she was able to meet her colleagues and those who work in a similar field. She even continued that through this, she has received generous job offers, while Twitter acts as inspiring content for her and improves her career. Dr. Choo [51] did not have many followers before, but after tweeting about her racism experience in 2017, she quickly gained more followers. She adds that in addition to the job aides she received on Twitter, it also helped her talk about annoying experiences and get support from others, while the world without those supports could be a hardly tolerated one.

Not all scientists are seen and heard as well as Choo on Twitter, but most have said that Twitter has been beneficial for their business. In fact, many researchers use Twitter to conduct their research and studies, especially in sociology. One example would be racism, while without Twitter space, it was not certain whether researchers could study this phenomenon and its extent.

Before Twitter, researchers had limitations on gaining a statistical population and seeing people's behavior and speech in everyday life [51].

One of the potent natural emotions of human beings that directly affects life satisfaction is happiness. That is why Mitchell et al. [52] studied the influence of words used on Twitter among adults in the US in correlation with happiness. They found that happiness in urban US areas has a strong negative relationship with rising poverty. Happiness can have many manifestations and causes, and research has shown that interpersonal and intercultural factors can also affect it. In the United States, for example, happiness is correlated with wealth, so that with increasing income, happiness advances, and with increasing poverty, happiness diminishes. However, it is clear that this cannot be the only reason for happiness; income has been seen to increase over time among the United States people, while happiness decreases in the long run. Besides, happiness has been found to be associated with education level and obesity, especially among people who have little self-control.

2.4 Racism on Social Media

We call these platforms "Social media," while media is responsible for any content published through them in its traditional form. If, by any chance, a journal prints a false statement and damages someone's reputation, they can sue the publisher. However, these platforms are not responsible for the content, and they sentence the user itself in this purpose.

Racism, anyhow, is a long-standing term that mutates into new forms by the emergence of social media. Early work on race and the Internet pointed to inequality in access levels and algorithmic visibility as essential factors in Internet inequality. Accordingly, Fernández [53] called it

'platformed racism' and researched the Internet's impact on racial identities' rapid increase. He believes that from another point of view, the Internet is both a chance to show racial identities, a place to exchange ideas, and a tool for social and anti-social uses.

He further refers to the hashtag movement #sosblackaustralia – which was coined by indigenous activists in 2015 to end discrimination against certain groups - and addresses that this organized a space on Twitter and Facebook to restore blacks' rights in Australia.

On the other hand, on Twitter, we also see hate speech and harassment, such as racial and gender-based abuse. Platforms like Twitter can lead to racist dynamism through their rules and algorithms. On Facebook, a community is formed based on pages and becomes more prevalent when many people like it. Some groups enable users to sell and advertise their goods to others. Once employers of a group were found to exclude African Americans and racial minorities from the list, this has led to the violation of labor laws, which do not allow discrimination based on race and gender [53].

Overall, while we expect these apps to be a place to initiate conversations for people all over the world and connect them together, it has become a place to settle racism.

In another paper "Thank You, Black Twitter," [54] Marc Lamont Hill stated some of the critical roles of Twitter in black movements and examined how this platform influenced the spread and visibility of these campaigns. His first case is the "BlackLivesMatter," hashtag, which he believes Twitter helped the story spread quickly after its emergence in 2014 when a police officer killed a young black man. The appearance of this hashtag promoted the tale and its consequences for followers. Succeeding is the hashtag "SayHerName" [55] which was created in 2015 after Sandra Bland [56] committed suicide. Police stopped her, arrested her, and later, after

three days, she hanged herself in her cell jail. Following the news of the suicide, the hashtag "SayHerName" was launched on Twitter. This hashtag makes many people aware of the incident all over the world, and protests took place.

2.5 Why study “Black lives matter” in academia

Racism in science can be very influential. In fact, it is more difficult for black researchers to reach higher levels. Black researchers and scientists face more hardship and discrimination than their white counterparts, and that can hold them back and prevent them from progressing.

On the other hand, this discrimination and racism make it difficult for these groups of people to progress. Still, the general belief is that they do not reach higher levels like whites because of their less talent and perseverance. Also, discrimination against a large segment of society is detrimental to science. In the absence of such discrimination, and in a situation where blacks would have an equal chance, more researchers would study phenomena [57].

Additionally, in the past, scholars attempted to discover ways to study people’s behavior in real situations and sometimes find it hard to achieve. However, social media removed this restriction, and as an instance, Twitter is a prominent place to see how people behave and what they think. At the same time, users do not know that their tweets are being observed in scientific studies, which could be another positive point for researchers to analyze actual conversations, reactions, and ideas.

With reference to the “BlackLivesMatter” hashtag, we have the possibility to study all aspects of black life and any discrimination they may endure by this. This hashtag activism was created in 2012 after the murder of Trayvon Martin [58] and later became a movement to fight injustice.

Today, this hashtag is not just about the murder and the police treatment of blacks, but its scope has expanded to include black subsistence quality features [59]. That is why we need to consider it in science and have more research to show new highlights of how we can improve these racialized groups' lives that certainly matter a lot.

2.6 Twitter mining for health data

Traditional databases have been used to study public health by many researchers for a long time, but now social media data like Twitter is taking its place to search on public health. It is what Sinnenberg et al. [60] highlighted in their study in 2017, which they applied for a systematic review by a broad examination of scientific search engines such as PubMed, Embase, Web of Science, Google Scholar, and CINAHL. They used peer-reviewed original papers that employed Twitter for their research on health topics. They determined a novel classification in order to define Twitter use in health research consisting of six categories: content analyzing; monitoring; engaging with tweets, recruiting, intervening, and network analyzing. They determined that many of these literatures did not report data elements in Twitter that are recognizable from a user's profile, particularly demographics. At the same time, some organizations funded Twitter-based health research because of its high significance knowledge. Such data can be used to identify the users whose data are used in these researches.

Edo-Osagie et al. (2020) [61] are on the same page with previous mentioned Sinnenberg et al. (2017) study [60], and assume that Twitter comes to pass former health research methods these days. Up to now, researchers in the field of public health were using traditional methods or conventional pathological knowledge banks to study the community's health. . However, these

days, the ubiquitous and accessible social media data, especially Twitter, has attracted more researchers' attention to use it for public health analysis purposes. Scientists are trying to dig deep on social media to access public opinions as they believe this novice invention had an extensive impact on people's lives. A case in point is Twitter, which is a platform broadly used by users worldwide. These data can help extract users' opinions and situations. They reviewed and combined the kinds of literature on public health that engaged by Twitter data. They used a scoping review methodology by searching four significant databases in the areas of health, computer science, and cross-disciplinary in 2020. They reviewed ninety-two articles meeting their criteria and classified outcomes into six areas that Twitter can be applied to as a platform for public health research. Their classifications include observation, identification of events, Pharmacovigilance, prediction, tracing the diseases, and determination of geography. They obtained a genuine picture on this topic and primarily found how these different domains were changed over time regarding the reputation. Additionally, they studied the various approaches to studying diseases and conditions that would lead to better ways of understanding them. Besides, they also end up with algorithms and techniques commonly used for each domain.

Al-Rawi et al. (2020) [62] tried a fancy subject on Emojis as his study and how their use altered during the COVID-19 pandemic. In this study, they tried to examine how people from various genders use Twitter emojis to discuss COVID-19. The online discussion about this outbreak had several aspects, so examining the complex ways online users express themselves is valuable. Emojis are the best way to convey ideas and feelings on social media like Twitter. They developed a mixed-method and categorized data into three sets, males, females, and gender minorities; then, they analyzed the use of emojis for each group. There were five topics in more than 50 million tweets addressing the hashtags #Covid-19 and #Covid19 during the two months:

disease anxieties, health affairs, financial issues, appreciation toward healthcare staff, and specific emojis for unique genders. The outcome of their study shows that most emojis are positive in different genders, but emojis used for conversations about sexual minorities showed an immense rate of negativity. In fact, this research indicates how COVID-19 impacted different genders and provides a dependable source of information about health crises in society.

McClellan et al. (2017) [25] performed a study to find the peak points that people are more interested in discussing mental health topics on Twitter. They inquired about the value of employing social media to follow up the discussions on mental health, while they aimed to develop an experimental model to recognize these periods. They employed a potentially valuable source of data about depression and suicide, 176 million tweets from a broad population between 2011 to 2014. They used the Autoregressive Integrated Moving Average (ARIMA) method for their analysis. Their findings show that this ARIMA model is valid for recognizing times of increased activity on Twitter that are associated with behavioral health. They showed that tweets associated with despair content mostly come next to unanticipated campaigns and even last longer than what is expected. Followed by this, awareness campaigns promote conversation between the community, health specialists, and fellows enduring mental health issues, especially on Twitter. Finally, they conclude that increasing tweet volume after a behavioral health event usually lasts less than two days.

According to the research of Leypunskiy et al. (2018) [63], changes in human physiology during the day affect behavior because of the body's daily routines, environmental cycles, and social timetables. Many scholars have researched the mechanism and function of circadian rhythms in steady conditions or environments where light and dark duration is ideal. However, they believe fewer studies are done on stress and pressures in society, such as the routine plans at work, or

school, or the effect of daily and seasonal human activity rhythms. Therefore Leypunskiy et al. decided to apply research about this issue; they tried to analyze people's Twitter activity in more than 1,500 US counties in 2012–2013 during 15-min intervals. They used tweets tagged geographically, representing about 0.1% of all the population each day. They understood that the regular periods of low Twitter activity are connected with adequate sleep measured by traditional surveys. They represented that the nighttime calm in Twitter activity is moved to later times on weekends, which is corresponding to weekdays; this phenomenon is called the jet lag on social media or Twitter. They concluded that regarding Twitter, the social messaging times could elaborate on users' regular behavioral patterns. Hence, Twitter might be an option to analyze sleep disorders in a large population by studying the relation between screen time and sleep.

Mohammadi, Thelwall, Kwasny, and Holmes in 2018 [64] searched 1,912 users' tweets related to journal articles to identify their Twitter practices regarding scholar works. They learned that almost 45% of the respondents were non-academic users, although some tweets were biased on academic topics. Mostly, Twitter was employed by users who had a background in social science or humanities. Users like to broaden their social ties on Twitter and search for information instead of searching for applicable tweets. Academics used Twitter to spread information about current events and expand their connections with other people. Users' tendency for using Twitter for scholarly purposes was to share academic information in different disciplines. Mohammadi et al. showed shreds of evidence to prove that Twitter has an essential role in finding scholarly information and spreading knowledge. Consequently, many users who are not in academia also trust the claims of those who use Tweeter as a proof for the outcomes of their research and being scholarly. They added that students use a wide range of social networks for their professional activities, and Twitter is among the most used ones. The reason that made the scientific

community interested in Twitter is its extended scope of information dissemination and promotion of research. It becomes a tool to facilitate these knowledge exchanges, rapidly establishing a trust relationship for users to consider it a reliable source of academic and non-academic impact.

2.7 Why Genes do not Interfere?

A study [65] in the United States found that white men lived five years and white women lived three years longer than African Americans on average. Even though most of these deaths were due to heart disease, these disorders were directly related to racial discrimination, giving blacks generally less life expectancy than whites. Extensive research has been done on the relationship between disease-related mortality and genes to investigate the role of genes in these deaths [66].

These researchers claim that: (1) Some say that it is more about finances than biology. As an instance, BiDil is a first-generation drug that affects genes and has been used for decades. In 2005, a company used the drug in a different way: They used the drug to treat heart defects only in a specific race of African Americans. The results showed that this drug was effective in treating heart disease in these people. However, on the other hand, Kahn (the author who wrote the book: Race in a Bottle) [67] stated that this medicine was effective because these people were humans, not just African Americans. (2) On the other hand, there are some diseases that occur in specific populations; Like Tay-Sachs[68], which is a genetic disorder more common in people of Ashkenazi Jewish descent. In such cases, researchers believe that the focus should be more on ancestral factors rather than race. Researchers state that when we address ancestral factors, we are actually referring to genes, which are a better explanation of diseases and not racial factors [66].

Gravlee [69], in his paper, tried to answer essential questions concerning racial discrimination. It is so important to have these debates because, first, racial discrimination affects physical and mental illness. Second, the debate over racism and health enhances our knowledge of science and leads to a greater understanding of race and human differences. Third, research into the relationship between racial factors and health makes us aware of the lack of information on racial factors in the fields of anthropology and other social sciences. Hence, it is highly worth considering and having researches on this topic.

Social researchers often observe racial factors as a cultural construct and not a biological reality. If racial factors are not biologically relevant, the question arises as to why there are significant differences in biological phenomena between particular racial groups. There are two main concepts in relation to race and biology. First, the sociocultural reality of race and racism has biological consequences for specific racial groups. As a result, biology can provide substantial evidence for racism and the influence of racial factors. Second, the epidemiological evidence for racial discrimination in our health suggests that race is a biological factor. All this has led to this cycle that social inequalities shaped racialized groups' biology, and embodied inequalities perpetuate a racialized view of human biology [69].

There is undeniable evidence of health inequalities among particular racial groups in societies in terms of epidemiological evidence for race and health. Epidemiological evidence in the United States shows undeniable racial discrimination in mortality and disease in biological systems. In fact, in the United States, the death rate among African Americans is much higher than among whites. The rates showed at least 30 percent more deaths among African Americans than whites in diabetes, cardiovascular disease, and kidney diseases [69].

While these difficulties are more prevalent among African Americans, there were also differences in infant mortality and life expectancy. Infant mortality was 2.4 percent higher among African Americans.

The race is not the human genetic variation since the old critique of race has emphasized three issues: (1) Most human genetic variations are clinal, meaning that there are clear genetic boundaries between communities. (2) Most human genetic variations are incompatible, meaning that the traits we use to differentiate between races may have no value in predicting other biological aspects. (3) Human genetic diversity is widely shared by our species [69].

Generally, racial criticism does not mean denying humans' biological diversity, nor does it claim that genes do not affect health. It wants to say that genes are not the only influencing factor; racial factors are even more influential.

The complement to the statement that "race is not biology" is the slogan that repeats "race is a cultural construct" while more research is needed in this area. Many research and science fields consider racial factors to be myths and do not believe in their influence. These fields believe that the main factors affecting humans and our health are specifically our genes. The amount of research done on the impact of racial factors is low and has not been very popular among researchers. By conducting scientific studies, valid experiments in this field and obtaining scientific results, we can clarify this myth. The reason for this work's importance is that this misconception harms a large number of people and does not recognize their suffering [69].

2.8 Racism Consequences

Racism impacts negatively upon poor health outcomes in the United States, particularly concerning low life expectancy [34]. Recently, a police brutality that leads to the demise of George Floyd - a black man living in Minneapolis - emerges to massive protests against Racism both on streets and social media [9]. Once again, the demand for "Black Lives Matter" got into the first topic of public conversation, which brings us to compare the rates of tweets published at this time with the states' life expectancy table. The objective was to see how significant these two are correlated.

On the other side of the story, we have Racism, which has adverse health outcomes for its victims, specifically Black people, in this case [34]. Williams et al.'s findings showed that discrimination in social conditions impacts variations in health [70]. Also, Lu et al. [71] believe that discrepancy in early life experiences gives rise to cumulative stress and eventually leads to poor birth outcomes. Compare the significant numbers of more inferior birth outcomes for Black infants with White infants in the United States [71]. The situation is even worse for men since our findings broadly support other studies' [72–77], which reported that black men face health consequences out of discrimination even when they reach commonly known success. However, due to Racism's complex nature, it is hard to measure it by traditional survey approaches.

This is precisely why we tend to analyze this question on Twitter data. Therefore, the observed correlation between "percentage of racist tweets" and "black life expectancy" might be explained in this way. We compared Twitter records in the first six months of 2020 with the "National Center for Health Statistics (NCHS)" [65] tables, which have records associated with Black people in 40 states. These findings are rather disappointing that in this modern world, still, some presume they can insult, judge, or irritate others based on skin color.

2.9 Sentiment Analysis

Zunic et al. (2020) [78], in their review about sentiment analysis, established a study to see the current condition of SA (sentiment analysis) linked to health and well-being. They performed a systematic review of recent literature and appended that sentiment analysis (SA) is a kind of natural language processing that aims to organize the text's sentiment. Although it has many applications across different societal contexts like marketing, economy, and politics, there are some difficulties. They believed that it is not clear why the results are not accurate, maybe because of the intrinsic difference between domain and their sublanguages, the size of datasets, the lack of domain lexica, or algorithms' choice. Therefore, they ended more studies required for this issue.

In various research on Twitter across various disciplines, data have been organized according to their goals. They measured tweets with nonexclusive categories like content analysis, sentiment analysis, event detection, user studies, prediction, and GIS analysis [79]. There are some content analysis methods like sentiment analysis in terms of texts, which studies the text of tweets as a basis for finding themes. Sentiment analysis is a kind of content analysis that uses psychological evidence about the meaning of words or symbols to show the text's emotional tone [80].

There are different tools for sentiment analysis. Although each one has its specific process, all of them try to quantify the text's affective dimensions. A standard method is a lexicon-based approach that uses a dictionary of words with affective meaning. Linguistic Inquiry and Word Count [81] could be used with Twitter data to score each tweet about positivity or negativity. Another lexicon-based tool that uses a particular word list for twitter to count slang, abbreviation and other symbols in Twitter format is the Affective Norms for English Words list [82]. Most of

these tools are small and cannot analyze a long text. Besides, sometimes we need more nuanced dimensions other than positivity and negativity [80]

Chapter 3

Data Extraction & Preprocessing

3.1 Overview

In the current chapter, an overview of how the data is extracted from Twitter and preprocessed to be ready for analysis is presented. The pros and cons of using the Twitter API platform, in addition to the alternative approaches is covered. Hence, how to extract retrospective data, in-depth discussion on Twitter as a social network, the tweet attributes, inclusion criteria, what can be derived from Twitter, Python, making data ready, and the steps involved in this are discussed.

3.2 About Twitter

Since 2006, the foundation year of Twitter, social media popularity has surged and identified itself as a platform for broadcasting the densely packed messages rapidly to the world while academies also sailed into its various horizons for various practical goals [83].

Besides, many use Twitter to make their academic research more apparent and understandable to the broader area. Announcing their newly published papers gives them this opportunity to have more exposure and gain popularity, and finally ending in more citations. "It turned out to be an important place to have conversations that had no equivalent elsewhere in the media," Professor Martha S. Jones says as a historian at Johns Hopkins University [84].

Furthermore, Twitter mentions are also an essential substitute metric, or altmetric, to strictly increase the non-scholarly focus on a paper. Simultaneously, a new term calling "Twitter impact

factor" follows the popular h-index factor that a researcher aims to gain as the impact of his publications. Even an author [85] claimed that scientists should spend more time on Twitter if they want a higher h-index. Another professor at Johns Hopkins - Jessica Johnson - who is a digital humanist analyzing how social media creates ideas concludes that: "In the beginning, I did not think that I was going to be on Twitter much, but there is exposure to new ideas and people doing interesting work that I may never have seen otherwise, and that really is amazing" [84].

3.3 Twitter Data Characteristics

Twitter originated as a text-based microblogging service limited to 140 characters. It now allows up to 280 characters, which is still short and brief but enables users to express their ideas in a compact format [86], [87], [88]. Twitter had 152 million daily active users worldwide by the end of 2019 [89]. The US Library of Congress began archiving tweets in 2011 and by January 2013 they had 133 terabytes of data for around 170 billion tweets [90].

We can reach Twitter data employing various techniques, where the differences could mainly impact the study time, samples, and budget.

First, a "Twitter Search" is openly available to all, and anybody can reach public tweets on the website regarding the search criteria, while it takes much time to shape a dataset this way. Next, the Twitter application programming interface (API) provided by Twitter [91] through a developer portal is also free yet has limitations on the free mode and provides just a small Twitter data sample. Following are the "Service Providers" that usually resell unlimited data streams with customizable data retrieval, storage, and analysis, yet are too costly. Sometimes, there are also existing Twitter datasets available that is a solution to overcome Twitter's public

API restrictions for historical records. However, it does not always meet the same criterion of our research requirements as it was prepared with different considerations. Finally, there is an archived resource on "Archive.org" [92] with a random sample of Twitter data for each hour and day of the month. Table 0.1 Comparison and the pros and cons of Twitter data providers. depicts the comparison and the pros and cons of each service [86], [93]. Also, refer to Figure 0.1 to see the main elements of a tweet data.

Table 0.1 Comparison and the pros and cons of Twitter data providers.

Data sources Characteristics	Twitter Search	Twitter API	Data resellers	Datasets	Archieve.org
Population	1500 tweets of the past week	1% of the population	100% of the population	Varies	1% of the population
Historical data	No	No	Varies by vendor	Maybe	Yes
Data retrieval mode	Manual	Semiautomated	Automated	Automated	Semiautomated
Data retrieval time	Long time	It depends on the user equipment	Varies by vendor	Fast	It depends on the user equipment
Data is cleaned	No	No	Maybe	Maybe	No
Customizable criteria	No	Yes	Yes	No	Yes
Data storage type	User Retrieval	User Retrieval	Varies by vendor	Dataset creator	JSON Files
Cost	Free	Free	Costly	Varies by provider	Free

Figure 0.1 The characteristics of a sample Twitter data.

(1) is the user's first and last name;(2) users' unique identifier; (3) date and time of the tweet posted; (4) tweets' contents; (5) tweet's hashtags

3.4 Dataset and Variables

We used the Twitter archiver service in order to be able to study the most recent 2020 tweets and a time before and after the death of George Floyd[94]. According to Twitter, the archive.org is a 1% sample from the Twitter stream API daily tweets and is the largest publicly available Twitter

dataset on the Internet. It contains multiple JSON files of tweets in a tar file for each day ordered by date for each month of the year.

We started getting TAR archived files for each month between January 1, 2020, to June 30, 2020, which was the only available data for 2020. Data are packed into ".bz2" and then ".tar" files to maximize the compression power. By decompressing these files, we have 24 files, named 00 to 23, indicating the hours for each day that these records are sampled for, and for each hour, we have 60 JSON files containing the tweets. Most tweet providers encode data using JavaScript Object Notation (JSON) [95] format. JSON files are text-based records constructed with key-value pairs, named attributes, and associated values. JSON uses these attributes and their values to represent objects. It is around 43,000 JSON files for each month, recording more than 180 million tweets that should be unzipped before starting the extraction process.

As shown in Figure 0.1, there are some basic variables for each tweet. However, there are more attributes for each tweet, which unfortunately is not complete for all records. Variables are divided into three categories, as shown in Table 0.2 or further in Figure 0.2 as an example of one record. The first one contains the basic elements and characteristics of a tweet, the second refers to the users' info, and the third is additional metadata like appended hashtags. We have 26, 39, and 4 different variables for each group accordingly, and as mentioned, not all of them are proper ones to study.

Table 0.2 Variable list of a tweet inside a JSON file.

	JSON		User		Entities
1	contributors	1	contributors_enabled	1	hashtags
2	coordinates	2	created_at	2	symbols
3	display_text_range	3	default_profile	3	urls
4	entities	4	default_profile_image	4	user_mentions
5	extended_tweet	5	description		
6	favorited	6	favourites_count		
7	favorite_count	7	followers_count		
8	filter_level	8	following		
9	geo	9	follow_request_sent		
10	in_reply_to_screen_name	10	friends_count		
11	in_reply_to_status_id	11	geo_enabled		
12	in_reply_to_status_id_str	12	id		
13	in_reply_to_user_id	13	id_str		
14	in_reply_to_user_id_str	14	is_translator		
15	is_quote_status	15	lang		
16	lang	16	listed_count		
17	place	17	location		
18	quote_count	18	name		
19	reply_count	19	notifications		
20	retweeted	20	profile_background_color		
21	retweet_count	21	profile_background_image_url		
22	source	22	profile_background_image_url_https		
23	text	23	profile_background_tile		
24	timestamp_ms	24	profile_banner_url		
25	truncated	25	profile_image_url		
26	user	26	profile_image_url_https		
		27	profile_link_color		
		28	profile_sidebar_border_color		
		29	profile_sidebar_fill_color		
		30	profile_text_color		
		31	profile_use_background_image		
		32	protected		
		33	screen_name		
		34	statuses_count		
		35	time_zone		
		36	translator_type		
		37	url		
		38	utc_offset		
		39	verified		

However, as mentioned not all of these data fits our study as long as not all of these variables have data at all. Hence, we end up with the variables listed in Table 0.3.

Table 0.3 Selected variable list of a tweet and their descriptions.

#	***Variable Name***	***Data Type***	***Description***
1	Tweet Id	Number	64-bit unsigned integers unique for each tweet as an identifier
2	Created At	Datetime	UTC date of the tweet posting time
3	Text	String	Main text
4	Hashtags	String List	Lists of hashtags appended by the user to the end of the tweet
5	User ID	Number	Identifier number for the user's profile
6	Name	String	The name of the user
7	Screen Name	String	The unique name of the account which is also used to refer to the user's profile
8	Location	String	A text indicating the location of the user's profile
9	GeoCoordinates	Two Numbers	Latitude and longitude of the user's location
10	Language	String	Tweet's language
11	Follower Count	Number	Number of followers of the user who posted that tweet

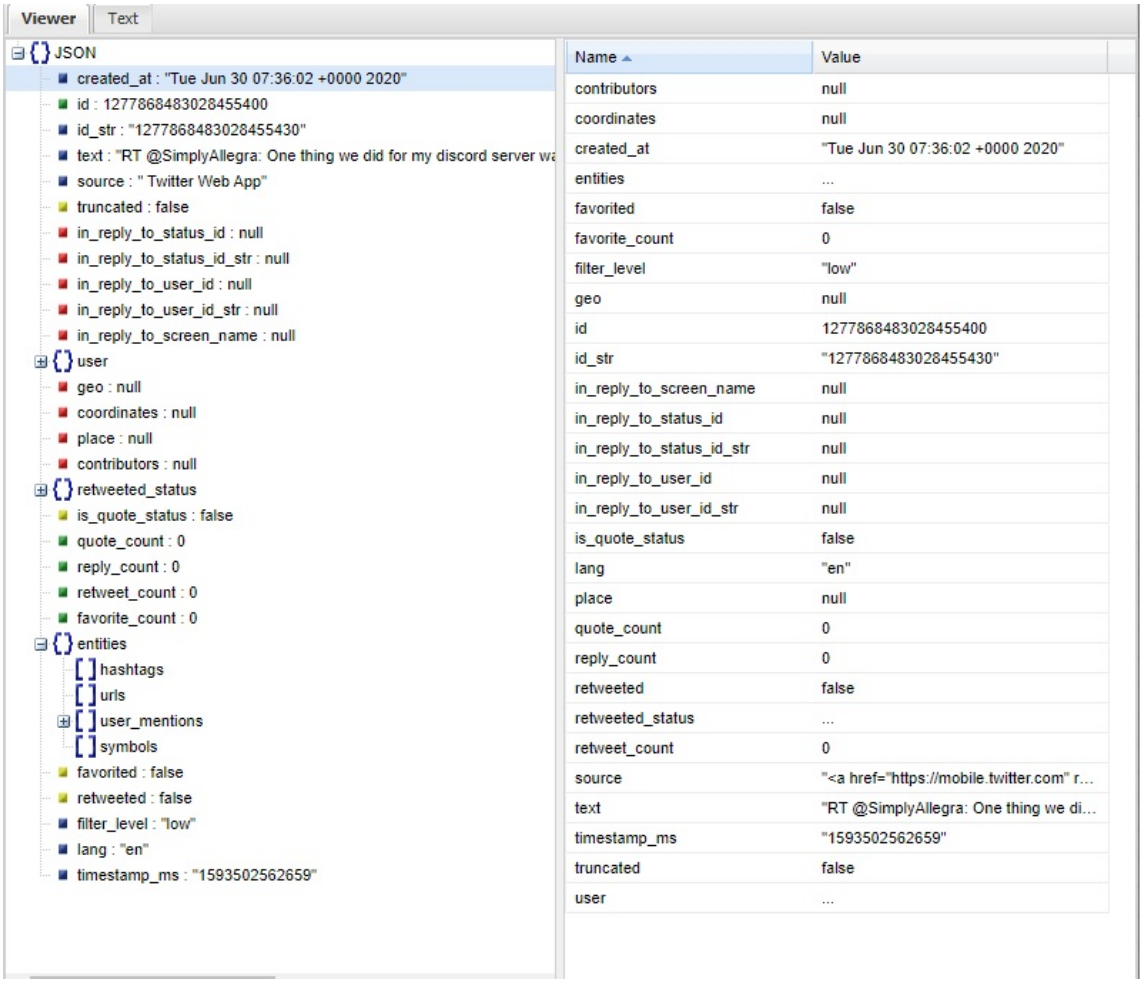

Figure 0.2 An Example of one record inside a file in a JSON viewer software.

3.5 Inclusion Criteria

Twitter studies originate by establishing a sampling rule using search terms or hashtags designated as "inclusion criteria" to retrieve associated data. This inclusion criterion has a high impact on the study sample size, and that is why it is one of the first considerations for each study. Researchers need to have careful search terms to avoid under-or overestimating data size that obscures the knowledge patterns they are trying to achieve [86].

Our inclusion criteria are about some specific words in text tweet or hashtags variable considering that we studied tweets dated in 2020 and having geo data located in the United States. We fetched the records based on three main groups "BlackLivesMatter (BLM) Movement," "Anti-BLM Movement," "Racist," and an extra one that is the ambiguous group, while each of these had its own search terms.

We discussed the story of the first group - BlackLivesMatter Movement - in chapter 1. We have seen that this movement was started some years ago; yet again, the killing of George Floyd makes it a top concern, and so many people used this hashtag to protest on social media. We get help from a top expert to shape our categorization rule, which is finally a set of word dictionaries for each group text or hashtag. Figure 0.3 shows the associated codes to create those dictionaries inside a python list variable for more manageable and transparent coding instead of hardcoding them inside the related function.

```
#BlackLivesMatter Group
antiracismHashtag = ['blacklivesmatter', 'blm', 'black_lives_matter']
antiracismDictionary = ['blacklivesmatter', 'black_lives_matter', 'black lives matter']

#Anti-BlackLivesMatter Group
antiBLMHashtag = ['alllivesmatter', 'all_lives_matter', 'alm', 'whitelivesmatter', 'white_lives_matter','wlm', 'bluelivesmatter', 'blue_lives_matter']
antiBLMDictionary = ['alllivesmatter', 'all lives matter', 'whitelivesmatter', 'white lives matter', 'bluelivesmatter', 'blue lives matter']

#Racist Group
racistDictionary = [' nigger ', ' niger ', ' nig ', ' nigor ', ' nigra ', ' nigre ', ' nigar ', ' niggur ', ' nigga ', ' niggah ', ' niggar ', ' niggress ', ' nigette ' , ' niglet ', ' negro ', ' niggas ', ' negros ', ' niggers ']
```

Figure 0.3 The code snippet for defining dictionary variable types of words to use as the study inclusion criteria.

First and foremost, we have our "BlackLivesMatter" or "**BLM**" in short, group, which are those who stand in line with this movement by their hashtags. In the BlackLivesMatter movement's face, some people were at odds with it and started reacting by opposite hashtags. We consider them the second group of the study named "**Anti-BLM**". Some minorities discussed this event by a tweet employing hashtags from both groups, which we considered a separate "**Ambiguous**" group.

Finally, we have the "**Racism**" group covering those who used a series of old, ugly, offensive words known as n-words in their tweets [96]. Use of these n-words can carry significant psychological effects, provoke severe harm on the audience; that is how Dr. Arthur K. Spears, professor at The City University of New York and specializes in Black languages of the Western Hemisphere, describes this word [97]. However, regarding the writing notation or pronunciation, there are multiple forms of these n-words that we referred to the dictionary enlisting [98] to shape our criteria and incorporate them in the study. Figure 0.4 shows the function developed to get the tweet based on the text and hashtag as input parameters and a grouping number as a return value for each tweet.

```
1.  def hashtagChecker(hashtagList,twText):
2.
3.      hashtagLower = []
4.      lenHashtag = len(hashtagList)
5.
6.      twTextLower = twText.lower()
7.
8.      if lenHashtag > 0:
9.          for h in range(0, lenHashtag):
10.                     hashAppend = hashtagList[h].lower()
11.                     hashtagLower.append(hashAppend)
12.
13.
14.     if 'blacklivesmatter' in  hashtagLower and 'alllivesmatter' in hashtagLower:
15.         whichCategory = 2
16.
17.     elif any(antiraceWord in hashtagLower for antiraceWord in antiracismHashtag):
18.         whichCategory = 0
19.
20.     elif any(antiraceWord in twTextLower for antiraceWord in antiracismDictionary):
21.         whichCategory = 0
22.
23.
24.     elif any(antiBLMWord in hashtagLower for antiBLMWord in antiBLMHashtag):
25.         whichCategory = 1
26.
27.     elif any(antiBLMWord in twTextLower for antiBLMWord in antiBLMDictionary):
28.         whichCategory = 1
29.
30.
31.     elif any(raceWord in twTextLower for raceWord in racistDictionary):
32.         whichCategory = 3
33.
34.     else :
35.         whichCategory = 4
36.
37. #Output is for racism categorization(0-Not,1-Yes,2-Ambiguous,3-Racist,4-Out of study)
38.     return whichCategory
```

Figure 0.4 The code snippet related to the function for grouping the tweets based on the study inclusion criteria.

3.6 About Python

There are more than 250 programming languages, and Python is one of the most attractive ones, described as an open-source, accessible, and portable language. The first use of Python was in February 1991, while now it is more than just a programming language and helps scientists in different areas. The reason for its popularity is because it is extensible and a kind of transportable programming language able to run on Unix, Mac, or Windows. It has no limitation for users; in addition to this, a wide range of community support enables a novice programmer to learn it quickly. Regarding science, some experts declared that Python is becoming the most popular coding language, and concerning the business, various companies and institutions prefer to use it for its flexibility features. Take as an example, Facebook production engineers respect Python as their first choice and even third most popular option at the whole company; in Google, it is one of the official languages; and in other companies such as Walt Disney Animation Studios, Industrial Light and Magic, Spotify, Netflix, Python is the preference for different purposes. People who are not professional in programming can also use Python because it supports multiple paradigms, including procedural, structured, object-oriented, and functional programming. Data scientists consider it as the best choice of language as it is multifaceted and flexible. Different libraries such as Pandas help clean up data and advanced manipulation since it is possible to organize row and column datasets, import CSV files, and much more in these libraries. Python is also an excellent start to extend data analysis using this popular programming language and its fascinating handy libraries [99]. Even now, in the year 2020, Statista [100] published new upshots showing that Python remained the trendiest language, according to GitHub and Google Trends. This position was for Java and JavaScript for a long time. See Figure 0.5 for reference.

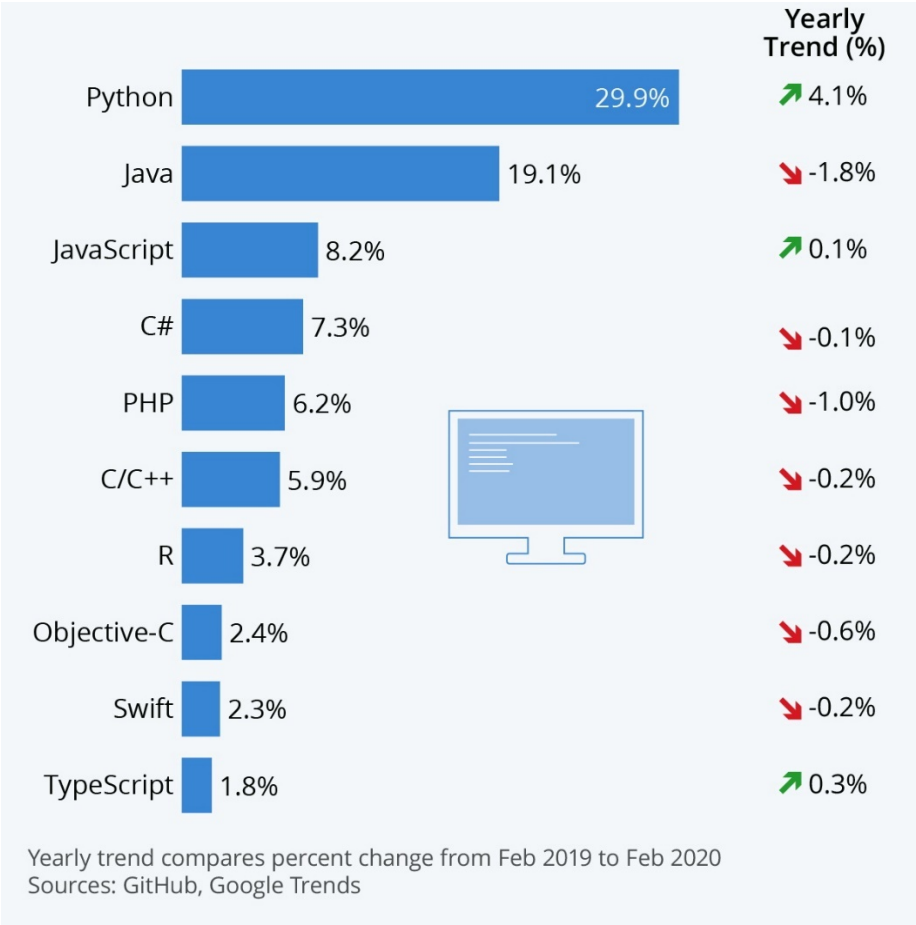

Figure 0.5 The top popular programming languages in reliance on their percentage of users that searched for instructions in Google.

3.7 Data Preprocessing

Data preprocessing steps have to be done to make sure we would get proper outcomes on our collected dataset. Preprocessing is the set of rounds passed to convert data into something understandable for the computer. One primary step is to remove ineffective data called stopwords in this term. Stopwords are useless data in natural language processing that are even ignored by search engines, such as "the," "it," or "in." Regarding the tweet text stopwords, we used NTLK Tokenize Package [101] and its English Stopwords Dictionary [102], while we also add our own dictionary to extract a more accurate outcome. The dictionary that we created consist of 25 punctuations or characters that we observed as unnecessary in our records. Figure 0.6 depicts the associated code for importing the NTLK package dictionary in addition to defining our own one.

```
stopWords = set(stopwords.words('english'))
myCleanDic = ['!', '#', '?', '@', ':', '.', "'", ',', '"', 'RT', 'I', 'https', '(', ')'
, '”', '&', 'amp', ';', '’', '..', "''", '“', '…', '…', '//t.co/']
```

Figure 0.6 Code snippets for the dictionaries used as stopwords

3.8 Data Preprocessing Methodology

We extended the preprocessing step and generated 17 new variables at this round. The first variable was the tweet's primary grouping number, which was described in section 3.5 as "inclusion criteria." This grouping consists of numbers between 0 and 3 applied to a tweet based on its content. We grouped our dataset into #BLM, anti-BLM, ambiguous, and racist.

Next, we used the Python Geopy library [103] to determine the associated state with the users' tweet profile locations across the US. We also obtained genders by using their username and by employing Gender-API [104], which includes 4 million names, and achieved the highest performance compared to its contenders [105].

As already stated, we do the text cleaning process based on a dictionary available in the NTLK library and one that we made since there were many HTML entities such as amp, which was embedded in the tweet. Hence, we required these entities to be removed from the text, and we eliminated them using this regular expression directly. There is also other information in our Twitter text fields, such as retweets, hashtags, usernames, and modified tweets, while all of this were ignored in the study.

We used TextBlob as our sentiment analysis to identify the polarity and subjectivity scores for each tweet. The next series of records is related to tweets' sentiment analysis and will be discussed in the next chapter. The TextBlob library that is a powerful python library for Natural Language Processing (NLP) developed using NLTK (Natural Language Toolkit) was used for the analysis.

The frequency of available tweets for each state was also calculated. We measured these sums of published tweets regarding our dataset location tags for each state to calculate the rates based on that. Table 0.4 shows the list of new variables in this study along with their descriptions and how the variable for acquired.

Table 0.4 List of new variables created in this study for each tweet in parallel to their descriptions and how that variable was formulated.

#	Variable Name	Data Type	Description / Values	The way acquired
1	What Group	Number	The classification of the group that tweet has resided (a number between 0 and 3)	Employing python codes and using the inclusion criteria
2	State	String	Name of a state	Using Python Geopy library [103]
3	Country	String	"United States of America" or "Other Countries"	Using Python Geopy library
4	Sentiment Polarity	Number	A number between -1 to 1 grading the sentiment polarity of the text	TextBlob Python library
5	Sentiment Subjectivity	Number	A number between 0 to 1 grading the sentiment subjectivity of the text	TextBlob
6	Polarity Score	String	Positive, Neutral, or Negative	Based on the "Sentiment Polarity" grade
7	Subjectivity Score	String	Subjective or Objective	Based on the "Sentiment Subjectivity" grade
8	Gender	String	Male, Female, or Null	Gender-API [105]
9	G-API Probability	Percentage	0 to 100%	Gender-API Probability for the gender
10	Is After Event	Boolean	If the tweet is before or after the George Floyd death (0 or 1)	Considering the creation date of the tweet
11	Is In USA	Boolean	If the tweet sender is located inside the US (0 or 1)	Considering the Geopy result
12	New Text Without Stop Words	String	The tweet text after removing the stopwords	Described in the content
13	Emojis	String (Emojis)	List of Emojis inside the tweet	We developed a function to do this
14	New Sentiment Polarity	Number	New score based on the text without stopwords	TextBlob
15	New Sentiment Subjectivity	Number	New score based on the text without stopwords	TextBlob

16	New Polarity Score	String	Positive, Neutral, or Negative	TextBlob
17	New Subjectivity Score	String	Subjective or Objective	TextBlob

3.9 Processing Time

We had about 32,000 Json files holding ~130 million tweets for each month processed individually using Python codes. The data was extracted and processed using several steps. In the first step, we obtain the tweets based on our inclusion criteria and if the user has a location tag inside the USA. We saved results in each of these steps and passed them to the next step for each month individually. Next, we used an API call to append state names in one similar shape to the data. Further, we developed separate codes to append gender values and doing the sentiment analysis. Overall, it took several weeks for us to extract all of these records and shape our data store.

3.10 Study Limitations

We had some limitations in our study regarding the data source. Twitter does not provide much information about users, such as gender, age, or exact location. Hence, if we had such information, we could compare the results by zip codes instead of states to have a more accurate outcome. The gender of the users was also unknown, and we had to use a third-party API to extract this information that reduced our sample size.

Chapter 4

Analysis of Data, Sentiment, and Statistics

4.1 Overview

We conducted an ecological data mining study utilizing Twitter data from archive.org in the first six months of 2020 to compare with national "Life Expectancy" as well as "Poverty Rate," "Educational Attainment," and “Race Composition” statistics. We aimed to see how these tweets categorized as racism or belong to the BLM movement are associated with these tables. As discussed in chapter 3, we gathered these data using an inclusion criterion, employing Python codes to extract tweets from the JSON files. We also did the sentimental analysis and tried to enhance data features utilizing geography and gender database APIs. We prepared and cleaned the data through different steps. Finally, we calculated the rates for all the states and compared them using Spearman Rho correlation analysis by IBM SPSS Statistics.

4.2 Study Design

4.2.1 Ecological Study

The term “Ecological Study” [106] refers to the research on a population for large-scale comparisons, usually to direct public health strategies. The difference between this population can be location or can be varied by time. Also, these people can be immigrants that are compared with natives or workers with diverse types of jobs [107]. The objective of such studies is to analyze the correlation linking inhabitants and the risk factors or diseases they are exposed to.

This is done especially when analyses at the individual level are unavailable or require aggregation of data.

Ecological studies are mostly defined by the level of depth they go to analyze the data, while the employed data belongs to a population rather than individuals. They are often used for rare diseases in which the study tries to determine the distribution rates of that particular disorder. The advantages of these studies are that they are low-cost and easy to perform, using straightforward ways of data collection. On the other hand, it has the disadvantage of being prone to bias and some confounders. These ecological studies are based on areas, and although individual studies are sometimes considered more accurate, the population-level research is more decisive in determining risk factors.

In particular, an ecological study aims to monitor population health and direct public health strategies by its outcomes. This method is developed by making largescale comparisons among those exposed to a specific health-risk and the rest of the population. It will show the contextual effects of risk factors by aggregating extensive data for this purpose

There are three types of ecological studies as defined by Levin [108]:

(1) Geographical: This is carried out by comparing geographical areas regarding the population health. Some might also measure the risk exposure for each area separately along with possible confounders such as demographic and socioeconomic variables.

(2) Longitudinal: Researchers monitor a cohort during a particular time to evaluate disease changes while also reporting the confounders in their analysis.

(3) Migration: Scholars analyze data associated with migrant populations in this research type, while time or place are ignored.

4.2.2 Statistical Analysis by Correlation Coefficients

Correlation means a deliberate relation with no chance of randomness between phenomena, things, mathematical elements, statistical variables that seems to have some associations, differences, or happening together, regarding the definition in Merriam-Webster dictionary [109]. There is a necessity to report correlation coefficient roots to show how the strength of variables is associated with the study. Instead of using categorical data in the form of strong, fair, weak, we use the numeral metrics in terms of correlation coefficients to qualify this relationship by scientific functions and show how a variable connection turns into a strong one [110]. The more these numerical metrics are close to 1 or -1, the deeper the level of association between them.

In math, correlation or more correctly, "correlation coefficient" is the metric that defines the dependency extent of one variable toward another. The requirement in this dependency is the fact that it should be in a predictive way. In a certain sense, if one variable changed, the other one should also fluctuate accordingly. This relation can be either positive or negative: it is defined as positive in an instance if one variable increased, the other one would rise too and it is identified as negative when a variable strengthens, the other one weakens [111], [112]. Figure 0.1 illustrates different kinds of these mathematical correlations by their degree level, considering that this relation could be positive, negative, or with no relation at all.

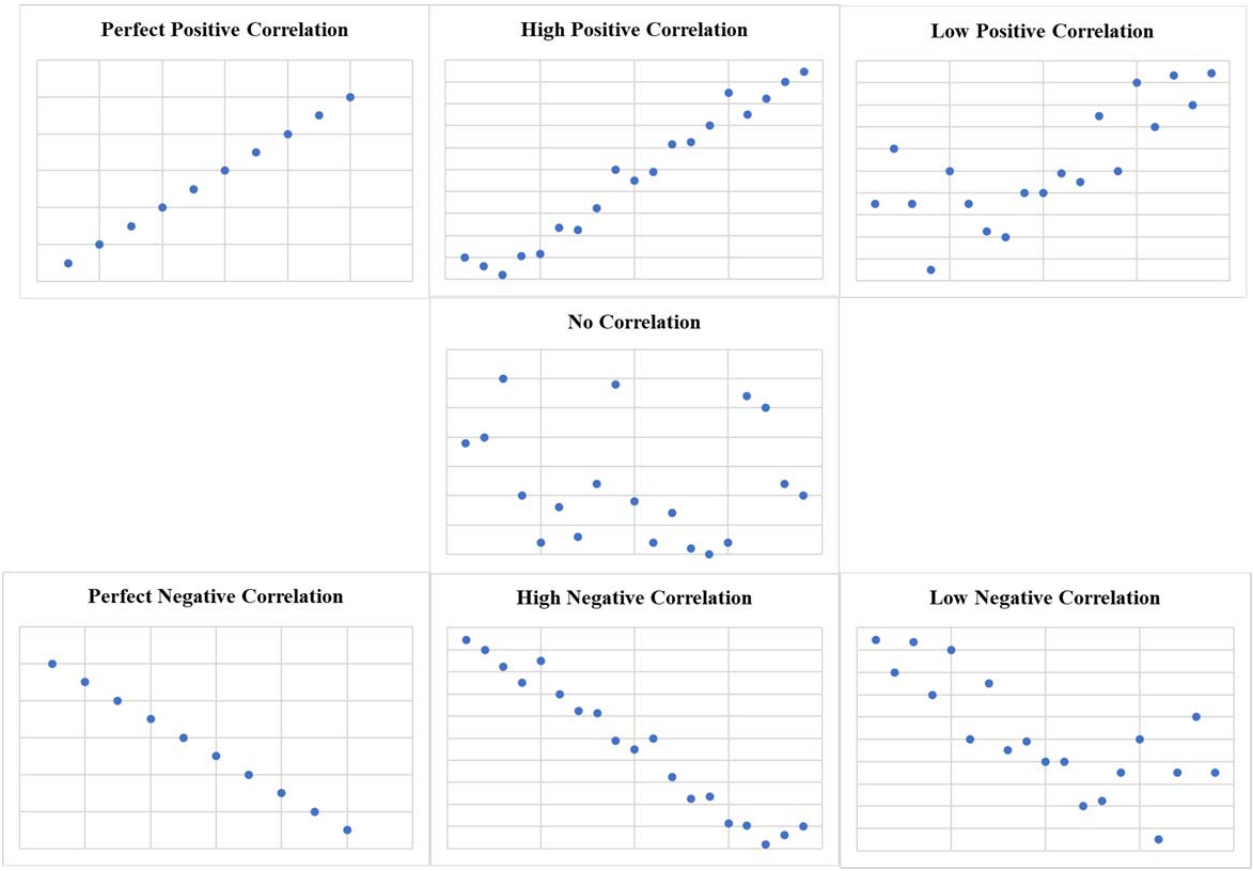

Figure 0.1 Various types of mathematical correlation regarding the degree level two variable are associated.

The outcome of the correlation coefficient labeled as R is a decimal number between -1.0 and 1.0. The more the number would be close to either 1 or -1, indicates that there is a stronger relationship between these two variables. As an instance, R= +1 implies that there is a perfect positive correlation between the two variables.

4.3 Natural Language Processing Toolkit (NLTK)

Natural Language Processing is a text processing technique that applies various analysis levels to transform linguistic objects into an appropriate format for further processing. NLP, anyhow, is sometimes a challenging task in computation as natural human language is hard to be converted into numbers. Computations are based on numbers, while language relies on words, and one word can get different meanings based on the content. Hence, the ambiguity of phrases, syntax, and semantics makes algorithm training a complicated process [113].

However, on the other side of the story and after this confusing process, we have a simplified outcome of a text analysis that reduces indexing size, improves information awareness, and accelerates message summarizing. These techniques also enhance the performance of data retrieval efficiency and robust documentation handling. These language handling procedures include “tokenization,” “stop word removal,” “lemmatization,” and “stemming,” which results in a sentiment polarity score [113].

There is a Python library called the “Natural Language Toolkit,” or NLTK in short, which is the best feasible solution for this purpose. NLTK is a package of various resources and libraries that enables researchers to process natural language, do computational linguistics, and perform artificial intelligence on their text. It makes the whole complicated job easy in Python coding steps, writing programs, classifying text, analyzing linguistic construction, among other things [114].

4.3.1 Word Tokenization

Tokens are the smallest elements in a text that could act independently and have a concrete structure and meaning. There are multiple paragraphs of text in every context, made by several

sentences that are divided into smaller components such as clauses, phrases, and words. There are two primary tokenization techniques for sentences and words that are used to fragment texts or corpus into their minimal elements. Tokenization can be referred to as the process of converting text into split diminutive meaningful words. The principal aim of tokenizing is to determine the actual words that shape that sentence's meaning [115], [116].

Natural Language Processing applications' first essential step is sentence tokenization as the corpus needs to be segmented into sentences prior to words. Each text corpus comprises paragraphs in the first level and sentences in the second that needs to be tokenized. Detecting these components is usually done by identifying delimiters such as a period (.), semicolon (;), or a new text line. In Python, we focus on specific sentence tokenizers from the NLTK library, which has several interfaces for this process [115]. Table 0.1 An example of a tokenized tweet. presents an instance of this process for tokenizing a tweet.

Table 0.1 An example of a tokenized tweet.

Title	*Text*
Original tweet	RT @bbyufo: say it with me: black lives matter is NOT a political movement! it is a HUMAN RIGHTS one.
Cleaned text	bbyufo say black lives matter NOT political movement HUMAN RIGHTS one
Tokenized value	[‘bbyufo’, ‘say’, ‘black’, ‘lives matter’, ‘NOT’, ‘political’, ‘movement’, ‘HUMAN’, ‘RIGHTS’, ‘one]

4.3.2 Stop Word Removal

The preprocessing step of removing “stop words” is an influential round in this term that is used by most NLP applications. The purpose is to get the text ready for the further process by eliminating the ineffective words as these words are regular across all corpus [117]. These words are mostly articles (such as a, an, the), determiners (such as this, that, these, those, my, his, her), and pronouns (such as these, those, mine, yours, its, ours) that are frequently used as a replacement for nouns to express them [118]. These phrases have an insignificant impact in the script, especially for the information extraction and analysis by NLP. Developers of the NLTK library created a carefully picked selection list of these words for at least 22 languages that occur most commonly. However, these are open-source dictionary files that can be modified based on the research needs or even the frequency in each context in order to end up with the most optimum list [117]. In our case, we added some words to the list, aside from the emojis. Since emojis were hugely used in our context, we preferred to clean them up and study them separately by different codes. Figure 0.2 shows a code snippet to remove stop words from the text.

```
1.  from textblob import TextBlob
2.  from nltk.corpus import stopwords
3.  from nltk.tokenize import word_tokenize
4.
5.
6.  stopWords = set(stopwords.words(‘english’))
7.  myCleanDic = [‘!’, ‘#’, ‘?’, ‘@’, ‘:’, ‘.’, “’”, ‘,’, ‘”’, ‘RT’, ‘I’, ‘https’, ‘(‘, ‘)’
    , ‘”’, ‘&’, ‘amp’, ‘;’, ‘’’, ‘..’, “’’”, ‘“’, ‘…’, ‘…’]
8.
9.
10. regrexPattern = re.compile(pattern = “[“
11.                 u”\U0001F600-\U0001F64F”  # emoticons
12.                 u”\U0001F300-\U0001F5FF”  # symbols & pictographs
13.                 u”\U0001F680-\U0001F6FF”  # transport & map symbols
14.                 u”\U0001F1E0-\U0001F1FF”  # flags (iOS)
15.                 u”\U00002702-\U000027B0”
16.                 u”\U000024C2-\U0001F251”
17.                 u”\U0001f926-\U0001f937”
18.                 u’\U00010000-\U0010ffff’
19.                 u”\u200d”
20.                 u”\u2640-\u2642”
21.                 u”\u2600-\u2B55”
22.                 u”\u23cf”
23.                 u”\u23e9”
24.                 u”\u231a”
25.                 u”\u3030”
```

```
                u”\ufe0f”
                “]+”, flags = re.UNICODE)


# Function to remove stop words
def removeStopWords(getTweetText):
    textTokens = word_tokenize(getTweetText)
    tokensWithoutSW = []
    tokensWithoutSW = [word for word in textTokens if not word in stopWords]

    cleanPunctuation = [x for x in tokensWithoutSW if x not in myCleanDic]

    if cleanPunctuation[-1].startswith(‘//t.co/’):
        cleanPunctuation[-1] = “”

    cleanEmojis = ‘ ‘.join(cleanPunctuation)
    cleanEmojis = regrexPattern.sub(r’’,cleanEmojis)

    return cleanEmojis


for i in range(2, sheet_obj.max_row+1) :
    thisTweetText = sheet_obj.cell(i,3).value
    cleanTweetText = removeStopWords(thisTweetText)

```

Figure 0.2 Code snippet to remove stop words and emojis.

4.3.3 Word Lemmatization and Stemming

Lemmatization is a method in which we transform a word into its raw form according to the meaning. It is the structural analysis of a word, intending to eliminate flexural endings. The algorithm gets the various kinds of a words and returns the base dictionary form of it, which is called a lemma. NTLK library uses the WorldNet’s built-in morph function for the Lemmatization process. In Table 4.2, we present an example of a tweet converted in this way.

Stemming is also a similar kind of processing on words in which it tries to convey words to the shorter form. It reduces the expressions in words to transform them to their root forms, such as converting to a stem even if it is not a real word. An instance of the NTLK module to do this process is the “PorterStemmer”. However, these two words have similar confusing meanings that are sometimes unclear, while there are differences between them. Lemmatization gets the words

and turns them into their meaningful root while stemming just removes the endings, responsible for the incorrect meanings or even spelling mistakes. Table 4.2 shows a sample outcome of these two functions [119], [120].

Table 0.2 An example of word lemmatization and stemming.

Title	*Text*
Original tweet	It's absolutely disgusting that you get mistreated because of the colour of your skin.
Tokenized value	['It', ''', 'absolutely', 'disgusting', 'get', 'mistreated', 'colour', 'skin', '.']
Stop Words removed	['It', 'absolutely', 'disgusting', 'get', 'mistreated', 'colour', 'skin']
Lemmatization	['It', 'absolutely', 'disgusting', 'get', 'mistreated', 'colour', 'skin']
Stemming	['It', 'absolut', 'disgust', 'get', 'mistreat', 'colour', 'skin']

4.4 Sentiment Analysis

Sentiment analysis examines the frequency of positive and negative opinions and emotions in natural language texts. According to social media's widespread usage for publishing news or information, sentiment analysis is a widely researched subject as well. Besides, businesses use this tool to follow public opinions toward their products. They distribute details about their products on social media, asking consumers for their feedback so that they would be able to analyze comments to see the positive and negative side effects. Some excellent software tools are available to automate sentiment analysis and examine large numbers of comments, but this is

expensive and time-consuming work. Aside from these, sometimes it is too hard to do an accurate sentiment analysis because language is subjective, complex, and full of creativity. In fact, a collaborative approach is necessary to produce a more advanced and precise tool for analyzing speech and measuring sentiment [121].

In this study, as mentioned in section 3.7, we used the TextBlob library for our analysis. The algorithm's outcome is two numbers for all the tweet texts in this study. The first quantity is that the sentiment polarity score is a float between -1.0 and 1.0. The second one is the subjectivity score, again a float number but within the range of 0.0 to 1.0. Consider the number x as the Polarity score and y as the subjectivity score, we categorized our definition of these number's meaning as below:

-1.0 <= x < -0.1 => Negative	y <= 0.5 => Objective
-0.1 <= x <= 0.1 => Neutral	y > 0.5 => Subjective
0.1 < x <= 1.0 => Positive	

4.5 Statistical Analysis

A key question is to select either parametric or nonparametric statistical tests [122]. A parametric test is a test that needs to assume specific parameters while we should be aware of the study population's distributions. The data in this type have a normal distribution; every group's samples are independent and the groups' variances are equal. The measure to calculate the central tendency is the mean value, while it is the median for the nonparametric test. The nonparametric tests do not make any assumptions, or in other words, it assumes that random samples are derived as observations of these groups. It is used when the collected data is ordinal

or does not fit a normal distribution curve. Researchers obtain ordinal data through experiments that use various rankings or orders. Hence, this type does not rely on the exact values like the way parametric tests employ it. However, usually, a parametric analysis is the preferred option for scientists compare to nonparametric tests. Consequently, if we have an unknown set of populations in our dataset, parametric tests cannot be applied, and the nonparametric types are the final choice [122], [123].

4.5.1 Spearman's correlation

We used Spearman's rank-order for analyzing our data as it is the nonparametric test for calculating the correlation. Correlation is the degree to which two variables are linearly related, and in bivariate data analysis, it is so critical. Spearman's coefficient values range between -1.0 and 1.0, while -1.0 shows a perfect negative correlation, inversely an outcome of 1.0 indicates a perfect positive correlation. Correlation can be any statistical relationship between two random variables in bivariate data, whether causal or not. We used Spearman, the nonparametric version of the Pearson correlation, as it shows how strong the link is between the variables. The input variables can be two ordinal intervals or ratios [124], which in our case are rates. Spearman's correlation determines the strength and direction of a monotonic relationship between two ranked variables instead of considering the linear relationship's strength and direction [125], [126]. A monotonic function's definition [127] is very similar to the linear one, while there are some differences. In a monotonic relationship, one of the following procedures happens:

(1) The increase in value of one variable, results in increase of the other variable.

(2) The increase of one variable's value, leads to the fall of value of the other variable.

While both monotonic and linear increase or decrease in the same direction, the difference is on the rate that they are similarly rising or falling. The monotonic variables do not always go upwards or downwards at the same rate. Unlike this, linear variables grow or fall at the same rate. Figure 0.2 shows an instance of a monotonic function [124].

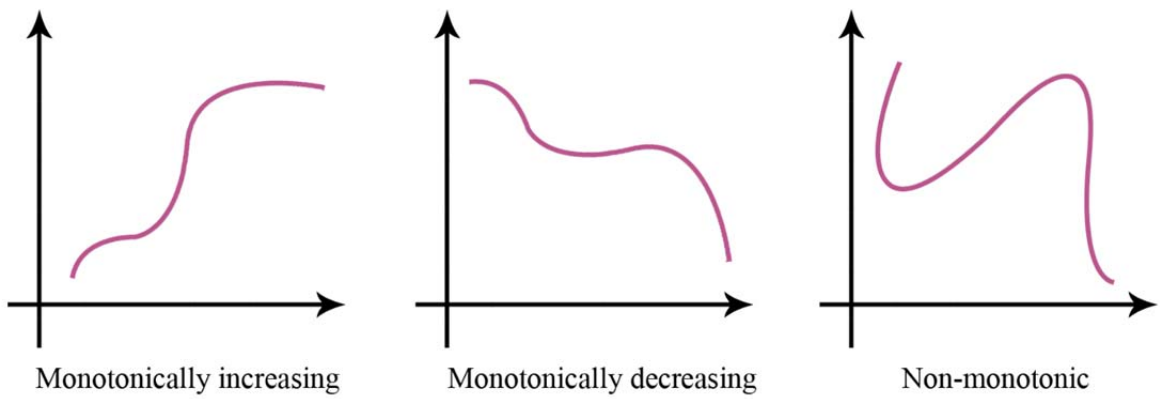

Figure 0.2 An instance of a monotonic function

The formula for the Spearman correlation coefficient [128]is shown below:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

ρ (rho) = Spearman's rank correlation coefficient
d_i *= difference between the two ranks in each observation*
n = number of observations

One of the positive points of monotonicity in Spearman's correlation is less limited than a linear relationship. We can run a Spearman's correlation on a non-monotonic relationship to determine if it is a monotonic component. Therefore, Spearman can be a better choice for broader types of data. For our purposes, as the variables do not have a normal distribution, and one of these

bivariates (states=52) are small [129], we had to use the Spearman rank correlation method. We used Spearman correlation analysis between our outcome rates and states, "Life Expectancy" tables, "Poverty Rates," 'Educational Attainment," and "Race Compositions," while a 2-sided $\alpha < 0.05$ was considered statistically significant. A two-tailed test and consideration of 0.05 as the significance level means that this test would share half of this significance level (0.025) to test the statistical analysis of one direction and the other half for testing the significance in the different direction [130]. We used IBM SPSS Statistics 26 for this purpose while it also clarifies that these coefficient numbers on the resulting tables are marked with ** for those results that could have a significant correlation at the 0.01 level (2-tailed) and * marks denote the findings that are significant at the 0.05 level (2-tailed).

4.5.2 Life Expectancy

The “life expectancy” term addresses the number of years that a newborn expects to be alive. While this rate cannot be identified at birth, researchers determine an average lifetime of years for members of a specific population to acquire this number. This factor is an essential and prominent measure in research and policy that is widely used as an indicator of health status [131], [132].

The results were compared with “life expectancy at birth” data [133] employing the “National Center for Health Statistics” (NCHS) records. They used the population counts census that happened in April 2000, separated by states. The death rates belonged to the three consecutive years starting 1999, and births from 1997 to 2001. They also employed former decennial life tables to estimate smaller population subgroups’ reliable rates where sufficient data is not

available. This method helped to append data for the black population of six more states which was not available before due to the low death toll. Most numbers associated with dying and life expectations in [109] are based on probabilities, hence they also provided the "standard errors" metrics for each record. These values would not indicate the accuracy of the calculations.

4.5.3 Poverty Rate

Many black Americans have a sad story at their birth, experiencing an environment where poverty is typical, educational failure dominates, and social deterioration overflows. Continuous exposure to such an environment diminishes the chance to have a social or economic success story for a black fellow [134].

There is a remarkable diversity in whites' typical lives, whereas blacks remain in the same economic class in a social environment. For instance, most white middle-class families live in average communities and send their children to typical schools. In contrast, many blacks in the same social class live in impoverished regions and foster their children at high-poverty schools [135]. Therefore, racial disparity, notably in poverty, persists in this modern age and influences many aspects of black life [136].

Here, we employed a list of U.S. states and the associated poverty rate [137] to compare and see its correlation with our dataset. We applied three available lists of poverty rates in the U.S. which are:

1. 2018 Poverty rate, percent of persons in poverty [138]
2. 2014 Poverty Rates that include unrelated children [138]

3. Supplemental Poverty Measure, an average measure of 2010–2014 with Geographically Adjusted [139]

4.5.4 Educational Attainment

Statisticians commonly use the term "educational attainment" to indicate the highest grade of education for an individual. It is mostly used as an average equivalent for a group or territory. The corresponding information is accomplished by a simple question that asks, "What is the highest grade of school you have completed?" and we derived these statistics using a survey done by the American Community Survey [140] from 2013 to 2017. This educational attainment list includes individuals at least 25 years or older after an extensive demographic survey collected by mail questionnaires, telephone interviews, and visiting representatives [141].

4.5.5 Race Composition

Learning the extent to which ethnic groups differ are combined regarding the BLM movement rates and racism in this study. Considering this, race and ethnicity study is complicated in this content as the United States of America has a vast, diverse population [142].We used statistics from "World Population Review" data [143] and specifically compared the white and black population rates in each state to the rates of tweets we acquired. These tables are statistics regarding the composition percentage of Black and White people in each state.

4.5.6 Emojis

As a supplementary part of the study, we also measured the top emojis used for the primary two groups: “Black Lives Matter” and “Racist.” For this purpose, we converted our primary classification variable into binary arguments and also created multiple new binary ones indicating each emoji occurrence by 0 or 1. Next, we used the Pearson Chi-Square test on these parameters to see the outcome and how they are related.

Pearson’s chi□square test is a widely used algorithm to measure the relationship between two categorical entities. Chi□square is often used to estimate the connection linking two continuous or ordered categorical variables. Pearson’s statistics is extensively applied in trials that aim to test the independence of the two ordinal variables that are also invariant to monotonic transformations of the primary variable [144]. Chi□square test is also named “goodness of fit” because of its feature that can observe how data distribution fits the expectations, in case we have independent variables [145]. Its primary task is to measure the independence of a bivariate table called cross-tabulation. This table is the input for the test presenting the distribution of two categorical variables to be analyzed for independence. The algorithm assesses how the patterns of value existence meet the expectations of being independent [146]. This type of test calculation determines how significantly different two observed records are toward each other. Chi-Square test statistics compares the cells’ diversity in tables to the expected values to reflect the association among these variables [146]. Pearson’s chi-square test is given by the formula below.

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Chapter 5

Results and Discussion

5.1 Overview

In this chapter results are presented and the findings are discussed. First, the details on the dataset are introduced by using the inclusion criteria and available tweets and followings; next, the statistics about states and the results are discussed considering the four tables implemented in this study. In this chapter, outstanding results will be presented in correlation with tweet rates and how racism rates affect Black people's lives. Also, a separate analysis of the results regarding the sentiment analysis will be presented. In addition, two supplementary sections of word clouds and some discussion about emojis are presented. Further the findings are summarized and the reasons for the results to be more consistent in males' correlation rates is explained. A reasoning for disregarding the reflecting genes as a confounder is discussed and finally, small talks with policymakers is discussed.

5.2 Tweets Review

The inclusion criteria yielded 306,925 tweets out of the 43,830,301 tweets with location values available inside the United States of America in the first six months of 2020 using archived data. The data groups are (1) BLM -supporting Black Lives Matter movement, (2) Anti-BLM – those who tried to confront this movement by their contentious words, (3) Ambiguous -having hashtags of both previous groups, and (4) Racists – those who used offensive n-words facing

Black people. Acknowledging before and after the George Floyd killing event and considering gender, different rates were obtained for each state and the total tweets published in the associated time.

The most extensive data group is the Racist with 229,924 records, and the following ones are BLM, Anti-BLM, and Ambiguous with 69,116; 7,617; and 267 records, respectively (See Table 0.1 for exact numbers for each month). Gender specification was achieved for 52.31% of the data, which is 160,539 records and is detailed in

.

Table 0.1 Number of tweets in each group detailed by month

Month	Black Lives Matter	Anti BLM	Ambiguous	Racist	Tweet Numbers	Sum of all GeoTagged Tweets
January	202	73	1	43950	44226	7,544,965
February	96	112	2	20740	20950	3,604,863
March	117	115	0	42638	42870	8,401,976
April	250	218	0	42245	42713	8,194,289
May Before 25	340	528	1	31554	32423	6,056,215
May After 25	15957	1588	186	9838	27569	1,985,889
June	52154	4983	77	38959	96173	8,042,104
Sum	69,116	7,617	267	229,924	306,925	43,830,301

The first visible outcomes are the impressive number of tweets publishing the #BlackLivesMatter movement by their tweets after George Floyd's demise on May 25, 2020 (Figure *0.1*). The numbers substantially rose to 68,111 in 35 days compared to 1,005 tweets in the first 145 days of the year. Taking into account that this hashtag movement's history gets back to 2013, this study's outcomes show a significant increase in tweets published after this sad event. Appendix 1 contains images as examples of tweets connected with the first group supporting the Black Lives Matter movement.

Similarly, those who stand on the other side with their Anti-BLM tweets were also more active after this event. Although this group number is small, it jumped to 6,571 from 1,046 comparing the one-month and four-day period after the event to the four months and 25 days before that. On the other hand, the Racist group had a slight decrease compared to the other groups. Taken all groups together regarding the event's impact, results are compared in Figure *0.2* to show the trends.

Table 0.2 Number of tweets in each group detailed by gender

Title	*Tweet Count*
All in USA	306,924
Males	88,280
Females	72,259
BLM Males	23,924
BLM Females	20,989
AntiBLM Males	2,649
AntiBLM Females	2,120
Racist Males	61,605
Racist Females	49,080

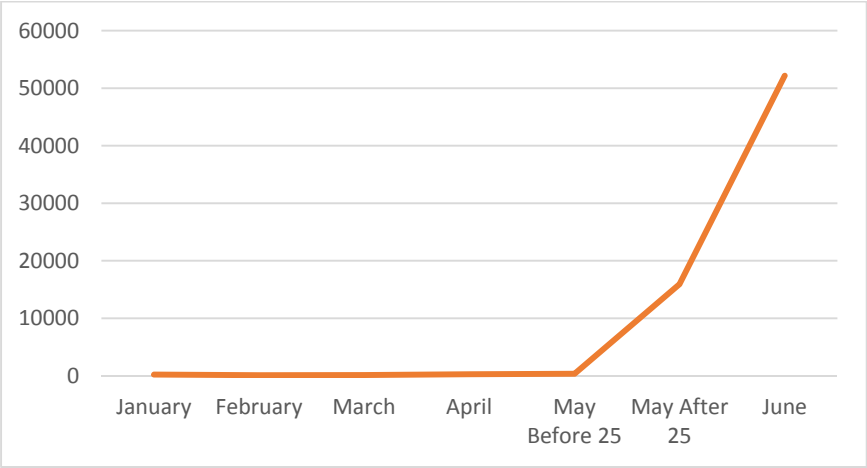

Figure 0.1. Increase in the number of tweets supporting Black Lives Matter movement in 2020

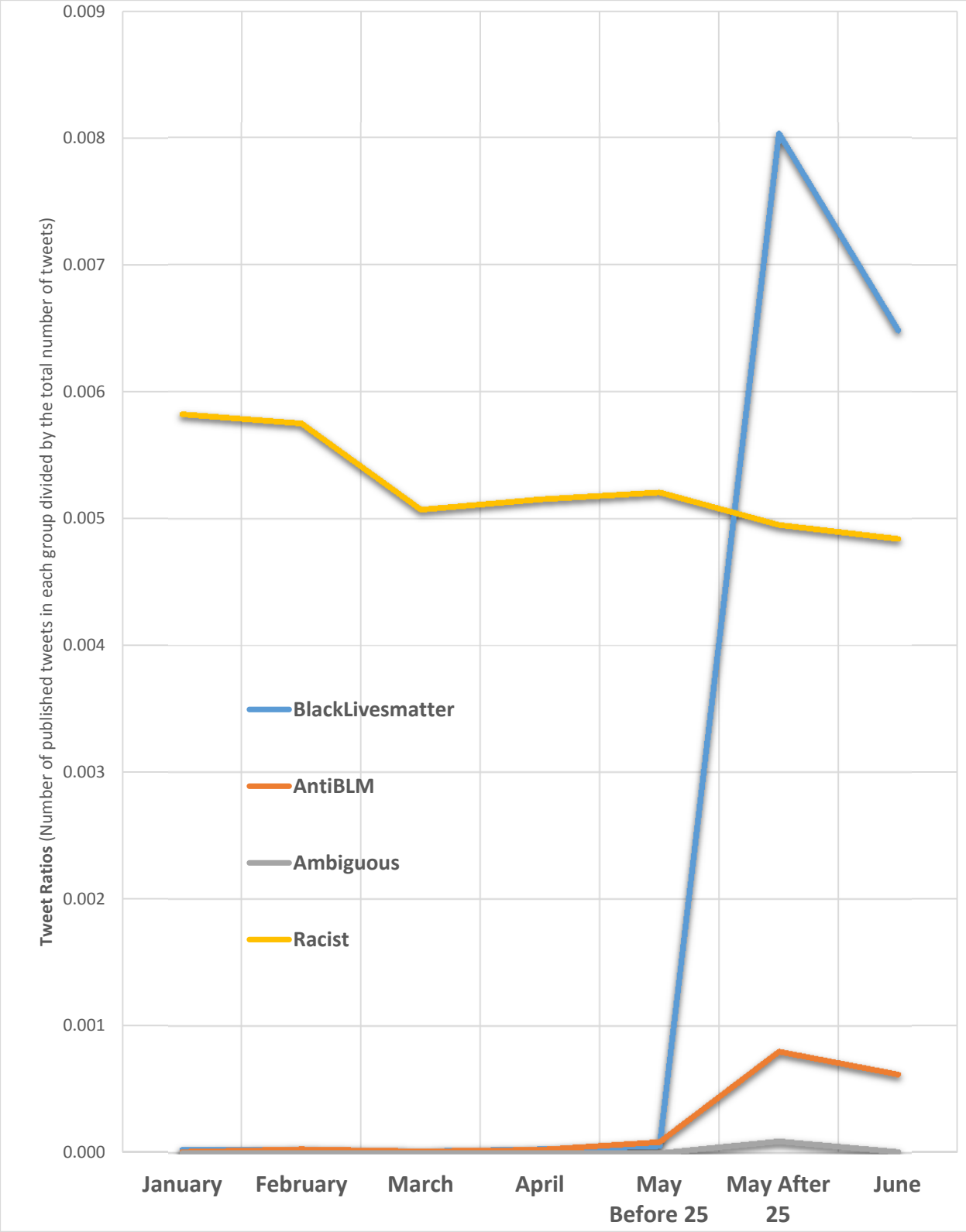

Figure 0.2. Comparing the trends of each group in 2020

5.3 States' Statistics

As described in previous chapters, tweets that have location tags in the USA were included. Further, the Geopy library was utilized to locate the users' areas and assign one of the 52 States names. Table 5.3 shows the number of tweets for each state by details on their group category. The last column in this table shows the total number of tweets for each area. It contains all the unrelated and related tweets published at this time in that specific area. We extract these numbers to calculate the rates for each area. Since different states have different populations, and as each group number could not be studied alone; hence, the outcome of each state group number was divided by the total number to end up with a percentage rate. Although this rate in most cases is a small float number with up to 20 decimal digits (Example: 0.0000267728213450957), the rates had to be studied this way to compare the portions of these topics to the whole conversation of the community. Figures 5.3, 5.4, and 5.5 depicts these rates for BLM, Anti-BLM, and Racist group coloring each state accordingly.

Additionally, see Table 5.4 for the number of tweets in two more extensive groups of BLM, and Racists detailed by the area, before and after the event of George Floyd' demise. Further, Table 5.5 shows the detailed numbers for tweets that show the gender data. This table has three main groups regarding the area. Also, see Appendix 2 for details on the number of tweets that had locations tags in each state regarding the month published.

Table 0.3. Number of tweets in each category regarding the area.

State	BLM	Anti-BLM	Racist	Ambiguous	All Tweets in This Area
Alabama	377	46	2252	0	291,698
Alaska	110	9	214	1	57,960
Arizona	1314	164	2409	3	749,954
Arkansas	314	36	1243	1	210,541
California	11684	1165	30819	39	6,732,314
Colorado	848	99	1317	2	531,722
Connecticut	278	35	1155	2	224,276
Delaware	180	20	731	0	130,147
District of Columbia	2185	175	3937	2	1,195,670
Florida	3508	384	16093	19	2,420,970
Georgia	2180	250	18294	11	1,492,226
Hawaii	182	21	417	0	110,308
Idaho	121	16	52	0	77,623
Illinois	3722	389	13199	20	2,348,862
Indiana	666	80	2114	1	449,515
Iowa	570	57	662	2	380,333
Kansas	527	80	866	2	351,748
Kentucky	422	50	1298	3	298,665
Louisiana	848	118	7824	2	650,554
Maine	23	1	28	0	21,351
Maryland	1715	203	11005	9	1,074,985
Massachusetts	1277	140	2617	4	841,753
Michigan	2051	258	8051	12	1,370,935
Minnesota	612	72	878	1	403,050
Mississippi	204	31	1312	1	185,143
Missouri	831	94	1528	3	550,886
Montana	63	8	35	1	42,761
Nebraska	275	41	395	0	205,073
Nevada	1177	134	3207	7	783,094
New Hampshire	66	10	85	0	54,283
New Jersey	516	56	3084	6	371,906
New Mexico	213	32	308	0	128,095
New York	7050	737	16030	20	4,474,808
North Carolina	1379	174	6312	8	917,125
North Dakota	36	1	41	0	39,419
Ohio	1498	162	7377	7	1,101,448
Oklahoma	709	88	1256	2	473,464

Oregon	1684	151	1072	6	827,140
Pennsylvania	1745	176	5994	7	1,225,688
Puerto Rico	27	1	80	0	56,633
Rhode Island	148	19	448	1	104,717
South Carolina	326	49	1370	0	246,991
South Dakota	34	0	26	0	33,419
Tennessee	971	113	4234	6	710,271
Texas	11074	1327	43182	42	6,987,472
Utah	256	39	206	0	167,547
Vermont	57	6	67	1	32,792
Virginia	521	66	1812	3	325,978
Washington	1952	151	1704	2	915,036
West Virginia	80	9	98	1	52,464
Wisconsin	443	57	1112	3	351,723
Wyoming	67	17	74	4	47,765

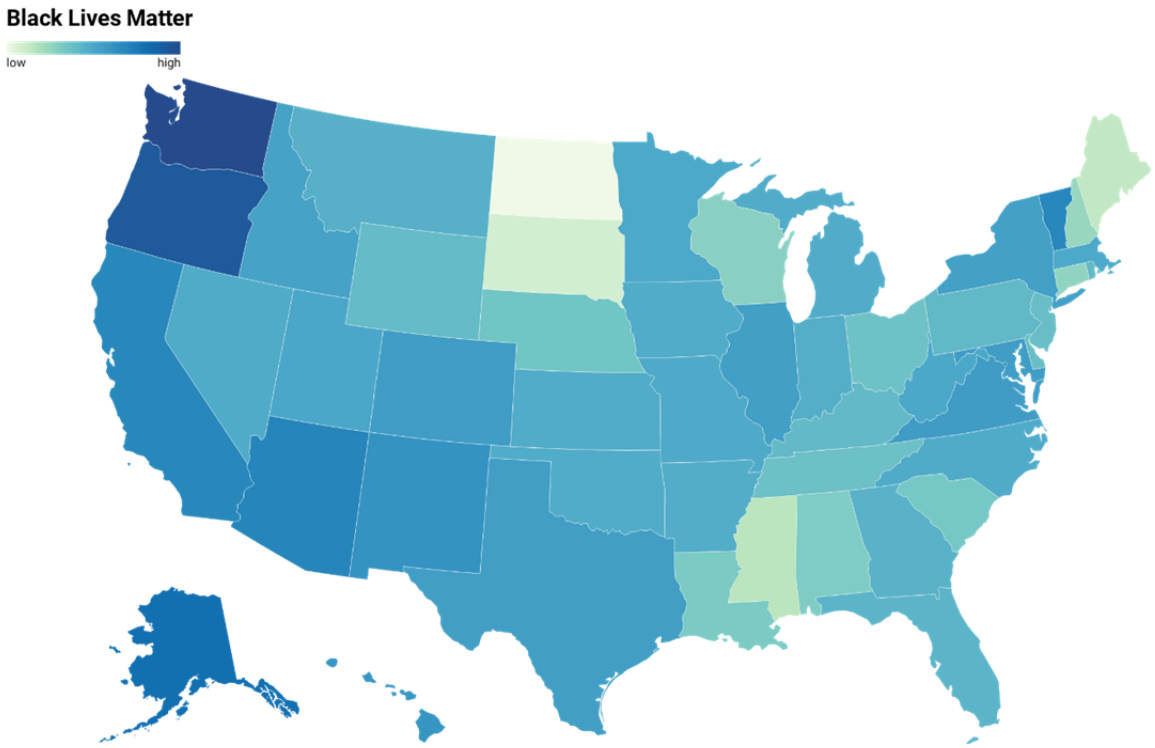

Figure 0.3. Comparing the rates of BLM tweets in each state on the map of the USA.

A darker colour shows a higher percentage of tweets joining this movement by their tweets. The rates are calculated utilizing the number of tweets in the BLM group divided by the sum

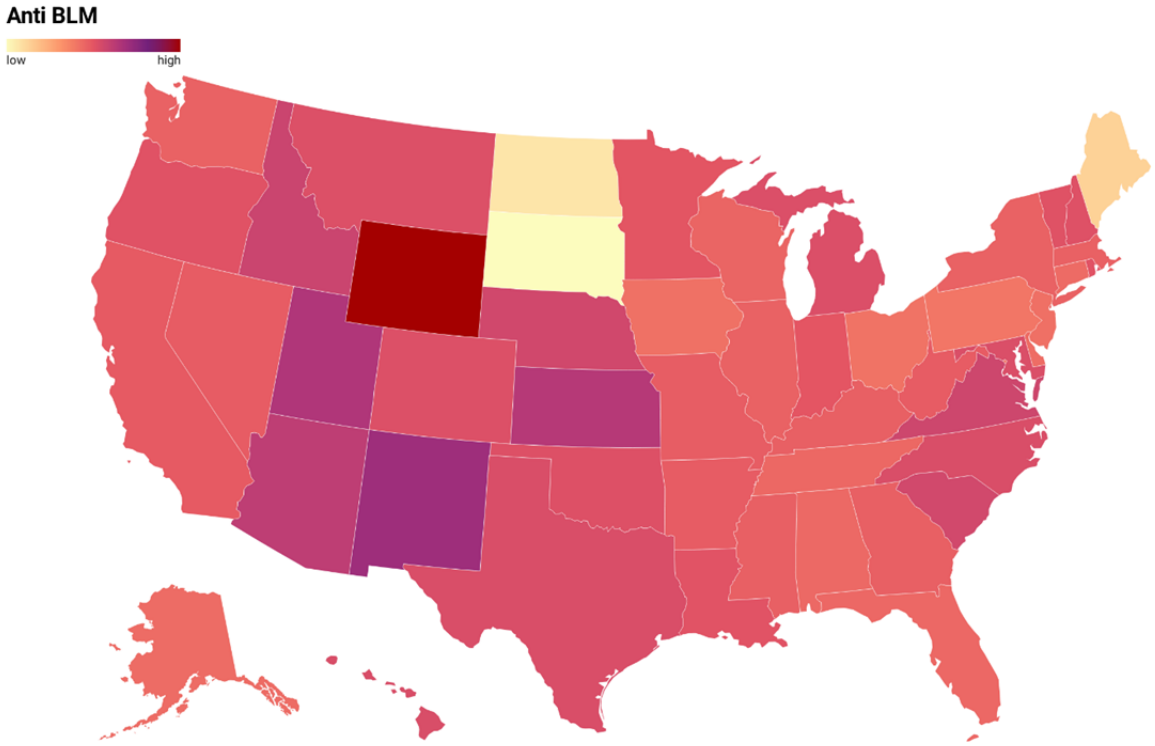

Figure 0.4. Comparing the rates of Anti-BLM tweets in each state on the map of the USA.

This group are those who faced the BLM movements by their counter-hashtags. A darker-contrasted colour shows a higher tweet rate regarding the sum of all irrelevant and relevant tweets published in each state.

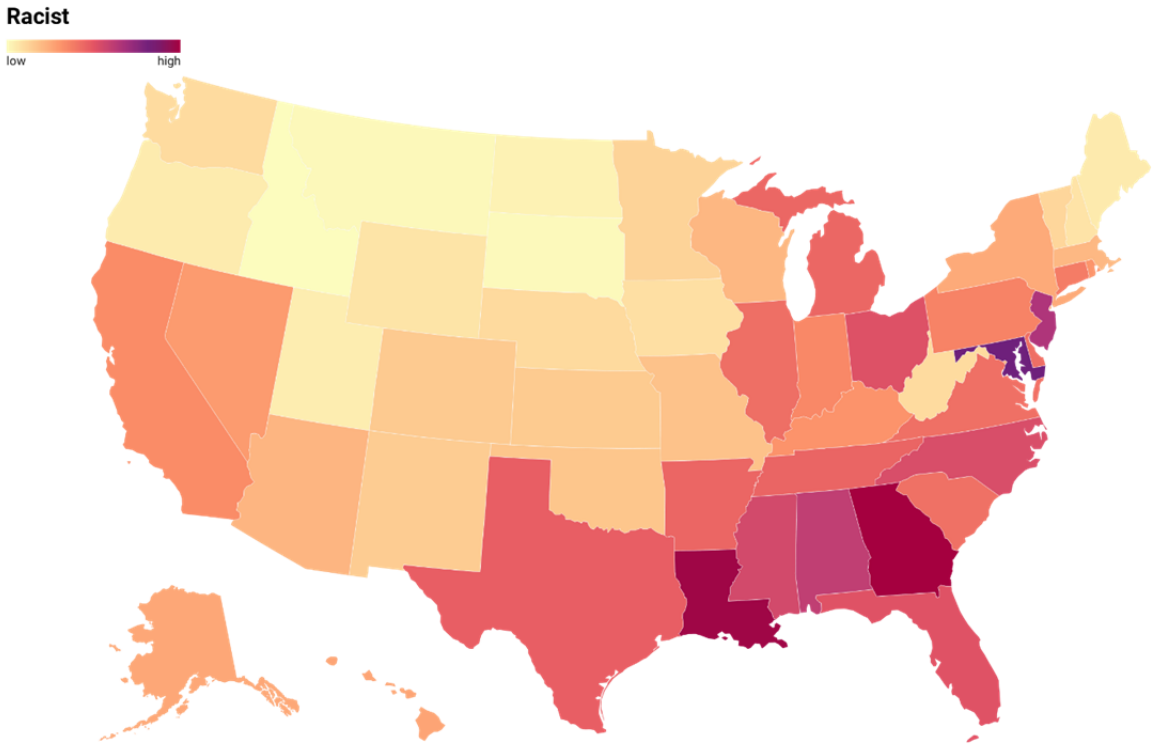

Figure 0.5. The rates of Racist tweets in each state on the map of the USA.

A darker-contrasted colour shows a higher rate of using offensive n-words compared to all conversation posted in tweets.

Table 0.4. Number of tweets in two main groups regarding the area and the event of George Floyd' demise.

State	BLM tweets before event	BLM tweets after event	BLM Overall Ratios	Racist tweets before event	Racist tweets after event	Racist Overall Ratios
Alabama	2	375	0.00129	1832	420	0.00772
Alaska	1	109	0.00190	167	47	0.00369
Arizona	19	1295	0.00175	1923	486	0.00321
Arkansas	1	313	0.00149	996	247	0.00590
California	148	11536	0.00174	24368	6451	0.00458
Colorado	6	842	0.00159	1034	283	0.00248
Connecticut	4	274	0.00124	901	254	0.00515
Delaware	5	175	0.00138	566	165	0.00562
District of Columbia	44	2141	0.00183	3073	864	0.00329
Florida	42	3466	0.00145	12565	3528	0.00665
Georgia	40	2140	0.00146	14230	4064	0.01226
Hawaii	5	177	0.00165	331	86	0.00378
Idaho	2	119	0.00156	49	3	0.00067
Illinois	60	3662	0.00158	10306	2893	0.00562
Indiana	12	654	0.00148	1672	442	0.00470
Iowa	9	561	0.00150	517	145	0.00174
Kansas	3	524	0.00150	667	199	0.00246
Kentucky	7	415	0.00141	1053	245	0.00435
Louisiana	10	838	0.00130	6189	1635	0.01203
Maine	0	23	0.00108	20	8	0.00131
Maryland	33	1682	0.00160	8710	2295	0.01024
Massachusetts	24	1253	0.00152	2056	561	0.00311
Michigan	32	2019	0.00150	6342	1709	0.00587
Minnesota	16	596	0.00152	666	212	0.00218
Mississippi	0	204	0.00110	1057	255	0.00709
Missouri	7	824	0.00151	1194	334	0.00277
Montana	1	62	0.00147	30	5	0.00082
Nebraska	5	270	0.00134	312	83	0.00193
Nevada	14	1163	0.00150	2524	683	0.00410
New Hampshire	1	65	0.00122	73	12	0.00157
New Jersey	9	507	0.00139	2450	634	0.00829

New Mexico	2	211	0.00166	244	64	0.00240
New York	122	6928	0.00158	12625	3405	0.00358
North Carolina	25	1354	0.00150	4956	1356	0.00688
North Dakota	1	35	0.00091	34	7	0.00104
Ohio	17	1481	0.00136	5765	1612	0.00670
Oklahoma	6	703	0.00150	1009	247	0.00265
Oregon	22	1662	0.00204	843	229	0.00130
Pennsylvania	41	1704	0.00142	4664	1330	0.00489
Puerto Rico	0	27	0.00048	56	24	0.00141
Rhode Island	1	147	0.00141	357	91	0.00428
South Carolina	4	322	0.00132	1098	272	0.00555
South Dakota	0	34	0.00102	21	5	0.00078
Tennessee	10	961	0.00137	3334	900	0.00596
Texas	142	10932	0.00158	34266	8916	0.00618
Utah	1	255	0.00153	162	44	0.00123
Vermont	1	56	0.00174	53	14	0.00204
Virginia	5	516	0.00160	1422	390	0.00556
Washington	34	1918	0.00213	1326	378	0.00186
West Virginia	0	80	0.00152	77	21	0.00187
Wisconsin	9	434	0.00126	885	227	0.00316
Wyoming	0	67	0.00140	57	17	0.00155

Table 0.5. Number of tweets that have gender data in three main groups regarding the area.

	BLM tweets Male	***BLM tweets Female***	***antiBLM tweets Male***	***antiBLM tweets Female***	***Racist tweets Male***	***Racist tweets Female***
Alabama	132	121	16	13	635	51
Alaska	38	31	2	4	49	486
Arizona	439	399	65	47	746	259
Arkansas	118	96	12	13	343	6461
California	4072	3604	405	330	8358	277
Colorado	304	244	36	26	362	227
Connecticut	100	85	13	12	296	182
Delaware	64	59	7	6	183	784
District of Columbia	721	698	67	27	890	3613
Florida	1169	1078	133	115	4169	3851
Georgia	756	596	71	82	4738	74
Hawaii	47	61	4	6	109	15
Idaho	49	43	8	3	14	2758
Illinois	1285	1125	145	109	3568	464
Indiana	250	208	27	27	519	172
Iowa	200	193	27	21	204	200
Kansas	186	162	25	23	247	277
Kentucky	160	122	18	14	383	1553
Louisiana	296	235	40	28	1713	0
Maine	10	7	1	0	13	2312
Maryland	605	503	70	45	2883	569
Massachusetts	502	414	56	28	747	1617
Michigan	705	626	80	70	2244	160
Minnesota	213	173	31	14	258	278
Mississippi	72	59	10	8	294	302
Missouri	294	237	33	29	459	6
Montana	20	15	4	2	8	104
Nebraska	115	83	14	15	123	653
Nevada	370	360	47	46	937	16
New Hampshire	30	17	4	3	23	578
New Jersey	199	159	26	11	844	79
New Mexico	80	64	13	8	90	3426
New York	2443	2114	276	177	4374	1384
North Carolina	495	435	57	50	1734	5
North Dakota	14	12	1	0	18	1618
Ohio	571	442	49	42	1965	261

Oklahoma	248	197	33	26	352	245
Oregon	582	491	47	54	332	1237
Pennsylvania	671	514	67	42	1484	15
Puerto Rico	10	7	0	0	25	80
Rhode Island	72	37	7	9	109	317
South Carolina	108	102	18	8	296	3
South Dakota	14	9	0	0	7	909
Tennessee	348	294	32	45	1100	9725
Texas	3504	3442	436	376	11944	28
Utah	98	74	18	11	65	14
Vermont	19	19	5	1	13	373
Virginia	180	165	18	24	513	320
Washington	723	593	49	41	468	23
West Virginia	30	25	4	5	32	224
Wisconsin	162	120	19	19	305	11
Wyoming	31	20	3	5	20	0

5.4 Principal Analysis

Subsequently, the outcomes of the study were analyzed to national "Life Expectancy," "Poverty Rate," "Educational Attainment," and "Race Composition" statistics. Significant outcomes will be shown, while there is a strong evidence of a relation between racist content rates and the life expectancy of male black people. Further, in the discussion section, the reasons why the outcomes show a more reliable relation with tables that are closely tied with black males will be addressed. The more surprising correlation is also with poverty rates, high school educational assessment, and race composition. The single most striking observation to emerge from the data comparison was a significant positive correlation between the number of bachelors or advanced degree people and the number of tweets in the BLM group. It means the higher the degree level, the more they participated in this movement defending blacks.

5.4.1 Correlation with Life Expectancy

We compared different rates of the tweets in each group to the life expectancy tables [133] from the "National Center for Health Statistics" (NCHS), and what stands out in the outcomes is a recurring positive correlation between the racist tweets and "Calculated Life Expectancy of White Male – Black Male". From this correlation, we can understand that there is a trend between these factors, and when the racist tweet numbers are more extensive, there is an enormous difference between White and Black's life expectancy. For instance, the result is significant at the $p < 0.01$ level by the Spearman correlation coefficient of 0.408 for 40 states with the available data for black life expectancy. See Table 5.6 for details on this purpose, while the related evidence is highlighted. Additionally, Appendix 3 contains the associated tables for life expectancy.

Moreover, the rate of BLM tweets has a positive correlation with black life expectancy. It alternatively implies that in areas where people tend to publish more BLM tweets and care more about this movement, black people have higher life expectancy rates. The rate of BLM tweets determined by dividing the sum of all BLM tweets by the sum of all tweets published over the study time. This rate has a significant correlation at the $p < 0.05$ level by the Spearman correlation coefficient of .357 with "black expectation of life at age 0" and at P-value < 0.05 level by the coefficient of .347 with "black male expectation of life at age 0". Results are highlighted in the first and second row of Table 5.7.

Table 0.6 Spearman's rho statistical analysis for percentage of racist tweets compared to life expectations of "White people minus Black" regarding gender.

		Number of Racist tweets Male	*Number of Racist tweets Female*	*Racist tweets Male Before event*	*Racist tweets Female Before event*	*Racist tweets Male After event*	*Racist tweets Female After event*	*Racist rates (Racist / Sum)*
Calculated life exp. Of Whites – Blacks	Correlation Coefficient	0.156	-0.064	0.162	0.177	-0.046	0.158	0.219
	Sig. (2-tailed)	0.336	0.695	0.318	0.276	0.778	0.329	0.174
	N	40	40	40	40	40	40	40
Calculated life exp. Of White Male – Black Male	Correlation Coefficient	.412**	-0.181	.416**	.435**	-0.162	.418**	.408**
	Sig. (2-tailed)	0.008	0.265	0.008	0.005	0.317	0.007	0.009
	N	40	40	40	40	40	40	40
Calculated life exp. Of White Female – Black Female	Correlation Coefficient	-0.223	0.227	-0.217	-0.202	0.225	-0.231	-0.093
	Sig. (2-tailed)	0.166	0.160	0.178	0.211	0.163	0.152	0.567
	N	40	40	40	40	40	40	40

Note: *Correlation for columns marked with ** is significant at the 0.01 level (2-tailed) and for those with * is significant at the 0.05 level (2-tailed). The last column contains the rate of all racist tweets in this six-month of 2020 divided by the sum of all tweets published over the mentioned period.*

Table 0.7 Spearman's rho statistical analysis for rate of BLM and Anti-BLM tweets regarding gender, compare to "life expectations," "Poverty Rates," "Educational Attainments," and "Race Composition" statistics.

		Rate of all BLM tweets (BLM / Sum)	Rate of BLM Male	Rate of BLM Female	Rate of Anti-BLM Male	Rate of Anti-BLM Female
Black expectation of life at age 0	Correlation Coefficient	.357*	.366*	0.287	0.245	0.107
	Sig. (2-tailed)	0.024	0.020	0.073	0.127	0.513
	N	40	40	40	40	40
Black male expectation of life at age 0	Correlation Coefficient	.347*	.391*	0.287	.338*	0.230
	Sig. (2-tailed)	0.028	0.013	0.073	0.033	0.154
	N	40	40	40	40	40
Calculated life exp. Of White Male – Black Male	Correlation Coefficient	-0.164	-0.178	-0.103	-0.231	-0.251
	Sig. (2-tailed)	0.311	0.272	0.526	0.152	0.118
	N	40	40	40	40	40
The composition percentage of White people in this state	Correlation Coefficient	-0.189	0.177	-0.177	.338*	0.140
	Sig. (2-tailed)	0.180	0.210	0.210	0.014	0.322
	N	52	52	52	52	52
The composition percentage of Black people in this state	Correlation Coefficient	-0.165	-.308*	-0.106	-.307*	-0.192
	Sig. (2-tailed)	0.243	0.026	0.457	0.027	0.174
	N	52	52	52	52	52
2014 Poverty Rates (includes unrelated children)	Correlation Coefficient	0.067	-0.111	0.083	-0.026	0.026
	Sig. (2-tailed)	0.639	0.434	0.561	0.855	0.856
	N	52	52	52	52	52
Supplemental Poverty Measure (2010–2014 average)	Correlation Coefficient	0.204	-0.004	0.225	0.003	0.048
	Sig. (2-tailed)	0.147	0.978	0.109	0.983	0.735
	N	52	52	52	52	52
High school graduate or higher	Correlation Coefficient	0.035	0.136	-0.005	0.094	-0.049
	Sig. (2-tailed)	0.804	0.335	0.972	0.508	0.730
	N	52	52	52	52	52
Bachelor's degree or higher	Correlation Coefficient	.367**	.327*	.275*	0.156	-0.157
	Sig. (2-tailed)	0.008	0.018	0.048	0.270	0.266
	N	52	52	52	52	52
Advanced degree	Correlation Coefficient	.444**	.440**	.366**	0.184	-0.069
	Sig. (2-tailed)	0.001	0.001	0.008	0.191	0.626
	N	52	52	52	52	52

5.4.2 Correlation with Poverty Rate

Looking at the poverty rates tables, the outcomes even apprised more prominently—the racist rates are positively correlated with "2014 Poverty Rates" and "Supplemental Poverty Measure" tables. In other words, there is a link that shows if the greater the number of Black people are suffering from racism content in a state, the more poverty rates is there in that state too. This rate has a significant correlation with P-value < 0.01 by the Spearman rho's correlation coefficient of 0.582 with "Supplemental Poverty Measure," and at P-value < 0.01 level by the coefficient of 0.460 with "2014 Poverty Rates", as shown in Table 5.8, results are highlighted in the last column. Other descendant variables like the number of racist tweets published by males, both male and female racist tweets before the event, and the number of racist tweets published by females after the event, all have a positive correlation with a p-value of < 0.01 and a Correlation Coefficient of 0.561, 0.566, 0.602, and 0.578 respectively with the "Supplemental Poverty Measure" statistics in each state (Table 5.8).

Table 0.8 Spearman's rho statistical analysis for percentage of Racist tweets compare to "Poverty Rates," "Educational Attainments," and "Percentage of Race Composition" in each state.

		Number of Racist tweets Male	Number of Racist tweets Female	Racist tweets Male Before event	Racist tweets Female Before event	Racist tweets Male After event	Racist tweets Female After event	Racist rates (Racist / Sum)
Composition of White people	Correlation Coefficient	-.578**	0.135	-.578**	-.610**	0.125	-.603**	-.665**
	Sig. (2-tailed)	0.000	0.341	0.000	0.000	0.376	0.000	0.000
	N	52	52	52	52	52	52	52
Composition of Black people	Correlation Coefficient	.711**	-0.060	.711**	.747**	-0.039	.736**	.848**
	Sig. (2-tailed)	0.000	0.673	0.000	0.000	0.785	0.000	0.000
	N	52	52	52	52	52	52	52
Composition of Native people	Correlation Coefficient	-.316*	-0.008	-.318*	-.325*	0.022	-.349*	-.423**
	Sig. (2-tailed)	0.022	0.954	0.022	0.019	0.876	0.011	0.002
	N	52	52	52	52	52	52	52
2018 Poverty rate	Correlation Coefficient	0.233	0.028	0.232	0.270	0.073	0.242	.304*
	Sig. (2-tailed)	0.096	0.846	0.098	0.053	0.606	0.084	0.029
	N	52	52	52	52	52	52	52
2014 Poverty Rates	Correlation Coefficient	.415**	0.070	.417**	.456**	0.092	.414**	.460**
	Sig. (2-tailed)	0.002	0.623	0.002	0.001	0.518	0.002	0.001
	N	52	52	52	52	52	52	52
Supplemental Poverty Measure	Correlation Coefficient	.561**	0.054	.566**	.602**	0.066	.578**	.582**
	Sig. (2-tailed)	0.000	0.703	0.000	0.000	0.641	0.000	0.000
	N	52	52	52	52	52	52	52
High school graduate or higher	Correlation Coefficient	-.530**	0.042	-.530**	-.556**	0.015	-.529**	-.573**
	Sig. (2-tailed)	0.000	0.769	0.000	0.000	0.916	0.000	0.000
	N	52	52	52	52	52	52	52
Bachelor's degree or higher	Correlation Coefficient	0.085	0.137	0.080	0.065	0.100	0.099	-0.086
	Sig. (2-tailed)	0.550	0.331	0.574	0.649	0.482	0.483	0.545
	N	52	52	52	52	52	52	52
Advanced degree	Correlation Coefficient	0.242	0.172	0.239	0.232	0.135	0.266	0.104
	Sig. (2-tailed)	0.084	0.224	0.087	0.099	0.340	0.057	0.463
	N	52	52	52	52	52	52	52

Note: *Correlation for columns marked with ** is significant at the 0.01 level (2-tailed) and for those with * is significant at the 0.05 level (2-tailed).*

5.4.3 Correlation with Educational Attainment

The most impressive findings to emerge from the data are how educational attainment is correlated with BLM and Racist groups. The more educated the people of an area are, the more they tend to support the BLM movement, and the less schooling level percentages lead to more racism rates in each state.

Referring to Table 5.7, the rate of BLM tweets calculated by dividing the sum of BLM tweets by the sum of published tweets in each state shows a positive correlation with states with a higher "Bachelor's Degree" or "Advanced Degree" numbers. The same story holds for descendant measures concerning males and females too. It can be put in two similar sentences: 1) Fellows with higher educational levels are more interested in this movement. Alternatively, 2) This movement was supported mostly by fellows with upper educational background.

The related findings show a Spearman's Rho Correlation Coefficient of 0.444 between "Rate of all BLM tweets" and "Advanced Degrees" educational level with a p-value of < 0.01 for 52 states. Similarly, we have the Coefficient of 0.367 between the same factor and the "Bachelor's Degree or Higher" variable, attaining a p-value of 0.008.

On the other side of the coin, there is an adverse correlation between the number of Racists and "High School Graduate or Higher" statistics. It means the higher the rate of high school graduates in each state, there is lower racist content posted in that area. A p-value of < 0.01 was obtained in this connection and the correlation coefficient of -0.573 for all Racist tweets' principal rate, as shown in the last column on Table 5.8. The theme identified for this one has similar recurring results for the descendant factors too. For instance, a p-value of < 0.01 was obtained for the published racist tweets by males and females before the event with the correlation coefficient of -0.530 and -0.556, respectively.

5.4.4 Correlation with Race Composition

Compared to the population compositions, the correlation with two Racist and Anti-BLM groups shows an outstanding sad novel finding. The rates of racist tweets are higher in areas with a higher number of black population and are lower in states having supremacy of white people. The Anti-BLM tweets, though, are more in states with more leading white compositions and lower in those with larger black residents.

With reference to Table 5.8, the statistics for racist groups in correlation with population compositions show a strong relationship with the population of black and white people. The analysis with Spearman's rho method shows a significant positive coefficient of 0.848 with the p-value of < 0.01 between all Racist tweet rates and "Composition of Black people", which means in areas where there are more black people living, there are greater number of tweets published with racism contents. On the contrary, states with a broader community of whites and natives tend to be more impartial toward racism, having negative correlation coefficients of -0.665 and -0.423 and p-values of < 0.01. Again, most descendant variables have identical rates.

Closer inspection of Table 5.7 shows an apparent correlation coefficient value between Anti-BLM tweets published by males and black and white residents' demographic determinants. In this connection, a p-value of < 0.05 is obtained with a positive coefficient of 0.338 toward "the composition percentage of White people in each state" and a negative coefficient of -0.307 with a p-value of < 0.05 getting significant with "Black people" population rates.

5.5 Analysis Based on the Sentiment Outcomes

Further, the user sentiments were calculated for three goals:

1) To see the associated percentages of sentiment polarity and subjectivity for each group

2) To see if these percentage rates are different regarding the event

3) To compare them with the "Life Expectancy," "Poverty Rate," "Educational Attainment," and "Race Composition" tables again and double-check their relationship with these statistics.

Results and figures regarding each are discussed in the following subsections.

5.5.1 Sentiment Details for Each Group

As described in the previous chapter, the process analysis of sentiment was done to identify tweets' expression polarity and subjectivity. Sentiment analysis applies computational examination of feelings, emotions, and attitudes in a written document. This computational process results in two decimal numbers for each tweet that are categorized as "negative," "neutral," or "positive" for the polarity score and "subjective," or "objective" for the subjectivity score. These categorizations show the users' attitude towards this particular topic, representing the intensity of sentiments expressed in the content.

By the first observation, it could be drawn that most tweets are written more neutrally and objectively in the three groups. Most polarity scores are neural, meaning most tweets are composed of neutral sentiment in the dataset. In a general sense, positive sentiment is usually equated to the words expressing happiness, joy, kindness, and similar sensations of this kind. The negative sentiment is those sentences that represent feelings like sadness, hate, brutality,

discrimination. Neutral sentiment then indicates more challenging literature to be identified as one of these two categories, and no significant emotion is detected in the text.

The bar chart in Figure 5.6 summarizes polarity percentages for each group in this study. The results show the highest negative rate among the racialized victim group of BLM, the most neutral one for Anti-BLM, and the most positive proportion for the racist.

Although all groups have at least 52 percent of the tweets lying on the neural side, the remaining portion's balance is different. Whereas the burden for the rest takes part in negative polarity for the BLM group with 36.86%, and 25.07% positive ones for racist groups. It seems that fellows who tend to argue the BLM movement with their tweets preferred to write it in a more neutral context, while both BLM and Anti-BLM groups have similar percentages of positive tweets with 14.56% and 13.02%, respectively. The Anti-BLM group shares a similar negative polarity rate with the racist group with 21.78% and 22.07% accordingly. However, what is highlighted is the uppermost rate of negative sentiment in tweets defending black people.

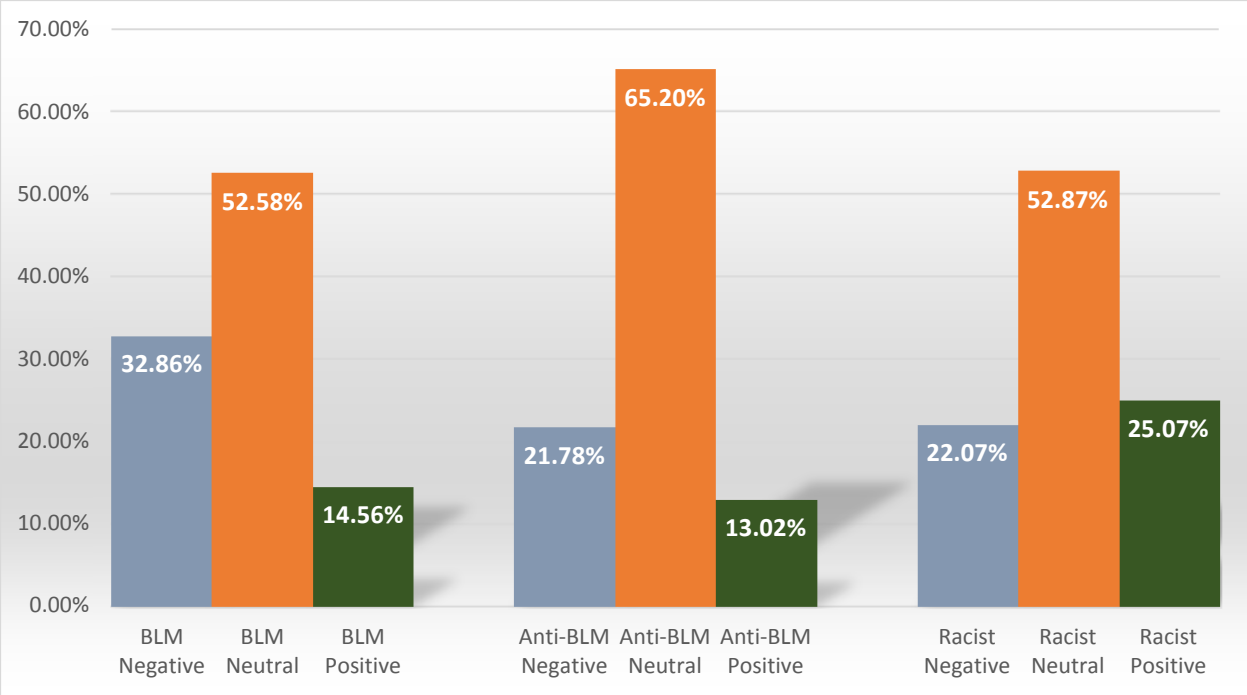

Figure 0.6 Comparing sentiment polarity percentages for each group.

Further, each group's subjectivity rates are shown in Figure 5.7 with similar rates for the BLM and Anti-BLM groups, yet with a different one for racists. The first two groups' rates are identical, around 80% objective writing and the rest ~20% for subjective tweets. In contrast, the racist group tends to be more subjective in their speech, with a rate of 34.17% in this manner.

A subjective statement is mostly a personal idea, belief, or opinion toward something that cannot be determined as right or wrong most of the time. On the other hand, the objective one is mostly a fact, concept, or evidence that some methods can verify its accuracy. Considering the meaning of the two terms of "subjectivity" and "objectivity" [147], it can imply that the BLM group and its direct opponent tend to write their messages in a factual manner, which can be declared a true or false judgment. On the other hand, a more subjective rate is noted for the racist groups, which means there are more personal expressions of belief, opinion, or feelings in these records.

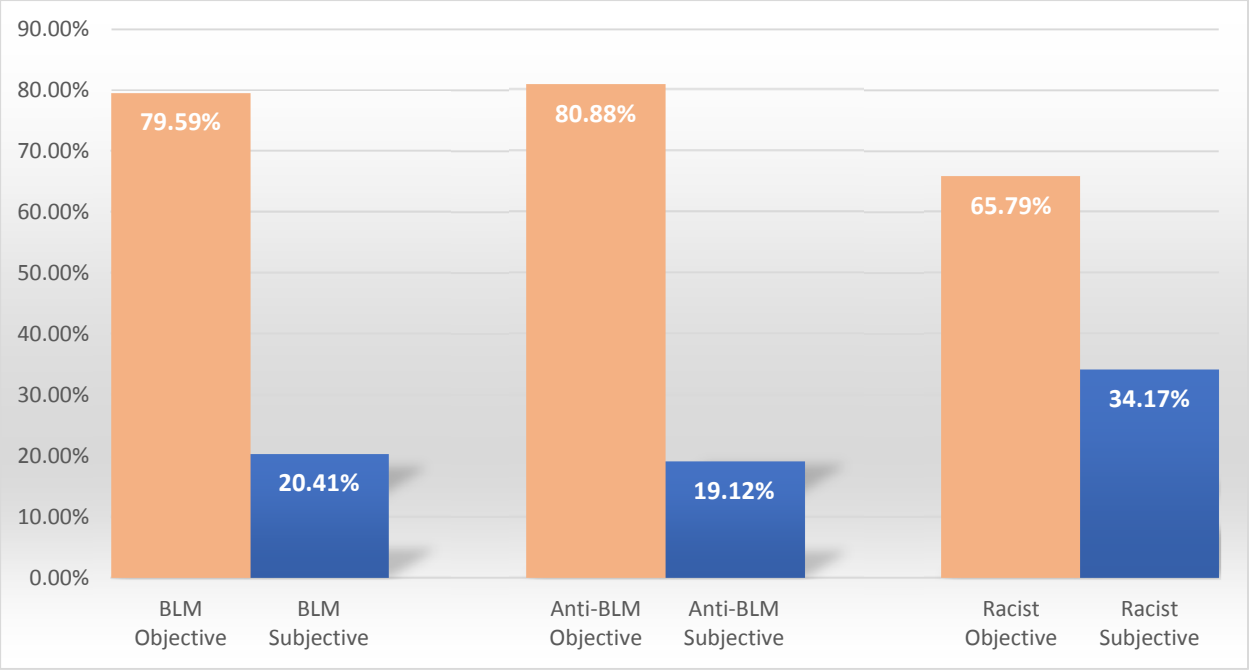

Figure 0.7 Sentiment subjectivity percentages for each group.

Figures 5.9, 5.10, 5.11, and 5.12 are scatter plots showing the four groups' sentiment polarity and subjectivity scores. Scatter plots are a graph known for diagraming the correlation. The scatter diagram is usually applied to show the covariance and correlation among two variables. This chart helps outline how similarly the two factors are related.

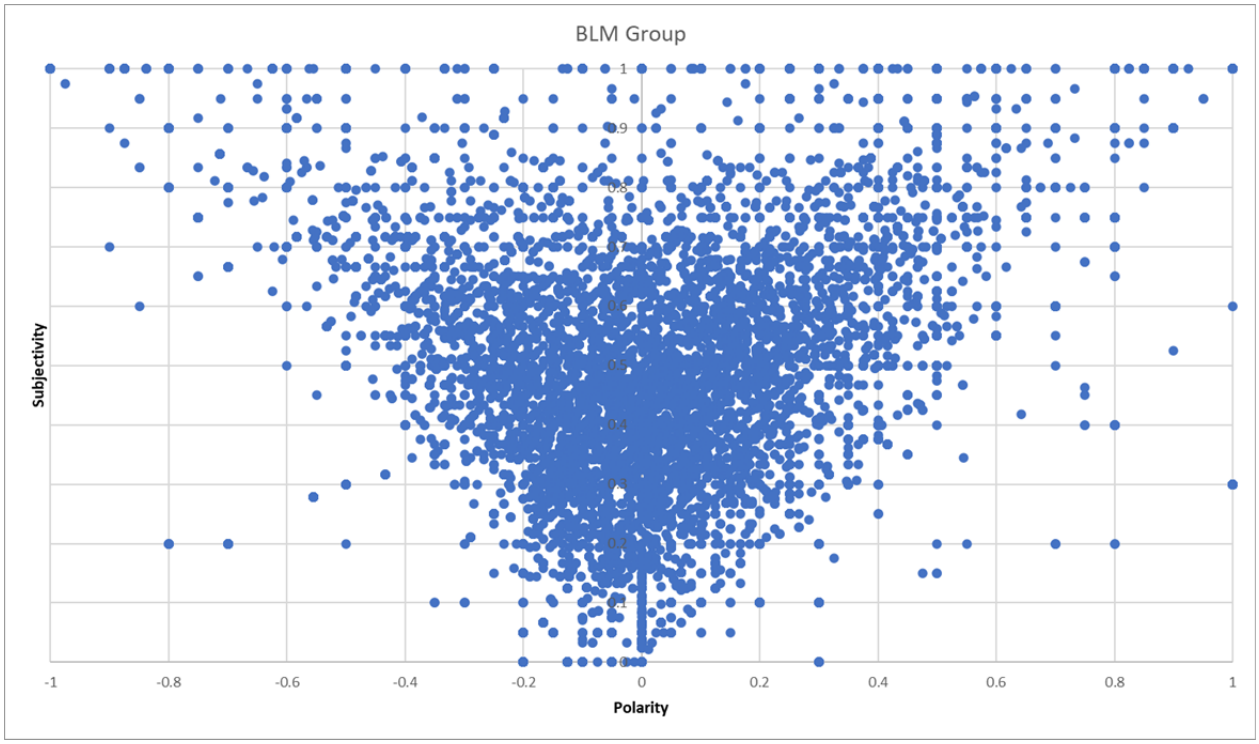

Figure 0.8 Scatter plot of sentiment polarity and subjectivity scores for BLM tweets.

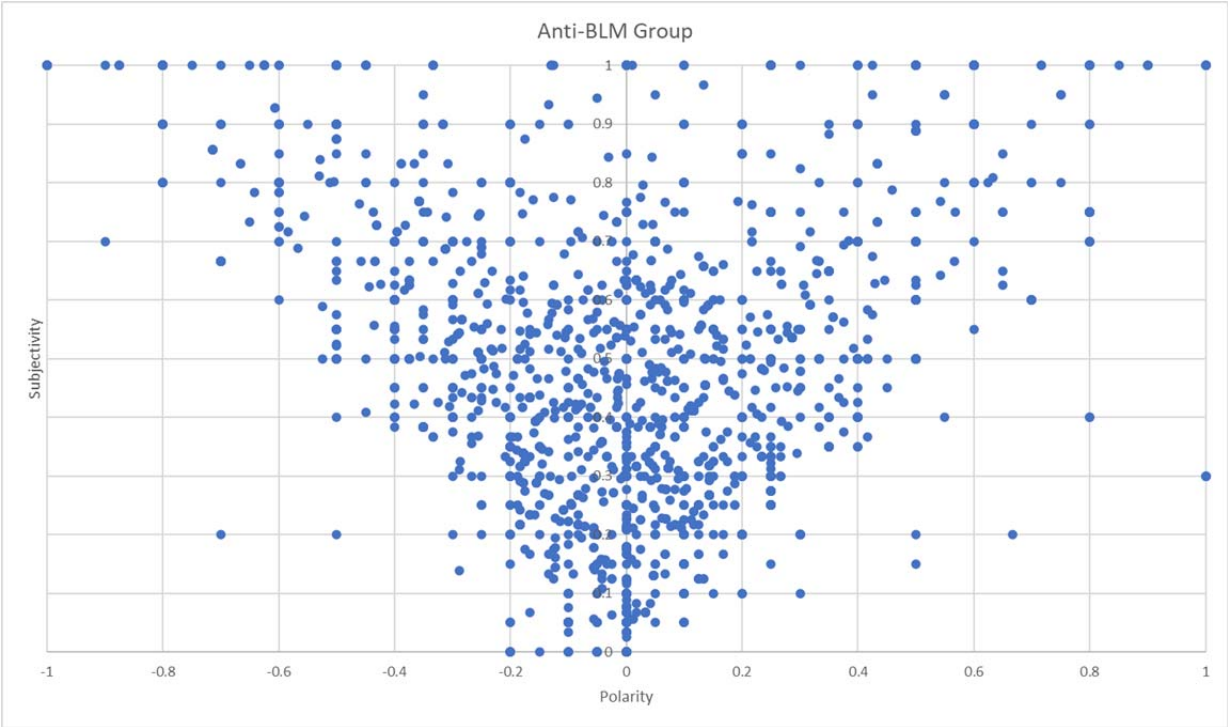

Figure 0.9 Scatter plot of sentiment polarity and subjectivity scores for Anti-BLM tweets.

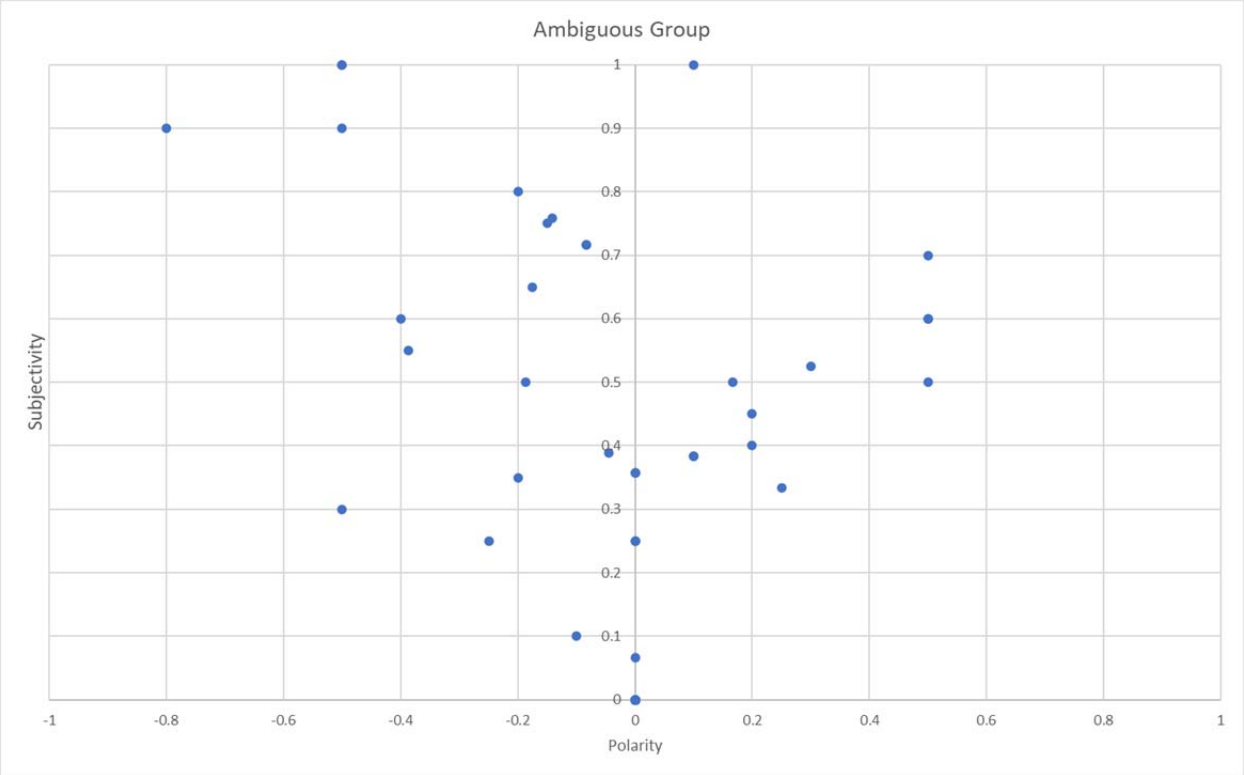

Figure 0.10 Scatter plot of sentiment polarity and subjectivity scores for ambiguous tweets.

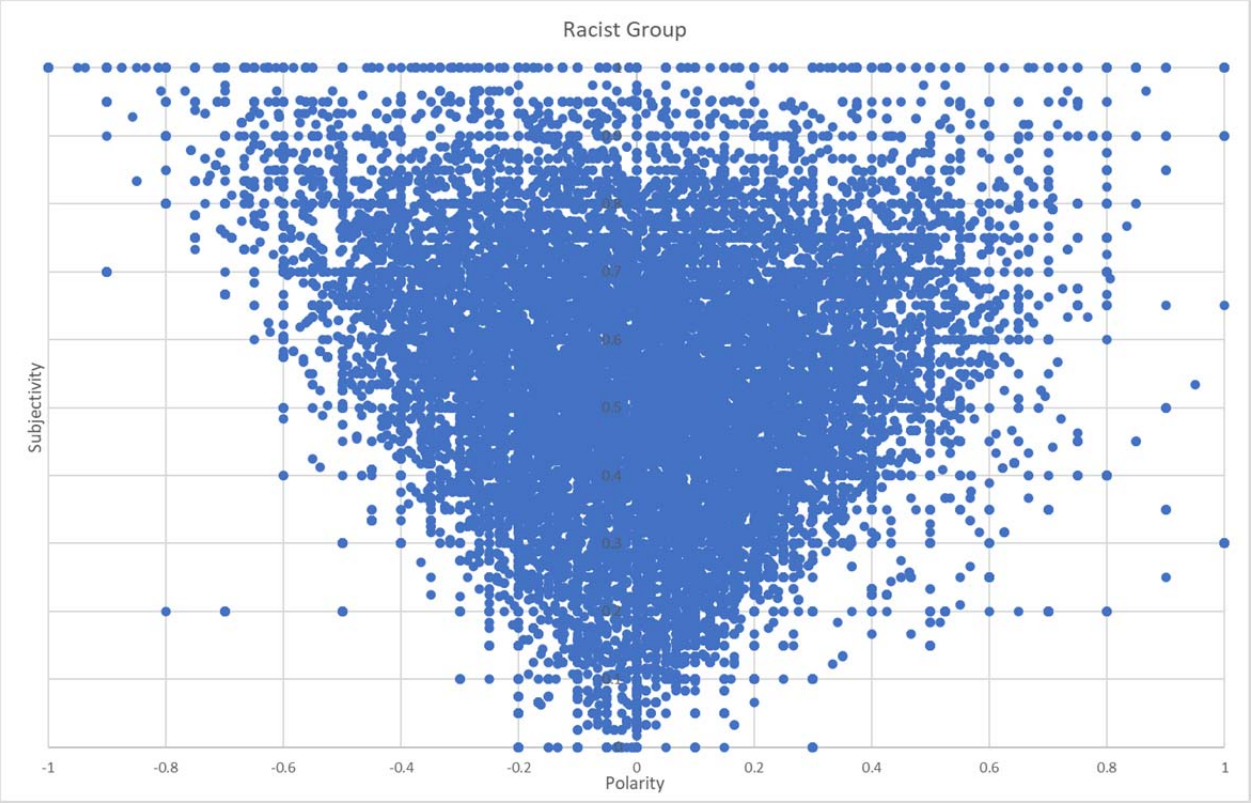

Figure 0.11 Scatter plot of sentiment polarity and subjectivity scores for racist tweets.

In Figures 5.12 and 5.13, this sentiment correlation is compared by considering gender. These is almost no difference between the sentiment polarity and subjectivity of men and women. The percentages were calculated to make sure this judgement is right. For the BLM group, the women percentages of negative, neutral, positive polarities were 32.1%, 53.4%, and 14.5% accordingly, while the men also showed similar values of 33%, 52.1%, and 14.9%. A similar story is observed concerning the racist group, while men seem to have a bit more neutral rate. The numbers are 22.8%, 51%, and 26.2% for women compared to 21.3%, 55.3%, and 23.4% for men regarding the negative, neutral, positive polarities, respectively.

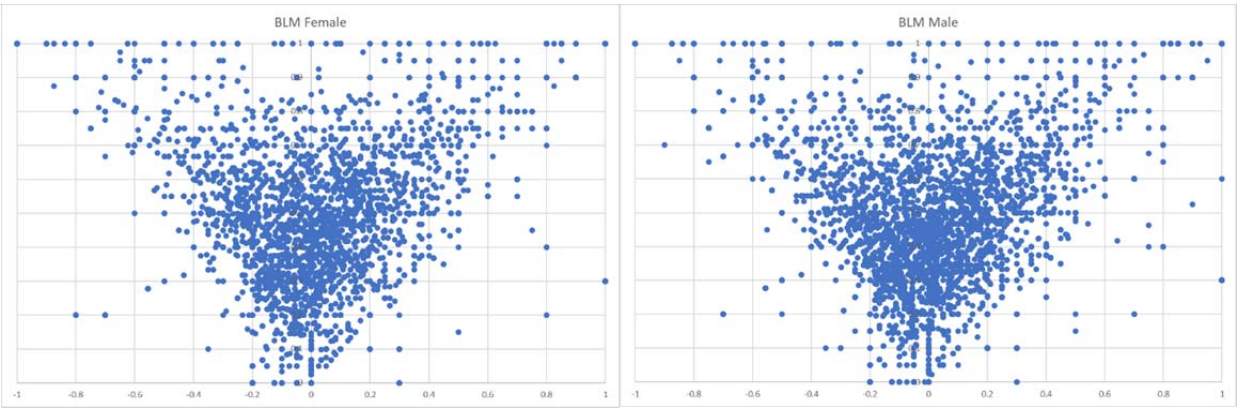

Figure 0.12 Scatter plot of sentiment polarity and subjectivity scores comparing male and female for BLM tweets.

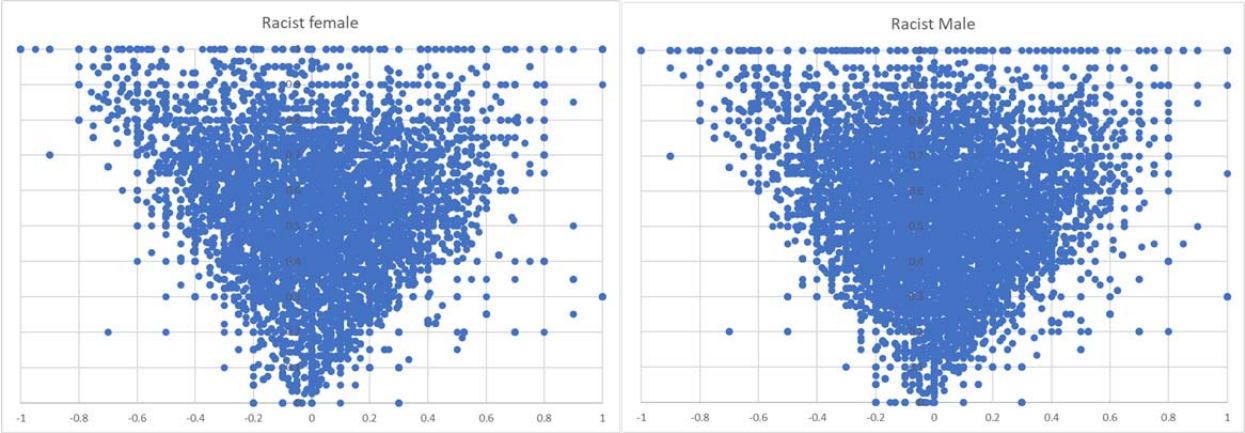

Figure 0.13 Scatter plot of sentiment polarity and subjectivity scores comparing male and female for racist tweets.

5.5.2 Sentiment Details Regarding the Event

In this section, the polarity and subjectivity of the sentiment results have been compared for the main groups. Figure 5.14 depicts the whole polarity rates in one bar chart. The first two BLM

and Anti-BLM groups have some fluctuations, yet the racist group seems to stay with small variations in these rates.

Regarding the BLM group, it has some reduction in negative messages, which decreased by ~14% from 46.17% to 32.66%, giving its place mostly to neutrals contents. Hence, unexpectedly the neutral content raised from 41.59% to 52.74%, and even positive polarities had a small 2.35% increase, comparing the before and after the George Floyd killing event.

Contrary to the BLM, the Anti-BLM group had an equal number in increasing the negative polarity tweets. The negative rates grew to 23.56% from 10.61% after the mentioned event. This negative tweets' rising of 12.95% took place instead of neutral content that decreased from 71.89% to 64.13% and even a slight decrease in positive ones from 17.50% to 12.31%.

In terms of the racist group, rates seem to stay almost the same before and after the event. There is only a small gain in neutral tweets and an equal decrease in positive ones.

Overall, most groups before and after the event keep having almost half of the messages in a neural sense. For negative tweets, the story is different; this proportion for the BLM group is more than double of the racist group regarding the time before the event. For positive ones, the racist group has double rates compare to the BLM group.

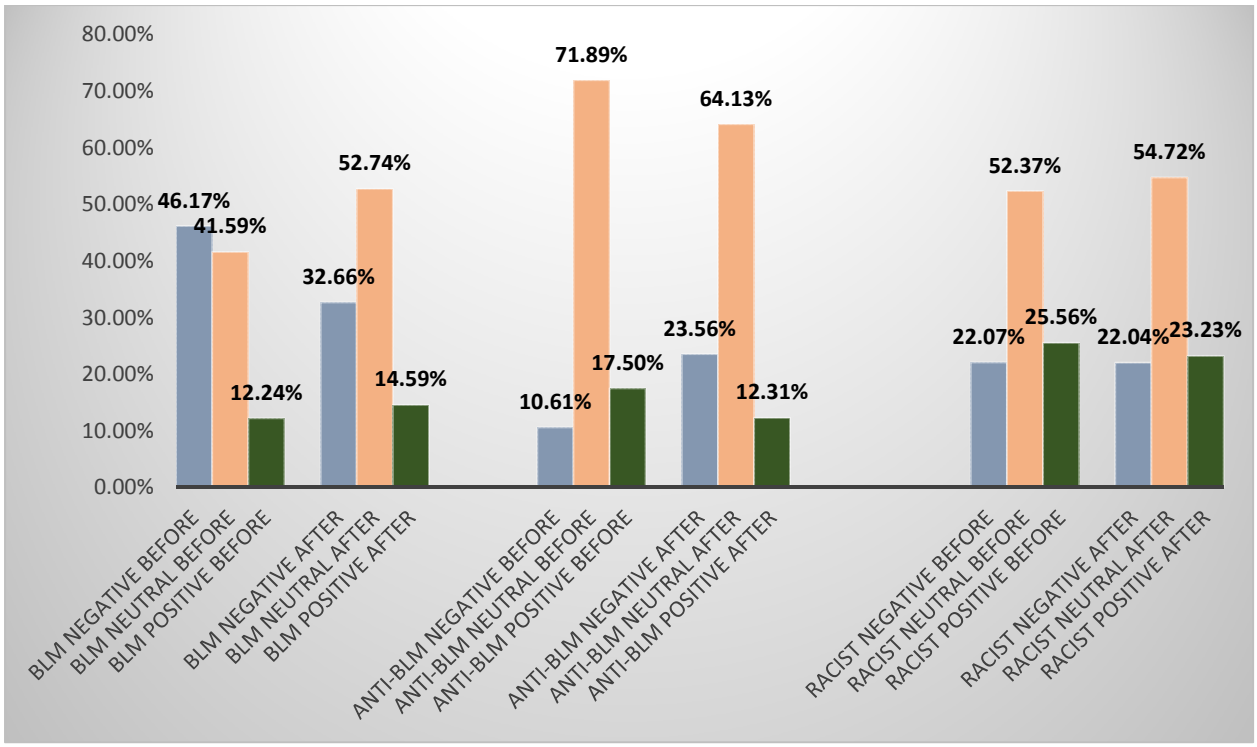

Figure 0.14 Comparing the sentiment polarity percentages before and after the event for each group.

In Figures 5.15 and 5.16, the BLM and racist groups' subjectivity scores are compared before and after the event. Both have an increase in objective contents taking the place of subjective ones. While the BLM group's rates of objectivity are upper, it also had a higher growth after the event with 5.03%. In this matter, the racist group had a minor 0.63% increase for objective contents.

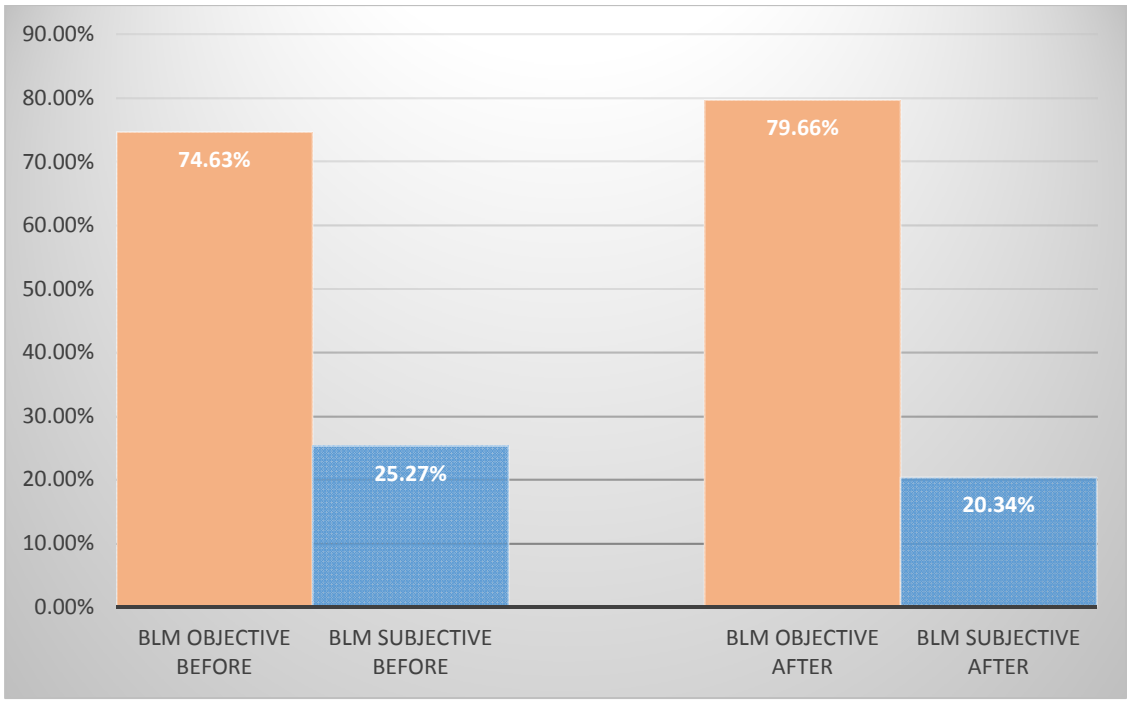

Figure 0.15 Comparing sentiment subjectivity percentages for BLM group regarding the before and after the event.

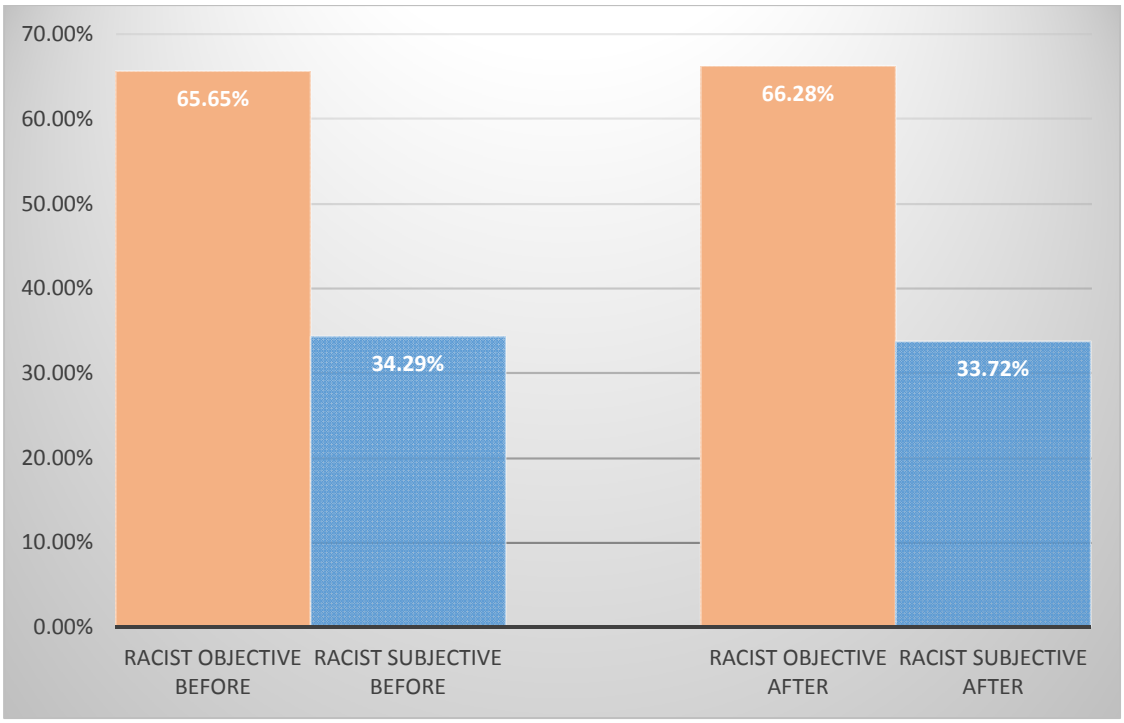

Figure 0.16 Comparing sentiment subjectivity percentages for Racist group regarding the before and after the event.

5.5.3 Sentiment Compared to Tables

Here, the negative and positive polarity rates of racist tweets are compared to the statistics tables to see the outcomes. The findings show equal treatment as before regarding the correlation coefficients with implemented tables in this study conducted in Table 5.9 and 5.10. Moreover, what is remarkable is that the rates associated with the positive polarity of racist contents are even having higher Spearman's rho correlation coefficient number compare to the negative ones regarding life expectancy tables. We have the same story in reaching those who published their racist tweets by having a positive sense of it toward the population of black people and the poverty rates. It seems this effect is only reverse for the educational rates that resemble logical.

Table 0.9 Spearman's rho statistical analysis for the percentage of racist tweets compared to life expectations tables regarding polarity score.

		Racist Tweets Negative Polarity Before Event	*Racist Tweets Positive Polarity Before Event*	*Rate of Negative Racist Tweets*	*Rate of Positive Racist Tweets*
Calculated life exp. of Whites - Blacks	Correlation Coefficient	0.191	0.172	0.192	0.211
	Sig. (2-tailed)	0.238	0.289	0.235	0.190
	N	40	40	40	40
Calculated life exp. of White Male - Black Male	Correlation Coefficient	.438**	.427**	.349*	.405**
	Sig. (2-tailed)	0.005	0.006	0.027	0.009
	N	40	40	40	40
Calculated life exp. of White Female - Black Female	Correlation Coefficient	-0.186	-0.211	-0.062	-0.117
	Sig. (2-tailed)	0.249	0.190	0.703	0.473
	N	40	40	40	40

Table 0.10 Spearman's rho statistical analysis for the percentage of Racist tweets compare to "Poverty Rates," "Educational Attainments," and "Percentage of Race Composition," considering the polarity scores of sentiment analysis.

		Racist Tweets Negative Polarity Before Event	Racist Tweets Positive Polarity Before Event	Rate of Negative Racist Tweets	Rate of Positive Racist Tweets
Composition of White people	Correlation Coefficient	-.604**	-.601**	-.626**	-.649**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000
	N	52	52	52	52
Composition of Black people	Correlation Coefficient	.740**	.732**	.811**	.838**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000
	N	52	52	52	52
Composition of Native people	Correlation Coefficient	-.332*	-.337*	-.430**	-.418**
	Sig. (2-tailed)	0.016	0.015	0.001	0.002
	N	52	52	52	52
2018 Poverty rate	Correlation Coefficient	0.259	0.241	.322*	.290*
	Sig. (2-tailed)	0.064	0.086	0.020	0.037
	N	52	52	52	52
2014 Poverty Rates	Correlation Coefficient	.456**	.448**	.466**	.471**
	Sig. (2-tailed)	0.001	0.001	0.001	0.000
	N	52	52	52	52
Supplemental Poverty Measure (2010–2014 average)	Correlation Coefficient	.609**	.610**	.596**	.597**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000
	N	52	52	52	52
High school graduate or higher	Correlation Coefficient	-.545**	-.535**	-.580**	-.551**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000
	N	52	52	52	52
Bachelor's degree or higher	Correlation Coefficient	0.072	0.084	-0.124	-0.080
	Sig. (2-tailed)	0.610	0.554	0.380	0.575
	N	52	52	52	52
Advanced degree	Correlation Coefficient	0.247	0.252	0.090	0.114
	Sig. (2-tailed)	0.078	0.071	0.524	0.420
	N	52	52	52	52

5.5.4 Sentiment Polarity Tweet Examples

Here, we are presenting sample tweets categorized as positive, negative and neutral tweets in the sentiment analysis. In racist ones we masked the hate words by three stars.

Table 0.11 Examples of tweet categorization in Sentiment Polarity process.

Polarity	*Group*	*Text*	*Date*
Positive	BLM	#BlackLivesMatterAtSchool leads to such amazing and honest conversations in my classroom. Kids want to talk about tough issues and it's on teachers to face the uncomfortable conversations so that everyone can learn and grow.	Feb 4
Negative	BLM	It's so crazy to me that one of my brothers could be running the neighborhood and end up being killed.. smh #BlackLivesMatter	May 20
Neutral	BLM	This is leadership in the community! #BlackLivesMatter	Jan 20
Positive	AntiBLM	all lives matter is trending. hilarious.	May 27
Negative	AntiBLM	Why would you use this moment to talk about how all lives matter, it's disrespectful and disgusting.	May 27
Neutral	AntiBLM	Vote Blue. Republicans need to be checked. All lives matter over and above money.	Apr 2
Positive	Racist	😂 it really depends on how it go down. *** can get trained too though...	Jun 24
Negative	Racist	Im not about to let no ***/ *** waste my time	May 27
Neutral	Racist	What this *** do now?	May 26

5.6 Supplementary Results

5.6.1 Word Cloud

Word clouds are an extensive visual portrayal of the recurrence of words in a context. It is a presentation of words in an image like shape with a simple rule; Words with more frequency will appear more prominent in this pool of phrases. Hence, the more a specific word occurs in a document, the bold and bigger it will resemble the word cloud.

In this section, we figured a depiction of words with the most frequency in each group. Nevertheless, as we were not interested in representing the offensive words of racist content, we put the other two groups in Appendix 4. Figure 0.17 depicts the word cloud for the BLM movement group representing the words with maximum frequencies. The two other images in Appendix 4 depicts the Anti-BLM, and Racist groups in the same manner.

Figure 0.17 Word Cloud of BLM group tweets.

5.6.2 Emojis

As a side part of the study, the top emojis used for the primary two groups were also measured and is available in Appendix 5. The apparent outcome of this examination was high usage of "broken heart" 💔 and "raised fist (black color)" ✊🏿 in those supporting the Black Lives Matter movement and a high volume of "skull" 💀 emoji in racist tweets. However, as there were other emojis in this content like "face with tears of joy," 😂 which seems to be a frequent one used in many tweets in different contents, this relation was calculated by a statistical method.

The data was converted into binary variables and multiple new binary ones were created indicating each emoji occurrence by 0 or 1. Next, the Pearson Chi-Square test was used on these variables and calculated them by IBM SPSS Statistics. A P-value of 0.000 was obtained for all these three cases mentioned with Phi and Cramer's V of 0.142946 for the correlation of Feast Emoji with BLM group, the number of 0.028306 for the correlation of Broken Heart Emoji with BLM group, and the outcome of 0.046827 for the Skull Emoji in relation with Racist group.

Hence, as expected, these outcomes signify that there a small but influential relation between using these emojis and the tweets posted on these groups. We compared the success rates between these metrics and found that the association is statistically significant. We considered Chi-square test results with $P < 0.05$ as statistically significant, while the outcomes were all zero. Having these results, the probability of obtaining such emojis in similar contents might be somehow small yet clearly reveals its contingency and shows that the observation of these emojis follows an explicit mathematical distribution.

5.7 Discussion: Summary of results

In this thesis, we studied the hashtag "BlackLivesMatter," and some of its adversary contentions regarding George Floyd's demise in 2020 on Twitter. Based on the extensive aftermath protests in the United States, we considered an area analysis to compare tweet rates in different groups to some previously studied health statistics.

This study's groups were defined in reference to the hashtag or words used in tweets' content. The first group was BLM comprising tweets endorsing the "BlackLivesMatter" movement. Following was the Anti-BLM group held tweets who encountered this movement concurrently by their opposed hashtags like "AllLivesMatter" or "BlueLivesMatter." Next, we had a small minority who embraced both of these criteria. Subsequently, the majority of our data was the Racist group, including users who posted texts with offensive n-words toward the black community.

The timeline of the tweets collected for this study was the first six months of 2020, regarding the availability of data, and the time of massive protests in the United States of America, defending black lives affairs. The dataset mined for this thesis has 307k records mined out of 43 million tweets with USA location data in this timeframe.

These data are categorized into four groups of BLM, Anti-BLM, Ambiguous, and Racist with around 69K, 7.6K, 267, and 230K records, respectively. The first analysis shows a sharp increase in tweet numbers for the first two groups regarding George Floyd's death after May 25, 2020, and a slight decrease in the racist content during the year.

Once Tweets were located and marked for each area, we calculated the rate variations for all states. Prior to initiating the analysis, some decent health tables were obtained from the previous

studies. These tables are (1) National "Life Expectancy" by the "National Center for Health Statistics" with data based on the black and white people. (2), (3) "Poverty Rates" and "Educational Attainment" in reference to each state derived from the "United States Census Bureau" which preserves regular surveys named "American Community Survey (ACS)" to gather this information and (4) the "Race Composition" statistics by the "world population review."

In summary, the findings in analysis with these tables showed a positive connection between the racist tweets' rate and the life expectancy of black males minus white males. Similar outcomes showed up for the rate of BLM tweets with Black's life expectations, which can imply that there is a connection between the number of BLM tweets published in each area and their health. It was slightly surprising that the rates of racist tweets extracted by our criteria – including offensive n-words – were noteworthily correlated with poverty measures numbers of each area. Another unanticipated finding is how the educational attainment links to the rates of BLM and Racist tweets in each state.

There was a direct connection between the upper percentages of bachelor or advanced degrees and the rates of support for the BLM hashtag regarding geography. On the other side, racism rates were also lower for states with more high school graduates. The population composition tables even revealed disconsolate outcomes that Racist and Anti-BLM tweets directly correlate with the black population. Therefore, it can be assumed that blacks are more prone to racism wherever they shaped a larger community.

5.8 Discussion: Why we studied “Black Lives Matter” with Respect to George Floyd

This study's findings show a significant relationship between online colour-based contents and some physical world indicators. It proves that these cyber-discriminations have real-world adverse link with the black community's mattering lives.

The year 2020 was a challenging year known mostly as the pandemic year. However, the notable event of George Floyd's killing broke many humans' hearts and made them protest on social media and streets as well. In this study, we analyzed the #BlackLivesMatter movement and Racism rates in reference to the injustice killing of this black man by a police officer. Then we compared these rates to the Life Expectancy, Poverty, Educational Attainment, and Race Compositions regarding the tweets published in the USA.

The tragedy of George Floyd's demise in 2020 was a justice failure that liberally impacted the number of related posts to anti-racism activity in the cyber world. He was an innocent black man killed by an American police officer on the street, coinciding with the murder of another black man in Atlanta, which attracted a lot of attention. These two events proved to everyone that racism still exists in the United States and has not disappeared.

The act of battling against racism with the idea of #BlackLivesMatter started in late 2012 as a movement, yet this event triggered it again. Racism is a hidden pain in our society, which is reflected again as a social media trend in this challenging year.

We believe that science and research should solve or reflect real-world problems; that is why we considered this event and racism in this thesis.

Now, this hashtag is a social movement to reflect endeavors, and we employed it in our research to present the highlights of these conversations compared to other statistical measures. The study's outcomes clearly show the changes regarding this event, while racist behaviors are still out there and a community is suffering.

5.9 Discussion: Why black males are more impacted?

The findings of this study considerably have a higher significance level when the results are focused on males. The primary outcome shows that the life expectancy of men, which are lower in default compared to black women, has a close connection to the rates of racism contents. It means the psychological effects of racism impact men more than the way it affects women.

However, to see why the findings concluded in this way, we considered the previous similar studies to perceive if they observe an identical outcome. Earlier surveys revealed that racial discriminations are the origin of physical health issues for black people, yet the effects for males go further [148]. It is widely known that depression, stress, and suicide, over and above the medical complaints, are the underlying causes of being exposed to racism. However, this status for men is far beyond be strangled by a police officer. This section will reference some previous similar findings that obtained identical outcomes on the further influence of racism on men compared to women.

In a study [149], scientists found that a “sense of mastery” can help black women reducing their stress, while the results were insignificant for men. Sense of mastery is a term referred to as the

ability of self-control on challenging life issues, whereas mastering this ability on men could not reflect the same rates as women in overcoming emotional distress.

In line with another study [150], scholars who sought to discover how educational attainment can defeat depressions concluded that women had conquered this battle better than men. They investigated 3.5k black adults' records to uncover how superior education is bound up with lowering psychological distress. This study also showed that women are more kept from the harms of these depressive symptoms apart from men. In other words, black men are more prone to the risks of depression, even with high educational achievement.

Overall, black men manifest in more severe forms of color separations that strikes their health [151]. They are unjustifiably shot or killed by police beyond black women [152], and black youth experience more notable prejudiced treatment annually [153]. Social inequalities accelerated in suicide, homicide, or sickness considerably more eminent than women's rates [154]. Finally, as a consequence of all these black men's life expectancy at birth is far less than other groups such as white men, black women, or white women [65].

5.10 Discussion: The Imputation of Social Media and Real-World

Our findings showed a strong correlation between social media content and real-world metrics. Similarly, Kolliakou et al. [155] intended to study if the daily posts on Twitter affect daily changes in mental health crisis incidents. They focused on finding the connection between everyday fluctuations of mental-related posts on Twitter and variations in mental health issues. They published the findings of this time-series regression analysis in Nature Scientific Reports in

February 2020. This study's outcomes show a positive association between the quantities of depression-related tweets and the number of same-day crisis episodes in two mental centers.

Unfortunately, we also observed that some users used the n-words for jokes or ridicule. The platform should label this inappropriate behavior as it might have a devastating effect on racialized communities' mental health.

We could infer how social media can have devastating effects on people's mental health. These effects can be even more destructive, especially if these individuals are susceptible to criticism, such as the racialized community of blacks or those with low self-esteem. There is a direct link between statements made on social media and self-criticism in sensitive people. Another research [156] has shown that these manifestations in the masses could even lead to suicide and self-destruction.

Some nations have their own principles and culture and can be collectively very sensitive to an issue, while other nationalities do not pay much attention to that. Therefore, publishing offensive content on social networking sites without being aware of other people and nations' sensitivities can hurt those people.

Overall, although social media emerge to make the world a better place in terms of communication, this research, parallel to the previous ones, implies that there are also adverse effects on some people's wellbeing.

5.11 A talk to policy makers

This study's combination of findings raises intriguing subjects of how higher education and wealth can reduce racism. Black lives must be mattered, and those in authority are requested to consider their health as a priority.

Before other types of racism is the institutional one that could allow any person in some position to enforce their color interest in their administrative domain. Moreover, for the rest types of racism, we expect policymakers to widen public education and instruction to inform how these behaviors can diminish a community's health. A small talk to the racist would also be to ask them to have a good look into a mirror before calling anybody by an offensive word.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis aimed to determine how racism content toward Black people is correlated with health, poverty, and education. We shaped a criterion in relation to the unfair death of a black man – George Floyd – to mine inside the published tweets and extract them in a dataset with proper classification of supporting and opponent groups to this story. We labeled dataset records with their area name, and gender values were also added by implementing some techniques as these two were impactful variables in our study. One of the potential points of this research was the extensive number of tweets gathered to make the accuracy rates more eminent.

BlackLivesMatter is a hashtag used from 2013, yet the social publishing rates seem to fluctuate during these years. This thesis has provided a more in-depth insight into how Twitter users reacted to support this movement with their amiable short messages. These rates showed that blacks are more welcome to be endorsed by areas with higher bachelor and advanced degrees. This outcome can imply that the academic family is among the closest supporters of this black rights movement. Hence, the science community is fully conscious that it is time to end the long-lasting color inequalities in this world to enhance their health.

On the other side of the coin, racist content continues to be publishing on social media, which indicated there are some racist conversations and discrimination in the real world out of cyberspace. This study showed how these discussions impact the health of blacks. When they go up, the life expectancy of black men decreases, and when they decline, we have a healthier,

wealthier, and educated population. We also employed the composition, and it was sad to see the racism rates are higher in areas with more black population.

The human being is social, and to make this being easier, we need to improve the lives of all people, and of course, blacks in priority as they are the topmost racialized fellows. Blacks are suffering from multiple psychological and physical issues because of some other part of the population's desire. They are exposed to these dangers and are vulnerable to health reductions. Racial ignorance is a terrible risk for all around the world's health, and not just blacks or black men.

Human life is precious, no matter what color it has. Racism is disastrous for the health of each and every member living in this globe, not just blacks or black men. The emotional impacts of social inequalities manifest in everyone's life adversely. Everyone will have a diminished lifetime in a community with racial ignorance cultivation. This is somehow because this culture of inequalities acts as a contagious psychological disease and contaminate all members.

The year 2020 was not just the COVID-19 pandemic; it was also the year for mankind to yell for the rights of part of its silent body. A black man's death in police's hands shed light on old distress for the world to see a group's pains. To see how there are suffering institutional racism in many aspects. To see the problem of this new decade is not just the outbreak rates, economic crises, or elections, but also systemic racism. This is why citizens decided to change this moment into a movement.

Social media can have devastating effects on racialized people's mental health. These effects can be even more destructive if these individuals are black people experiencing discrimination in the real world, too.

Blacks are susceptible to social inequalities, as this study showed the direct link between racism and their life expectancy. Some previous research has shown how these disparities could lead to preterm birth, psychological issues, or self-destruction. Some users who are distributing in this language might be unaware of how these conversations impact some other users' mental lives.

Overall, the most prominent finding to emerge from this study is that the conversation discussed on Twitter can be imputed as a sample speech of the offline world and directly influences black people's lives when going nasty. This study showed that racism content is associated with the lower life expectancy of Black men that could be due to its psychological consequences. The analysis also showed a strong association in areas with immense racism content and the more poverty rates. We also found that higher education, is correlated with lower offensive language.

I would like to end my notes with a poem from a famous Persian poet – Saadi Shirazi – [157] who believes:

"Human beings are like different organs of one body, which are originated by one essence. If a limb of this body suffers pain, the other parts will suffer too. In case anybody neglects the misery of others, not so appropriate to call them a human."

6.2 Future Works

Racism is a broad subject in psychology and is a fruitful area for further works. Considerably more research will need to be done to discover the added effects of social inequality on black people since there are enormous questions still remaining to be answered.

A further study could assess the risk exposures for specific areas in detail or study it in longer-term effects. A greater focus on geography could provide valuable findings that account for vulnerability details for particular regions.

Further research might explore to redo this study when data for longer times become available to examine how racism changed regarding the pre- and post-COVID times. Some might also investigate the ambiguous tweets and their contents. Some might also be interested in showing how users applying these offensive words can be alerted on social media by labelling tools to remind them what they are doing.

Future research can also study the effect of some concurrent incidents on the way racism tweet rates changed. An instance of this can be lootings or other outbreaks of violence that happened in the real-world during the study.

Similarly, some might like to study the same topic in Canada or about the indigenous people suffering comparable issues in their lives.

References

[1] Lopes AR. The Impact of Social Media on Social Movements: The New Opportunity and Mobilizing Structure. J. Polit. Sci. Res. 2014; 1–23
URL:
https://www.creighton.edu/fileadmin/user/CCAS/departments/PoliticalScience/Journal_of_Political_Research__JPR_/2014_JSP_papers/Lopes_JPR.pdf

[2] Chang HC, Iyer H. Trends in Twitter hashtag applications: Design features for value-added dimensions to future library catalogues. Libr. Trends 2012; 61:248–258; DOI: 10.1353/lib.2012.0024
URL: https://muse.jhu.edu/article/485537

[3] Vessey R. Zappavigna, M. (2012). Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web. London: Bloomsbury. 2015; 295–299; DOI: 10.1007/978-3-319-17948-3_13
URL: https://www.bloomsbury.com/us/discourse-of-twitter-and-social-media-9781441141866/

[4] van den Berg JA. The story of the hashtag(#): A practical theological tracing of the hashtag(#) symbol on twitter. HTS Teol. Stud. / Theol. Stud. 2014; 70:1–6; DOI: 10.4102/hts.v70i1.2706
URL: http://www.hts.org.zahttp//dx.doi.org/10.4102/hts.v70i1.2706

[5] Olofinlua T. Twitter: social communication in the twitter age. Information, Commun. Soc. 2019; 22:2037–2038; DOI: 10.1080/1369118x.2019.1620824
URL: https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1620824

[6] Fasching-Varner KJ, Martin LL, Albert KA, et al. Trayvon Martin, Race, and American Justice. Trayvon Martin, Race, Am. Justice Writ. Wrong 2014; ; DOI: 10.1007/978-94-6209-842-8
URL: http://link.springer.com/10.1007/978-94-6209-842-8

[7] Ransby B. Making All Black Lives Matter : Reimagining Freedom in the Twenty-First

Century. 2018;
URL: https://wrlc-hu.primo.exlibrisgroup.com/discovery/fulldisplay?context=L&vid=01WRLC_HOW:Law_Library&search_scope=LawHL&tab=Lawlibrary&docid=alma99185892422704109

[8] Everbach T, Clark M, Nisbett GS. #IfTheyGunnedMeDown: An Analysis of Mainstream and Social Media in the Ferguson, Missouri, Shooting of Michael Brown. Electron. News 2018; 12:23–41; DOI: 10.1177/1931243117697767
URL: http://journals.sagepub.com/doi/10.1177/1931243117697767

[9] Williams O, Ovbiagele B. Stroking Out While Black—The Complex Role of Racism. JAMA Neurol. 2020; 77:1343; DOI: 10.1001/jamaneurol.2020.3510; PMID: 10936998
URL: https://jamanetwork.com/journals/jamaneurology/fullarticle/2769834

[10] Mackenbach JP. The persistence of health inequalities in modern welfare states: The explanation of a paradox. Soc. Sci. Med. 2012; 75:761–769; DOI: 10.1016/j.socscimed.2012.02.031; PMID: 22475407

[11] Donkin AJM. Social Gradient. Wiley Blackwell Encycl. Heal. Illness, Behav. Soc. 2014; 2172–2178; DOI: 10.1002/9781118410868.wbehibs530
URL: http://doi.wiley.com/10.1002/9781118410868.wbehibs530

[12] Kröger H, Pakpahan E, Hoffmann R. What causes health inequality? A systematic review on the relative importance of social causation and health selection. Eur. J. Public Health 2015; 25:951–960; DOI: 10.1093/eurpub/ckv111; PMID: 26089181
URL: https://pubmed.ncbi.nlm.nih.gov/26089181/

[13] Shavers VL, Shavers BS. Racism and health inequity among Americans. J. Natl. Med. Assoc. 2006; 98:386–96; PMID: 16573303
URL: http://grants.nih.gov/grants/guide/pa-

[14] Lawrence K, Keleher T. Chronic Disparity: Strong and Pervasive Evidence of Racial Inequalities POVERTY OUTCOMES Structural Racism. Race Public Policy Conf. 2004; 1
URL: https://www.racialequitytools.org/resourcefiles/Definitions-of Racism.pdf

[15] chrc-ccdp.gc.ca. What is Discrimination? | Canadian Human Rights Commission. 2020 2020;
URL: https://www.chrc-ccdp.gc.ca/eng/content/what-discrimination

[16] Vernellia Randall. What is institutional racism? 2008;
URL: https://academic.udayton.edu/race/2008electionandracism/raceandracism/racism02.htm#_edn1

[17] BRIDGES KM. How Institutional Racism Kills Black People | Time.
URL: https://time.com/5851864/institutional-racism-america/

[18] Bailey ZD, Krieger N, Agénor M, et al. Structural racism and health inequities in the USA: evidence and interventions. Lancet 2017; 389:1453–1463; DOI: 10.1016/S0140-6736(17)30569-X; PMID: 28402827
URL: http://www.thelancet.com/article/S014067361730569X/fulltext

[19] Klassen M, Russell MA. Mining the Social Web, 3rd Edition. 2019;
URL: https://www.oreilly.com/library/view/mining-the-social/9781491973547/

[20] Criss S, Michaels EK, Solomon K, et al. Twitter Fingers and Echo Chambers: Exploring Expressions and Experiences of Online Racism Using Twitter. J. Racial Ethn. Heal. Disparities 2020; ; DOI: 10.1007/s40615-020-00894-5
URL: http://link.springer.com/10.1007/s40615-020-00894-5

[21] Ferrara E, Yang Z. Measuring Emotional Contagion in Social Media. PLoS One 2015; 10:e0142390; DOI: 10.1371/journal.pone.0142390
URL: https://dx.plos.org/10.1371/journal.pone.0142390

[22] Capurro D, Cole K, Echavarría MI, et al. The Use of Social Networking Sites for Public Health Practice and Research: A Systematic Review. J. Med. Internet Res. 2014; 16:e79; DOI: 10.2196/jmir.2679; PMID: 24642014
URL: https://pubmed.ncbi.nlm.nih.gov/24642014/

[23] Alshaikh F, Ramzan F, Rawaf S, et al. Social Network Sites as a Mode to Collect Health Data: A Systematic Review. J. Med. Internet Res. 2014; 16:e171; DOI:

10.2196/jmir.3050; PMID: 25048247
URL: https://pubmed.ncbi.nlm.nih.gov/25048247/

[24] Laranjo L, Arguel A, Neves AL, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J. Am. Med. Informatics Assoc. 2015; 22:243–256; DOI: 10.1136/amiajnl-2014-002841
URL: https://academic.oup.com/jamia/article/22/1/243/833940

[25] McClellan C, Ali MM, Mutter R, et al. Using social media to monitor mental health discussions − evidence from Twitter. J. Am. Med. Informatics Assoc. 2017; 24:496–502; DOI: 10.1093/jamia/ocw133; PMID: 27707822
URL: /pmc/articles/PMC7651954/?report=abstract

[26] Joseph-Shehu EM, Ncama BP, Mooi N, et al. The use of information and communication technologies to promote healthy lifestyle behaviour: a systematic scoping review. BMJ Open 2019; 9:e029872; DOI: 10.1136/bmjopen-2019-029872; PMID: 31662364
URL: /pmc/articles/PMC6830587/?report=abstract

[27] Park BK, Calamaro C. A Systematic Review of Social Networking Sites: Innovative Platforms for Health Research Targeting Adolescents and Young Adults. J. Nurs. Scholarsh. 2013; 45:n/a-n/a; DOI: 10.1111/jnu.12032
URL: http://doi.wiley.com/10.1111/jnu.12032

[28] Pew. Social Media Fact Sheet | Pew Research Center. Pew Res. Cent.
URL: https://www.pewresearch.org/internet/fact-sheet/social-media/

[29] NCHS. Use of the Internet for Health Information: United States, 2009. 2011;
URL: https://www.cdc.gov/nchs/products/databriefs/db66.htm

[30] MarketingCharts. 61% of Americans Seek Online Health Info - Marketing Charts.
URL: https://www.marketingcharts.com/industries/pharma-and-healthcare-9464

[31] Pershad Y, Hangge P, Albadawi H, et al. Social Medicine: Twitter in Healthcare. J. Clin. Med. 2018; 7:121; DOI: 10.3390/jcm7060121
URL: http://www.mdpi.com/2077-0383/7/6/121

[32] Chretien K, Azar J, Kind T. Physicians on Twitter. JAMA 2011; 305:566; DOI:

10.1001/jama.2011.68
URL: https://hsrc.himmelfarb.gwu.edu/smhs_peds_facpubs/5441

[33] Choo EK, Ranney ML, Chan TM, et al. Twitter as a tool for communication and knowledge exchange in academic medicine: A guide for skeptics and novices. Med. Teach. 2015; 37:411-4161. Choo EK, Ranney ML, Chan TM, et al. Twit; DOI: 10.3109/0142159X.2014.993371; PMID: 25523012
URL: https://www.tandfonline.com/doi/abs/10.3109/0142159X.2014.993371

[34] Chae DH, Clouston S, Martz CD, et al. Area racism and birth outcomes among Blacks in the United States. Soc. Sci. Med. 2018; 199:49–55; DOI: 10.1016/j.socscimed.2017.04.019; PMID: 28454665
URL: /pmc/articles/PMC5640467/?report=abstract

[35] Kung HC, Hoyert DL, Xu J, et al. Deaths: final data for 2005. Natl. Vital Stat. Rep. 2008; 56:1–120; PMID: 18512336

[36] Levine RS, Foster JE, Fullilove RE, et al. Black-White Inequalities in Mortality and Life Expectancy, 1933–1999: Implications for Healthy People 2010. Public Health Rep. 2001; 116:474–483; DOI: 10.1093/phr/116.5.474
URL: http://journals.sagepub.com/doi/10.1093/phr/116.5.474

[37] (US) NC for HS. Health, United States, 2005. Natl Vital Stat Rep 2008; ; PMID: 20698066
URL: https://www.ncbi.nlm.nih.gov/books/NBK20990/

[38] Hicken MT, Kravitz-Wirtz N, Durkee M, et al. Racial inequalities in health: Framing future research. Soc. Sci. Med. 2018; 199:11–18; DOI: 10.1016/j.socscimed.2017.12.027
URL: https://linkinghub.elsevier.com/retrieve/pii/S0277953617307578

[39] LaVeist TA, Gaskin D, Richard P. Estimating the Economic Burden of Racial Health Inequalities in the United States. Int. J. Heal. Serv. 2011; 41:231–238; DOI: 10.2190/HS.41.2.c
URL: http://journals.sagepub.com/doi/10.2190/HS.41.2.c

[40] Chae DH, Clouston S, Hatzenbuehler ML, et al. Association between an Internet-Based Measure of Area Racism and Black Mortality. PLoS One 2015; 10:e0122963; DOI:

10.1371/journal.pone.0122963
URL: https://dx.plos.org/10.1371/journal.pone.0122963

[41] Anthopolos R, James SA, Gelfand AE, et al. A spatial measure of neighborhood level racial isolation applied to low birthweight, preterm birth, and birthweight in North Carolina. Spat. Spatiotemporal. Epidemiol. 2011; 2:235–246; DOI: 10.1016/j.sste.2011.06.002; PMID: 22748223
URL: https://linkinghub.elsevier.com/retrieve/pii/S1877584511000232

[42] Tackett KJ, Jenkins F, Morrell DS, et al. Structural racism and its influence on the severity of atopic dermatitis in African American children. Pediatr. Dermatol. 2020; 37:142–146; DOI: 10.1111/pde.14058
URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/pde.14058

[43] Trent M, Dooley DG, Dougé J. The Impact of Racism on Child and Adolescent Health. Pediatrics 2019; 144:e20191765; DOI: 10.1542/peds.2019-1765
URL: http://pediatrics.aappublications.org/lookup/doi/10.1542/peds.2019-1765

[44] Greevy E, Smeaton AF. Classifying racist texts using a support vector machine. Proc. Sheff. SIGIR - Twenty-Seventh Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. 2004; 468–469; DOI: 10.1145/1008992.1009074

[45] Chaudhry I. "Not So Black and White": Discussions of Race on Twitter in the Aftermath of #Ferguson and the Shooting Death of Mike Brown. Cult. Stud. ↔ Crit. Methodol. 2016; 16:296–304; DOI: 10.1177/1532708616634814
URL: http://journals.sagepub.com/doi/10.1177/1532708616634814

[46] Telafaro S. A death to justice (The shooting of Michael Brown—Summer 2014). J. Hum. Behav. Soc. Environ. 2016; 26:251–253; DOI: 10.1080/10911359.2015.1125197
URL: http://www.tandfonline.com/doi/full/10.1080/10911359.2015.1125197

[47] Chaudhry I. #Hashtagging hate: Using Twitter to track racism online. First Monday 2015; ; DOI: 10.5210/fm.v20i2.5450
URL: https://journals.uic.edu/ojs/index.php/fm/article/view/5450

[48] Chaudhry I. The Twitter Racism Project. MacEwan Univ. Press 2015;

URL: https://roam.macewan.ca/islandora/object/gm:180

[49] Bartlett J, Reffin J, Rumball N, et al. Anti-social media. Anti-Social Media 2014;
URL: https://demos.co.uk/project/anti-social-media/

[50] Stephens M. The Geography of Hate. humboldt.edu
URL: http://www.floatingsheep.org/2013/05/hatemap.html

[51] Wetsman N. How Twitter is changing medical research. Nat. Med. 2020; 26:11–13; DOI: 10.1038/s41591-019-0697-7
URL: http://www.nature.com/articles/s41591-019-0697-7

[52] Mitchell L, Frank MR, Harris KD, et al. The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. PLoS One 2013; 8:e64417; DOI: 10.1371/journal.pone.0064417
URL: https://dx.plos.org/10.1371/journal.pone.0064417

[53] Matamoros-Fernández A. Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. Information, Commun. Soc. 2017; 20:930–946; DOI: 10.1080/1369118X.2017.1293130
URL: https://www.tandfonline.com/doi/full/10.1080/1369118X.2017.1293130

[54] Hill ML. "Thank You, Black Twitter": State Violence, Digital Counterpublics, and Pedagogies of Resistance. Urban Educ. 2018; 53:286–302; DOI: 10.1177/0042085917747124
URL: http://journals.sagepub.com/doi/10.1177/0042085917747124

[55] Borda JL, Marshall B. Creating a space to #SayHerName: Rhetorical stratification in the networked sphere. Q. J. Speech 2020; 106:133–155; DOI: 10.1080/00335630.2020.1744182
URL: https://www.tandfonline.com/doi/full/10.1080/00335630.2020.1744182

[56] Klein GC. On the Death of Sandra Bland: A Case of Anger and Indifference. SAGE Open 2018; 8:215824401875493; DOI: 10.1177/2158244018754936
URL: http://journals.sagepub.com/doi/10.1177/2158244018754936

[57] Nature-Methods. Why Black lives matter in science. Nat. Methods 2020; 17:645–645; DOI:

10.1038/s41592-020-0908-7; PMID: 32616929
URL: http://www.nature.com/articles/s41592-020-0908-7

[58] Wiki. Shooting of Trayvon Martin.
URL: https://en.wikipedia.org/wiki/Shooting_of_Trayvon_Martin

[59] Ray R, Brown M, Laybourn W. The evolution of #BlackLivesMatter on Twitter: social movements, big data, and race. Ethn. Racial Stud. 2017; 40:1795–1796; DOI: 10.1080/01419870.2017.1335423
URL: https://www.tandfonline.com/doi/full/10.1080/01419870.2017.1335423

[60] Sinnenberg L, Buttenheim AM, Padrez K, et al. Twitter as a Tool for Health Research: A Systematic Review. Am. J. Public Health 2017; 107:e1–e8; DOI: 10.2105/AJPH.2016.303512; PMID: 27854532
URL: /pmc/articles/PMC5308155/?report=abstract

[61] Edo-Osagie O, De La Iglesia B, Lake I, et al. A scoping review of the use of Twitter for public health research. Comput. Biol. Med. 2020; 122:103770; DOI: 10.1016/j.compbiomed.2020.103770
URL: https://linkinghub.elsevier.com/retrieve/pii/S0010482520301426

[62] Al-Rawi A, Siddiqi M, Morgan R, et al. COVID-19 and the Gendered Use of Emojis on Twitter: Infodemiology Study. J. Med. Internet Res. 2020; 22:e21646; DOI: 10.2196/21646; PMID: 33052871
URL: /pmc/articles/PMC7647473/?report=abstract

[63] Leypunskiy E, Kıcıman E, Shah M, et al. Geographically Resolved Rhythms in Twitter Use Reveal Social Pressures on Daily Activity Patterns. Curr. Biol. 2018; 28:3763-3775.e5; DOI: 10.1016/j.cub.2018.10.016
URL: https://linkinghub.elsevier.com/retrieve/pii/S0960982218313459

[64] Mohammadi E, Thelwall M, Kwasny M, et al. Academic information on Twitter: A user survey. PLoS One 2018; 13:e0197265; DOI: 10.1371/journal.pone.0197265
URL: https://dx.plos.org/10.1371/journal.pone.0197265

[65] Wei R, Curtin LR. National VItal Statistics Reports, Volume 60, Number 9, 09/14/2012.

1999;
URL: https://www.cdc.gov/nchs/data/nvsr/nvsr60/nvsr60_09.pdf

[66] SILVERSTEIN J. Genes Don't Cause Racial-Health Disparities, Society Does - The Atlantic. 2015;
URL: https://www.theatlantic.com/health/archive/2015/04/genes-dont-cause-racial-health-disparities-society-does/389637/

[67] Kahn J. Race in a Bottle | Columbia University Press. Columbia Univ. Press 2014;
URL: http://cup.columbia.edu/book/race-in-a-bottle/9780231162982

[68] Hussein N, Weng SF, Kai J, et al. Preconception risk assessment for thalassaemia, sickle cell disease, cystic fibrosis and Tay-Sachs disease. Cochrane Database Syst. Rev. 2018; ; DOI: 10.1002/14651858.CD010849.pub3
URL: http://doi.wiley.com/10.1002/14651858.CD010849.pub3

[69] Gravlee CC. How race becomes biology: Embodiment of social inequality. Am. J. Phys. Anthropol. 2009; 139:47–57; DOI: 10.1002/ajpa.20983
URL: http://doi.wiley.com/10.1002/ajpa.20983

[70] Williams DR, Yu Y, Jackson JS, et al. Racial differences in physical and mental health. Socio-economic status, stress and discrimination. J. Health Psychol. 1997; 2:335–351; DOI: 10.1177/135910539700200305; PMID: 22013026
URL: https://pubmed.ncbi.nlm.nih.gov/22013026/

[71] Lu MC, Halfon N. Racial and ethnic disparities in birth outcomes: a life-course perspective. Matern. Child Health J. 2003; 7:13–30; DOI: 10.1023/A:1022537516969; PMID: 12710797
URL: https://pubmed.ncbi.nlm.nih.gov/12710797/

[72] Hammond WP, Mattis JS. Being a Man About It: Manhood Meaning Among African American Men. Psychol. Men Masc. 2005; 6:114–126; DOI: 10.1037/1524-9220.6.2.114
URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/1524-9220.6.2.114

[73] Assari S. Combined Racial and Gender Differences in the Long-Term Predictive Role of Education on Depressive Symptoms and Chronic Medical Conditions. J. Racial Ethn.

Heal. Disparities 2017; 4:385–396; DOI: 10.1007/s40615-016-0239-7; PMID: 27270925
URL: https://link.springer.com/article/10.1007/s40615-016-0239-7

[74] Assari S, Caldwell CH. High Risk of Depression in High-Income African American Boys. J. Racial Ethn. Heal. Disparities 2018; 5:808–819; DOI: 10.1007/s40615-017-0426-1; PMID: 28842841
URL: https://link.springer.com/article/10.1007/s40615-017-0426-1

[75] Assari S, Lee DB, Nicklett EJ, et al. Racial Discrimination in Health Care Is Associated with Worse Glycemic Control among Black Men but Not Black Women with Type 2 Diabetes. Front. Public Heal. 2017; 5:; DOI: 10.3389/fpubh.2017.00235
URL: http://journal.frontiersin.org/article/10.3389/fpubh.2017.00235/full

[76] Assari S, Moazen-Zadeh E, Caldwell CH, et al. Racial Discrimination during Adolescence Predicts Mental Health Deterioration in Adulthood: Gender Differences among Blacks. Front. Public Heal. 2017; 5:; DOI: 10.3389/fpubh.2017.00104
URL: http://journal.frontiersin.org/article/10.3389/fpubh.2017.00104/full

[77] Assari S, Lankarani M, Caldwell C. Does Discrimination Explain High Risk of Depression among High-Income African American Men? Behav. Sci. (Basel). 2018; 8:40; DOI: 10.3390/bs8040040
URL: http://www.mdpi.com/2076-328X/8/4/40

[78] Zunic A, Corcoran P, Spasic I. Sentiment Analysis in Health and Well-Being: Systematic Review. JMIR Med. Informatics 2020; 8:e16023; DOI: 10.2196/16023
URL: https://medinform.jmir.org/2020/1/e16023

[79] Zimmer M, Proferes NJ. A topology of twitter research: Disciplines, methods, and ethics. Aslib J. Inf. Manag. 2014; 66:250–261; DOI: 10.1108/AJIM-09-2013-0083
URL: https://www.emerald.com/insight/content/doi/10.1108/AJIM-09-2013-0083/full/html

[80] Oscar N, Fox PA, Croucher R, et al. Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer's Disease Stigma on Twitter. Journals Gerontol. Ser. B 2017; 72:742–751; DOI: 10.1093/geronb/gbx014

URL: http://academic.oup.com/psychsocgerontology/article/72/5/742/3059641/Machine-Learning-Sentiment-Analysis-and-Tweets-An

[81] Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count (LIWC). Univ. Texas 2001;
URL: https://www.researchgate.net/publication/246699633_Linguistic_inquiry_and_word_count_LIWC

[82] Nielsen FÅ. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. CEUR Workshop Proc. 2011; 718:93–98
URL: http://arxiv.org/abs/1103.2903

[83] Lee J-SM. How to use Twitter to further your research career. Nature 2019; ; DOI: 10.1038/d41586-019-00535-w
URL: http://www.nature.com/articles/d41586-019-00535-w

[84] Dickinson EE. The promise and peril of academia wading into Twitter | Hub. 2019;
URL: https://hub.jhu.edu/magazine/2019/spring/more-academics-turn-to-twitter/

[85] Sturges L. Altmetrics reveal insights into the impact of scientific knowledge.
URL: https://www.elsevier.com/connect/altmetrics-reveal-insights-into-the-impact-of-scientific-knowledge

[86] Kim AE, Hansen HM, Murphy J, et al. Methodological Considerations in Analyzing Twitter Data. JNCI Monogr. 2013; 2013:140–146; DOI: 10.1093/jncimonographs/lgt026
URL: https://academic.oup.com/jncimono/article-lookup/doi/10.1093/jncimonographs/lgt026

[87] Sarah Perez. Twitter's doubling of character count from 140 to 280 had little impact on length of tweets | TechCrunch.
URL: https://techcrunch.com/2018/10/30/twitters-doubling-of-character-count-from-140-to-280-had-little-impact-on-length-of-tweets/

[88] Twitter-Developer. Counting characters | Docs | Twitter Developer.
URL: https://developer.twitter.com/en/docs/counting-characters

[89] Statista. • Twitter: most users by country | Statista.
URL: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[90] Osterberg G. Update on the Twitter Archive at the Library of Congress | Library of Congress Blog. 2013;
URL: https://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/

[91] Twitter. About Twitter's APIs.
URL: https://help.twitter.com/en/rules-and-policies/twitter-api

[92] Twitter. Archive.org.
URL: https://archive.org/search.php?query=collection%3Atwitterstream&sort=-publicdate

[93] Littman J. Where to get Twitter data for academic research • Social Feed Manager. github.io
URL: https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data

[94] Barrie C. Searching Racism after George Floyd. Socius Sociol. Res. a Dyn. World 2020; 6:237802312097150; DOI: 10.1177/2378023120971507
URL: http://journals.sagepub.com/doi/10.1177/2378023120971507

[95] Twitter. Introduction | Docs | Twitter Developer.
URL: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview

[96] Flynn J. Say What's the Word? An Exploration of the Controversy around the 'N-words'. 2012;
URL:
https://www.academia.edu/8941584/Say_Whatt_the_Word_An_Exploration_of_the_Controversy_around_the_N_words_

[97] Spears AK. Perspectives: A View of the 'N-Word' from Sociolinguistics - Higher Education.
URL: https://diverseeducation.com/article/6114/

[98] Ayto J, Simpson J. The Oxford Dictionary of Modern Slang. 2008; ; DOI: 10.1093/acref/9780199543700.001.0001

URL:
http://www.oxfordreference.com/view/10.1093/acref/9780199543700.001.0001/acref-9780199543700

[99] datascience@berkeley. Python for Data Science - Blog. online Master Inf. Data Sci. from UC Berkeley
URL: https://ischoolonline.berkeley.edu/blog/python-data-science/

[100] Roper W. Chart: Python Remains Most Popular Programming Language | Statista. 2020;
URL: https://www.statista.com/chart/21017/most-popular-programming-languages/

[101] NTLK.org. nltk.tokenize package — NLTK 3.5 documentation.
URL: https://www.nltk.org/api/nltk.tokenize.html

[102] Dutta A. Removing stop words with NLTK in Python - GeeksforGeeks.
URL: https://www.geeksforgeeks.org/removing-stop-words-nltk-python/

[103] GeoPy. Welcome to GeoPy's documentation! — GeoPy 2.0.0 documentation.
URL: https://geopy.readthedocs.io/en/stable/

[104] Gender API. Gender API - Determines the gender of a first name. https://gender-api.com/
URL: https://gender-api.com/en/

[105] Santamaría L, Mihaljević H. Comparison and benchmark of name-to- gender inference services. PeerJ Comput. Sci. 2018; 2018:e156; DOI: 10.7717/peerj-cs.156
URL: https://peerj.com/articles/cs-156

[106] Coggon D, Rose G, Barker D. Chapter 6. Ecological studies.
URL: https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated/6-ecological-studies

[107] National-Cancer-Institute. Definition of ecological study - NCI Dictionary of Cancer Terms - National Cancer Institute. Natl. Institutes Heal.
URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/ecological-study

[108] Levin KA. Study Design VI - Ecological Studies. Evid. Based. Dent. 2006; 7:108–108; DOI: 10.1038/sj.ebd.6400454; PMID: 17187048

URL: http://www.nature.com/articles/6400454

[109] . Correlation | Definition of Correlation by Merriam-Webster.
URL: https://www.merriam-webster.com/dictionary/correlation

[110] Akoglu H. User's guide to correlation coefficients. Turkish J. Emerg. Med. 2018; 18:91–93; DOI: 10.1016/j.tjem.2018.08.001
URL: https://linkinghub.elsevier.com/retrieve/pii/S2452247318302164

[111] Cruzan J. What is correlation? Xaktly.com 2019;
URL: https://xaktly.com/ProbStat_Correlation.html

[112] MathsIsFun-Advanced. Correlation. 2018;
URL: https://www.mathsisfun.com/data/correlation.html

[113] Yogish D, Manjunath TN, Hegadi RS. Review on Natural Language Processing Trends and Techniques Using NLTK. Commun. Comput. Inf. Sci. 2019; 1037:589–606; DOI: 10.1007/978-981-13-9187-3_53
URL: https://doi.org/10.1007/978-981-13-9187-3_53

[114] . Natural Language Toolkit. 2015;
URL: https://www.nltk.org/

[115] Sarkar D. Text Analytics with Python. 2019; ; DOI: 10.1007/978-1-4842-4354-1
URL: http://link.springer.com/10.1007/978-1-4842-4354-1

[116] Tailor C, Patel B. Sentence Tokenization Using Statistical Unsupervised Machine Learning and Rule-Based Approach for Running Text in Gujarati Language. 2019; 319–326; DOI: 10.1007/978-981-13-2285-3_38
URL: http://link.springer.com/10.1007/978-981-13-2285-3_38

[117] Hardeniya N. NLTK Essentials | Guide books. 2015;
URL: https://dl.acm.org/doi/book/10.5555/2846245

[118] NSW-Government. Identifying and using correct articles and pronouns. 2020;
URL: https://education.nsw.gov.au/teaching-and-learning/student-assessment/smart-teaching-strategies/literacy/language-conventions/stage-2/noun-groups/identifying-and-

using-correct-articles-and-pronouns

[119] . Stemming and Lemmatization with Python NLTK.
URL: https://www.guru99.com/stemming-lemmatization-python-nltk.html

[120] Prabhakaran S. Lemmatization Approaches with Examples in Python. machinelearningplus.com
URL: https://www.machinelearningplus.com/nlp/lemmatization-examples-python/

[121] Gohil S, Vuik S, Darzi A. Sentiment Analysis of Health Care Tweets: Review of the Methods Used. JMIR Public Heal. Surveill. 2018; 4:e43; DOI: 10.2196/publichealth.5789
URL: http://publichealth.jmir.org/2018/2/e43/

[122] Kitchen CMR. Nonparametric vs Parametric Tests of Location in Biomedical Research. Am. J. Ophthalmol. 2009; 147:571–572; DOI: 10.1016/j.ajo.2008.06.031
URL: https://linkinghub.elsevier.com/retrieve/pii/S0002939408005199

[123] Fields H. Difference Between Parametric and Nonparametric Difference Between | Difference Between. differencebetween.net 2020;
URL: http://www.differencebetween.net/science/difference-between-parametric-and-nonparametric/

[124] Ellis and Victoria. Spearman's correlation. 2011; 1–7
URL: http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf

[125] Laerd-Statistics. Spearman's Rank-Order Correlation - A guide to when to use it, what it does and what the assumptions are.
URL: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

[126] David Sarmento. Chapter 22: Correlation Types and When to Use Them. R-Language 2020;
URL: https://ademos.people.uic.edu/Chapter22.html

[127] Glen S. Monotonic Relationship: Definition - Statistics How To.
URL: https://www.statisticshowto.com/monotonic-relationship/

[128] Oxford-College. Spearman's Rank Sum Correlation Test. Oxford 2019;
URL: http://mathcenter.oxford.emory.edu/site/math117/rankCorrelation/

[129] Hauke J, Kossowski T. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. Quaest. Geogr. 2011; 30:87–93; DOI: 10.2478/v10117-011-0021-1
URL: https://content.sciendo.com/doi/10.2478/v10117-011-0021-1

[130] UCLA. FAQ: What are the differences between one-tailed and two-tailed tests?
URL: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-the-differences-between-one-tailed-and-two-tailed-tests/

[131] Ortiz-Ospina E. "Life Expectancy" – What does this actually mean? - Our World in Data. Oxford Martin Sch. 2017;
URL: https://ourworldindata.org/life-expectancy-how-is-it-calculated-and-how-should-it-be-interpreted

[132] . Health status - Life expectancy at birth - OECD Data.
URL: https://data.oecd.org/healthstat/life-expectancy-at-birth.htm

[133] National-Center-for-Health-Statistics. NVSS - Life Tables - State Life Tables - LEWK4.
URL: https://www.cdc.gov/nchs/nvss/mortality/lewk4.htm

[134] Massey DS. American apartheid: segregation and the making of the underclass. Am. J. Sociol. 1990; 96:329–357; DOI: 10.1086/229532
URL: https://www.jstor.org/stable/2781105?seq=1

[135] Quillian L. Segregation and Poverty Concentration. Am. Sociol. Rev. 2012; 77:354–379; DOI: 10.1177/0003122412447793
URL: http://journals.sagepub.com/doi/10.1177/0003122412447793

[136] O'Connell HA. The impact of slavery on racial inequality in poverty in the contemporary U.S. South. Soc. Forces 2012; 90:713–734; DOI: 10.1093/sf/sor021
URL: https://go-gale-com.librweb.laurentian.ca/ps/i.do?p=AONE&sw=w&issn=00377732&v=2.1&it=r&id=GALE%7CA294505509&sid=googleScholar&linkaccess=fulltext

[137] United-States-Census-Bureu. 2018 Poverty Rate in the United States. census.gov 2019;
URL: https://www.census.gov/library/visualizations/interactive/2018-poverty-rate.html

[138] . U.S. Census Bureau QuickFacts: United States.
URL: https://www.census.gov/quickfacts/US

[139] Bureau UC. Income, Poverty and Health Insurance Coverage in the U.S.: 2014.
URL: https://www.census.gov/newsroom/press-releases/2015/cb15-157.html

[140] . Census 2000 Gateway - U.S. Census Bureau.
URL: https://www.census.gov/main/www/cen2000.html

[141] United-States-Census-Bureu. Educational Attainment by State.
URL: https://worldpopulationreview.com/state-rankings/educational-attainment-by-state

[142] Schaefer RT. Schaefer, Race and Ethnicity in the United States [RENTAL EDITION] | Pearson. DePaul Univ. 2019;
URL: https://www.pearson.com/us/higher-education/product/Schaefer-Race-and-Ethnicity-in-the-United-States-RENTAL-EDITION-9th-Edition/9780134732824.html

[143] . US States by Race 2020.
URL: https://worldpopulationreview.com/states/states-by-race

[144] Shih JH, Fay MP. Pearson's chi-square test and rank correlation inferences for clustered data. Biometrics 2017; 73:822–834; DOI: 10.1111/biom.12653
URL: http://doi.wiley.com/10.1111/biom.12653

[145] Ling-300. Tutorial: Pearson's Chi-square Test for Independence.
URL: https://www.ling.upenn.edu/~clight/chisquared.htm

[146] . Using Chi-Square Statistic in Research - Statistics Solutions.
URL: https://www.statisticssolutions.com/using-chi-square-statistic-in-research/

[147] . Objective and Subjective Claims - TIP Sheet - Butte College.
URL: http://www.butte.edu/departments/cas/tipsheets/thinking/claims.html

[148] TheConversation. Black men face high discrimination and depression, even as their education and incomes rise.

URL: https://theconversation.com/black-men-face-high-discrimination-and-depression-even-as-their-education-and-incomes-rise-141027

[149] Assari S. High sense of mastery reduces psychological distress for African American women but not African American men. Arch. Gen. Intern. Med. 2019; 3:5–9; PMID: 31289793
URL: http://www.ncbi.nlm.nih.gov/pubmed/31289793

[150] Assari S. Educational Attainment Better Protects African American Women than African American Men Against Depressive Symptoms and Psychological Distress. Brain Sci. 2018; 8:182; DOI: 10.3390/brainsci8100182
URL: http://www.mdpi.com/2076-3425/8/10/182

[151] Assari S. This is the toll that everyday racism takes on black men in America.
URL: https://www.weforum.org/agenda/2020/07/george-floyd-racism-opportunities-life-expectancy/

[152] Chaney C, Robertson R V. Racism and Police Brutality in America. J. African Am. Stud. 2013; 17:480–505; DOI: 10.1007/s12111-013-9246-5
URL: http://link.springer.com/10.1007/s12111-013-9246-5

[153] Seaton EK, Caldwell CH, Sellers RM, et al. The prevalence of perceived discrimination among African American and Caribbean Black youth. Dev. Psychol. 2008; 44:1288–1297; DOI: 10.1037/a0012747
URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0012747

[154] Assari S, Moghani Lankarani M, Caldwell C. Discrimination Increases Suicidal Ideation in Black Adolescents Regardless of Ethnicity and Gender. Behav. Sci. (Basel). 2017; 7:75; DOI: 10.3390/bs7040075
URL: http://www.mdpi.com/2076-328X/7/4/75

[155] Kolliakou A, Bakolis I, Chandran D, et al. Mental health-related conversations on social media and crisis episodes: a time-series regression analysis. Sci. Rep. 2020; 10:1342; DOI: 10.1038/s41598-020-57835-9
URL: http://www.nature.com/articles/s41598-020-57835-9

[156] Ueda M, Mori K, Matsubayashi T, et al. Tweeting celebrity suicides: Users' reaction to prominent suicide deaths on Twitter and subsequent increases in actual suicides. Soc. Sci. Med. 2017; 189:158–166; DOI: 10.1016/j.socscimed.2017.06.032
URL: https://linkinghub.elsevier.com/retrieve/pii/S0277953617304082

[157] Zarrintan S, Ranjbar F, Aslanabadi S, et al. Bani Adam: Saadi Shirazi (AD 1184–1283/1291) and the concept of empathy. Child's Nerv. Syst. 2015; 31:1211–1212; DOI: 10.1007/s00381-014-2517-1
URL: http://link.springer.com/10.1007/s00381-014-2517-1

Appendix 1
Tweet Examples

Picture of some tweets by users supporting Black Lives Matter movement


Frederick Joseph @FredTJoseph · May 30
New York belongs to the people. #BlackLivesMatter
THE MET 150
2
49
256


aaron @aaronesparza14 · Jun 7
let's remain persistent #BlackLivesMatter #Revolution #Not #Evolution
JUSTICE FOR THEM ALL
it's a MOVEMENT not a moment
12
9

Appendix 2
Tweet statistics table

Table 0.1 Sum of all tweets that had locations regarding the month published.

State	January	February	March	April	May Before the George Event	May After the George	June	Sum of all before event	Sum of all after event	Sum of all
California	1,098,623	537,497	1,297,707	1,270,551	947,403	316,565	1,263,968	5,151,781	1,580,533	6,732,314
Indiana	82,731	38,052	85,707	82,523	61,061	19,190	80,251	350,074	99,441	449,515
Arizona	123,904	61,650	145,252	141,036	103,834	35,222	139,056	575,676	174,278	749,954
Maryland	187,135	86,669	202,097	197,806	152,071	48,568	200,639	825,778	249,207	1,074,985
Michigan	240,868	112,671	257,695	259,625	188,958	61,080	250,038	1,059,817	311,118	1,370,935
Illinois	399,307	190,644	446,883	437,558	329,149	108,086	437,235	1,803,541	545,321	2,348,862
Ohio	197,607	93,124	210,258	205,013	148,786	48,937	197,723	854,788	246,660	1,101,448
Hawaii	16,801	9,285	21,715	21,633	15,575	4,862	20,437	85,009	25,299	110,308
Georgia	260,696	122,606	278,401	282,779	203,780	70,092	273,872	1,148,262	343,964	1,492,226
Colorado	88,122	42,794	104,838	101,846	72,419	24,642	97,061	410,019	121,703	531,722
North Carolina	160,721	76,603	176,230	170,183	125,024	41,670	166,694	708,761	208,364	917,125
Florida	417,528	195,410	450,162	456,004	340,275	110,658	450,933	1,859,379	561,591	2,420,970
Tennessee	133,095	59,190	137,595	130,821	94,415	30,370	124,785	555,116	155,155	710,271
Nevada	131,617	64,096	147,818	147,095	110,055	36,179	146,234	600,681	182,413	783,094
New York	747,073	360,262	855,898	830,781	636,877	203,520	840,397	3,430,891	1,043,917	4,474,808
Pennsylvania	209,813	103,627	235,399	231,453	167,408	55,290	222,698	947,700	277,988	1,225,688
Texas	1,255,174	585,852	1,322,877	1,290,749	955,294	311,116	1,266,410	5,409,946	1,577,526	6,987,472
Minnesota	68,890	32,063	78,198	71,783	53,205	22,853	76,058	304,139	98,911	403,050
Oregon	136,014	66,727	163,621	155,428	114,504	38,171	152,675	636,294	190,846	827,140
District of Columbia	200,991	98,295	235,708	222,126	165,143	54,132	219,275	922,263	273,407	1,195,670
Washington	147,410	74,678	182,780	171,234	126,775	42,692	169,467	702,877	212,159	915,036
Oklahoma	85,747	41,730	91,474	87,357	63,758	19,820	83,578	370,066	103,398	473,464
Massachusetts	142,755	71,633	168,150	157,845	114,740	35,945	150,685	655,123	186,630	841,753
Alabama	52,545	26,502	56,233	53,932	39,159	12,084	51,243	228,371	63,327	291,698
Wisconsin	62,778	28,342	68,364	66,353	47,488	15,455	62,943	273,325	78,398	351,723
Nebraska	35,866	16,396	39,253	39,964	27,744	9,053	36,797	159,223	45,850	205,073
Puerto Rico	10,552	4,341	11,305	11,069	7,749	1,934	9,683	45,016	11,617	56,633
Louisiana	121,140	52,464	123,622	123,158	86,942	28,143	115,085	507,326	143,228	650,554
Kansas	64,573	30,119	67,775	65,173	47,236	14,818	62,054	274,876	76,872	351,748
Iowa	68,096	32,676	75,868	72,009	50,244	15,598	65,842	298,893	81,440	380,333
Connecticut	40,081	19,155	43,437	42,959	29,660	9,662	39,322	175,292	48,984	224,276
South Carolina	44,306	21,444	47,753	45,436	32,871	11,155	44,026	191,810	55,181	246,991
Wyoming	8,140	4,182	9,135	9,232	6,521	2,017	8,538	37,210	10,555	47,765

Virginia	58,857	26,877	63,163	59,685	44,150	14,548	58,698	252,732	73,246	325,978
Missouri	99,830	46,848	106,497	99,999	73,906	24,950	98,856	427,080	123,806	550,886
Arkansas	38,863	19,003	39,956	39,183	27,988	8,780	36,768	164,993	45,548	210,541
Kentucky	53,357	24,627	56,344	56,723	40,144	13,663	53,807	231,195	67,470	298,665
New Jersey	63,069	31,513	70,187	70,709	51,950	16,264	68,214	287,428	84,478	371,906
Mississippi	32,645	16,209	36,610	34,925	25,068	7,309	32,377	145,457	39,686	185,143
Idaho	12,673	6,789	15,304	14,313	10,847	3,425	14,272	59,926	17,697	77,623
New Mexico	20,850	9,884	25,583	25,306	17,046	6,190	23,236	98,669	29,426	128,095
Delaware	21,911	10,893	24,847	24,320	18,475	5,613	24,088	100,446	29,701	130,147
West Virginia	8,991	4,490	10,682	10,349	6,815	2,161	8,976	41,327	11,137	52,464
Utah	27,675	14,361	34,368	31,077	22,233	7,800	30,033	129,714	37,833	167,547
Vermont	5,449	3,120	6,633	5,966	4,547	1,265	5,812	25,715	7,077	32,792
Rhode Island	18,245	8,852	21,073	19,479	14,225	4,309	18,534	81,874	22,843	104,717
North Dakota	6,907	3,309	8,219	8,132	4,853	1,573	6,426	31,420	7,999	39,419
Montana	6,675	3,393	9,101	8,178	5,844	1,863	7,707	33,191	9,570	42,761
South Dakota	5,448	3,102	7,722	7,153	3,952	1,045	4,997	27,377	6,042	33,419
Alaska	9,991	4,634	12,329	11,894	7,113	2,443	9,556	45,961	11,999	57,960
New Hampshire	9,268	4,689	10,382	10,192	7,756	2,120	9,876	42,287	11,996	54,283
Maine	3,562	1,491	3,768	4,192	3,180	989	4,169	16,193	5,158	21,351
SUM	7,544,965	3,604,863	8,401,976	8,194,289	6,056,215	1,985,889	8,042,104	33,802,308	10,027,993	43,830,301

Appendix 3
Life expectation table

Table 0.1 Life expectation table of white and black people in addition to the genderLife expectation table of white and black people in addition to the genderTable 0.2

	Total	***White***	***Male***	***Female***	***Black***	***Male***	***Female***
Alabama	74.8000	75.9642	71.3201	78.3426	70.8186	66.4218	74.9403
Alaska	76.6288	77.6065	74.1816	79.4119			
Arizona	78.1458	78.4888	75.2475	81.1637	74.0265	70.9476	77.6951
Arkansas	75.4302	76.2759	72.0481	78.9874	70.5853	67.2992	73.5778
California	78.8013	78.7054	76.0222	81.6259	73.2280	69.9700	76.6734
Colorado	78.7160	78.7342	76.2879	81.1588	74.1347	71.7096	76.5915
Connecticut	78.9032	79.3627	76.1314	81.6297	74.8087	71.7295	77.7078
Delaware	77.0405	77.8019	74.2416	79.7806	72.6914	70.2685	75.2057
District of Columbia	73.0886	81.5404	68.5670	77.5943	69.6091	64.5944	74.4599
Florida	78.1022	78.7860	74.9703	81.4045	72.2161	68.9794	75.5338
Georgia	75.2711	76.6781	72.2836	78.2157	72.3599	68.2935	76.1635
Hawaii	80.2278	80.6442	77.1717	83.6513			
Idaho	78.2871	78.4371	76.1834	80.5011			
Illinois	77.0647	78.0548	73.9106	80.2628	70.6209	66.8102	74.2010
Indiana	76.4659	76.9708	73.5467	79.4460	71.9756	67.4711	76.5532
Iowa	78.7595	78.7702	76.1115	81.3913	72.9071	70.8084	75.1629
Kansas	77.7824	78.1847	74.8394	80.8849	71.6973	68.4692	75.0250
Kentucky	75.2036	75.5851	72.2459	78.2027	71.7143	69.0097	74.4576
Louisiana	74.2834	76.3780	71.1171	77.4372	70.7743	66.4519	75.2974
Maine	77.4570	78.2310	75.2250	79.6277			
Maryland	76.3600	78.1325	73.5485	79.0812	72.2037	68.4091	75.7815
Massachusetts	78.7592	79.0068	75.7906	81.6785	76.3551	73.1398	79.3311
Michigan	76.9029	77.9181	73.9804	79.8324	71.6243	67.3757	75.9538
Minnesota	79.2556	79.7494	76.7358	81.8006	74.0646	71.5891	76.6047
Mississippi	73.8762	75.5975	70.3002	77.6175	70.3771	66.7216	73.6784
Missouri	76.5156	77.0758	73.5922	79.4625	70.9372	67.2173	74.4959
Montana	77.7409	77.9448	75.1762	80.5649			
Nebraska	78.3693	78.6302	75.9980	80.7773	71.8780	69.1496	74.6663
Nevada	76.0527	76.0302	73.3439	79.2383	72.5658	70.5581	74.7503
New Hampshire	78.7936	78.8823	76.2394	81.4018			
New Jersey	77.5785	78.5879	74.7713	80.3217	72.3046	68.8485	75.5391
New Mexico	77.2632	77.8946	74.5156	80.0564	72.9656	71.6266	74.3682
New York	78.1963	75.7828	75.1302	81.1609	74.2426	70.1258	77.8400

North Carolina	76.2684	77.2744	73.0478	79.5591	71.4492	66.3339	76.6832
North Dakota	79.0623	79.5287	75.9597	82.6147			
Ohio	76.4930	77.3096	73.9378	78.9509	71.8554	68.5985	74.9145
Oklahoma	75.6058	75.8173	72.7474	78.5949	71.7352	68.9679	74.3408
Oregon	78.0894	77.9564	75.8150	80.3745	74.0282	70.6655	78.2417
Pennsylvania	77.0177	77.8667	74.0919	79.9025	71.3205	67.2570	75.3411
Puerto Rico							
Rhode Island	78.6486	78.8474	75.8317	81.4193	74.8422	72.1803	77.3944
South Carolina	75.0357	76.6952	71.6821	78.4880	71.4490	67.3387	75.4088
South Dakota	78.3437	79.3323	75.1949	81.7941			
Tennessee	75.2935	76.2203	71.9808	78.6630	70.6803	66.8633	74.1563
Texas	77.0443	77.5103	74.1180	80.0463	72.0649	69.1820	74.7641
Utah	78.8897	78.9346	76.8385	80.9488			
Vermont	78.2409	78.5667	76.1777	80.2881			
Virginia	76.9505	78.1664	74.4822	79.3357	72.7875	69.3708	76.2015
Washington	78.6418	78.5100	76.1788	81.1373	74.2919	71.9048	77.1100
West Virginia	75.2773	75.5131	72.7451	77.8412	71.3418	69.8676	72.6493
Wisconsin	78.5587	78.9263	75.6058	81.6358	71.5057	68.4115	74.3439
Wyoming	76.6364	77.7379	74.8348	78.5535			

Appendix 4
Word Clouds


worried
always
weird
talkin
saying
music
bitches
play
female
first
get
really
money
trying
straight
home
like
life
last
dick
work
another
fucked
negro
wrong
fucking
talking
something
different
anything
nothing
getting
cause
make
baby
beat
phone
know
ass
right
said
back
black
dead
next
call
take
talk
game
put
lil
gon
hoes
feel
mad
n't
much
come
big
boy
dumb
head
nigga
got
find
bout
tell
girl
tryna
dont
shit
real
ion
try
yall
aint
seen
new
used
hard
even
see
one
bitch
lol
bro
hit
hoe
ask
way
want
niggas
fuck
told
done
say
wan
take
type
hood
better
think
love
good
leave
hate
wit
never
let
man
ever
crazy
whole
broke
gone
day
time
give
Niggas
look
stop
keep
need
bad
still
post
pussy
finna
going
swear
around
everything
Nigga
lmao
year
nobody
NIGGAS
every
people
damn
start
females
women
NIGGA
school
young
white
thought
friends

Appendix 5
Emojis

	#BLM	Emoji Name	Occurrence	Racist	Emoji Name	Occurrence
1	😂	face with tears of joy	549	😂	face with tears of joy	61725
2	💔	broken heart	488	😭	loudly crying face	38592
3	🔥	fire	415	💀	skull	4663
4	😭	loudly crying face	252	😩	weary face	3583
5	💜	purple heart	251	🔥	fire	3000
6	💯	hundred points symbol	247	💯	hundred points symbol	2910
7	❗	white exclamation mark ornament	216	😍	smiling face with heart-shaped eyes	1888
8	🚨	police cars revolving light	205	💔	broken heart	1743
9	👇	white down pointing backhand index	187	💦	splashing sweat symbol	1325
10	😄	face with look of triumph	185	😒	unamused face	1309
11	✊	raised fist	184	😈	smiling face with horns	915
12	😩	disappointed face	170	😪	tired face	849
13	✨	sparkles	157	😄	face with look of triumph	779
14	😩	pouting face	154	🍆	aubergine	708
15	💥	collision symbol	113	😌	relieved face	706