Twitter Corpus of the #BlackLivesMatter Movement and Counter Protests: 2013 to 2021

Salvatore Giorgi,[1,2] **Sharath Chandra Guntuku,**[2] **McKenzie Himelein-Wachowiak,**[1]
Amy Kwarteng,[1] **Sy Hwang,**[2] **Muhammad Rahman,**[1] **Brenda Curtis**[1]

[1] National Institute on Drug Abuse
[2] University of Pennsylvania
sal.giorgi@nih.gov, brenda.curtis@nih.gov

Abstract

Black Lives Matter (BLM) is a decentralized social movement protesting violence against Black individuals and communities, with a focus on police brutality. The movement gained significant attention following the killings of Ahmaud Arbery, Breonna Taylor, and George Floyd in 2020. The #BlackLivesMatter social media hashtag has come to represent the grassroots movement, with similar hashtags counter protesting the BLM movement, such as #AllLivesMatter, and #BlueLivesMatter. We introduce a data set of 63.9 million tweets from 13.0 million users from over 100 countries which contain one of the following keywords: *BlackLivesMatter*, *AllLivesMatter*, and *BlueLivesMatter*. This data set contains all currently available tweets from the beginning of the BLM movement in 2013 to 2021. We summarize the data set and show temporal trends in use of both the *BlackLivesMatter* keyword and keywords associated with counter movements. Additionally, for each keyword, we create and release a set of Latent Dirichlet Allocation (LDA) topics (i.e., automatically clustered groups of semantically co-occuring words) to aid researchers in identifying linguistic patterns across the three keywords.

Introduction

The murder of George Floyd, an unarmed Black man, at the hands of police started a wave of global protests across the second half of 2020. In the U.S., the number of locations holding protests related to this event, as well as other killings of unarmed Black individuals such as Breonna Taylor and Ahmaud Arbery, outnumbered any other demonstration in U.S. history (Putnam, Chenoweth, and Pressman 2020). Notably, demonstrations were not limited to larger, urban areas, with protests occurring in all 50 states. An overwhelming number of these events were associated with the Black Lives Matter (BLM) movement (Kishi and Jones 2020), a decentralized grass roots movement protesting police brutality and violence against Black individuals. The global response to George Floyd's murder was in part due to the lose network of BLM related organizations, as well as previous demonstrations dating back to the movement's origins following the killing of unarmed Black teenager Trayvon Martin and the subsequent acquittal of perpetrator George Zimmerman. While support for BLM has fluctuated since its inception, support during the summer of 2020 had increased across all ethnic and racial groups (Horowitz 2020) and increased attention towards other Black victims of police violence (Wu et al. 2021).

Central to the BLM movement is advocacy against police violence toward Black individuals, which perpetuates negative health and psychological repercussions in Black individuals and communities. Multiple studies have shown the presence of racial bias in police violence (Ross 2015; Collaborators et al. 2021), with police use of force being one of the leading causes of death for Black men between 25 and 29 years of age (Edwards, Lee, and Esposito 2019). There is also evidence showing that police killings have negative effects on mental health in Black populations (Bor et al. 2018; Williams et al. 2018). Increases in depression and post-traumatic stress disorder in Black individuals are often higher than White individuals as a result of shared community violence (Galovski et al. 2016). Similar negative mental health effects in Black and Latin adolescents have been shown to be related to online exposure to traumatic events, such as widely shared videos of police killings (Tynes et al. 2019). The emotional impact of the murder of George Floyd was immediately felt: more than a third of the US population reported sadness and anger, with increased rates among Black Americans (Eichstaedt et al. 2021). Furthermore, a sentiment analysis of tweets showed that May 31, 2020, six days after the death of George Floyd, was the saddest day in Twitter's history (Schwartz 2020).

Given the global reach of the BLM movement, as well as the mental and physical health impacts of violence on Black communities (a central theme of the movement), we open-source a large-scale data set to facilitate associated research in the areas of computational social science, communications, political science, natural language processing, and machine learning. In the past, similarly themed, though much smaller in scope, BLM data sets have been used for studying discourse in protest and counter protest movements (Gallagher et al. 2018; Blevins et al. 2019), predicting retweets (Keib, Himelboim, and Han 2018), examining the role of social media in protest movements (Mundt, Ross, and Burnett 2018; Ince, Rojas, and Davis 2017; Wilkins, Livingstone, and Levine 2019), examining changes in implicit and explicit racial attitudes (Sawyer and Gampa 2018), and

	Tweets	Users	Retweets	Replies	Geotagged	User Location	Top Languages
All	63,884,799	13,061,316	47,083,420	3,266,120	86,641	10,820,854	en, fr, es, pt, ja
BlackLivesMatter	56,693,715	12,322,212	42,693,046	2,590,724	77,257	9,552,502	en, fr, es, pt, ja
AllLivesMatter	4,343,704	1,845,937	2,287,247	564,714	9,928	698,252	en, es, nl, ja, fr
BlueLivesMatter	5,075,833	1,224,933	3,494,159	306,711	2,329	938,335	en, fr, es, ja, de

Table 1: Descriptive counts for the entire data set and each keyword. Note that tweets can contain more than one keyword and can therefore be included in more than one row. ISO 639-1 Language codes: en = English, fr = French, es = Spanish, pt = Portuguese, ja = Japanese, nl = Dutch, de = German.

exploring narrative agency (Yang 2016). Research has also shown that Russian disinformation campaigns have infiltrated the BLM conversation on social media (Aceves 2018), in which case this data set could be used to study these campaigns. The current data set has been used for evaluating automatic event extraction systems in the context of sociopolitical events (Giorgi et al. 2021; Hürriyetoğlu et al. 2021).

These data are useful because they showcase the entire timeline of a large, ongoing social movement (Black Lives Matter) and its counter protests (All Lives Matter and Blue Lives Matter). To our knowledge, no other Twitter data sets exist that cover the entire span of the Black Lives Matter movement to date.

All researchers interested in systemic racism, social movements, grassroots campaigns, racial inequality, police brutality and counter protests, especially those working in the fields of computational social science, computational linguistics, communications, and political science, can benefit from this data.

Data Description

All data is available through Zenodo (Giorgi et al. 2022).

Tweets

Tweets containing the keywords *BlackLivesMatter*, *AllLivesMatter*, and *BlueLivesMatter* were collected through the Twitter API from January 2013 to December 31, 2021. Table 1 contains counts of total number of tweets and users for the entire data set and each keyword. It also includes counts for the following: retweets (original tweets which are shared by other users on the platform), replies (tweets which directly respond to another tweet), geotagged (latitude/longitude coordinates associated with the tweet), user location (free text field which we were able to map to U.S. counties; see details below) and top languages (automatically detected language of the tweet). Retweets may or may not contain additional content created by the user doing the retweeting.

Tweets also contain a large number of other metadata, such as user profile data and place information. User profiles contain information such as user handles, free text descriptions (often called "bios"), and profile pictures. Places are named locations users decide to associate with a tweet. While Places describe physical locations, they do not necessarily imply that the tweet originated from this location. Twitter users may manually tag a location when their tweet is about that Place, regardless of the user's location at the time of posting. Due to the large number of additional fields available for each tweet, we do not provide counts for any additional content.

The monthly volume of each keyword is plotted in Figure 1. Here we plot the seven day running average of the total count (logged) of all tweets containing one of our keywords. We also label high profile events (e.g., deaths, court related events, and viral videos) which resulted in an increase in BLM related activity. All labels marked with a single name indicate the date of police brutality-related killings.

In Figure 2 we visualized the spread of tweets across the United States over three equal time intervals: 2013 to 2015, 2016 to 2018, and 2019 to 2021. Tweets are mapped to U.S. counties using tweet level latitude and longitude coordinates and self-report location information via a free text field in the user's profile. First, if a tweet object contains latitude and longitude coordinates, then the tweet can be trivially mapped to a U.S. county. Next, we examine the location free text field in the user profile and use a rule-based system to match this text to a list of unambiguous U.S. cities (i.e., New York City as opposed to Springfield) which can then be mapped to U.S. counties. This process is described in Schwartz et al. (2013).

Our data set consists of monthly csv files which contains a single row for each tweet. Rows consist of the numeric tweet id (status_id; as given by the Twitter API) and three binary indicators for whether or not the tweet contains a *BlackLivesMatter*, *AllLivesMatter*, or *BlueLivesMatter* related keyword (for four columns total).

LDA Topics

The topic sets for each keyword contain three values: *topic*, *term*, and *weight*. The *topic* column is a numeric indicator for each topic: 1 through 100 for *BlackLivesMatter*, 1 through 50 for *AllLivesMatter*, and 1 through 25 for *BlueLivesMatter*. The *term* column is the word within the topic. The *weight* column is the conditional probability of the topic given the term, as derived through the LDA process. For each LDA topic set, we visualize the most prevalent topics across each corpus. To do this, we extract the relative frequency of single words (i.e., tokens) from each tweet in the no retweet, no reply, single keyword data sets described above: 10,881,298 *BlackLivesMatter* tweets, 976,244 *AllLivesMatter* tweets, and 1,069,362 *BlueLivesMatter* tweets. For each tweet we calculate the conditional probability of the topic given the tweet:

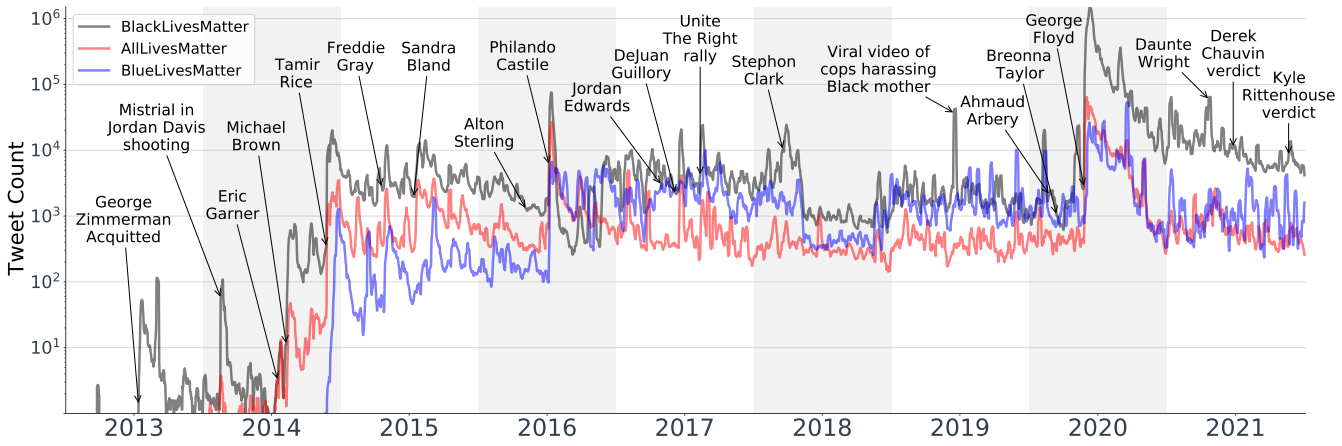


Figure 1: Seven day moving average of monthly tweet count from 2013 to 2021 of all three keywords. We include annotations for high profile events associated with the BLM movement.

$$P(\text{topic}|\text{tweet}) = \sum_{\forall \text{token} \in \text{topic}} P(\text{topic}|\text{token}) \times P(\text{token}|\text{tweet}). \quad (1)$$

Here $P(\text{topic}|\text{token})$ is the conditional probability of the topic given the topic, which is estimated through the LDA process. We estimate $P(\text{token}|\text{tweet})$ as the relative frequency of the token given the tweet. Then for each tweet we have a conditional probability of each topic, which we average for each topic. Figure 3 shows the 5 topics for each keyword with the highest average condition probability.

Data Set Creation

Data Collection

On July 14, 2016, we set up a data puller using the Python package TwitterMySQL[1] to collect tweets matching at least one of our keywords: *BlackLivesMatter*, *AllLivesMatter* and *BlueLivesMatter*. This package uses the official Twitter Application Programming Interface (API) to stream tweets in real time. The data puller continuously collected tweets from the Twitter stream until December 31, 2021. In total we collected 67,336,447 tweets. While the Twitter API was queried using the keywords *BlackLivesMatter*, *AllLivesMatter* and *BlueLivesMatter*, the API delivers a more robust set of matching tweets. For example, a tweet might contain the phrase "black lives matter", "blm" or "#blacklivesmatter", among other variations, instead of the exact keyword *BlackLivesMatter*.

We note that the Twitter API limits such streams to 1% of the total Twitter volume at any given moment. To see if our keyword data set was limited at any point, we compared the monthly keyword volume to a full 1% monthly pull (not limited to any single keyword, location, etc.) Our keyword data set pulled in a monthly average of 1,463,835 tweets (4,630,450 SD) as compared to a monthly average of 96,385,502 tweets (27.146,801 SD) from the 1% pull. Since our data set is much smaller than the 1% sample we do not believe our data set was limited by the Twitter API.

Due to server maintenance, there were periods when we were unable to collect data. These include: October 17 through November 23, 2016; January 1 through January 21, 2017; March 11 through March 16, 2017; May 2 through December 18, 2018; and March 16 through March 20, 2019; June 1 through June 3, 2021; and November 19 through November 21, 2021. Additionally, the Black Lives Matter movement began in 2013, roughly three years before the beginning of our data collection. In order to fill these gaps, we used the Python package GetOldTweets3[2], which pulls historical tweets containing a given keyword. Using this method, we collected 4,276,423 historical tweets across the dates listed above (i.e., the gaps in our data).

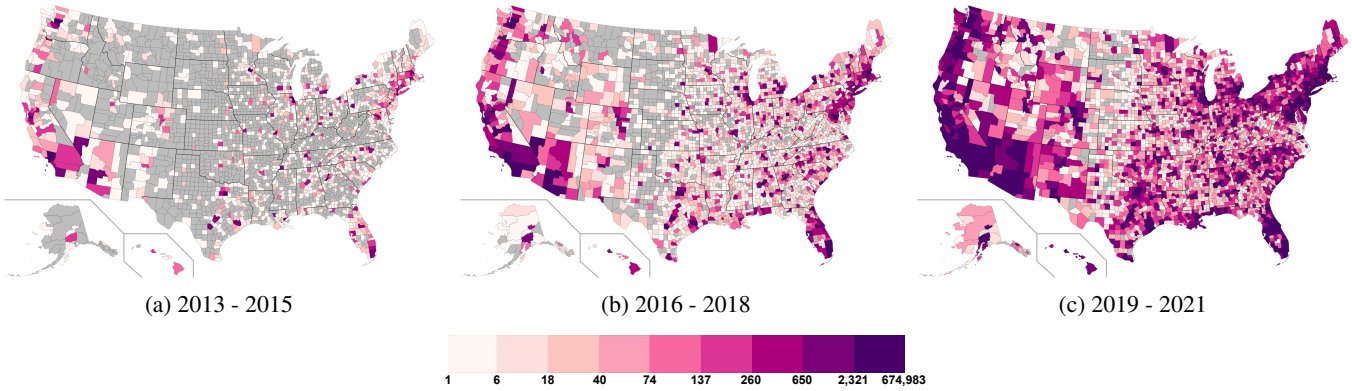
While Twitter data is publicly available, at any point a user may delete a tweet, delete their account, or set their account to private. Thus, when pulling prospective data, we collected tweets which may have been deleted or made private at some point after the initial pull. On the other hand, deleted or private tweets cannot be pulled with a retrospective collection. Thus, the number of tweets pulled prospectively or retrospectively can be very different, especially as one goes further back in time. In order to ensure the data set only contained presently available tweets, we executed a one-time historical. As a result, any tweet deleted after our initial pull will not be made available. Our final data set consisted of 63,884,799 tweets.

Topic Modeling

For each keyword we created a set of topics using Latent Dirichlet Allocation (LDA; Blei, Ng, and Jordan 2003). LDA is a Bayesian mixture model which automatically groups together words that frequently appear in similar contexts. Topic models such as LDA are often used to statistically derive categories in a data driven fashion, rather than manually assigning words to predetermined categories.

[1]https://github.com/dlatk/TwitterMySQL

[2]https://github.com/Mottl/GetOldTweets3



(a) 2013 - 2015 (b) 2016 - 2018 (c) 2019 - 2021

Figure 2: Distribution of *BlackLivesMatter* tweets across the United States for three 3-year periods: 2013 to 2015, 2016 to 2018, and 2019 to 2021. Tweets are mapped to U.S. counties from latitude/longitude coordinates or self-reported user location. Counties grouped into 9 quantiles: darker shades indicate higher tweet density, lighter shades indicate lower tweet density, and grey areas contain no tweet data.

Given the highly specific nature of our data set, we created Content Specific LDA (CSLDA) topics (Zamani et al. 2020). CSLDA is a method used for generating topics across a thematically narrow corpus (i.e., tweets about BLM as opposed to a random selection of tweets) and has successfully been used to model excessive drinking (Giorgi et al. 2020), diabetes (Griffis et al. 2020), and COVID-19 (Zamani et al. 2020) discussion on Twitter. In particular, CSLDA uses a text pre-processing step, executed before topic modeling, which identifies words most associated with the theme (i.e., Black Lives Matter). CSLDA does not assume that words frequently appearing in the keyword tweets are associated with the keyword. For example, the retweet keyword "RT" appears in a large number of our tweets, but is more associated with the Twitter platform than *BlackLivesMatter*. The CSLDA pipeline is briefly described below. Further details on CSLDA can be found in Zamani et al. (Zamani et al. 2020).

In order to find words that are most associated with each keyword, we first built a corpus comprised of a random sample of tweets containing our keywords and a matched sample of tweets that do not. For each keyword, we randomly select 500,000 tweets that are neither replies nor retweets. We also select tweets containing only a single keyword, that is, no tweet contains some combination of *BlackLivesMatter*, *AllLivesMatter*, and *BlueLivesMatter* keywords. For the matched sample, we randomly selected 500,000 tweets that do not contain any of the three keywords. These tweets were selected from a random 1% stream of publicly available data and are also (1) not replies nor retweets, (2) written in English (as reported by Twitter's API), and (3) match the same temporal distribution as the random 500,000 *BlackLivesMatter* tweets above (i.e., the number of tweets per year for the non-keyword set matches the number of tweets per year in the 500,000 *BlackLivesMatter* tweets). We then created three sets of one million tweets separately combining the random 500,000 random tweets with our three sets of 500,000 keyword tweets.

Next, we broke up each tweet into its constituent words, in a process called tokenization. As opposed to splitting up the tweets by white space, we use a tokenizer built specifically for social media text that can pick emojis, hashtags, and misspellings (Schwartz et al. 2017). For each tweet, we then created a feature vector which records relative frequency of the words appearing in the tweet. We also created a binary outcome for each tweet in the three matched data sets above: 1 if the tweet contains one of the keywords and 0 otherwise. Using this binary outcome and the feature vector, we calculated a weighted log odds ratio using an Informative Dirichlet prior (Jurafsky et al. 2014). This calculation estimates the difference in word frequency across the keyword data sets and their matched samples, using a prior which shrinks each keyword word frequency towards known frequencies in the matched sample. In the end, we took the top 3,000, 2,500, and 2,500 associated unigrams (i.e., largest weighted log odds rations) for the *BlackLivesMatter*, *AllLivesMatter*, and *BlueLivesMatter* keywords, respectively. A larger number of unigrams associated with *BlackLivesMatter* were chosen because the data set contains significantly more *BlackLivesMatter* tweets than the *AllLivesMatter* and *BlueLivesMatter* data sets. Finally, we considered all tweets that are neither retweets nor replies and only contain a single keyword (i.e., no combination of *BlackLivesMatter*, *AllLivesMatter*, and *BlueLivesMatter* keywords). These data sets contain 9,758,272; 1,386,087; and 1,167,006 tweets for *BlackLivesMatter*, *AllLivesMatter*, and *BlueLivesMatter*, respectively. We then filtered these tweets to contain only the unigrams derived in the previous step, removing words that do not differ in log odds frequency vs. the matched set. We also removed urls, @-mentions, and all variations of the keywords (e.g., black, lives, matter, blm, #blm, blacklivesmatter, and #blacklivesmatter) The LDA process was then run over these filtered data sets.

We used the Mallet software package,[3] which uses Gibbs sampling (Gelfand and Smith 1990) to estimate the latent variables of the topic. All default Mallet parameters are used

[3]http://mallet.cs.umass.edu

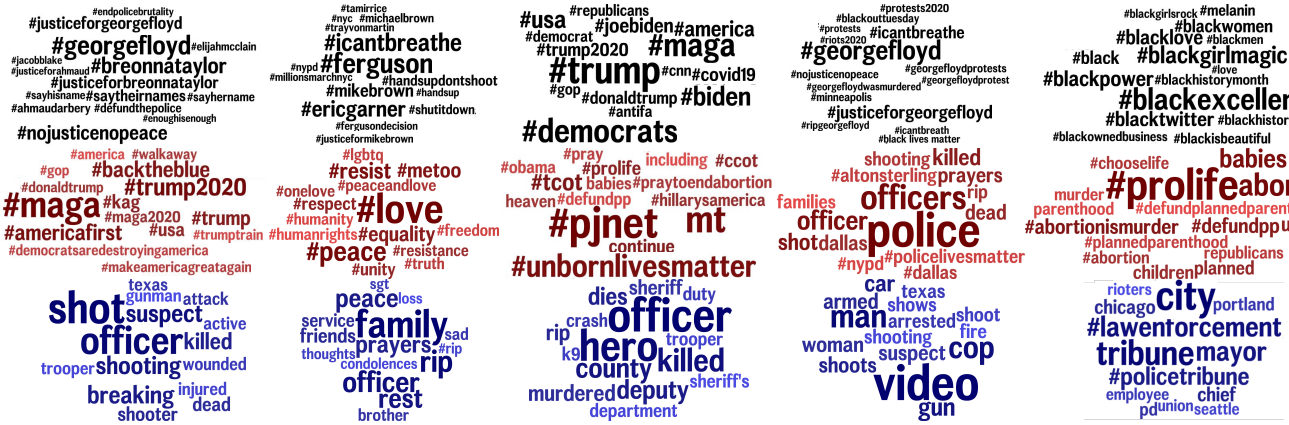


Figure 3: Word cloud visualizations of the top five most prevalent topics in the *BlackLivesMatter* (colored in black), *AllLivesMatter* (colored in red), and *BlueLivesMatter* (colored in blue) data sets, respectively. Topics are ordered left to right by descending prevalence (average conditional probability of topic given tweet). Word clouds contain the 15 most prevalent words within the topic and words are sized according to their prevalence relative to other words in the topic.

except for α, the prior on the expected number of topics per document. Here $\alpha = 2$, since our tweets are shorter than typical documents (such as newspaper articles or blog posts), and thus contain fewer topics. This value has previously been used in CSLDA over Twitter data (Giorgi et al. 2020). For each keyword, we created three sets of LDA topics, varying the number of topics per set. For *BlackLivesMatter* we created sets of 25, 50, and 100 topics. For both *AllLivesMatter* and *BlueLivesMatter* we created sets of 10, 25, and 50 topics, again noting that the total number of *BlackLivesMatter* tweets are more than 10 times the number of tweets in the other two keywords data sets. Other studies which have created LDA topics over thematically similar tweet data sets (e.g., COVID-19, excessive alcohol consumption, and maternal health) have used similarly sized topic sets (Zamani et al. 2020; Giorgi et al. 2020; Abebe et al. 2020). Figure 4 shows the full pipeline.

In order to pick the final topic set, three of the authors qualitatively ranked the three sets. All raters were asked to consider (1) breadth of themes, (2) single topics contain single themes, and (3) themes are not repeated across a large number of topics. A similar qualitative process was used to identify Twitter topics associated with maternal morality (Abebe et al. 2020). All three raters agreed on the 100 topic set for *BlackLivesMatter* and the 50 topic set for *AllLivesMatter*. The three raters did not initially agree on the *BlueLivesMatter* topic set (with raters split between the 25 and 50 topic sets), but eventually settled on the 25 topic set since themes repeated across a large number of topics in the 50 topic set.

To evaluate the quality of the topics, we use two metrics. The first metric, Topic Uniqueness (TU), is a measure of diversity (Nan et al. 2019). TU is inversely proportional to the number of times a set of L representative keywords is repeated across the set of K topics, with high TU meaning the words are rarely repeated, and thus the topics are unique. Traditionally, TU scores are between 1 and $1/K$, where K is the number of topics. Since the number of topics across our three keywords varies, we normalize the TU score to be between 0 and 1. We set $L = 30$ and K equal to 100, 50, and 25, for the *BlackLivesMatter*, *AllLivesMatter*, and *BlueLivesMatter* topic sets, respectively. This produces TU scores of .79 for *BlackLivesMatter*, .97 for *AllLivesMatter*, and 1.0 for *BlueLivesMatter*. We note that as L increases, TU scores should decreases since the probability of a given word appearing in more than one topic will increase. Traditionally, L is set to 10, which we increase to 30 in order to give a more conservative estimate (Nan et al. 2019).

The second metric measures coherence, or the semantic similarity between the words in the topic, using Normalized Pointwise Mutual Information (NPMI) (Syed and Spruit 2017). Coherence is calculated for each topic and then averaged across all topics within each keyword topic set. Scores range between 0 and 1, where topics with high semantic similarity between words having scores closer to 1. We use the Gensim Python package to calculate these scores (Řehůřek and Sojka 2010) and see scores of 0.64 for *BlackLivesMatter*, 0.73 for *AllLivesMatter*, and 0.70 for *BlueLivesMatter*.

Usage Notes

Due to Twitter's Terms of Service, only numeric tweet IDs can be publicly shared. The numeric IDs can be used to pull the full tweet set using the Twitter API. There are a number of open source software packages which allow researchers to

	Number of Topics	Topic Uniqueness	Coherence
BlackLivesMatter	100	1.00	0.64
AllLivesMatter	50	0.97	0.73
BlueLivesMatter	25	0.79	0.70

Table 2: Topic quality as measured by Topic Uniqueness and coherence.

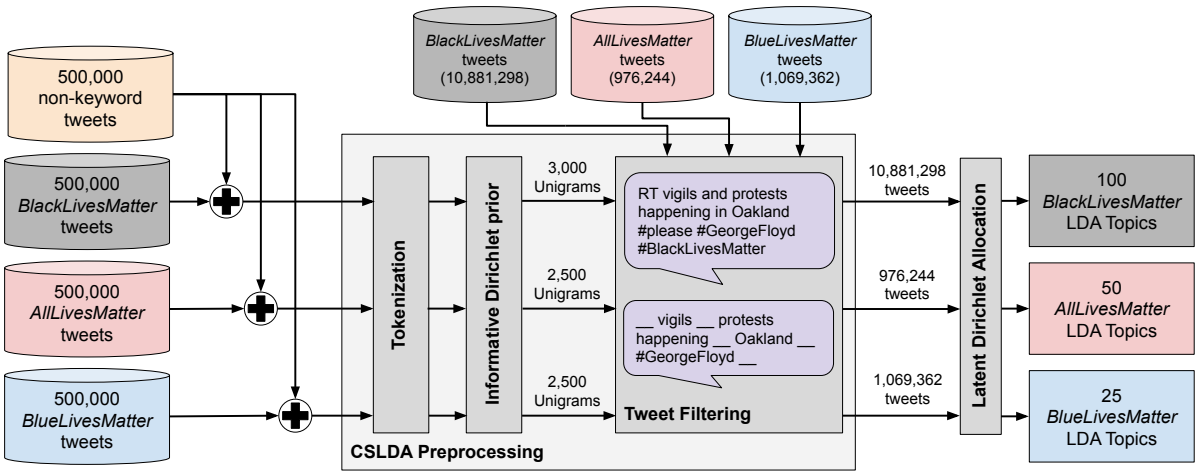


Figure 4: Content Specific LDA pipeline used for creating the three LDA topic sets. All tweets contained a single keyword (e.g., *AllLivesMatter* keywords and not *BlackLivesMatter* and *BlueLivesMatter* keywords) and no retweets or replies were used.

easily interface with the API. The authors used the Python package TwitterMySQL[1], which saves tweet information in a MySQL database. Other packages exist which do not rely on relational databases, such as the Python package twarc[4], which saves tweets to text files in JSON format, or Hydrator[5], which relies on an easy to use GUI and saves tweets to both JSON and CSV formats. Regardless of which tool is used to download the Twitter data, researchers need an active Twitter Developer account in order to access their API.

Code Availability

The data sets have been created using Python 3.5 and MySQL 5.5. The code is available through GitHub[6]. A Python script and short tutorial have been developed to aid in obtaining this data from the Twitter API.

Conclusions

To date, this data set is the largest publicly available collection of Black Lives Matter related social media posts. This data set was created to aid researchers in studying social media activity and discourse around the Black Lives Matter movement, with no specific task in mind. It is our hope that a central repository for this large, multi-year data base will give researchers easier access to the data, especially those researchers less comfortable using open source APIs or who lack computational bandwidth or storage capacity.

We believe there are a number of applications and potential use cases for this data, which could include analyses of conversations, temporal shifts, and spatial trends. From this data set, one could pull conversations associated with each tweets (i.e., replies and retweets) via the Twitter API. Using this larger data set one could examine conversations within and between BLM and the counter protests. Additionally, one could compare conversations before and after the murder of George Floyd, since there was a major increase in Twitter activity after that event.

Using location data associated with the tweets and Twitter users, one could examine how BLM social media activity and conversations have evolved spatially. Figure 2 shows how the movement has grown within the U.S. and we note that the data set contains tweets from over 100 countries. One could attempt to look at spatial or cultural components of BLM conversations.

Additionally, one could combine the Twitter data with other geolocated data such as demographics, socioeconomics, and voting behavior. Past studies have examined BLM related protests and police-caused deaths (Williamson, Trump, and Einstein 2018), both of which could be combined with this Twitter data. Within the U.S., there are a number of publicly available data sets which measure racial attitudes at the population level. One such data set is from Project Implicit[7] which has publicly released data from the race implicit association test (IAT; Greenwald, McGhee, and Schwartz 1998). This and other similar data sets can be mapped to U.S. locations such as counties and then correlated with the BLM Twitter data. A similar analysis was carried out by Sawyer and Gampa (2018), who examined implicit and explicit attitudes during various phases of the BLM movement.

Ethics Statement

Here we consider the ethical questions outlined in Datasheets for Datasets (Gebru et al. 2021). As with any Twitter data set, there are a number of ethical concerns, especially due to the nature of this data set (i.e., racial justice and grassroots social movements). First, it is possible to identify individuals within the data set, though we note

[4]https://github.com/DocNow/twarc

[5]https://github.com/DocNow/hydrator

[6]https://github.com/sjgiorgi/blm_twitter_corpus

[7]https://osf.io/y9hiq/

that this is only possible after rehydrating the tweet set and only because the data is publicly available (i.e., public Twitter profiles with public tweets). Thus, an individual would have to have posted something identifying on their account at some point in time that remains public at the time of rehydration. Along with any Twitter data, there are other possibly sensitive attributes such as images, location information, and friend/follower networks. Due to the nature of the data set, this could also potentially identify a person's support or opposition to a political movement. Finally, no individuals within the data set have consented to the authors to have their data shared, though, again, we note that the data used in this article is publicly available and distributed within Twitter's Terms of Services (i.e., only numeric tweet IDs are distributed). While there is no official way to "revoke consent", Twitter users may delete tweets, delete accounts, or set accounts to private, at which point any tweet in our data set would no longer be available. It was our hope that repulling all data at the time of writing would ensure the numbers reported within reflect the most current publicly available version of the data set.

While we have released tweets related to counter protests (All Lives Matter and Blue Lives Matter), the authors do not intend to draw equivalences between the counter protests and Black Lives Matter. Despite the fact that all three protests are given equal space in the analysis, we believe the numbers reported show a different story: an overwhelming majority of tweets are related to *BlackLivesMatter* as opposed to the other protests.

It is our hope that this data set is used for social good, though there are a number of questionable use cases such as monitoring or forecasting of protests by law enforcement. As such, we limit the distribution of this data to non-commercial, research entities in hopes of limiting surveillance type tasks by for-profit entities.

Acknowledgments

This research was supported in part by the Intramural Research Program of the NIH, National Institute on Drug Abuse (NIDA).

References

Abebe, R.; Giorgi, S.; Tedijanto, A.; Buffone, A.; and Schwartz, H. A. A. 2020. Quantifying Community Characteristics of Maternal Mortality Using Social Media. In *Proceedings of The Web Conference 2020*, 2976–2983.

Aceves, W. J. 2018. Virtual hatred: How Russia tried to start a race war in the united states. *Mich. J. Race & L.*, 24: 177.

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

Blevins, J. L.; Lee, J. J.; McCabe, E. E.; and Edgerton, E. 2019. Tweeting for social justice in# Ferguson: Affective discourse in Twitter hashtags. *new media & society*, 21(7): 1636–1653.

Bor, J.; Venkataramani, A. S.; Williams, D. R.; and Tsai, A. C. 2018. Police killings and their spillover effects on the mental health of black Americans: a population-based, quasi-experimental study. *The Lancet*, 392(10144): 302–310.

Collaborators, G. . P. V. U. S.; et al. 2021. Fatal police violence by race and state in the USA, 1980–2019: a network meta-regression. *The Lancet*, 398(10307): 1239–1255.

Edwards, F.; Lee, H.; and Esposito, M. 2019. Risk of being killed by police use of force in the United States by age, race–ethnicity, and sex. *Proceedings of the National Academy of Sciences*, 116(34): 16793–16798.

Eichstaedt, J. C.; Sherman, G. T.; Giorgi, S.; Roberts, S. O.; Reynolds, M. E.; Ungar, L. H.; and Guntuku, S. C. 2021. The emotional and mental health impact of the murder of George Floyd on the US population. *Proceedings of the National Academy of Sciences*, 118(39).

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

Galovski, T. E.; Peterson, Z. D.; Beagley, M. C.; Strasshofer, D. R.; Held, P.; and Fletcher, T. D. 2016. Exposure to violence during Ferguson protests: Mental health effects for law enforcement and community members. *Journal of Traumatic Stress*, 29(4): 283–292.

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

Gelfand, A. E.; and Smith, A. F. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410): 398–409.

Giorgi, S.; Guntuku, S. C.; Himelein-Wachowiak, M.; Kwarteng, A.; Hwang, S.; Rahman, M.; and Curtis, B. 2022. Twitter Data of the #BlackLivesMatter Movement And Counter Protests: 2013 to 2021.

Giorgi, S.; Yaden, D. B.; Eichstaedt, J. C.; Ashford, R. D.; Buffone, A. E.; Schwartz, H. A.; Ungar, L. H.; and Curtis, B. 2020. Cultural Differences in Tweeting about Drinking Across the US. *International Journal of Environmental Research and Public Health*, 17(4): 1125.

Giorgi, S.; Zavarella, V.; Tanev, H.; Stefanovitch, N.; Hwang, S.; Hettiarachchi, H.; Ranasinghe, T.; Kalyan, V.; Tan, P.; Tan, S.; Andrews, M.; Hu, T.; Stoehr, N.; Re, F. I.; Vegh, D.; Atzenhofer, D.; Curtis, B.; and Hürriyetoğlu, A. 2021. Discovering Black Lives Matter Events in the United States: Shared Task 3, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, 218–227. Online: Association for Computational Linguistics.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6): 1464.

Griffis, H.; Asch, D. A.; Schwartz, H. A.; Ungar, L.; Buttenheim, A. M.; Barg, F. K.; Mitra, N.; and Merchant, R. M.

2020. Using Social Media to Track Geographic Variability in Language About Diabetes: Infodemiology Analysis. *JMIR diabetes*, 5(1): e14431.

Horowitz, J. 2020. *Amid protests, majorities across racial and ethnic groups express support for the Black Lives Matter movement*. Pew Research Center.

Hürriyetoğlu, A.; Tanev, H.; Zavarella, V.; Piskorski, J.; Yeniterzi, R.; Mutlu, O.; Yuret, D.; and Villavicencio, A. 2021. Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021): Workshop and Shared Task Report. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, 1–9. Online: Association for Computational Linguistics.

Ince, J.; Rojas, F.; and Davis, C. A. 2017. The social media response to Black Lives Matter: how Twitter users interact with Black Lives Matter through hashtag use. *Ethnic and racial studies*, 40(11): 1814–1830.

Jurafsky, D.; Chahuneau, V.; Routledge, B. R.; and Smith, N. A. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*.

Keib, K.; Himelboim, I.; and Han, J.-Y. 2018. Important tweets matter: Predicting retweets in the# BlackLivesMatter talk on twitter. *Computers in Human Behavior*, 85: 106–115.

Kishi, R.; and Jones, S. 2020. Demonstrations & Political Violence in America: New Data for Summer 2020. *The Armed Conflict Location & Event Data Project (ACLED)*.

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

Nan, F.; Ding, R.; Nallapati, R.; and Xiang, B. 2019. Topic Modeling with Wasserstein Autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6345–6381.

Putnam, L.; Chenoweth, E.; and Pressman, J. 2020. The Floyd Protests are the broadest in US history—and are spreading to white, small-town America. *Washington Post*, 6.

Řehůřek, R.; and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Ross, C. T. 2015. A multi-level Bayesian analysis of racial bias in police shootings at the county-level in the United States, 2011–2014. *PloS one*, 10(11): e0141854.

Sawyer, J.; and Gampa, A. 2018. Implicit and explicit racial attitudes changed during Black Lives Matter. *Personality and Social Psychology Bulletin*, 44(7): 1039–1059.

Schwartz, C. 2020. Is Everybody Doing ... OK? Let's Ask Social Media. *The New York Times*. Accessed: 2021-07-24.

Schwartz, H.; Eichstaedt, J.; Kern, M.; Dziurzynski, L.; Lucas, R.; Agrawal, M.; Park, G.; Lakshmikanth, S.; Jha, S.; Seligman, M.; et al. 2013. Characterizing geographic variation in well-being using tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 583–591.

Schwartz, H. A.; Giorgi, S.; Sap, M.; Crutchley, P.; Ungar, L.; and Eichstaedt, J. 2017. DLATK: Differential language analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 55–60.

Syed, S.; and Spruit, M. 2017. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)*, 165–174. IEEE.

Tynes, B. M.; Willis, H. A.; Stewart, A. M.; and Hamilton, M. W. 2019. Race-related traumatic events online and mental health among adolescents of color. *Journal of Adolescent Health*, 65(3): 371–377.

Wilkins, D. J.; Livingstone, A. G.; and Levine, M. 2019. Whose tweets? The rhetorical functions of social media use in developing the Black Lives Matter movement. *British Journal of Social Psychology*, 58(4): 786–805.

Williams, M. T.; Metzger, I. W.; Leins, C.; and DeLapp, C. 2018. Assessing racial trauma within a DSM–5 framework: The UConn Racial/Ethnic Stress & Trauma Survey. *Practice Innovations*, 3(4): 242.

Williamson, V.; Trump, K.-S.; and Einstein, K. L. 2018. Black lives matter: Evidence that police-caused deaths predict protest activity. *Perspectives on Politics*, 16(2): 400–415.

Wu, H. H.; Gallagher, R. J.; Alshaabi, T.; Adams, J. L.; Minot, J. R.; Arnold, M. V.; Welles, B. F.; Harp, R.; Dodds, P. S.; and Danforth, C. M. 2021. Say Their Names: Resurgence in the collective attention toward Black victims of fatal police violence following the death of George Floyd. arXiv:2106.10281.

Yang, G. 2016. Narrative agency in hashtag activism: The case of #BlackLivesMatter. *Media and communication*, 4(4): 13.

Zamani, M.; Schwartz, H. A.; Eichstaedt, J.; Guntuku, S. C.; Ganesan, A. V.; Clouston, S.; and Giorgi, S. 2020. Understanding Weekly COVID-19 Concerns through Dynamic Content-Specific LDA Topic Modeling. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*. Association for Computational Linguistics.