
BLMTWITTER: THE BLACK LIVES MATTER (BLM) TWITTER CORPUS

TRACKING AN INCREASINGLY GLOBAL SOCIAL MOVEMENT ON SOCIAL MEDIA

• **A. Maurits van der Veen**
Department of Government
William & Mary
Williamsburg, VA 23185
maurits@wm.edu

August 19, 2022

ABSTRACT

In May 2020, the murder of George Floyd led to protests throughout the United States and around the world. Along with these protests, social media platforms saw a dramatic increase in the volume of discussion regarding Black Lives Matter (BLM). Two years on, online discussion of BLM continues to be lively, not just in the United States but also internationally. This paper introduces one of the most extensive collections of Black Lives Matter tweets available, covering from 2010 through July 2022 and including more than 50 million tweets overall. Parallel corpora for All Lives Matter and Blue Lives matter are similarly collected. I describe the corpus construction process, including selection criteria for tweets using the BLM acronym. The paper provides a summary description of key dimensions of the three corpora, including patterns in tweet counts over time as well as the number of distinct users and the volume of tweets in languages other than English. The corpus is publicly available and updated regularly.

Keywords Black Lives Matter · BLM · social movements · social media · Twitter

1 Introduction

The Black Lives Matter movement has grown to be global in scope. Much of its activity takes place online, on social media such as Twitter (Freelon et al., 2016). While a number of studies have looked at the networks and trends on social media associated with the movement, using bespoke corpora and including a number of related hash tags and search strings, the present corpus is the first to make available, from the earliest days of the movement to the present, all retrievable tweets explicitly referencing Black Lives Matter (BLM). In addition, accompanying the BLM Twitter corpus are parallel corpora for the related phrases “All Lives Matter” and “Blue Lives Matter”. The BLM corpus contains 50.1 million tweets through July 2022; the other two are much smaller, at 3.9 and 2.3 million tweets, respectively.

Scholars and journalists have studied Black Lives Matter discussions on social media from a number of different angles: how the volume of posts or tweets varies with and after real-life events (Dunivin et al., 2022; Center et al., 2020); how hashtags associated with Black Lives Matter and All Lives Matter have been used strategically (Gallagher et al., 2018); which hashtags are most commonly used in conjunction with references to BLM and how this helps frame the movement (Ince et al., 2017); how discussions on Twitter facilitate the making of connections and real-world mobilization (Mundt et al., 2018); and which groups are prominent contributors to the online debates, and how this has changed over time (Freelon et al., 2016), among many others.

Most of these studies analyze a comparatively brief period of time, while including a wide range of hashtags related to Black Lives Matter. This helps improve our understanding of the wider movement at a specific moment. However, it risks obscuring patterns over time. Moreover, in some cases tweets explicitly mentioning Black Lives Matter represent only a minority of all tweets studied (Freelon et al., 2016), which may obscure specific characteristics of the discourse

surrounding BLM, as opposed to a broader discourse on policing, for example. The present corpus focuses specifically on BLM, flagging the specific phrases or terms used, and going back to the first tweets to mention the term, in 2010.

In accordance with Twitter regulations, only tweet ids are made available. These can be easily “hydrated” using a variety of applications developed for that purpose, such as Hydrator (<https://github.com/DocNow/hydrator>). The ids are divided into chronological chunks, to facilitate analyzing just a particular period. For each tweet id, the files also indicate which of the relevant search phrases are found in the associated tweet. This makes it possible to analyze individual phrases separately. The corpora are updated at monthly intervals, and are available on Zenodo¹ and Github.²

The remainder of this document has two parts. The first describes the corpus creation process; the second provides a brief overview of the three corpora, in terms of total tweets, number of distinct users, and the main languages (other than English) represented in the data.

2 Tweet collection

The main method of tweet collection for the present corpora is retrospectively, using the python module `twint` (<https://github.com/twintproject/twint>). The tweets thus collected are supplemented from a comparable corpus collected by Giorgi et al. 2020; 2022. There are some important differences between the present corpus and that collected by Giorgi et al., which are outlined below. The first part of this section describes the main tweet collection process and criteria, following which I discuss the supplementation from the Giorgi et al. corpus.

2.1 Retrospective collection

I collected all retrievable direct references to Black/All/Blue Lives Matter in tweets. The collection process started during the late Summer of 2021 and is ongoing. Specifically, tweet collection is done using the search strings shown in Table 1 (capitalization is ignored in all searches).

Table 1: Corpus search strings

Black Lives Matter	All Lives Matter	Blue Lives Matter
“black lives matter” ^a	“all lives matter”	“blue lives matter”
blacklivesmatter	alllivesmatter	bluelivesmatter
blkivesmatter ^b	allivesmatter (2 Ls)	
blm		

^a Twitter’s search function appears to accept non-alphanumeric characters in addition to or in lieu of the specified spaces. Therefore the corpus also contains tweets with phrases such as “Black. Lives. Matter” or “#Black #Lives #Matter”. In addition, the corpus also contains a few thousand tweets that misspell “matter” as “mater”.

^b This is also the Twitter user id of the Black Lives Matter organization. It is used much less often than the other hashtags or key phrases.

For the acronym search string “blm”, a straightforward search is not possible, for several reasons. First, in Indonesian tweets, “blm” is generally used to mean “not yet,” rather than to refer to Black Lives Matter. Accordingly, I exclude Indonesian-language results that contain “blm” but not any of the longer Black Lives Matter search strings. I do so by collecting “blm” tweets by language code, excluding Indonesian (‘in’). Specifically, I collect language codes for all languages in the Black Lives Matter corpus (other than Indonesian) that use one of the longer search strings at least once, and then search for tweets in those languages containing “blm”. This includes a total of 64 languages other than English, plus the language code ‘und’, used by Twitter to indicate an undetermined language (usually because a tweet is too short to enable confident language identification).³ As an added filter, I also remove any tweets that contain only a sequence of usernames (each preceded by the @ sign), followed by, as the last word, “blm”. More often than not, these are also Indonesian tweets, but the lack of any words in the tweet other than the “blm” makes it difficult for Twitter to correctly code the language.

¹<https://doi.org/10.5281/zenodo.6628275>.

²<https://github.com/amaurits/BLMtwitter/>.

³The approach adopted possibly misses a few tweets posted in languages that are otherwise not present at all in the corpus. However, it is unlikely that more than a few dozen tweets in total are missed. As it stands, the corpus includes 65 languages (plus ‘undetermined’), and for the least common language, Uyghur (language code ‘ug’), there is just a single tweet.

A second issue with “blm” is that it forms part of many Twitter usernames. Some of these are Black Lives Matter organizations, but most are not. Erring on the side of restrictiveness, I exclude all tweets that only contain “blm” as part of a username (i.e. preceded by the initial character “@”), with the exception of @blm_to, which is the account of the Canadian Black Lives Matter organization, and @ukblm, which is the United Kingdom (UK) account. Of course, any tweets which additionally include one of the search strings that is not directly preceded by “@” are included as well.

Lastly, the acronym BLM can stand for other things than Black Lives Matter. Most notably, it also refers to the United States Bureau of Land Management. That organization was engaged in a standoff with the Bundy family in 2014 that produced a large number of tweets referencing BLM. To eliminate most of these, I remove from the corpus any tweet that contains “blm” along with either “Bureau of Land Management”, “Cliven”, or “Bundy”, but does not contain any of the longer Black Lives Matter search strings.

The tweet collection process has certain limitations. First, tweets that have since been deleted are no longer retrievable. With highly politicized topics, the tweet deletion rate tends to be quite high (Bastos, 2021). In addition, tweets by accounts that have been suspended by Twitter are also eliminated: they are retrieved in the initial collection process, but the tweet text is replaced by Twitter with: “<username>’s account is temporarily unavailable because it violates the Twitter Media Policy.”⁴ Second, on days where a particular phrase is used a lot — most obviously, in the immediate aftermath of the murder of George Floyd — not all tweets may be retrieved successfully later. To fill in any such gaps as much as possible, I supplement the collection from the BLM twitter corpus collected by Giorgi et al. 2022, which was largely collected contemporaneously and has been updated through the end of 2021.

2.2 Supplementation from prior contemporaneous collection

The Giorgi et al. corpus incorporates tweets related to Black Lives Matter as well as All Lives Matter and Blue Lives Matter. The original corpus, which ran through June 19, 2020, contained 41.8 million tweet ids across the three terms. Looking up those tweets (“hydrating”) by tweet id in the late Spring of 2021 recovered just 29.3 million tweets: a loss of 30%. Especially for a sensitive issue like BLM, deletions both by users themselves (regretting intemperate remarks) and by Twitter (for violations of norms and rules) tend to be quite common (Zubiaga et al., 2018). In fact, Giorgi et al. themselves encountered a loss rate of about 20% when they tried to retrospectively recapture tweets they had initially captured live between 2016 and 2020 (Giorgi et al., 2020, p. 3). A recent update to the corpus extends it through the end of 2021 (Giorgi et al., 2022).

I separate the corpus into subcorpora for Black, Blue, and All Lives Matter, assigning individual tweets to more than one corpus if more than one key term was mentioned.⁵ Table 2 shows the degree of overlap between the BLM tweets collected for the present dataset (identified as ‘vdV’) and those in the Giorgi corpus. Comparable tables for Blue and All Lives Matter appear in the appendix. To make file sizes more manageable, I break the corpus down by year through 2019 and by month thereafter, with the exception of the peak months of May through August 2020, which are broken down by week.

The data in Table 2 illustrate the value of combining tweets both contemporaneously and retrospectively. The contemporaneous collection method primarily used by Giorgi et al. netted many tweets that did not get captured when requesting tweets after the fact, even though they still exist (i.e. they have not been deleted nor are they associated with a suspended account). Conversely, retrospectively collecting tweets garners many tweets that the contemporaneous collection process missed. Indeed, the relative contribution of the retrospective collection process is greater in 88% (42 of 48) of the time periods in the table (through 2021), at times significantly so. The problem of incomplete collection may be greatest when tweet volumes are high. If we look, for example, at the first week of June 2020 in Table 2 — after tweet volumes had exploded following George Floyd’s murder — the two separate collection methods overlap for less than 10% of the total collected, with each uniquely contributing more than 40% of the count for that week.

These patterns indicate that it is impossible to guarantee completeness of a corpus. The present corpus is considerably more complete than that collected by Giorgi et al., since it uniquely contributes more than half of the total volume of tweets in the corpus for the period where they overlap, i.e. through the end of 2021 (24.7 mn. out of 48.0 mn. total, versus 15.4 mn. tweets uniquely contributed by the Giorgi corpus). At the same time, the corpus is also more strictly defined: tweets not explicitly mentioning a search phrase are eliminated, even though many such tweets (including

⁴Such tweets are removed from the general dataset, though they are kept in a separate file since the accounts could be unlocked again in the future.

⁵The original part of the hydrated Giorgi corpus also includes 13.0 mn tweets (44.38% of the total hydrated corpus) not mentioning any of these three. Most likely these were collected because they were part of a discussion: responding to tweets containing the search terms, for instance. They are not included in the present corpus. The hydrated update to the corpus, running through the end of 2021, similarly includes 9.7 mn tweets (48.4% of the total) not explicitly mentioning one of the three phrases/hashtags. These are also excluded.

Table 2: Overlap between vdV and Giorgi corpora: Black Lives Matter

Period	vdV only	%	Both	Giorgi only	%	Total BlackLM
2010-2012	11	100.0	0	0	0.0	11
2013	216	10.6	1,804	17	0.8	2,037
2014	225,253	33.6	411,462	33,431	5.0	670,146
2015	777,750	35.3	1,305,587	121,661	5.5	2,204,998
2016	2,332,569	65.6	1,029,809	195,104	5.5	3,557,482
2017	1,065,298	46.5	397,625	829,194	36.2	2,292,117
2018	476,952	46.9	314,109	226,272	22.2	1,017,333
2019	322,616	49.2	176,416	156,420	23.9	655,452
January 2020	27,084	58.2	9,607	9,827	21.1	46,518
February 2020	28,416	44.4	15,012	20,643	32.2	64,071
March 2020	23,688	59.8	9,344	6,601	16.7	39,633
April 2020	27,745	40.7	11,779	28,637	42.0	68,161
May 1-7, 2020	15,754	23.9	10,248	39,999	60.6	66,001
May 8-14, 2020	13,324	35.0	10,333	14,381	37.8	38,038
May 15-21, 2020	11,917	58.3	4,647	3,862	18.9	20,426
May 22-28, 2020	1,065,692	66.8	156,991	373,033	23.4	1,595,716
May 29-31, 2020	1,088,439	49.3	220,015	900,637	40.8	2,209,091
June 1-7, 2020	2,915,791	48.5	550,237	2,546,918	42.4	6,012,946
June 8-14, 2020	1,003,660	23.1	574,746	2,773,104	63.7	4,351,510
June 15-21, 2020	694,437	35.8	303,319	941,004	48.5	1,938,760
June 22-28, 2020	613,582	46.1	175,152	543,663	40.8	1,332,397
June 29-30, 2020	174,490	34.1	42,717	294,533	57.6	511,740
July 1-7, 2020	536,179	45.4	129,962	491,816	41.6	1,180,927
July 8-14, 2020	403,238	48.1	100,500	335,062	39.9	838,800
July 15-21, 2020	336,303	48.1	91,009	272,048	38.9	699,360
July 22-28, 2020	365,920	44.6	89,083	365,276	44.5	820,279
July 29-31, 2020	111,837	45.5	34,526	99,636	40.5	245,999
August 1-7, 2020	262,303	53.7	67,334	159,169	32.6	488,806
August 8-14, 2020	233,503	54.5	51,646	143,542	33.5	428,691
August 15-21, 2020	261,151	58.6	46,686	137,857	30.9	445,694
August 22-28, 2020	481,800	42.0	186,087	479,004	41.8	1,146,891
August 29-31, 2020	239,558	53.6	42,487	165,276	36.9	447,321
September 2020	1,312,229	63.0	223,108	548,460	26.3	2,083,797
October 2020	701,807	58.3	145,848	356,892	29.6	1,204,547
November 2020	604,442	66.0	102,695	208,899	22.8	916,036
December 2020	464,158	66.6	86,279	146,221	21.0	696,659
January 2021	1,399,962	80.7	120,828	214,485	12.4	1,735,275
February 2021	405,616	64.3	67,944	157,404	24.9	630,964
March 2021	397,420	66.7	84,733	113,524	19.1	595,677
April 2021	530,194	53.9	174,248	278,880	28.4	983,322
May 2021	431,043	66.9	74,640	138,251	21.5	643,934
June 2021	429,369	73.2	63,596	93,553	16.0	586,518
July 2021	423,759	74.1	52,258	95,626	16.7	571,643
August 2021	240,121	71.5	24,310	71,182	21.2	335,613
September 2021	236,427	73.6	15,831	69,063	21.5	321,321
October 2021	220,033	68.1	18,947	83,946	26.0	322,926
November 2021	479,330	81.2	28,413	82,662	14.0	590,405
December 2021	263,543	75.9	21,393	62,429	18.0	347,365

Tweet counts, with percentages of the period total uniquely supplied by each source.

those included in the Giorgi corpus but excluded here) are likely also about BLM. Moreover, the corpus can still not be said to be complete. First, as already noted, many tweets have been deleted and can no longer be captured. More significantly, the very fact that each collection method adds so many tweets compared to the other strongly suggests that there are additional tweets meeting the criteria that fail to be returned by either method. Relatedly, tweet collection, whether contemporaneous or retrospective appears to be non-deterministic: it will not return the same set of tweets each time. Fortunately, scholarship on the persistence of different Twitter corpora over time suggests that the textual content of the present collection is likely to be both representative of and similar to the full set of tweets that met or meet the search criteria (Zubiaga et al., 2018, p. 982).

3 Corpus overview

The tweet corpus is made available in chronological chunks, to facilitate retrieving only part(s) of the full corpus. Specifically, the data is split up by year through 2019, and by month starting in January 2020. The four months of May-August 2020, which saw the highest activity so far, are further subdivided into weeks (although those are aggregated to the monthly level below). Table 3 shows the breakdown of the tweet counts by period, for each corpus, noting each time period’s share in the total corpus volume.

To provide a better sense of trends and patterns in tweet counts across the three corpora, Figure 1 provides a graphical overview of tweet rates over time. Rates are logged, in order to make the scales comparable. The figure makes clear that tweet volumes broadly move together, especially at major moments, but that smaller movements are often independent of one another. While the correlation of daily tweet tallies between the BLM corpus and All Lives Matter is a robust 0.83, the correlation between the former and Blue Lives Matter is a more modest 0.44.

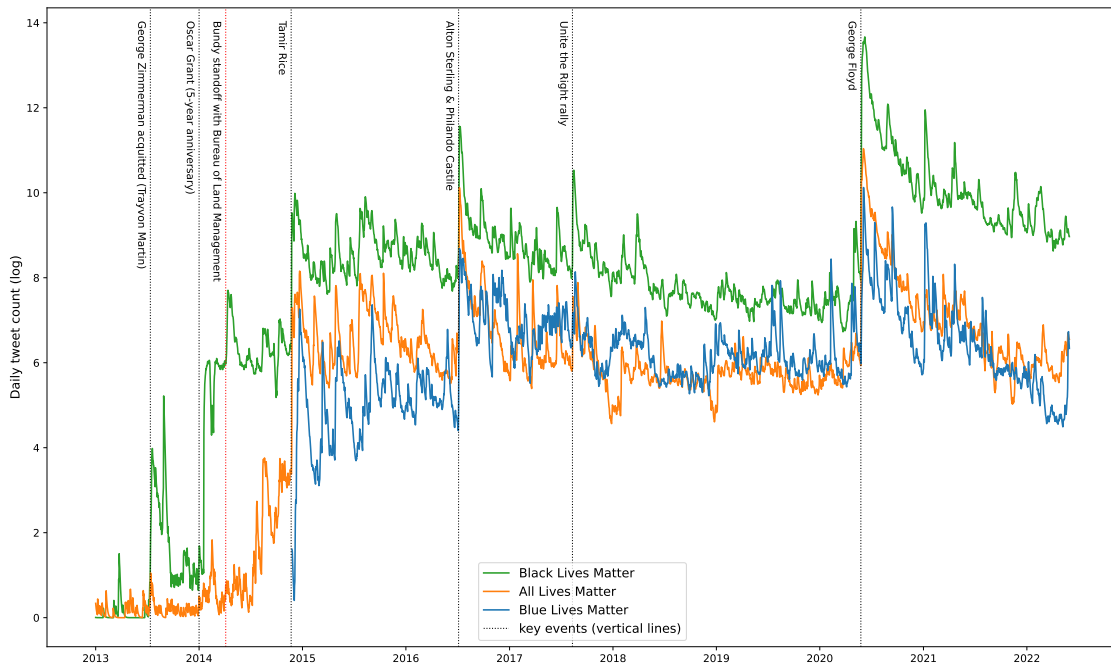


Figure 1: Daily tweet counts over time (log scale), 7-day exponential moving average, 2013-2022

It is worth noting that the daily tweet counts in this corpus are rather lower than some of the very high figures presented in the media. For example, a Pew Research Center report (2020) indicated that the `#blacklivesmatter` hash tag was used 8.8 million times on May 28, 2020 alone, after the murder of George Floyd. However, in the present dataset, the number of tweets explicitly using the phrase “blacklivesmatter” on that date is far less, at 1,209,669 tweets. If we limit the analysis to English-language tweets, the top day for the Black Lives Matter corpus is actually June 8, 2020, with 812,482 tweets total; for All Lives Matter it was the day of George Floyd’s murder, with 114,644 tweets, and for Blue Lives Matter the top day fell mid-way in between, on June 3rd, with 59,191 tweets.

Table 3: Total corpus size, by time period, through May 2022

Period	BlackLM	%	AllLM	%	BlueLM	%
2010-2012	11	0.0	33	0.0	0	0.0
2013	2,037	0.0	76	0.0	0	0.0
2014	670,146	1.3	73,572	1.9	14,150	0.6
2015	2,204,998	4.4	352,842	9.1	74,113	3.3
2016	3,557,482	7.2	567,190	14.6	385,948	17.1
2017	2,292,117	4.6	231,515	6.0	294,774	13.1
2018	1,017,333	2.0	106,071	2.7	172,726	7.7
2019	655,452	1.3	114,114	2.9	195,872	8.7
January 2020	46,518	0.1	8,582	0.2	11,288	0.5
February 2020	64,071	0.1	8,613	0.2	32,970	1.5
March 2020	39,633	0.1	10,254	0.3	9,022	0.4
April 2020	68,161	0.1	13,701	0.4	20,876	0.9
May 2020	3,929,272	7.9	313,530	8.1	42,472	1.9
June 2020	14,147,353	28.5	880,794	22.7	229,095	10.2
July 2020	3,785,365	7.6	315,812	8.1	109,604	4.9
August 2020	2,957,403	6.0	186,359	4.8	81,001	3.6
September 2020	2,083,797	4.2	131,716	3.4	131,830	5.8
October 2020	1,204,547	2.4	76,920	2.0	33,718	1.5
November 2020	916,036	1.8	50,246	1.3	23,319	1.0
December 2020	696,659	1.4	39,393	1.0	12,565	0.6
<i>2020 total</i>	<i>29,938,815</i>	<i>60.3</i>	<i>2,035,920</i>	<i>52.5</i>	<i>737,760</i>	<i>32.7</i>
January 2021	1,735,275	3.5	42,356	1.1	123,714	5.5
February 2021	630,964	1.3	30,866	0.8	38,879	1.7
March 2021	595,677	1.2	48,580	1.3	22,028	1.0
April 2021	983,322	2.0	42,718	1.1	42,031	1.9
May 2021	643,934	1.3	48,627	1.3	24,373	1.1
June 2021	586,518	1.2	27,212	0.7	19,260	0.9
July 2021	571,643	1.2	24,525	0.6	23,421	1.0
August 2021	335,613	0.7	16,326	0.4	18,249	0.8
September 2021	321,321	0.6	11,114	0.3	9,583	0.4
October 2021	322,926	0.7	15,495	0.4	10,321	0.5
November 2021	590,405	1.2	9,777	0.3	9,395	0.4
December 2021	347,365	0.7	13,315	0.3	8,978	0.4
<i>2021 total</i>	<i>7,664,963</i>	<i>15.4</i>	<i>330,911</i>	<i>8.5</i>	<i>350,232</i>	<i>15.5</i>
January 2022	351,745	0.7	10,501	0.3	9,079	0.4
February 2022	552,653	1.1	15,830	0.4	5,886	0.3
March 2022	253,950	0.5	12,200	0.3	3,615	0.2
April 2022	214,682	0.4	8,487	0.2	3,222	0.1
May 2022	265,270	0.5	18,466	0.5	8,508	0.4
June 2022	257,662	0.5	19,972	0.5	6,749	0.3
July 2022	171,391	0.3	12,073	0.3	5,411	0.2
<i>2022 total (Jan-Jul)</i>	<i>2,067,353</i>	<i>4.1</i>	<i>97,529</i>	<i>2.5</i>	<i>42,470</i>	<i>1.9</i>
<i>Cumulative total</i>	<i>50,070,707</i>		<i>3,909,773</i>		<i>2,268,045</i>	

Tweet counts, with percentages of the corpus (column) total in parentheses.

3.1 Distinct users

The increasing reach of the Black Lives Matter movement is also illustrated by the large and growing number of individual users who have used the phrase, hash tag, or acronym on Twitter. Table 4 shows the number of unique users, by year, for each of the corpora, starting in 2014. In addition, it shows the relationship of that number to the total number of tweets in the corpus for that year, measured by dividing the latter by the former.

Table 4: Unique users, by year and overall

Period	BlackLM	Avg. tweets/user	AllLM	Avg. t/u	BlueLM	Avg. t/u
2014	256,777	2.6	18,447	4.0	8,052	1.8
2015	564,384	3.9	92,713	3.8	22,794	3.3
2016	1,087,164	3.3	109,325	5.2	64,647	6.0
2017	504,018	4.5	84,639	2.7	33,570	8.8
2018	312,796	3.3	45,549	2.3	39,130	4.4
2019	224,210	2.9	59,871	1.9	37,526	5.2
2020	6,006,371	5.0	914,810	2.2	275,061	2.7
2021	2,215,283	3.5	169,908	1.9	175,720	2.0
2022	856,918	2.4	67,715	1.4	31,860	1.3
<i>Overall total</i>	<i>8,776,752</i>	<i>5.7</i>	<i>1,376,858</i>	<i>2.8</i>	<i>584,308</i>	<i>3.9</i>

More than 8 million unique users are represented in the BLM corpus. Many of them contribute to the corpus over multiple years, which explains why the overall number of tweets per user is larger than the figures for individual years. For the other two corpora, the number of unique users is rather smaller. The average number of tweets per user, however, was highest overall in the Blue Lives Matter corpus, from 2016-2019, with a peak in 2017 of nearly 9 tweets/user, suggesting a more active counter-discourse to BLM than is the case today (Gallagher et al., 2018).

3.2 Tweet language

While the Black Lives Matter movement arose first in the United States, and most of its activities in that country take place in English, the global nature of Twitter means that neither the geographic nor the linguistic scope of the corpus is limited. In particular, there are many Spanish-language tweets about BLM that originate in the United States, while many English-language tweets originate in other English-speaking countries and beyond.

Twitter attempts to automatically identify the language of each tweet. The usual approach to automatic language detection is to assess the frequencies of different ngrams in the text, by language (Cavnar and Trenkle, 1994). For tweets, this is a non-trivial challenge, since tweets are short, and often contain comparatively few “normal” words (as opposed to URLs, hashtags, names, etc.). As a result, Twitter’s assigned language codes are not perfect. However, a sampling of tweets in the corpus suggests the algorithm does a very good job. In addition, the python module `langdetect` (<https://pypi.org/project/langdetect/>) produces essentially identical results.

English is the primary language in multiple countries with a significant online presence: the United States, the United Kingdom, Canada, Australia, Ireland, New Zealand, etc. In addition, English is a lingua franca of sorts in online media. Even if the Black Lives Matter movement were not of U.S. origin, therefore, one would expect English-language tweets to be dominant, as indeed they are. However, it is clear from the data that the movement has reached well beyond the English language sphere. Table 5 shows the top 10 languages after English, by corpus, with the total number of tweets in each of these languages. While no individual language accounts for a very large share of the total BLM discussion on Twitter, taken together the top 10 account for about 1 of every 12 tweets in the total corpus. Moreover, the top individual languages each account for several hundred thousand tweets overall — a considerable volume in and of itself.

It is also important to note that this table shows only foreign-language tweets that use the English-language phrase, hashtag, or acronym: “Black Lives Matter”, `blacklivesmatter`, or `BLM`. The corpus does not include tweets that only include local-language translations of these, such as (for instance) the Spanish “las vidas negras importan.”

4 Conclusion

Black Lives Matter’s fortunes as a social movement are inextricably linked to its emergence and development online. As Freelon et al. put it, it can be seen as an “online struggle for offline justice” (Freelon et al., 2016). For this reason alone, studying its presence on social media such as Twitter is of considerable importance (Jackson et al., 2020). In

Table 5: Tweet counts for the top 11 languages, by corpus

Black LM			All LM			Blue LM		
Language	Count	%	Language	Count	%	Language	Count	%
English	41,364,190		English	3,452,737		English	2,011,349	
Spanish	709,919	1.7	Spanish	29,768	0.9	Spanish	15,013	0.7
French	659,143	1.6	Dutch	20,969	0.6	French	9,085	0.5
Japanese	543,146	1.3	Japanese	19,309	0.6	Portuguese	3,769	0.2
Portuguese	499,561	1.2	Portuguese	19,156	0.6	German	2,554	0.1
Dutch	276,798	0.7	French	17,376	0.5	Japanese	2,432	0.1
German	264,758	0.6	German	16,805	0.5	Italian	2,150	0.1
Thai	182,028	0.4	Italian	6,803	0.2	Dutch	2,123	0.1
Italian	174,480	0.4	Hindi	6,141	0.2	Indonesian	1,997	0.1
Indonesian	142,383	0.3	Indonesian	4,442	0.1	Tagalog	969	0.0
Turkish	108,717	0.3	Tagalog	3,590	0.1	Haitian Creole	915	0.0

Percentages are relative to the English count.

addition, BLM serves as an instance of a broader category: ideas or phrases that have taken on a life of their own on social media and are at the heart of ongoing online discussions that are global in reach. Finally, online discussions surrounding Black Lives Matter allow us to learn more about the nature and shape of such discussions in general: What types of actors drive the discussion? How can we characterize the tweet-retweet-response network? Does it make sense to talk about a single discussion, or are there multiple, linked, contemporaneous discussions taking place (Shugars et al., 2021; Terechshenko et al., 2020)?

The present dataset is intended to make it easier for scholars and researchers to investigate these and other questions. It makes available the most complete collection available of Tweets about Black Lives Matter, from its origins to the present, and across many languages. The criterion of explicit references to BLM greatly reduces the likelihood that analyses based on the corpus will be biased by the inclusion of tweets that are not actually about Black Lives Matter (because, for example, they form part of a discussion that started out being about BLM but veered in a different direction). In addition, the focus on explicit references makes it possible to learn more about how people use the phrase, hashtags, and acronym associated with BLM and how such use may have changed over time (Ince et al., 2017).

The discourse on Twitter is but a small part of larger ongoing public discourses, and Twitter users are not representative of the overall population in any particular country (Wojcik and Hughes, 2019). At the same time, people use Twitter in many different ways: to pursue justice, for self-expression, to frame broader societal debates, and to rally or rile audiences, to name but a few. Much remains to be learned about how Black Lives Matter, along with similar movements or ideas, has found expression and has developed over time on social media (Jackson et al., 2020). The BLM Twitter corpus can assist in that effort.

5 Data availability

The dataset is available on Zenodo (<https://doi.org/10.5281/zenodo.6628275>). A datasheet for the dataset, along the lines suggested by (Gebru et al., 2021) is available on Github (<https://github.com/amaurits/BLMtwitter>), along with monthly updates for the current calendar year (Black Lives Matter corpus only).

References

- Bastos, M. (2021). This Account Doesn’t Exist: Tweet Decay and the Politics of Deletion in the Brexit Debate. *American Behavioral Scientist*, 65(5):757–773.
- Cavnar, W. B. and Trenkle, J. M. (1994). N-Gram-Based Text Categorization. In *Proceedings of SDAIR-94*, pages 161–175, Las Vegas, NV.
- Center, P. R., Anderson, M., Barthel, M., Perrin, A., and Vogels, E. a. (2020). #BlackLivesMatter surges on Twitter after George Floyd’s death. Technical report, Pew Research Center, Washington, DC.
- Dunivin, Z. O., Yan, H. Y., Ince, J., and Rojas, F. (2022). Black Lives Matter protests shift public discourse. *Proceedings of the National Academy of Sciences*, 119(10):e2117320119.

- Freelon, D., McIlwain, C. D., and Clark, M. D. (2016). Beyond the hashtags: #Ferguson, #Blacklivesmatter, and the online struggle for offline justice. Technical report, Center for Media & Social Impact, American University, Washington, DC.
- Gallagher, R. J., Reagan, A. J., Danforth, C. M., and Dodds, P. S. (2018). Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter. *PLOS One*, 13(4):e0195644.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.
- Giorgi, S., Guntuku, S. C., Himelein-Wachowiak, M., Kwarteng, A., Hwang, S., Rahman, M., and Curtis, B. (2022). Twitter Corpus of the #BlackLivesMatter Movement And Counter Protests: 2013 to 2021. page 8, Atlanta, GA. AAAI.
- Giorgi, S., Guntuku, S. C., Rahman, M., Himelein-Wachowiak, M., Kwarteng, A., and Curtis, B. (2020). Twitter Corpus of the #BlackLivesMatter Movement And Counter Protests: 2013 to 2020. arXiv: 2009.00596.
- 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.
- Jackson, S. J., Bailey, M., and Welles, B. F. (2020). *#HashtagActivism: Networks of Race and Gender Justice*. MIT Press, Cambridge, MA, USA.
- 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):205630511880791.
- Shugars, S., Gitomer, A., McCabe, S., Gallagher, R. J., Joseph, K., Grinberg, N., Doroshenko, L., Foucault Welles, B., and Lazer, D. (2021). Pandemics, Protests, and Publics: Demographic Activity and Engagement on Twitter in 2020. *Journal of Quantitative Description: Digital Media*, 1.
- Terechshenko, Z., Kates, S., Linder, F., Tucker, J. A., Vakilifathi, M., and Nagler, J. (2020). Influential Users in the Common Core and Black Lives Matter Social Media Conversation. Technical report, CSMaP, NYU, New York, NY.
- Wojcik, S. and Hughes, A. (2019). U.S. adult Twitter users are younger and more likely to be Democrats than the general public. Most users rarely tweet, but the most prolific 10% create 80% of tweets from adult U.S. users. Technical report, Pew Research Center, Washington, DC.
- Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and Resolution of Rumours in Social Media: A Survey. *ACM Computing Surveys*, 51(2):1–36. arXiv: 1704.00656.

6 Appendix

Table 6 and Table 7 show analogous data to that in Table 2 about overlap between the two collection methods for the All Lives Matter and Blue Lives Matter corpora, respectively.

Table 6: Overlap between vdV and Giorgi corpora: All Lives Matter

Period	vdV only	%	Both	Giorgi only	%	Total AllLM
2010-2012	33	100.0	0	0	0.0	33
2013	64	84.2	10	2	2.6	76
2014	2,871	3.9	20,724	49,977	67.9	73,572
2015	17,380	4.9	121,106	214,356	60.8	352,842
2016	106,396	18.8	53,836	406,958	71.7	567,190
2017	73,489	31.7	52,860	105,166	45.4	231,515
2018	23,892	22.5	38,700	43,479	41.0	106,071
2019	60,080	52.6	33,317	20,717	18.2	114,114
January 2020	4,915	57.3	2,449	1,218	14.2	8,582
February 2020	4,475	52.0	2,639	1,499	17.4	8,613
March 2020	5,380	52.5	3,300	1,574	15.4	10,254
April 2020	7,635	55.7	3,903	2,163	15.8	13,701
May 1-7, 2020	3,053	61.7	1,206	687	13.9	4,946
May 8-14, 2020	2,367	60.8	1,032	493	12.7	3,892
May 15-21, 2020	1,627	54.1	747	633	21.1	3,007
May 22-28, 2020	35,282	23.3	10,383	105,547	69.8	151,212
May 29-31, 2020	96,992	64.5	6,993	46,488	30.9	150,473
June 1-7, 2020	318,115	77.0	40,018	54,990	13.3	413,123
June 8-14, 2020	97,091	45.6	44,842	71,158	33.4	213,091
June 15-21, 2020	74,508	56.0	26,418	32,188	24.2	133,114
June 22-28, 2020	63,909	64.5	17,553	17,590	17.8	99,052
June 29-30, 2020	13,639	60.9	4,337	4,438	19.8	22,414
July 1-7, 2020	59,310	61.5	18,641	16,882	17.5	96,424
July 8-14, 2020	53,808	58.1	17,923	20,865	22.5	92,596
July 15-21, 2020	32,845	54.4	12,764	14,774	24.5	60,383
July 22-28, 2020	27,946	56.9	10,477	10,701	21.8	49,124
July 29-31, 2020	9,936	57.5	3,136	4,213	24.4	17,285
August 1-7, 2020	19,496	49.1	7,053	13,155	33.1	39,704
August 8-14, 2020	22,509	58.4	8,569	7,442	19.3	38,520
August 15-21, 2020	23,386	63.1	8,119	5,559	15.0	37,064
August 22-28, 2020	32,611	61.7	12,383	7,877	14.9	52,871
August 29-31, 2020	12,093	66.4	3,582	2,525	13.9	18,200
September 2020	79,367	60.3	26,125	26,224	19.9	131,716
October 2020	42,749	55.6	17,053	17,118	22.3	76,920
November 2020	34,969	69.6	9,488	5,789	11.5	50,246
December 2020	28,351	72.0	6,878	4,164	10.6	39,393
January 2021	29,738	70.2	7,079	5,539	13.1	42,356
February 2021	19,914	64.5	5,961	4,991	16.2	30,866
March 2021	27,529	56.7	7,981	13,070	26.9	48,580
April 2021	25,943	60.7	8,193	8,582	20.1	42,718
May 2021	29,019	59.7	10,551	9,057	18.6	48,627
June 2021	17,478	64.2	4,970	4,764	17.5	27,212
July 2021	14,381	58.6	3,303	6,841	27.9	24,525
August 2021	7,060	43.2	87	9,179	56.2	16,326
September 2021	5,618	50.5	70	5,426	48.8	11,114
October 2021	9,287	59.9	138	6,070	39.2	15,495
November 2021	1,964	20.1	26	7,787	79.6	9,777
December 2021	7,447	55.9	2,120	3,748	28.1	13,315

Tweet counts, with percentages of the period total uniquely supplied by each source.

Table 7: Overlap between vdV and Giorgi corpora: Blue Lives Matter

Period	vdV only	%	Both	Giorgi only	%	Total BlueLM
2014	2,281	16.1	11,134	735	5.2	14,150
2015	11,764	15.9	33,693	28,656	38.7	74,113
2016	55,785	14.5	95,064	235,099	60.9	385,948
2017	40,772	13.8	22,458	231,544	78.5	294,774
2018	41,940	24.3	33,954	96,832	56.1	172,726
2019	34,584	17.7	31,842	129,446	66.1	195,872
January 2020	2,236	19.8	4,188	4,864	43.1	11,288
February 2020	2,743	8.3	7,074	23,153	70.2	32,970
March 2020	1,935	21.4	4,164	2,923	32.4	9,022
April 2020	3,586	17.2	4,969	12,321	59.0	20,876
May 1-7, 2020	7,209	52.0	3,182	3,465	25.0	13,856
May 8-14, 2020	1,263	38.5	1,139	880	26.8	3,282
May 15-21, 2020	1,849	44.8	1,153	1,127	27.3	4,129
May 22-28, 2020	5,982	75.7	1,003	916	11.6	7,901
May 29-31, 2020	10,686	80.3	1,019	1,599	12.0	13,304
June 1-7, 2020	68,377	47.6	19,453	55,787	38.8	143,617
June 8-14, 2020	8,305	28.0	8,124	13,243	44.6	29,672
June 15-21, 2020	9,023	28.1	9,900	13,214	41.1	32,137
June 22-28, 2020	6,662	39.1	4,504	5,877	34.5	17,043
June 29-30, 2020	1,802	27.2	1,132	3,692	55.7	6,626
July 1-7, 2020	5,791	42.1	3,592	4,101	29.8	13,751
July 8-14, 2020	9,946	17.3	14,229	33,451	58.0	57,626
July 15-21, 2020	8,366	40.6	5,193	7,047	34.2	20,606
July 22-28, 2020	6,263	47.1	3,382	3,639	27.4	13,284
July 29-31, 2020	2,215	51.1	1,222	900	20.8	4,337
August 1-7, 2020	5,318	48.7	2,818	2,782	25.5	10,918
August 8-14, 2020	4,632	49.6	2,451	2,259	24.2	9,342
August 15-21, 2020	13,525	51.4	5,836	6,951	26.4	26,312
August 22-28, 2020	13,637	55.5	5,716	5,202	21.2	24,555
August 29-31, 2020	4,609	46.7	2,462	2,803	28.4	9,874
September 2020	32,352	24.5	30,795	68,683	52.1	131,830
October 2020	16,843	50.0	7,971	8,904	26.4	33,718
November 2020	14,890	63.9	4,011	4,418	18.9	23,319
December 2020	8,094	64.4	2,836	1,635	13.0	12,565
January 2021	89,132	72.0	17,785	16,797	13.6	123,714
February 2021	23,381	60.1	6,455	9,043	23.3	38,879
March 2021	14,688	66.7	3,774	3,566	16.2	22,028
April 2021	22,284	53.0	6,744	13,003	30.9	42,031
May 2021	14,123	57.9	5,214	5,036	20.7	24,373
June 2021	11,821	61.4	3,362	4,077	21.2	19,260
July 2021	14,332	61.2	4,147	4,942	21.1	23,421
August 2021	433	2.4	1	17,815	97.6	18,249
September 2021	5,311	55.4	26	4,246	44.3	9,583
October 2021	4,857	47.1	63	5,401	52.3	10,321
November 2021	4,985	53.1	38	4,372	46.5	9,395
December 2021	3,886	43.3	655	4,437	49.4	8,978

Tweet counts, with percentages of the period total uniquely supplied by each source.