Final paper
For Big Data, Small Data
-

How did the discourse of #blacklivesmatter on twitter evolved after the murder of Georges Floyd?

-

Danique de Rijk
daniquederijk@gmail.com
Student number: 2668000

&

Auriane Vez
Auriane.vez@gmail.com
Student number: 2711415

Vrije Universiteit
Research Master Societal Resilience, Period 1 & 2, 2020

Number of words: 13 pages

Inhoudsopgave

1. Introduction 2

1.1 Adaptation/appropriation 2

1.2 Influence of the 2020 protests 3

2. Methodology 4

2.1 Mixed methods 4

2.2 Subject 4

2.3 Quantitative methods - topic modeling 5

2.4 Qualitative methodology – content analysis 6

2.5 Co-production of knowledge - interviews with people of color 8

3. Results 9

3.1 Topic Modeling 9

3.1.1 Data set - Before the murder of George Floyd 9

3.1.2 Data set - After the murder of George Floyd 9

3.2 Content Analysis 10

3.3 Combining quantitative and qualitative results 11

4. Discussion 13

5. Bibliography 15

6. Appendix 17

6.1 Appendix 1 : R Code 17

6.2 Appendix 2: Code Schemes 28

6.3 Appendix 3 : Interview guide 30

1. Introduction

In 2013, Alicia Garza, Patrisse Cullors and Opal Tometi created #BlackLivesMatter in response to the trial that was hold posthumously against Trayvon Martin, regarding his own murder and during which the killer, George Zimmerman, was acquitted (Garza, 2016). This hashtag was created "as a call to action for Black people" (Garza, 2016, 23). Garza explains that, very quickly, the hashtag expanded outside the internet sphere and materialized into the real world through protest and conferences (Garza, 2016). The hashtag was subsequently used after several murders on Black People, like Michal Brown, who was killed by the police Officer Darren Wilson, in 2015. The Pew Research Center, a nonpartisan fact tank, published an analysis of published tweets containing the #BlackLivesMatter (see figure 1) (Anderson et al., 2020). If peaks in the use of #Backlivesmatter occurs after racist and unjustified killing, Alicia Garza emphasizes that the aims of the movement are broader. Indeed, she writes: "Black Lives Matter affirms the lives of Black queer and trans folks, disabled folks, Black undocumented folks, folks with records, women, and all Black lives along the gender spectrum. […]. When we say Black Lives Matter, we are talking about the ways in which Black people are deprived of our basic human rights and dignity" (Garza, 2016, 25). Overall, the movement is fighting and protesting against structural racism and inequalities that result in state violence (Garza, 2016).

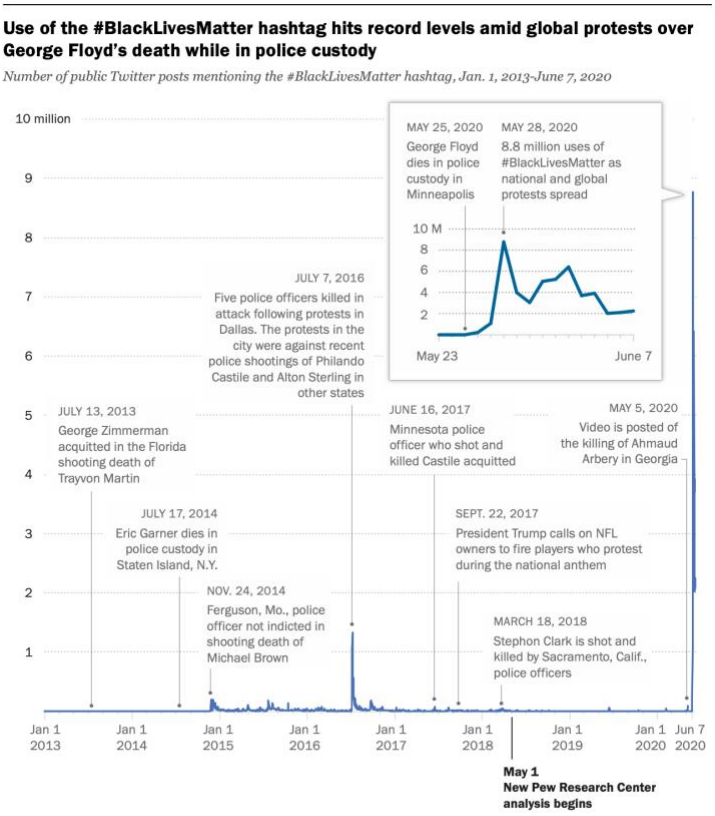


Figure 1 Pew Research Center Analysis (Anderson et al., 2020)

1.1 Adaptation/appropriation

Black Lives Matter is a decentralized movement, this makes it very likely to be appropriated by other actors for their own benefits (Ince et al., 2017). Moreover, as this movement is evolving on social media, this enables individuals to interact with the movement but also make their own claim about the movement. Indeed, Ince et al. (2017) write: "the public's interaction with a movement via social media is of wider importance because it is one process, among many, that shapes a movement's development

and final outcomes" (p.1817). This follows among the lines of Alicia Garza testimony about the appropriation of the movement by mainstream media, corporations and various actors. She writes "we began to come across various adaptations of our work – all lives matter, migrant lives matter, women's lives matter and on and on" (Graza, 2016, 24). Ince et al. (2017) argue that it is the "open nature of social media" (p.1818) that enables counter movements to take form (for example #AllLivesMatter and #BlueLivesMatter in the context of this study). If countering a movement is clear and obvious reframing of the concept, Garza explains that some actors, which were considered as ally, reappropriated the #BlackLivesMatter in a more subtle manner. Indeed, she writes that errors have been made when arguing for unity (all lives) and therefore letting on the side "differences in context, experience, and, oppressions" (Graza, 2016, 27). In other words, ignoring systemic oppression that results in state violence, which is the core element upon which the Black Lives Matter movement protests.

1.2 Influence of the 2020 protests

The Pew Research Center report (Anderson et al., 2020) mentions that as a response to the murder of George Floyd, from the 26th of May to the 7th of June, an average of 3.7 million tweets were written containing the #BlackLivesMatter. The report mentions that this has been the highest number since the center started studying this hashtag. The Pew Research Center also found that after the murder of George Floyd, the number of members of the congress tweeting about #BlackLivesMatter has doubled since 2015 (Shah et al., 2020). Those new members were mostly democrats, which illustrates how new people engaged into the debate. Therefore, this paper aims to study the influence of this new event, because compared to the high number of tweets, it is likely that new people engaged into the debate who could have an influence on the final outcome of the movement. To do so, this paper focuses on the #BlackLivesMatter and not on the obvious counter movements. Indeed, it aims to discover if the subtle appropriation made by some actors has had an influence on the overall outcome and framing of the movement. Therefore, our research question is the following: how did the discourse of #BlackLivesMatter on twitter evolved after the murder of Georges Floyd? In order to answer this question, we are using a quantitative and qualitative methodology. First of all, a Topic modelling (LDA) is conducted on two sets of tweets made from random sampling with tweets before and after the murder of George Floyd. This is supplemented with a discourse analysis on tweets. The classification and coding of the discourse analyses will be compared and based on Ince et al. (2017) findings. They analyzed tweets from 2014 with #BlackLivesMatter. Since according to Ince et al. (2017) also argue that a movements outcome depend on the public interaction with it will be interesting to see if the results of the analysis of the tweets after the murder of George Floyd vary with the general trend found by Ince et al. (2017). The first section of the paper is focused on the methodology followed by the results and finally the discussion.

2. Methodology

2.1 Mixed methods

To answer the proposed research question, a combination of mixed methods will be used. There are several reasons to conduct a mixed method research approach. One of those is the fact that a mixed methods approach might offset the limitations of just a quantitative or qualitative method by including the other one that has its own strengths (Bryman, 2012). For the quantitative analysis in this study, topic modeling will be used. Why this is useful for this study and how the analysis is done will be discussed later on in this chapter. However, it is known that the interpretation of topics is a weakness of this approach. Even though the topics are automatically assigned without the interference of any human being, the topics in the end still need to be interpreted by the researcher (Jacobi, Van Atteveldt & Welbers, 2015). Without any in-depth knowledge of the research subject it might be hard to give a good interpretation of these topics. However, by using a qualitative method complementary to this quantitative method, the limitation of interpreting the topic might be offset by the knowledge gained from this qualitative analysis. As will become clear in the results section, the results of the qualitative analysis made it possible to correctly interpret the results of the quantitative analysis. A mixed method approach in this study is thus not only used to offset the limitations of one of the two approaches but a mixed method approach also helped to understand and explain the results of the quantitative analysis. Besides this, the combination of quantitative and qualitative methods might give a more complete answer to the proposed research question. Where a quantitative approach could give a more objective analysis of the evolution of the movement to make clear that a change actually did happen. Adding a qualitative perspective on this adds other insightful information on what this change than actually means to finally give a more complete answer on the question: How did the discourse of #BlackLivesMatter on twitter evolved after the murder of Georges Floyd? In this mixed method-based study, quantitative and qualitative methods thus work for several reasons complementary to each other.

2.2 Subject

As a subject of our study we used tweets on #BlackLivesMatter, posted between January 1st 2019 up until June 30th 2020. First a dataset was received which consisted of all the tweets containing #BlackLivesMatter, #AllLivesMatter or #BlueLivesMatter from the beginning of 2013, when the movement began, up until June 30th 2020 (Giorgi, Guntuku, Rahman, Himelein-Wachowiak, Kwarteng & Curtis, 2020). A Twitter API was used to collect these data. The dataset contained 48,801,153 tweets in total with 36,892,99 tweets containing #BlackLivesMatter (Giorgi, 2020). This dataset, however, only contained information on message-id's and the hashtags used. Additional information, like dates

and messages, was needed to answer our research question. A twitter API, Hydrator, was used to collect this information using the corresponding message-ids. Collecting all this information on more than 48 million tweets would have been impossible. Before using the twitter API, we thus cleaned the dataset and took a random sample. We removed tweets containing #AllLivesMatter and #BlueLivesMatter. This means that tweets containing one of those hashtags together with #BlackLivesMatter were removed as well. This choice was made because tweets using both hashtags could be just representing the counter movements instead of the Black Lives Matter movement. This left us with 35,664,304 tweets. After this we took a sample of 1 million tweets which we then used to collect more information about them with the twitter API. A bit more than half of the tweets, 502,103, were returned containing additional information, including messages and dates.

2.3 Quantitative methods - topic modeling

For the quantitative part of this study a topic model has been performed using RStudio (see Appendix 1 for the R code). Topic models identify patterns of words that occur using the distribution of words in a collection of documents (Jacobi, Van Atteveldt & Welbers, 2015). For this a Latent Dirichlet Allocation (LDA) approach for probabilistic topic modeling has been used. LDA "automatically creates topics based on patterns of (co-)occurrence of words in the documents that are analysed" (Jacobi, Van Atteveldt & Welbers, 2015). In performing a topic model, topics are thus automatically assigned to documents using the distribution of words. Thus, in this case, the use of a topic model could make clear which patterns of words occurred in the tweets on #BlackLivesMatter. By performing a topic model on tweets before and after the murder of George Floyd, differences in topics, if existent, would emerge through the results of the LDA.

To perform two separate topic models on both tweets before and after the murder on George Floyd, two separate datasets had to be created. When inspecting the dataset, it was found that there were quite a few messages in the dataset that did not contain #BlackLivesMatter. Therefore, first all the tweets without this hashtag were removed, leaving us with 30,241tweets. In order to compare two time slots, two new datasets were created based on dates. At first, we wanted to use tweets posed in the month before the murder of George Floyd because the tweets in the dataset also only ran up to one month after his death. This would keep the time slots even. However, since we had to delete quite a few tweets that did not contain the hashtag, and because of the fact that the use of the hashtag exploded after the murder of George Floyd, the number of tweets before the murder were very low. Therefore, we decided to take a bigger time slot, using the tweets from the beginning of 2019 up until may 25th 2020, the day George Floyd was murdered. This time slot is chosen because since the movement is evolving over time, we did not want to go too far back with comparing tweets but we also had to make sure that

enough tweets were in the sample to perform a topic model. In the end the dataset we used on tweets before the murder contained 1,116 tweets. The dataset on tweets after the murder, between may 25th 2020 and 30th of June 2020, contained 14,042 tweets.

After this, the topic models could be created. For both datasets a DTM (Document-Term-Matrix) was created because in this, texts are treated as individual words making topic modeling possible. In this DTM several aspects were removed from the tweets like all punctuations, stop words, words shorter than three characters, hashtags, replies (@...), URLs and special symbols. Emojis often consist of a combination of special symbols and characters. Because it is not possible to interpret these "words", they were removed from the tweets. The same applies to the other removed aspects. It would return meaningless words which makes it impossible to interpret the topics.

Several LDA topic models have been made with different numbers of topics. According to Jacobi, Van Atteveldt & Welbers (2015) perplexity and interpretability are both important in choosing the number of topics when using large amounts of text. A low perplexity shows you that the probability distribution is good at predicting the sample. The interpretability on the other hand takes a look at how good the topics contribute to answering the questions and are thus interpretable (Jacobi, Van Atteveldt & Welbers, 2015). This theory of Jacobi, Van Atteveldt & Welbers (2015) however, is used for topic models on large amounts of texts. Since for example the dataset used for topic modeling on tweets before the murder of George Floyd only contained 1,116 tweets, the statistical measure for perplexity was not applicable in this study. To still get to the right number of topics, the number of tweets that were characterized by each topic has been used. Besides this, for both datasets it made sense, regarding the interpretability of the topics, to choose five topics. This number however had only been set after the content analysis because these results helped us understand and interpret the topics. At first the topics that emerged were not interpretable. A better understanding of what the tweets were about enabled a better understanding of the topics that emerged from the topic modeling. The content analysis thus worked complementary to the topic models because it helped to interpret the topics. Besides the number of topics, an alpha of 0.1 has been chosen. A low alpha has been chosen because this "leads to a higher concentration of topic distributions … meaning that documents score high on a few topics rather than low on many" (Jacobi, Van Atteveldt & Welbers, 2015, p. 6).

2.4 Qualitative methodology – content analysis

Besides topic modeling, this study makes use of content analysis as well. According to Bryman (2012) a content analysis is an "approach to the analysis of documents and texts that seeks to quantify content in terms of predetermined categories and in a systemic and replicable manner" (p. 290). There are different reasons why a content analysis is useful but one of those is when a research is interested in

what gets reported (Bryman, 2012). This research is interested in what people tweeted about before and after the murder of George Floyd in post using #BlackLivesMatter in order to see how the movement is shaped. Therefore, a content analysis is useful in giving an answer to this. Using a topic model already gave the topics which were mostly used in the tweets. However, this content analysis can be used to analyze the full content of the tweet in order to better understand the meaning of the emerged topics.

The same two datasets that were used for the topic modeling were used for this content analysis. In contrast to the tweets used for the topic modeling, the two original datasets were used without any transformations in the texts. This choice was made because the aim of a content analysis is to uncover the apparent content of an item, in this case tweets. Changing something in this tweet would thus also change the content that will be analyzed giving different results. Besides this similar, returning tweets like retweets could appear more than once in the dataset. Although the content of those tweets would be the same, it was still important to keep those tweets in the dataset. As explained before, Ince et al. (2017) argue that the publics interaction on social media with a movement changes the outcome of a movement. Tweets that occur often thus really affect and determine the discourse of a movement. From both datasets a random sample had been taken of 300 tweets. In the end 300 tweets posted after the murder on George Floyd had been analyzed to reach saturation and only 250 tweets were needed to determine the content of tweets posted before the murder.

Since the definition of a content analysis contains the use of structural and replicable manners (Bryman, 2012), analyzing the tweets has been done through three phases. According to Van Maanen (1979) it is important to structure qualitative data in two phases: first and second order concepts. First order concepts are seen as "facts", something that is really present in the data. So, first order concepts are things that are noticeable in the tweets. For these codes it is important to stay as close to the original data as possible. This phase gave us many different codes. In order to give more structure to the data we moved to second order concepts. These are more theoretical concepts (Van Maanen, 1979) which could explain patterns in first order concepts by comparing them. This was done separately for both samples so different second order concepts were formed (see appendix 2).

However, it was still important to be able to analyze the differences in discourses between the two time slots. Therefore, we came up with a set of 'aggregate dimensions'. Gioia, Corley & Hamilton (2012) add this as a third phase to the two concepts of Van Maanen (1979). Aggregate dimensions seek to emerge even further into abstract theoretical concepts. Such abstract theoretical concepts were useful to in the end compare the two different content analyses. Again, patterns were sought but now in the second order themes to make sure that the final concepts to analyze tweets were the same for both datasets (see appendix 2). In forming these aggregate themes, we made use of the study of Ince et al. (2017). In looking at social media response to Black Lives Matter through hashtag use on twitter, they

found five elements in the framing of Black Lives Matter: Ferguson, movement tactic, police violence, movement solidarity and counter protests. Two of those elements did not come back in our dataset. First, the sample of tweets on Black Lives matter used in the study of Ince et al. (2017) consisted of tweets from the beginning of 2014 until November 2014. This includes the event of the killing of Michael Brown in Ferguson and the social unrest that started after this event in the streets of Ferguson. Therefore, it is explicable why Ferguson as an element came up in their study but not in ours. Beside this, hashtags of counter movements were removed from the dataset to capture the discourse of the Black Lives Matter movement instead of capturing the discourse of the broader discussion that is started. Therefore, countermovement did not come up as such a big element in this study compared to the study of Ince et al. (2017). In contrast, the other three, movement tactic, police violence and movement solidarity, were recognized as themes in the tweets in both datasets. Besides this some new elements were found which will be further described in the result section.

2.5 Co-production of knowledge - interviews with people of color

To verify the results gained from the topic modeling and content analysis this study aimed to conduct interviews with people of color. As white researchers doing research about a movement to empower the black community, it has to be acknowledged that we as researchers might miss certain inside information or have biases. Besides that, according to anti-racism research methodology it is important to "seek to empower the subjects of research rather than simply seeking information from them in a disinterested or domineering way" (Okolie, 2015, p. 249). In order to make sure this study not just seeks information from people of color or about #BlackLivesMatter, it is important to make sure the subject of research is empowered as well. So, to empower the subject of research and to restrict the influence of our own biases it is important to verify the results with people of color.

To succeed in this, this study aimed to conduct expert interviews with people of color who are activist in the Black Lives Matter movement. According to Meuser & Nagel (2009) expert interviews are useful to find more information on socio-cultural conditions, consisting of knowledge generated outside the scientific world. Activist would thus in this study be considered experts of the socio-cultural context of the Black Lives Matter movement. Because these interviews were going to be used to verify the results of the topic modeling and content analysis, these interviewees could be considered co-producers of knowledge. This means that as co-producers they bring experienced knowledge to the field of science, combining two types of knowledge as co-production of knowledge (Markkanen & Burgess, 2015). By conducting expert interviews in which people of color are made co-producers of knowledge, this study seeks not only to get information from them as a subject of research but also to empower them. However, due to the short time span of this research we unfortunately were unable to conduct

these interviews. The results that will be presented below will thus have to be considered provisional research results. The limitations that are added to our study because of this will be discussed in the discussion. An interview guide had already been made and has been added to the appendix (see appendix 3)

3. Results

3.1 Topic Modeling

3.1.1 Data set - Before the murder of George Floyd

Figure 2 bellow displays the results of the topic modeling on the data set containing the tweets before the murder of George Floyd. Each topic represents the following number of tweets: Topic 1 - 205, Topic 2 – 210, Topic 3 – 182, Topic 4 – 155 and Topic 5 – 150. So, each topic contains approximately the same proportion of tweets.

Figure 2 Results Topic Modeling - Before Data Set

3.1.2 Data set - After the murder of George Floyd

The figure 3 bellow displays the results of the topic modeling on the data set containing the tweets since the murder of George Floyd. Each topic represents the following number of tweets: Topic 1 – 1439, Topic 2 – 2412, Topic 3 – 1478, Topic 4 – 3729 and Topic 5 – 2424. The distribution of tweets is a bit less proportionate compared to the before data set with some topics a bit more dominant than others.

As stated in the method section, it is difficult at this stage of the analysis to define the dimensions and the names for each topic created by the LDA. Indeed, in the data set before the murder of George Floyd (figure 2) topic 3 contains surprising words such as anomelodious, paper and trail. The same goes for the data set after the murder of George Floyd. One might wonder why, in figure 3, topic

1 has TikTok and fuck as the most used terms. Topic 3 in this same figure also appears to be difficult to interpret. Therefore, the results of the content analysis will be exposed in order to define the dimensions of the LDA.

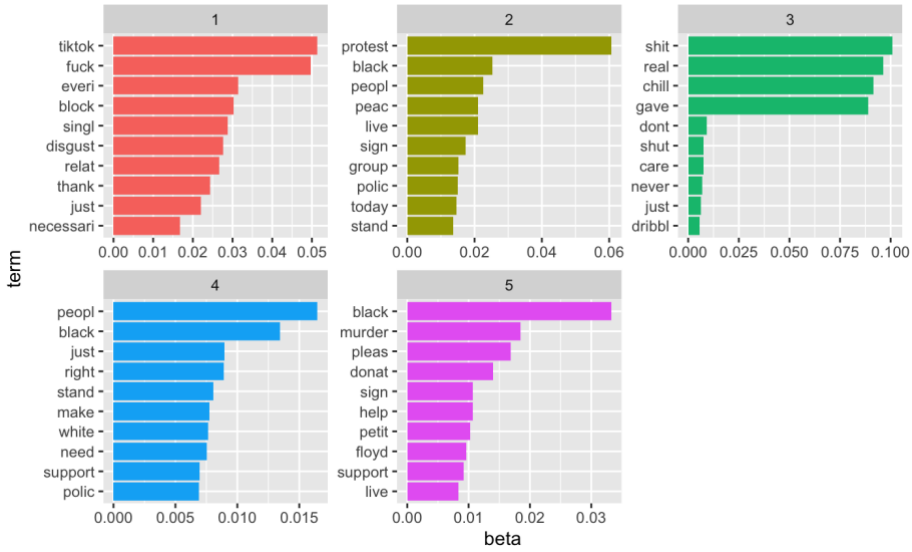


Figure 3 Results Topic Modeling - After Data Set

3.2 Content Analysis

The content analysis made on a sample of tweets from both data sets were informative. First of all, it enabled us to determine the different dimensions present in the data sets. The graphs in appendix 2 illustrate the second order classification as well as the aggregate dimensions that have been defined and synthesized in order to obtain the same classification for both data sets. The nine aggregate dimensions obtained are the following: intersectionality, support, police violence, systemic change, tactic, I am black and I am proud, TikTok and denunciation. As illustrated in the graphs, the intersectionality dimension clusters tweets about the necessity of intersectional activism which reflects one of the objectives of the movement, that is to include "Black queer and trans folks, disabled folks, undocumented folks, folks with records, women, and all Black lives along the gender spectrum" (Black Lives Matter, n.d.). The support dimension represents all the tweets that support the movement and the cause of Black Lives Matter. The tweets coded with police violence most of the time referred to the killings of people of color by police officers as well as to defund the police which is also a recurring theme within the movement (Black Lives Matter, 2020). The systemic change dimension refers to tweets mentioning the systemic dimension of racism and the need for structural change. The tweets classified within the tactics dimension referred to protest and peaceful protest but also rising awareness about racism and black culture. The dimension I am Black and I am proud comes from a single tweet, where a black army officer at a protest repeats with the crowd: I am Black and I am proud. The TikTok

dimension also relate to one tweet that mention TikTok banning any mention related to Black Lives Matter and the murder of George Floyd. Finally, the denunciation dimension represents tweets that denounce white people not only of inaction but also of being complicit to inequality.

Four dimensions were present in both data sets: support, police violence, systemic change and tactics. The data set containing the tweets before the murder of George Floyd also had the intersectionality dimension. The second data was characterized by two recurring tweets. The first being the tweet I am Black and I am proud. This tweet appears 705 times in the 14,042 sample (5% of the sample) and 12 times in the 300 sample (4% of the sample). The second was the tweet characterized by the TikTok dimension. This tweet appears more than 260 times in the 14,042 sample. One more category characterizes the after data set, denunciation. Indeed, the second data set contains a lot more tweets that denounce white people's behavior.

In order to get a clear vision of the coded tweets, the results of the qualitative analysis were quantified and graphed (Figure 4 and 5). Not only it illustrates the dimension of the elements but also how this evolved over time.

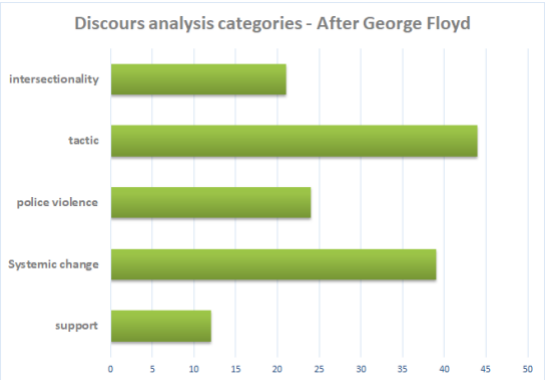


Figure 5 Results of qualitative analysis – Before Data Set

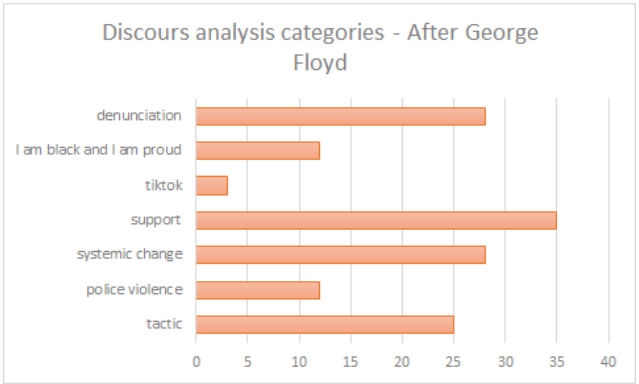


Figure 4 Results of qualitative analysis - After Data Set

The tactic dimension is more important in the data set before the murder of George Floyd but it is also still an important dimension of the after data set. Support evolved in a significant way within the after data set. Regarding systemic change, the proportion is quite similar within the two data sets. Finally, police violence is a bit more mentioned in the before data set than in the after data set.

3.3 Combining quantitative and qualitative results

The content analysis enabled us to understand the topics from the topic modeling. Indeed, topic 3 from the before data set is characterized by a tweet talking about the latest Anomelodious paper trail, being a website posting updates on Black Lives Matter and events that might be of interest for the movement. The data set after the murder of George Floyd also included topics that related to recurrent tweets. Topic 1 refers to the tweet that mentions the ban conducted by TikTok. Topic 3 refers to the tweet of I am Black and I am proud, where the soldier stands with "his brothers and sisters". Therefore, most of the words in those topics are the words present in those tweets. For examples, see figure 6 to 8:

Figure 6 I am black and I am proud Tweet

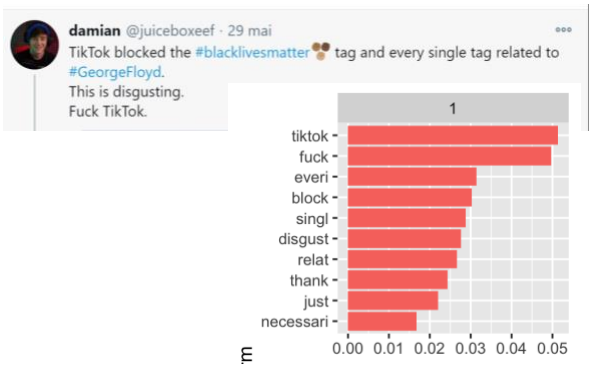


Figure 7 TikTok Tweet

Figure 8 Anomelodious Paper

Therefore, after identifying recurrent tweets, the results from the topic modeling started to make sense. This highlights the importance of mix methods in order to answer our research question. Defining the dimensions of the content analysis enabled us the name the topics from the topic models as presented in table 1 and table 2.

Before Data Set	
Topic 1	Tactic
Topic 2	Support
Topic 3	News on BLM (paper trail)
Topic 4	Systemic change
Topic 5	Police Violence

Table 1 : Topic before data set

After Data Set	
Topic 1	TikTok
Topic 2	Tactic
Topic 3	I am Black and I am proud
Topic 4	Systemic Change
Topic 5	Support

Table 2 : Topic after data set

The results obtained with the topic modeling align with the results from our qualitative analysis. However, the intersectionality dimension which was an important dimension from the qualitative

analysis is not highlighted by the topic modeling. Indeed, only the term woman appears in topic 4 of the before data set which is not enough information to link it to intersectional activism. Apart from this dimension, the topic modeling is aligned with the findings from the qualitative analysis.

4. Discussion

Our research question was the following: how did the discourse of #BlackLivesMatter on twitter evolved after the murder of Georges Floyd? Our findings show that support to the Black Lives Matter movement, police violence, claims about systemic change as well as tactics such as peaceful protest and rising awareness are recurring themes within the two data sets. Apart from these similarities, the discourse of #blacklivesmatter evolved after the murder of George Floyd. Indeed, less mentions were made about intersectional activism. Moreover, the discourse of the data set after the murder of George Floyd was more confrontational in the sense that it was denunciating more intensely. Both data sets contained a lot of tweets that were related to a specific moment or element, such as the Anomelodious paper or the TikTok ban and therefore appeared only in the data set that correspond to the time span of these events. All in all, this study shows that the movement discourse evolved after the murder of George Floyd, specifically regarding the decline of the intersectionality aspect and the increase of denunciation.

Overall, this analysis of the #BlackLivesMatter discourse thus illustrates how the murder of George Floyd induced a shift. Indeed, the data set of tweets before the murder has more of an inward focus. Activist were tweeting about the movement, black history, empowerment of the black community, education, black art, businesses owned by people of color etc. So, this shows that the movement was focused towards their own community, really promoting and empowering "Black Lives". However, after the murder, the focus had shifted to more outward focus as the tweets were denouncing police violence and structural inequalities such as white supremacy. According to Bagguley (2002) social movements that are in a moment of insurgency, movements that gain popularity and achieve mass mobilization are keener to adopt a critical outward focus than movements that are in a moment of abeyance. Movements that are in a moment of insurgency often confront the system and demand structural, fundamental changes. This study thus fits within a broader academic debate on the focus of social movements because this study shows that as #BlackLivesMatter gained popularity, the overall discourse changed from empowering one's own community to confronting and denouncing other people in order to change systemic racism. According to this theory, however, the focus of a movement might change back when popularity is lost and the movement ends up in a moment of abeyance. Future research might have a look at if indeed the Black Lives Matter movement returned to a more inward focus when this sudden popularity declined to see if this movement then still fits this theory. Besides that, this might explain why the dimension of intersectionality disappeared from the

discourse because this is related to a more inward focus however more research is needed on this as well.

Our mixed methods approach enabled us to obtain complementary results. The topic modeling proceeded to a first aggregation of the enormous amount of information contained in the more than 15,000 tweets that were analyzed. The qualitative analysis, namely the content analysis, enabled us to refine the topic modeling findings as well as to shed light on unexpected elements. Even though these two methods are complementary there are, however, also some limitations. Indeed, our analysis focused only on the text of the tweets. Yet, some important tweets did not contain any text but only pictures or links. For example, a tweet containing a picture of a pacifist activist giving flowers to soldiers and then being arrested appeared in the data set more than 790 times but was not taken into consideration when performing a topic model. This reflects however, an important message that the movement wished to vehiculate: nonviolence. In order to overcome this limitation, a more in-depth content analysis or an analysis of images used could be added alongside the topic modeling. However, peaceful protest and nonviolence were elements that, because of the qualitative analysis, were found in the text of tweets as well. This once more highlights the importance mixed methods

Figure 9 Tweet with only pictures

As previously stated, the two authors of this paper are white women. The time constraint for this assignment did not allow us to conduct interviews with people of color in order to verify our findings and interpretation. Therefore, until further interviews are conducted, these results must be considered as provisional. Indeed, one of the main objectives of this study was to highlight the voice of the Black community, hence without their input the results cannot be considered as conclusive and complete.

5. Bibliography

Anderson, M., Barthel, M., Perrin, A. & Vogels, E.A., (2020). *#BlackLivesMatter surges on Twitter after George Floyd's death*. Pew Research Center. https://www.pewresearch.org/fact-tank/2020/06/10/blacklivesmatter-surges-on-twitter-after-george-floyds-death/ [Retrieved 16.12.2020]

Bagguley, P. (2002). Contemporary British Feminism: A social movement in abeyance? *Social Movement Studies, Vol.1*(2), 169-185, DOI: 10.1080/1474283022000010664

Black Lives Matter (2020, July 6). *What Defunding the Police Really Means*. https://blacklivesmatter.com/what-defunding-the-police-really-means/ [retrieved 14.12.2020]

Black Lives Matter (2020, July 6). *What Defunding the Police Really Means*. https://blacklivesmatter.com/what-defunding-the-police-really-means/ [retrieved 14.12.2020]

Black Lives Matter (n.d.). *About Black Lives Matter*. https://blacklivesmatter.com/about/ [retrieved 14.12.2020]

Bryman, A. (2012). *Social research methods*. Oxford: Oxford University Press.

Garza, A. (2016). A Herstory of the #BlackLivesMatter Movement. In Hobson, J. Editor, *Are all the women still white? RETHINKING RACE, EXPENDING FEMINISM* (pp.23-28). State University of New York.

Gioia, D., Corley, K.G. & Hamilton, A.L. (2012). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational Research Methods, Vol.16*(1), 15-31. DOI:10.1177/109428112352151

Giorgi, S., Guntuku, S.C., Rahman, M., Himelein-Wachowiak, M., Kwarteng, A. & Curtis, B. (2020). Twitter Corpus of the #BlackLivesMatter Movement And Counter Protests: 2013 to 2020. arVix:2009.00596

Ince, J., Rojas, F., 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, Vol. 40*(11), 1814-1830 DOI: 10.1080/01419870.2017.1334931

Jacobi, C., Van Atteveldt, W. & Welbers, K. (2015). Quantitaive analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*. DOI:10.1080/21670811.2015.1093271

Markkanen, S. & Burgess, G. (2015). Introduction to co-production in research: summary report. Chambridge: Cambridge Centre for Housing and Planning Research

Meuser, M. & Nagel, U. (2009). The Expert Interview and Changes in Knowledge Production. In: Bogner A., Littig B., Menz W. (eds) *Interviewing Experts. Research Methods Series*. Palgrave, Macmillan, London.

Okolie, A.C. (2005). Toward an Anti-racist Research Framework: THE CASE FOR INTERVENTIVE IN-DEPTH INTERVIEWING. *Counterpoints, Vol.252*, 241-267

Shah, S. & Regina, W. (2020). *Posts mentioning 'Black lives matter' spiked on lawmakers' social media accounts after George Floyd killing*. Pew Research Center. https://www.pewresearch.org/fact-tank/2020/07/16/posts-mentioning-black-lives-matter-spiked-on-lawmakers-social-media-accounts-after-george-floyd-killing/ [Retrieved 16.12.2020]

Van Maanen, J. (1979). The Fact of Fiction in Organizational Ethnography. *Administrative Science Quarterly*, *Vol.24*(4), 539-550

6. Appendix

6.1 Appendix 1 : R Code

1. Extracting Tweets from tweet ID (p. 17 – 19)
2. Selecting Tweets (p. 20 - 21)
3. Topic Modeling (p. 22 - 26)

1. Extracting tweets from tweet ID

Following Peter's instruction to retrieve tweets from tweets ID for final project on #BlackLivesMater

0) Setting working directory

```
setwd()

library(tidyverse)
```

reading csv file of blm corpus

```
#importing the data
blm_corpus <- read_csv("blm_corpus.csv.gz") %>%
  #turning message_id into char in order to prevent scientific notation
  mutate(message_id = as.character(message_id))
```

1) Exploring the dataset

```
blm_corpus
```

```
## # A tibble: 41,801,153 x 5
##    message_id          user_id blacklivesmatter alllivesmatter blueliv
esmatter
##    <chr>                 <dbl>            <dbl>          <dbl>
<dbl>
##  1 347173160686321664 184163609                0              1
0
##  2 316973754951540736  82960228                1              0
0
##  3 315462982803009536  15164565                1              0
0
##  4 357143306540564480 406428810                1              0
0
##  5 357306157330735104  86224590                1              0
0
##  6 357259883743150080  51546100                1              0
0
##  7 316242983945129984  18380495                1              0
0
##  8 357305491447230464 164976199                1              0
0
##  9 356945513100230656  33100499                1              0
0
## 10 356294346645053440  66502631                1              0
0
## # ... with 41,801,143 more rows
```

2) Counting the number of tweet that only have #balcklivesmatter –> 35'664'304 tweets

```
nbr_blm <- blm_corpus %>%
  subset(blacklivesmatter == 1) %>%
```

```
  subset(alllivesmatter == 0) %>%
  subset(bluelivesmatter == 0) %>%
  narrow()

nbr_blm
```

```
## [1] 35664304
```

Counting the number of tweets that have #Blacklivesmatter to make sure the results are coherent

```
nbr_blm_2 <- blm_corpus %>%
  subset(blacklivesmatter == 1)

nbr_blm_2
```

```
## # A tibble: 36,892,699 x 5
##    message_id           user_id blacklivesmatter alllivesmatter blueliv
esmatter
##    <chr>                  <dbl>            <dbl>          <dbl>
<dbl>
##  1 316973754951540736  82960228                1              0
0
##  2 315462982803009536  15164565                1              0
0
##  3 357143306540564480 406428810                1              0
0
##  4 357306157330735104  86224590                1              0
0
##  5 357259883743150080  51546100                1              0
0
##  6 316242983945129984  18380495                1              0
0
##  7 357305491447230464 164976199                1              0
0
##  8 356945513100230656  33100499                1              0
0
##  9 356294346645053440  66502631                1              0
0
## 10 355451390400802816  57498443                1              0
0
## # ... with 36,892,689 more rows
```

3) filtering only BLM and keeping only the message ID

```
blm <- blm_corpus %>%
  filter(blacklivesmatter == 1, alllivesmatter == 0, bluelivesmatter == 0)
%>%
  select(1)

blm
```

```
## # A tibble: 35,664,304 x 1
##    message_id
##    <chr>
```

```
##  1 316973754951540736
##  2 315462982803009536
##  3 357143306540564480
##  4 357306157330735104
##  5 357259883743150080
##  6 316242983945129984
##  7 357305491447230464
##  8 356945513100230656
##  9 356294346645053440
## 10 355451390400802816
## # ... with 35,664,294 more rows
```

4) selecting randomly 200'000 tweet id with set.seed() that enable replication

```
# making sure the selection can be replicated use set.seed
set.seed(1)

selection <- blm %>%
  sample_n(400000)

selection
```

```
## # A tibble: 400,000 x 1
##    message_id
##    <chr>
##  1 1273680531721342976
##  2 1269324894577348608
##  3 1267635641011109888
##  4 1269478582654504960
##  5 1270162818051051520
##  6 1183688301087248384
##  7 1270105630586466304
##  8 1255097325803864064
##  9 1275570230249439232
## 10 751755485838053376
## # ... with 399,990 more rows
```

6.1.1.1 creating the file that contains the tweet id we want to extract with the API

putting id back to numeric because hydrator does not allow ""

```
#deactivation scientific notation in R
options(scipen = 999)

#transforming message_id back to numeric
selection <- selection %>%
  mutate(message_id = as.numeric(message_id))
```

creating the file for hydrator

```
write.table(selection, col.names = FALSE, row.names = FALSE, "tweetids_2.t
xt")
```

2. Selecting tweets

0) setting working directory

```
setwd()
```

1) libraries

```
library(tidyverse)
library(dplyr)
library(lubridate)

## Warning: package 'lubridate' was built under R version 4.0.3
```

2) reading CSV files

```
data <- read_csv("tweets_blm_3.csv")

## Warning: Duplicated column names deduplicated: 'user_screen_name' =>
## 'user_screen_name_1' [30]

## Warning: 2 parsing failures.
##   row       col expected        actual                file
## 40865 user_name          embedded null 'tweets_blm_3.csv'
## 93575 user_name          embedded null 'tweets_blm_3.csv'
```

3) Deleting tweets with no hashtag and that does not contain black*

```
data <- data %>%
  drop_na(hashtags) %>%
  filter(grepl("blackli", hashtags, ignore.case = F))
```

4) Deleting unnecessary columns

```
data <- data %>%
  select(id, created_at, hashtags, text, tweet_url)
```

5) Selecting based on the date

```
# making sure the created_at class is char
class(data$created_at)

## [1] "character"

# making sure the format will be read
Sys.setlocale("LC_TIME", "C")

## [1] "C"

# checking format of date

head(data$created_at)

## [1] "Tue Dec 15 17:56:21 +0000 2015" "Thu Jun 18 18:14:37 +0000 2020"
## [3] "Tue Jun 09 01:16:28 +0000 2020" "Tue Dec 16 05:33:07 +0000 2014"
## [5] "Mon Apr 30 08:36:16 +0000 2018" "Mon Jun 08 19:40:15 +0000 2020"
```

```
# cleaning date

data$date <- format(as.Date(data$created_at, format = "%a %b %d %H:%M:%S %
z %Y" ), "%Y-%m-%d")
data$year <- format(as.Date(data$date, format = "%Y-%m-%d"),"%y")
data$month <- format(as.Date(data$date, format = "%Y-%m-%d"),"%m")
```

selecting tweets from 2020

```
data <- data %>%
  filter(year >= "19")
```

data set before the murder of George Floyd - 1 January 2020 - 24 May 2020

```
before_GF <- data %>%
  filter(date < as.Date("2020-05-25"))
```

data set from the murder of George Floyd 25 may

```
after_GF <- data %>%
   filter(date >= as.Date("2020-05-25"))
```

creating csv files

```
write_csv(before_GF, "before_gf.csv", col_names = T)
write_csv(after_GF, "after_gf.csv", col_names = T)
```

3. Topic modeling

0. setting working directory

```
setwd("~/Documents/RMA Societal Resilience/Year 1/BDSD/Final/R")
```

1. Load packages + import dataset

```
library(tidyverse)
library(quanteda)
library(topicmodels)
library(ggplot2)
library(dplyr)
library(tidytext)

before_GF = read.csv("before_gf.csv")
after_GF = read.csv("after_gf.csv")
```

2. create dtm before_GF and clean it up

```
corpus_before = corpus(before_GF, text_field = "text")
dtm_before = corpus_before %>% dfm(tolower = TRUE, stem = TRUE, remove = s
topwords('en'), remove_punct = TRUE)
dtm_before = dfm_select(dtm_before, pattern = c("\\b\\w{1,3}\\b", "[^[:aln
um:]]", 'http\\S+\\s*', "#\\w+", "@\\w+"), selection = "remove", valuetype
= "regex")
dtm_before
```

```
## Document-feature matrix of: 1,116 documents, 2,037 features (99.7% spar
se) and 7 docvars.
##        features
## docs    climat justic mean black girl like world alreadi feel effect
##   text1      2      1    1     1    1    1     1       1    1      1
##   text2      0      0    0     2    0    0     0       0    0      0
##   text3      0      0    0     0    0    0     0       0    0      0
##   text4      0      0    0     0    0    0     0       0    0      0
##   text5      0      2    0     0    0    0     0       0    0      0
##   text6      0      0    0     0    0    0     0       0    0      0
## [ reached max_ndoc ... 1,110 more documents, reached max_nfeat ... 2,02
7 more features ]
```

3. Run LDA before_gf

```
before_2 = convert(dtm_before, to = "topicmodels")
set.seed(1)
model_before = LDA(before_2, method = "Gibbs", k = 5, control = list(alpha
= 0.1))

terms(model_before, 10)
```

```
##       Topic 1  Topic 2   Topic 3     Topic 4   Topic 5
##  [1,] "black"  "need"    "thank"     "black"   "polic"
##  [2,] "peopl"  "know"    "latest"    "back"    "white"
##  [3,] "matter" "come"    "daili"     "white"   "real"
##  [4,] "live"   "year"    "phone"     "histori" "offic"
##  [5,] "life"   "just"    "trail"     "peopl"   "kill"
```

```
##  [6,] "need"   "pleas"   "block"     "month"   "state"
##  [7,] "just"   "share"   "america"   "defend"  "fuck"
##  [8,] "white"  "justic"  "anomelodi" "endors"  "murder"
##  [9,] "time"   "help"    "paper"     "woman"   "like"
## [10,] "like"   "support" "polit"     "public"  "famili"
```

3a. Tweets per topic before

```
topics_bf <- topics(model_before)
table(topics_bf)

## topics_bf
##   1   2   3   4   5
## 205 210 182 155 150
```

4. create dtm after_GF and clean it up

```
corpus_after = corpus(after_GF, text_field = "text")
dtm_after = corpus_after %>% dfm(tolower = TRUE, stem = TRUE, remove = sto
pwords('en'), remove_punct = TRUE)
dtm_after = dfm_select(dtm_after, pattern = c("\\b\\w{1,3}\\b", "[^[:alnum
:]]", 'http\\S+\\s*', "@\\w+", "#\\w+"), selection = "remove", valuetype =
"regex")
dtm_after

## Document-feature matrix of: 14,042 documents, 8,429 features (99.9% spa
rse) and 7 docvars.
##        features
## docs    sunkiss chef know wonder antin davi taught group milwauke prote
st
##   text1       1    1    0      0     0    0      0     0        0
0
##   text2       0    0    1      1     0    0      0     0        0
0
##   text3       0    0    0      0     1    1      1     1        1
1
##   text4       0    0    0      0     0    0      0     0        0
0
##   text5       0    0    1      0     0    0      0     0        0
0
##   text6       0    0    0      0     0    0      0     0        0
0
## [ reached max_ndoc ... 14,036 more documents, reached max_nfeat ... 8,4
19 more features ]
```

5. Run LDA after_gf

```
after_2 = convert(dtm_after, to = "topicmodels")
set.seed(1)
model_after = LDA(after_2, method = "Gibbs", k = 5, control = list(alpha =
0.1))

terms(model_after, 5)
```

```
##      Topic 1  Topic 2   Topic 3 Topic 4 Topic 5
## [1,] "tiktok" "protest" "shit"  "peopl" "black"
## [2,] "fuck"   "black"   "real"  "black" "murder"
## [3,] "everi"  "peopl"   "chill" "just"  "pleas"
## [4,] "block"  "peac"    "gave"  "right" "donat"
## [5,] "singl"  "live"    "dont"  "stand" "sign"
```

5a. Tweets per topic after

```
topics_af <- topics(model_after)
table(topics_af)

## topics_af
##    1    2    3    4    5
## 1439 2412 1478 3729 2424
```

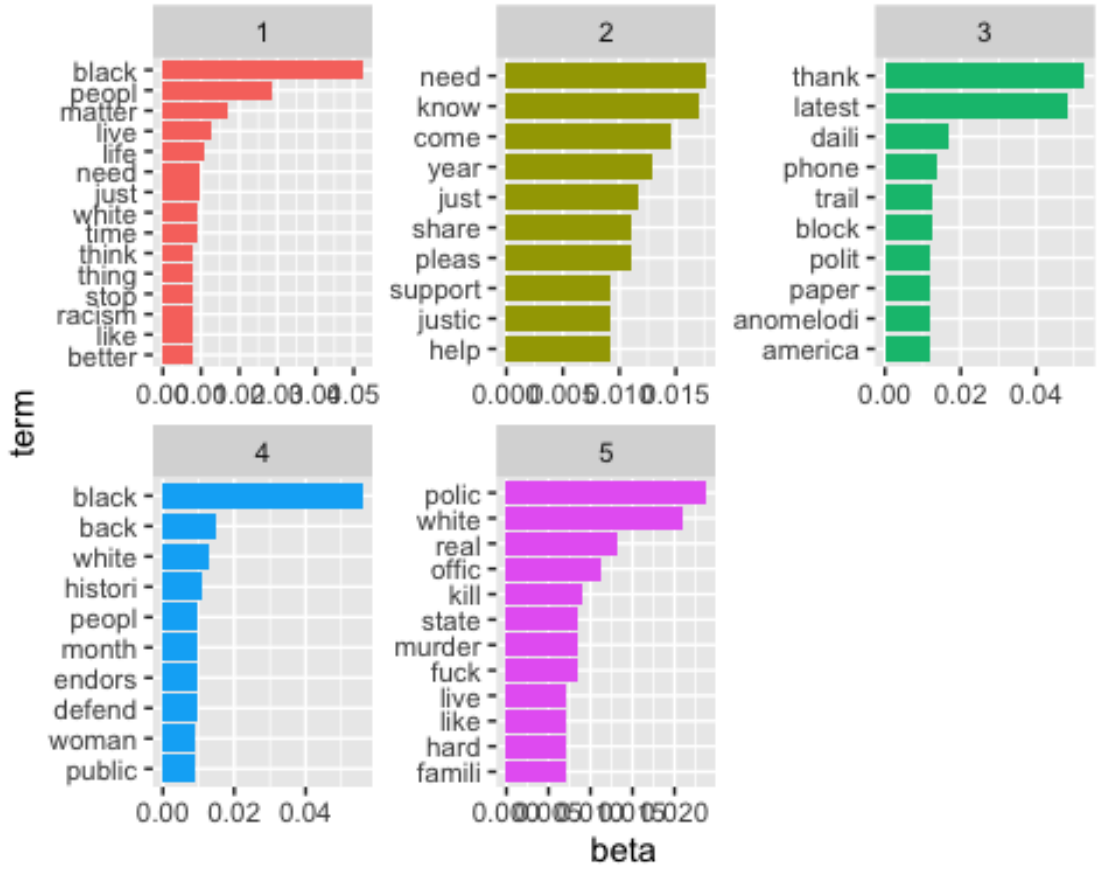
6. top terms per topic before

```
before_topics = tidy(model_before, matrix = "beta")
before_topics

## # A tibble: 10,185 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     1 climat 0.0000594
##  2     2 climat 0.00128
##  3     3 climat 0.0000788
##  4     4 climat 0.0000708
##  5     5 climat 0.0000697
##  6     1 justic 0.0000594
##  7     2 justic 0.00918
##  8     3 justic 0.0000788
##  9     4 justic 0.0000708
## 10     5 justic 0.00146
## # … with 10,175 more rows

before_top_terms = before_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

before_top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered()
```



7. top terms per topic after

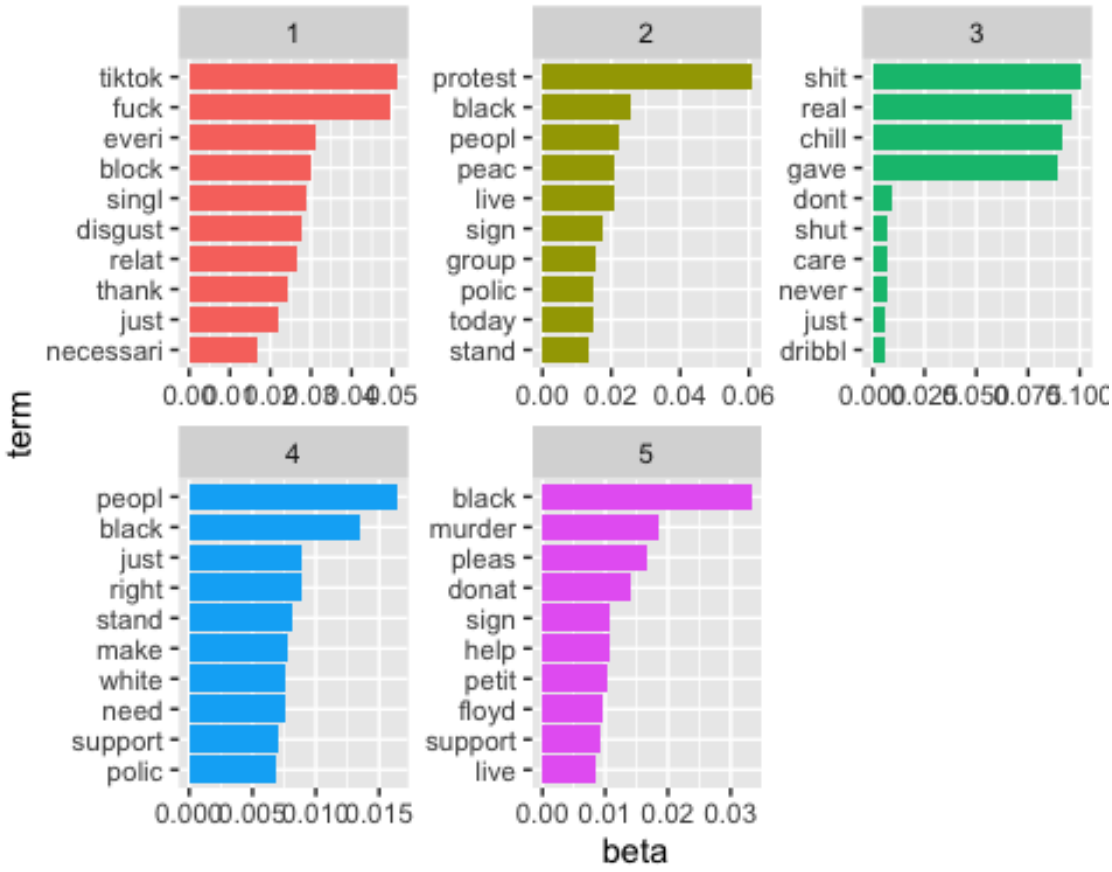
```
after_topics = tidy(model_after, matrix = "beta")
after_topics
```

```
## # A tibble: 42,145 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     1 sunkiss 0.000117
##  2     2 sunkiss 0.00000678
##  3     3 sunkiss 0.0000125
##  4     4 sunkiss 0.00000350
##  5     5 sunkiss 0.00000578
##  6     1 chef    0.0000106
##  7     2 chef    0.00000678
##  8     3 chef    0.000138
##  9     4 chef    0.00000350
## 10     5 chef    0.00000578
## # … with 42,135 more rows
```

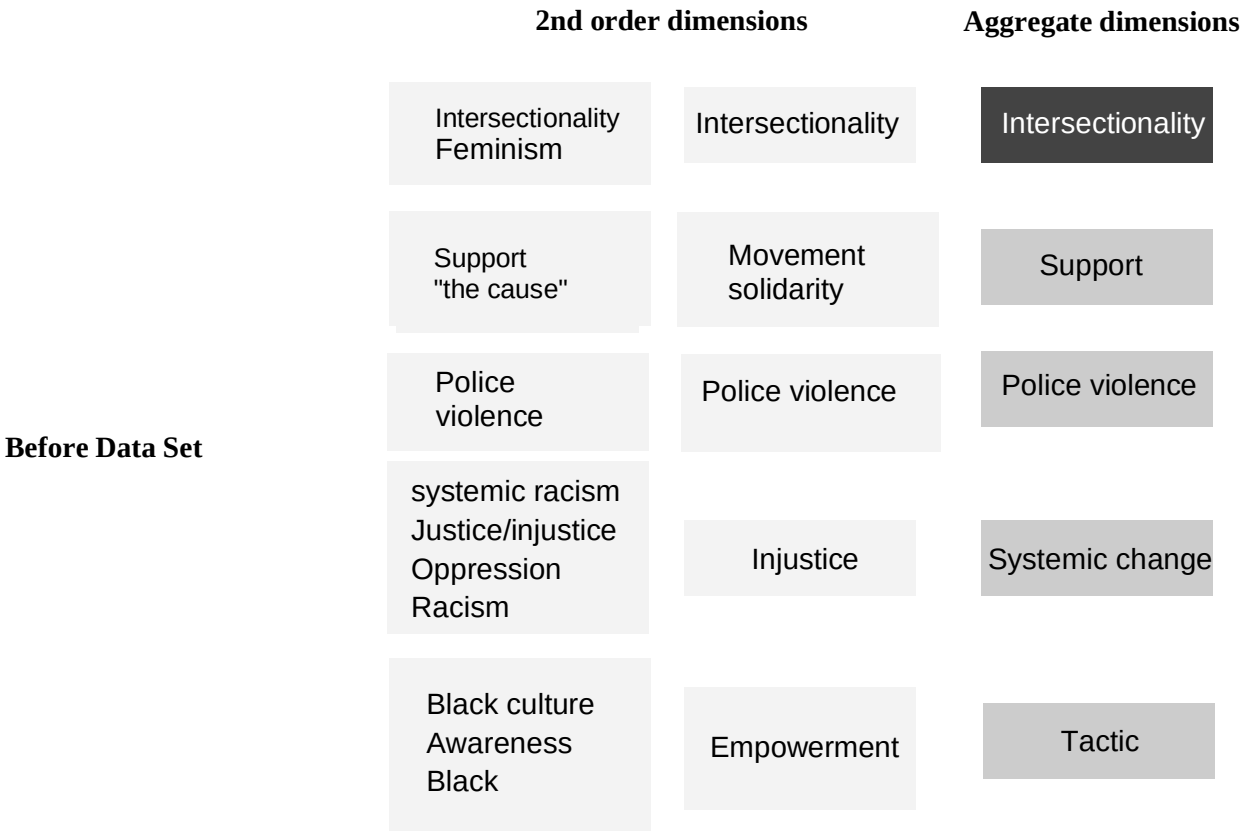
```
after_top_terms = after_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

after_top_terms %>%
```

```
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered()
```



6.2 Appendix 2: Code Schemes



	2nd order dimensions		Aggregate dimensions
After Data Set	I am black and I am proud	I am black and I am proud	I am black and I am proud
	TikTok	TikTok	TikTok
	Denunciation	Denunciation	Denunciation
	Support	Support	Support
	Denunciation Police	Denunciation Police	Police violence
	Systemic racism History Justice	Systemic change	Systemic change
	Peaceful protest Protest	Peaceful protest Protest	Tactic

6.3 Appendix 3 : Interview guide

/General questions /

Can you tell me how you relate to the movement #BlackLivesMatter?

How is your activism expressed through the movement of #BlackLivesMatter?

Are you using twitter as a tool of activism?

Do you think twitter is an impactful tool for a social movement?

Do you think that twitter can have an influence on the overall discourse of a social movement?

/More specific question/

In your opinion, does the discourse of #BlackLivesMatter evolve on twitter over time?

After the murder of George Floyd, tweets about #BlackLivesMatter grew exponentially. Why is that the case in your opinion?

After the murder of George Floyd, did you feel that the content of the tweets regarding #BlackLivesMatter changed?

/Questions about the findings/

~ showing the results of topic modeling

When looking at the results of topic modeling, what comes to your mind?

~ showing file with the sample from both data set and code

In your opinion does this sample reflect the type of tweets you have seen on your twitter feed?

Does the code scheme we have created seem relevant to you? Would you have used other words or categories?

~ showing the final analysis and results

Is our analysis lacking important points? and can you see any biases?

Would you like to add anything?