The Louisiana-Minnesota-Dallas crisis across media and time: a big data exercise

Risto Hinno & Pablo Bernabeu

July 2016

1 Introduction

Racism has long been ingrained in human societies. Ancient Greek Aristotle already claimed that non-Greeks were slaves by nature, as they easily submitted to despotic government (Reilly, Kaufman, & Bodino, 2002). This study focuses on racism in the United States, which extends from the foundation of the country, when black people were generally born into slavery, and were at any rate regarded as an inferior people. US racism stands out globally for two reasons. First, the country has played a hegemonic part in the World since soon after its foundation. Second, the US is regarded as the most advanced society technology-wise, as it sets the minutes for the technology sector worldwide. In spite of these advantages, the country has long suffered the plague of widespread racism. Indeed, the abolition of slavery in the mid-nineteenth century did not grant equal citizen rights to the black population. Over time, the black population started to confront this situation. Especially the mid-nineteenth century saw large uprisings and a patent division of different societal sectors, as reflected in literary works such as Ellison's 'Invisible Man' (1952). Inequality and confrontation about racism has extended to date, and the costs thereof have been large in terms of lives and otherwise (Feagin, 2004).

With the era of global communication, what happens in the World's most powerful country is quickly and largely spread overseas—so too with racism matters. The last major such event related to racism happened during the first weeks of July 2016. Within five days, two cases of dubious, lethal police intervention with black citizens were followed by the killing of five policemen by a black youth. The specific course of events was as follows. On July 5th, Alton Sterling was killed by police officers in Louisiana. Next day, Philando Castile was also killed by police. In this case, the presence of Castile's girlfriend during the tragedy likely determined the following events, because she described the event to the media, underscoring how gratuitous the killing was. During the following hours, outrage escalated within the already-wary population of the US. Yet the crisis would not stop there. During one of the various demonstrations held across the country, a dozen policemen were shot by a sniper, leaving five of them dead. The attacker was a black youth linked to black militant groups which target the Establishment on the grounds of patent racial discrimination. We will refer to this concatenation of events as the Louisiana-Minnesota-Dallas (LMD) crisis.[1]

In the welter of events in Louisiana, Minnesota and Dallas, some journalists warned of the return of social divisions such as those from the mid-nineteenth century. Such divisions might lead some people to an incomplete perception of the situation, and thus hinder the achievement of any solutions. However, President Obama denied such divisions as he spoke at the funeral for the policemen killed. At the same, he addressed each of the different groups in the problem, including Establishment institutions and black protesters, advising them all to exercise greater open-mindedness towards the other aspects and bands in the problem (read here: https://goo.gl/BmjFSC (https://goo.gl/BmjFSC)).

It must be noted that this small study is primarily a way for us to practise data analysis at a course. Neither the background nor the analyses make a realistic study of racism or the LMD crisis.

2 Goals

We wanted to look at these developments from the scope of online data. For this purpose, we scraped online discussions on these developments from a variety of media sources, and within defined time frames in the crisis. We then probed for any noticeable fluctuations in the topics throughout the course of events, and also for any differences across the different media. In this analysis, we took an exploratory approach by means of topic modeling. We wanted to check, first, whether topic modeling would be sensitive and useful at all within such a compact time scale. Were it to allow us, we would analyze how the journalistic and the social media reflected any fluctuations based on the live developments in Louisiana, Minnesota and Dallas. As such, the dependent variable (DV) in this study is the overall topic under discussion, which we measured via topic modeling. So, the language we analyze are messages related to the LMD crisis. Two factors are checked as potentially affecting the DV, namely Media and Time.

The Media factor regarded the three different sources of information from which we retrieved LMC content. These sources were: (1) the New York Times (NYT), (2) public tweets related to the NYT, and (3) public comments on the NYT's Facebook posts.

The Time factor was based on the following periods. The first period, from 2 to 4 July, was selected as a baseline during which no remarkable events racism-wise happened. The second period, including 5 and 6 July, contains the days when the two black citizens were killed by policemen. The third period, from 7 to 11 July, contains the aftermath of the crisis overall.

3 Hypotheses

We had several hypotheseses for our planned analyses. For the Media factor, we hypothesized a greater objectivity and formality overall for the NYT articles compared to the other two sources.

We did not have any hypotheses about the Time factor, i.e., the nature of any potential topic changes. In fact, we had considerable reservations as to whether any fluctuations would present, given the fact that the latent feeling of such a crisis might stays negative, critical and fearful from the start, regardless of particular events.

With respect to the interaction between the two factors, we hypothesized that the NYT articles would present the lowest degree of thematic variation, due to the fact that such journal pieces require time to investigate and write up —even if they are published online. Comparatively, popular comments on Twitter and Facebook would present more emotionality and subjectivity, and likely they would also present greater influence of immediate events. Furthermore, Twitter should be yet more immediate than Facebook.

Last, with respect to the DV, we did not actually have any hypotheses about the nature of possible topic fluctuations.

4 Methods

Online reactions to the LMD developments were scraped from various online sources. This content was constrained to language, bearing no extensions such as pictures or videos. In order to narrow the scope of the information, all scraped sources were related to The New York Times journal. The sources were, first, the NYT online edition (nytimes.com); second, public tweets related to NYT (@nytimes); and, third, public comments posted on the NYT page (@nytimes). Crucially, these sources are different in nature. Whereas the articles in the journal's online edition broadly follow the standard article form of mainstream journals, Facebook comments on the

page are aligned with the Facebook standards, that is, comparatively informal and outspoken. In accord, tweets referring to @nytimes follow the Twitter conventions, characterized by the 140-character restriction, and the relative immediacy of their information (Oh & Syn, 2015; Wang, He & Zhao, 2014; Josephson & Miller, 2015).

The method to scrape content related to the LMD crisis was through keywords. For the three media, the following keywords were entered, such that articles containing *any* of those words would be returned: 'black' OR 'racism' OR 'police' OR 'dallas' OR 'alton' OR 'sterling' OR 'philando' OR 'castile.' Further particulars are provided in turn.

New York Times. This scraping pipeline started from the official API site for NYT (https://developer.nytimes.com (https://developer.nytimes.com)). Metadata was downloaded for 2,000 articles adjusting to the abovementioned keywords. This returned articles dating back to the start of the year. After preprocessing, 29 articles were returned for period 1; 53 for period 2; and 275 for period 3.

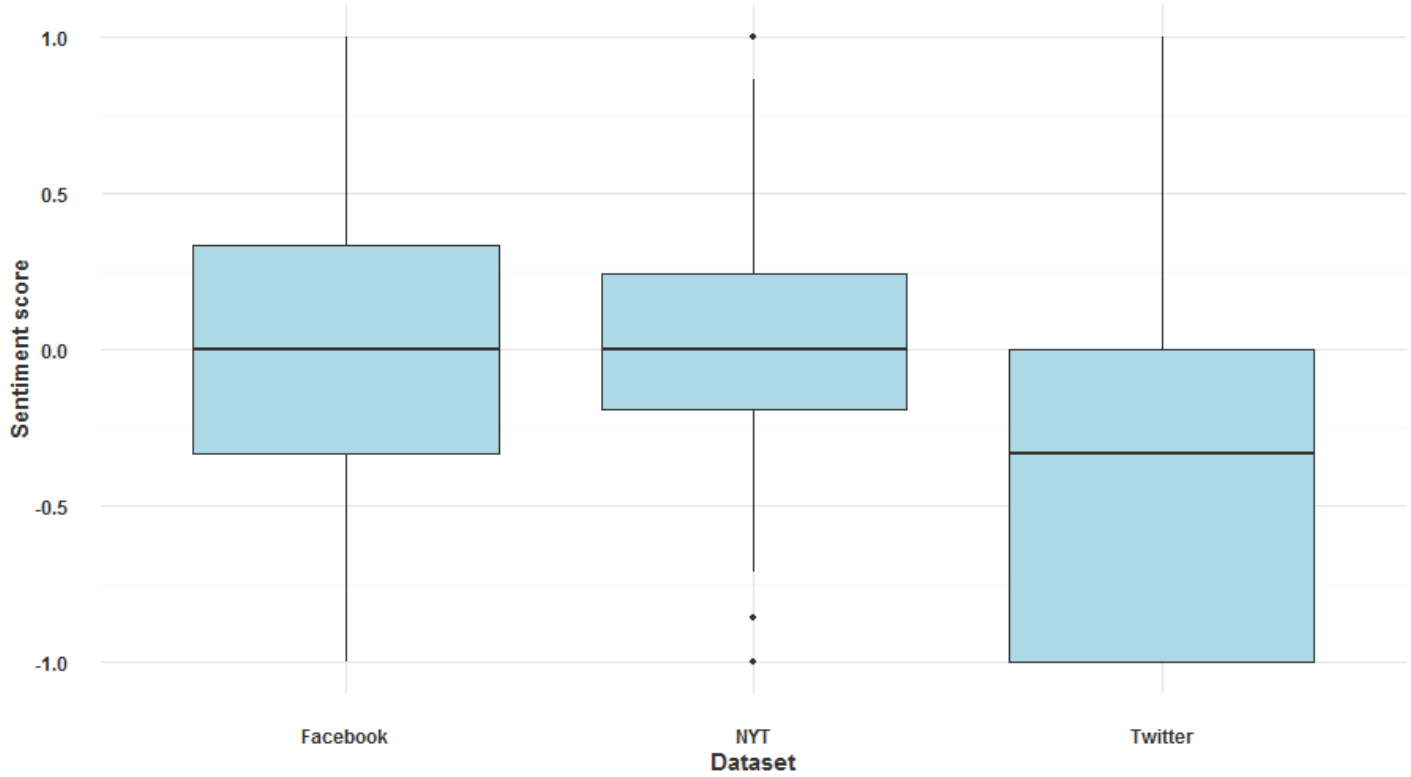
Twitter. This scraping was performed through Geoff Jentry's R package 'twitteR' (https://github.com/geoffjentry/twitteR (https://github.com/geoffjentry/twitteR)). Here tweets were selected based on the same keywords, in addition to '@nytimes' (mailto:'@nytimes'). Due to the ten-day maximum range of Twitter's API, no time range was entered. Retweets were removed. With the naked eye we realized that the tweets contained considerable information on Saudi events, we entered the word 'saudi' as a negative keyword. After preprocessing, 157 articles were returned for period 1; 489 for period 2; and 5025 for period 3.

Facebook. The 'RFacebook' R package, by Pablo Barbera (https://github.com/pablobarbera/Rfacebook (https://github.com/pablobarbera/Rfacebook)), was used for this scraping. The Facebook API currently allows for the download of all or any posts from one page, with no time restrictions (broader-search functions seem to have been deprecated). We downloaded any comments on the pages' posts which adjusted to our keywords. After preprocessing, 195 articles were returned for period 1; 1107 for period 2; and 8724 for period 3.

Preprocessing was performed equally for all sources—as standard in topic modeling, by removing non-relevant ('stop-words') and non-linguistic elements. Removed items included the names of the media, as well as numbers, punctuation, links, and technical signs such as @.

5 Results

Our hypothesis about the Media factor was only partially confirmed. First, sentiment analysis showed that NYT posts were very tempered, with an average sentiment score near 0. In contrast, Twitter presented a rather negative sentiment (Thelwall, Buckley, & Paltoglou, 2011; Saif, He, Fernandez, & Alani, 2016). Yet, to our surprise, Facebook came out with a neutrality close to that of NYT articles, even if there was greater variance among the scores of the Facebook posts. These overall tendencies are illustrated in the plot below. Caution must recommended, however, when considering this sentiment analysis, as this technique is arguably fuzzy generally, and especially so with data under such a tight time frame. This is the case because sentiment analysis, as other big data techniques, capitalizes on the size of data. What it lacks on the precision aspect, compared to null-significance hypothesis testing, for instance, it compensates with the size of the samples, in which the noise is suppressed by thousands of cases. In this case, however, the sampling within only nine days of unusual circumstances calls for circumspection.



Sentiment comparison across sources

Next, topic modeling was conducted on each source separately. The parameters for topic discovery were entered based on several attempts with different numbers of topics (K) and internal-coherence thresholds (alpha). Finally, topics were selected alike for every source, K = 3, alpha = .2. Below, the first ten words for each topic in each source are shown (note that columns are aligned rightward).

New York Times articles

```
##    Foreign policy Shooting Elections
## 1            said   police       new
## 2         percent     said       one
## 3            vote officers    people
## 4            year    black       can
## 5          brexit   dallas      like
## 6            will  officer      york
## 7           since shooting     trump
## 8        european      two      july
## 9         british     shot      even
## 10        britain     says      just
```

Facebook comments

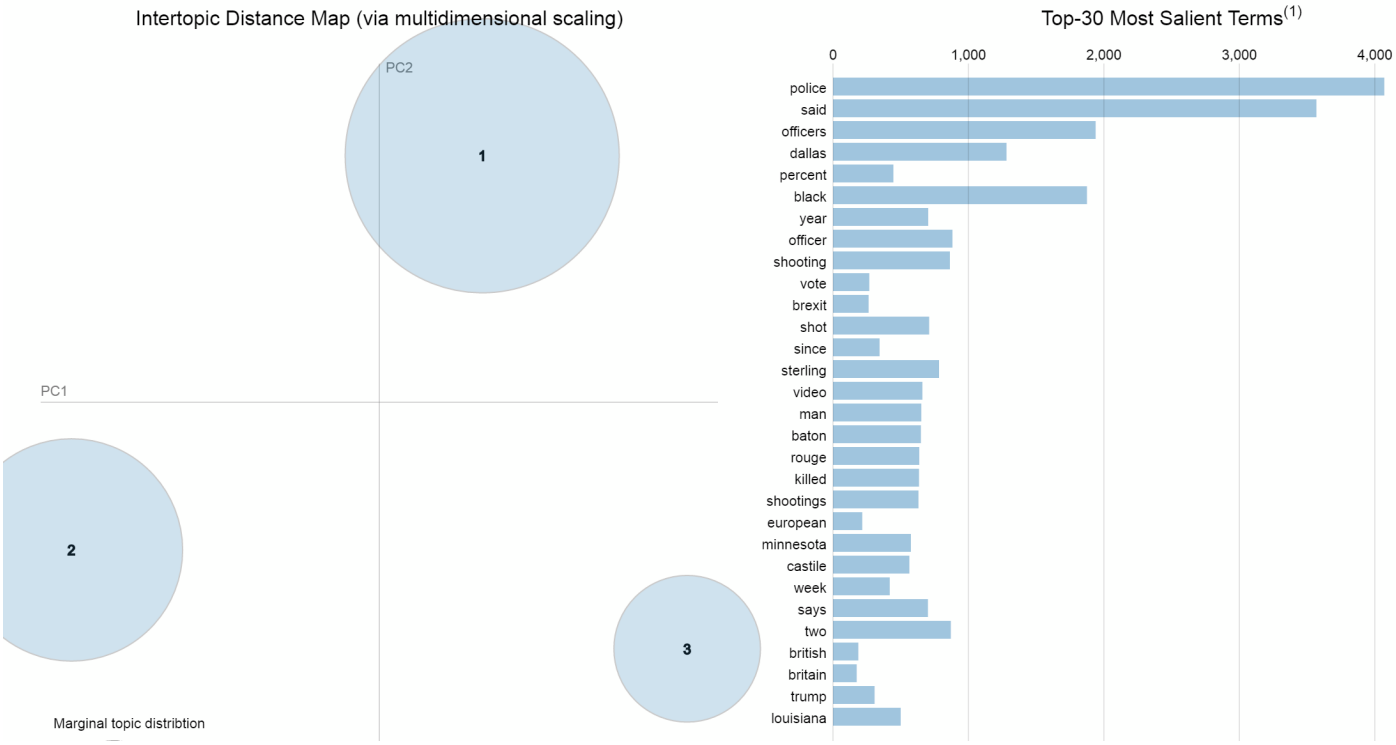
```
##      Police People Black lives matter
## 1    police people              black
## 2       gun   will              lives
## 3      cops    can             matter
## 4    people   like              white
## 5  officers    get             people
## 6      dont    one             racist
## 7      just   just                blm
## 8       man   need             police
## 9       get   dont             blacks
## 10  officer police              obama
```

Tweets

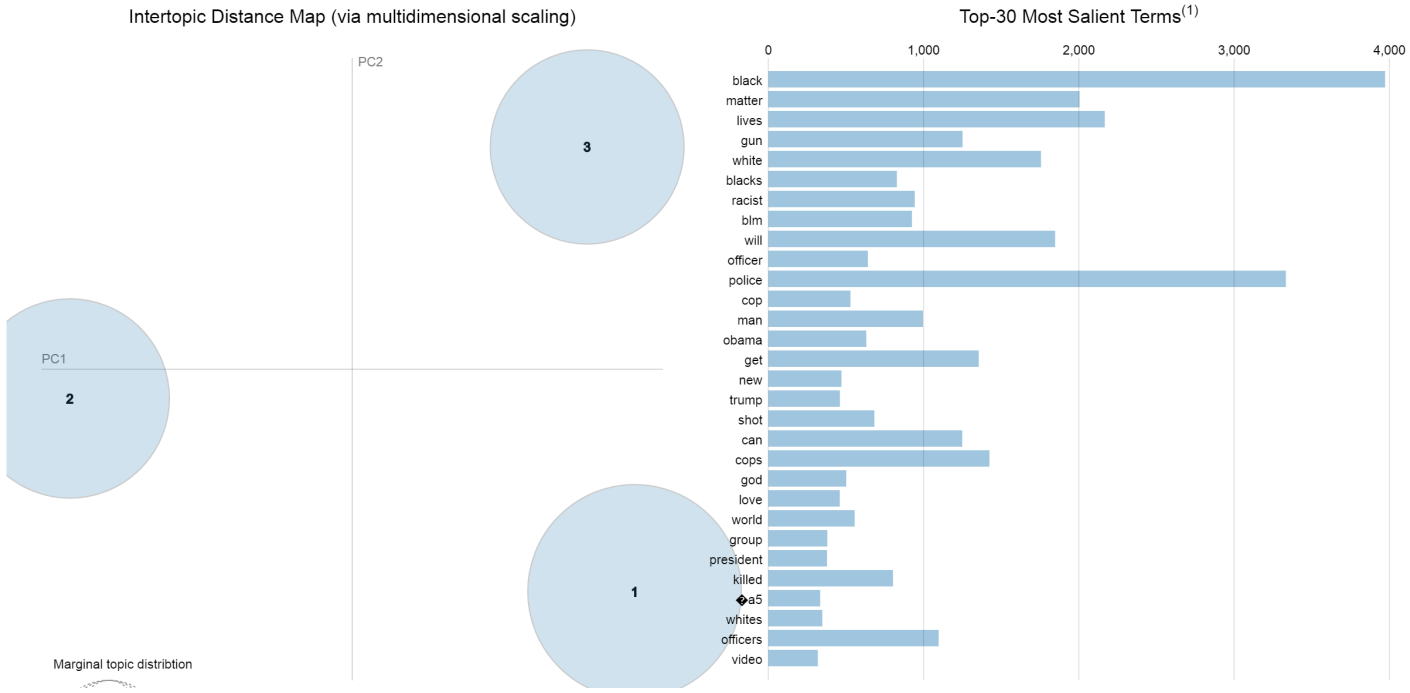
```
##    Racism Dallas shooting Philando shooting
## 1   black          police            police
## 2  police          dallas            blacks
## 3   white        officers               new
## 4   lives          killed          shooting
## 5  people         protest          philando
## 6  racism        shooting             force
## 7     amp       shootings           castile
## 8  matter           alton               use
## 9    stop        sterling              says
## 10   cops           baton          violence
```

To start, it may stand out that different topics appear across sources, all the while some are indeed shared. This is perfectly normal for topic modeling on different sources, even when the same topic is being studied. Indeed, it is very relevant for us to remark on the inclusion of foreign affairs and election matters within the NYT articles, but not within people's tweets and Facebook comments. This makes sense for several reasons. To start, the space a journalist counts on in a NYT article is considerable, compared to tweets, and also compared to ruling conventions of Facebook posts (users may write further, but the average simply will not). Second, the breadth of relation in NYT articles likely responds to the expectations from renown journalists to enrich the news with a broader contextualization. Furthermore, this extension of topics might correspond to the tacit but doubtless alignment of journals to concrete political agendas. While people commenting on Twitter or Facebook are plausibly characterized by just the same virtues and vices, their online reactions could be driven by more emotion and immediacy of focus than those of mass media journalists.

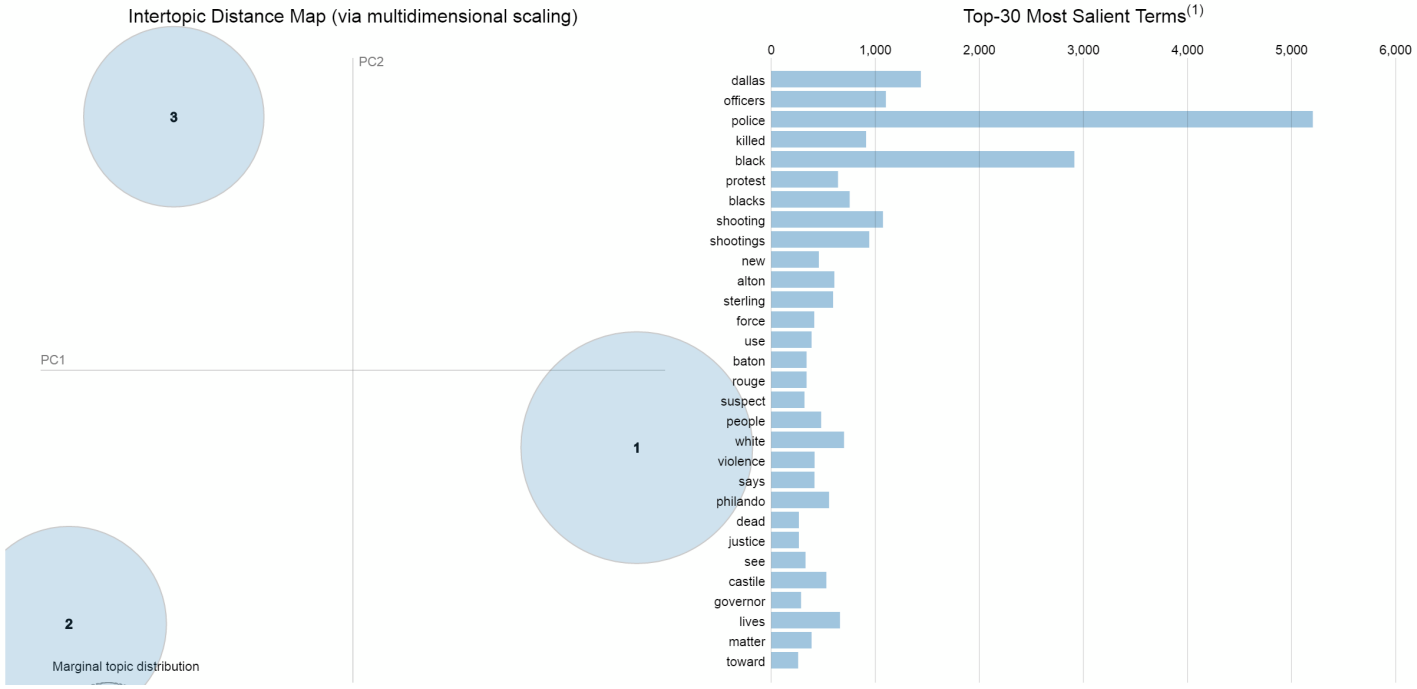
For greater visualization, we also provide some captions from the interactive LDAvis tool below. Please click on the figure titles to enjoy the full visualization.



LDAvis visualization of NYT articles (http://rristo.github.io/NYT/index.html)

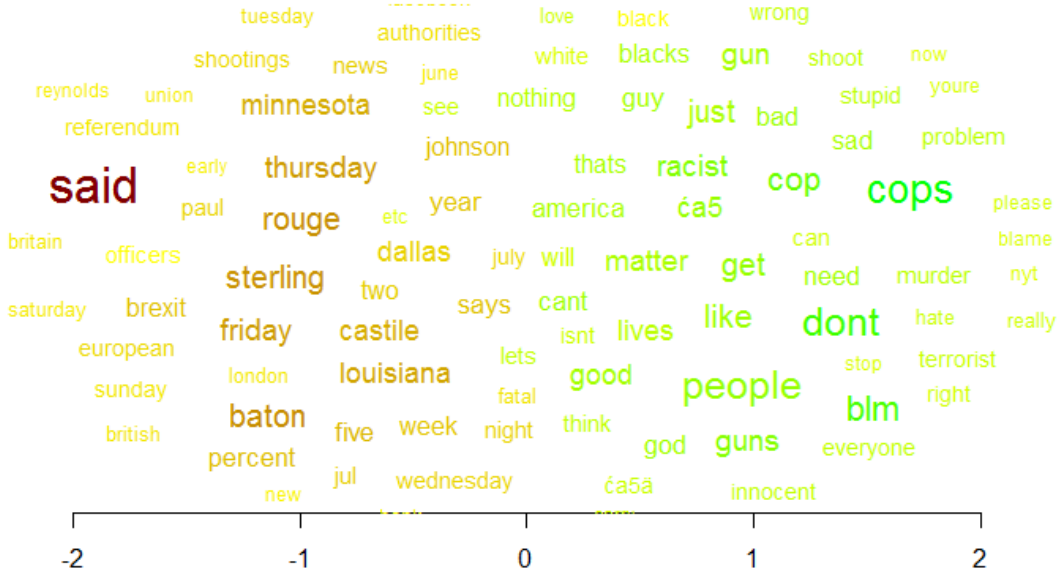


LDAvis visualization of Facebook comments
(http://rristo.github.io/Facebook/index.html#topic=0&lambda=1&term=)

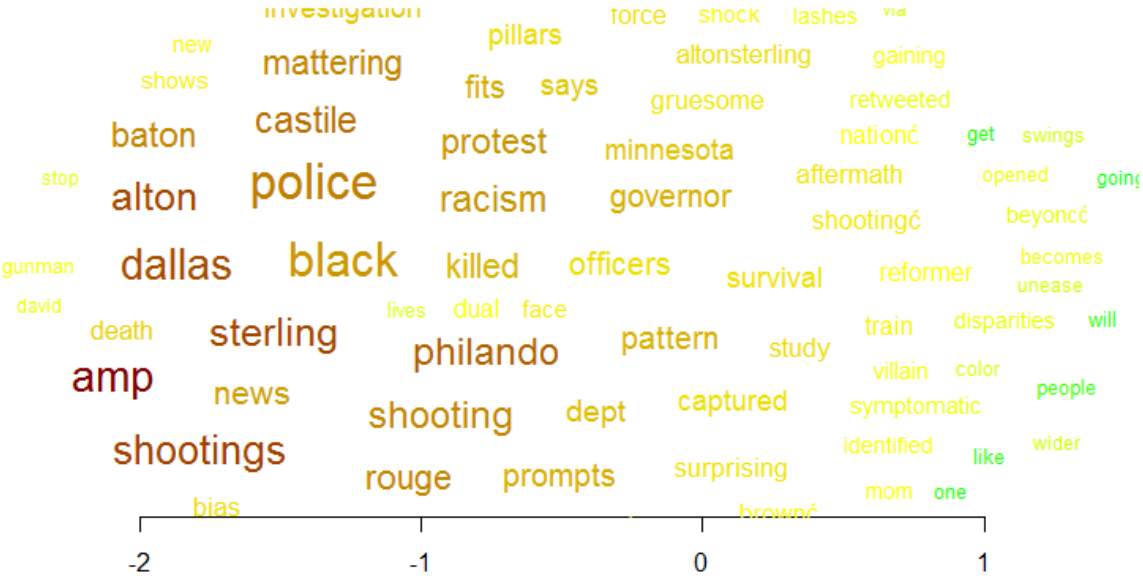


LDAvis visualization of tweets (http://rristo.github.io/Twitter/index.html#topic=0&lambda=1&term=)

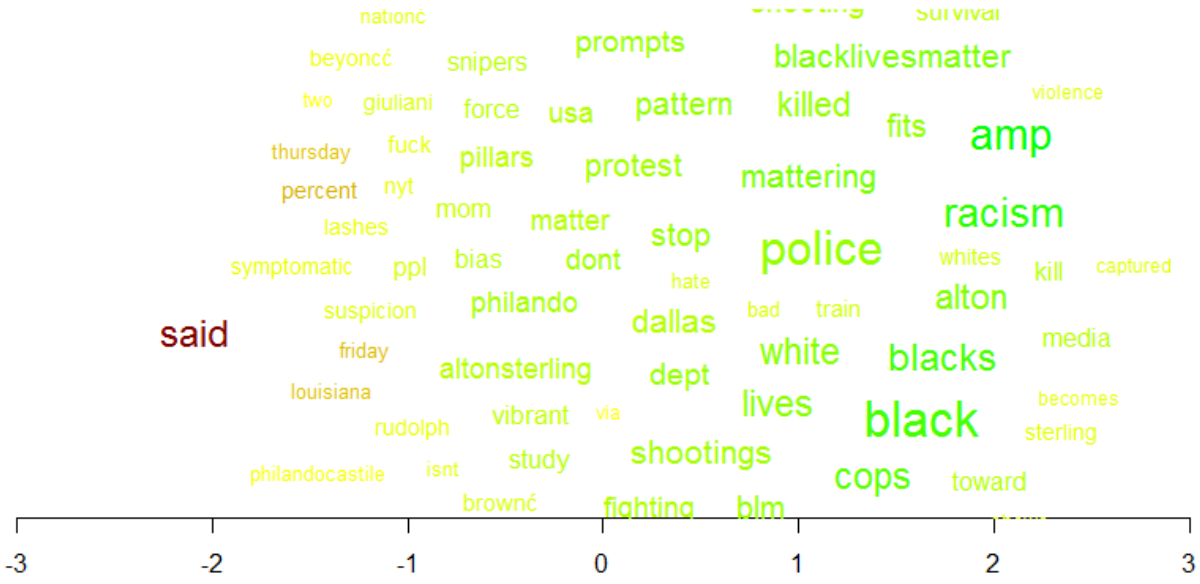
In order to specifically compare different content sources, we plotted the major language from two sources on the same plot, with an axis spanning from one source to the other, as shown below. The size of the words indicates the frequency of use, and the colour is essentially parallel with the axis, with specific different colours for different corpora, and darker hues for greater association.



Facebook comments and NYT articles

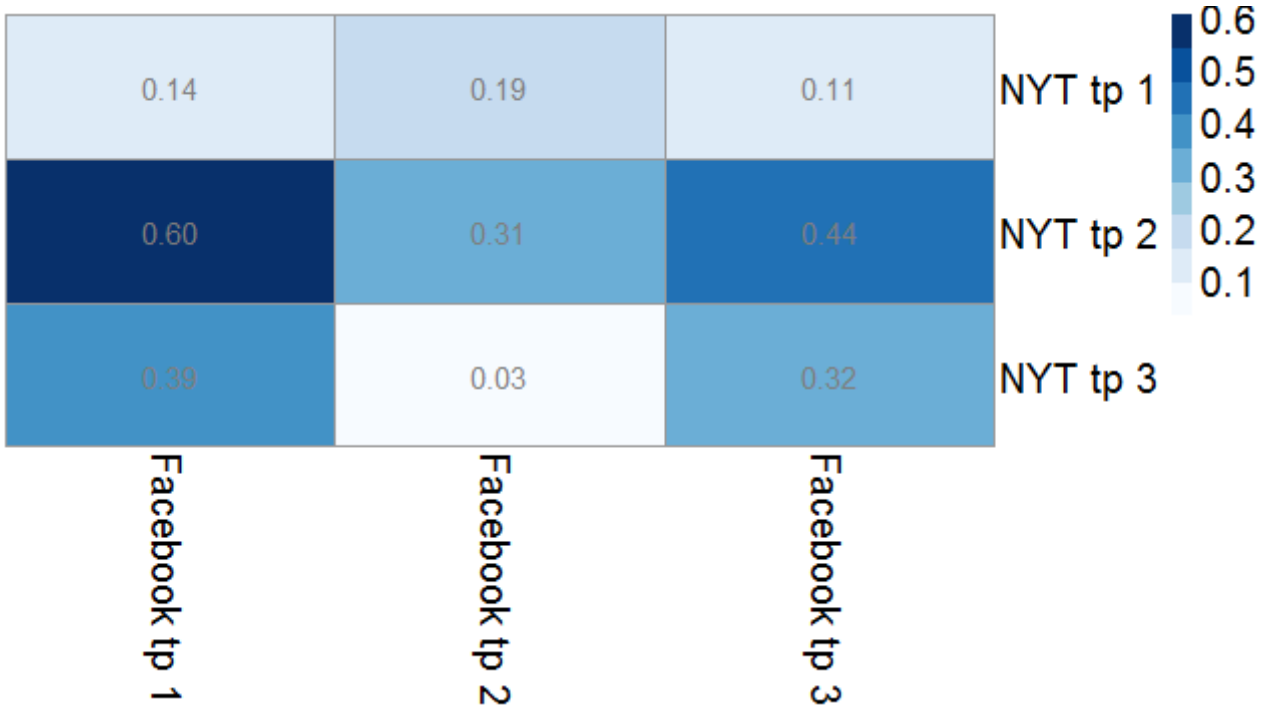


Facebook comments and tweets

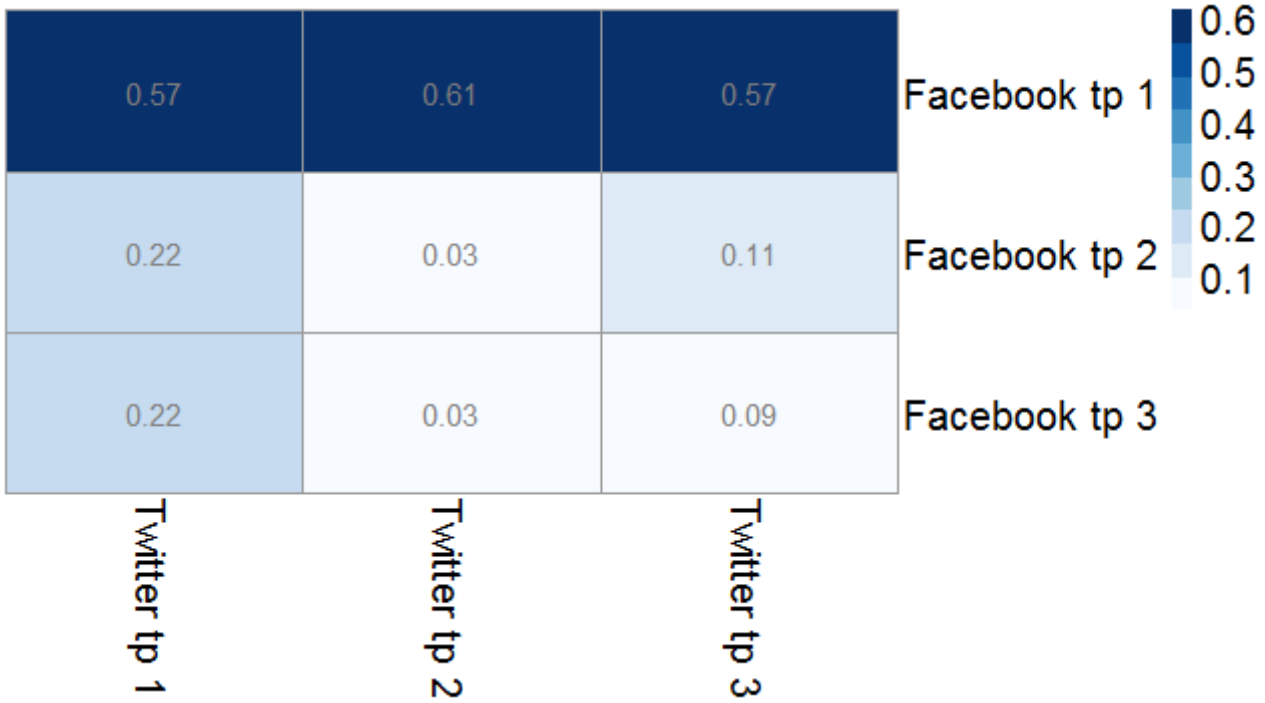


Tweets and NYT articles

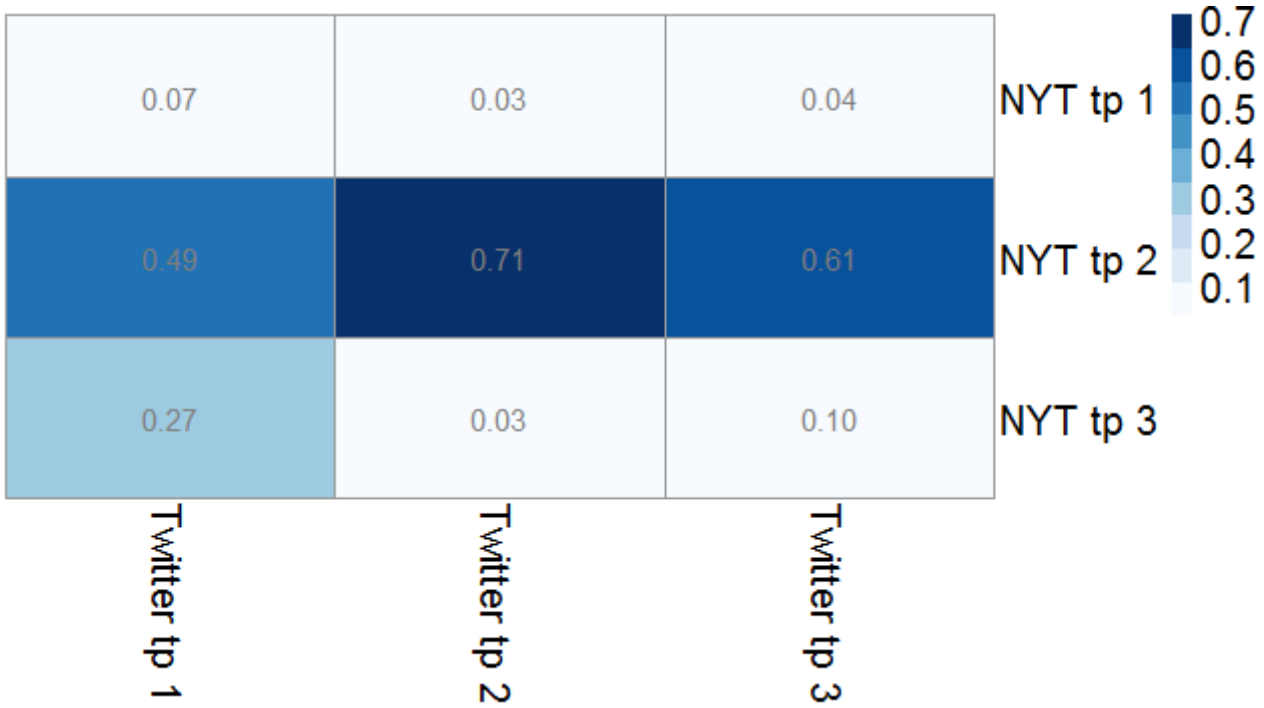
We went on to analyze the overlap in topics across journals, in order to quantitatively check whether some topics were indeed shared across sources, even if in different positions (for instance, topic 1 in some source and topic 3 in some other). We did this by means of cosine similarity scores. This scores represent the degree of similarity of two sources on a continuous scale from 0 to 1, where 1 would mean identical. The plots illustrate this comparisons in turn.



Similarity between Facebook comments and NYT articles

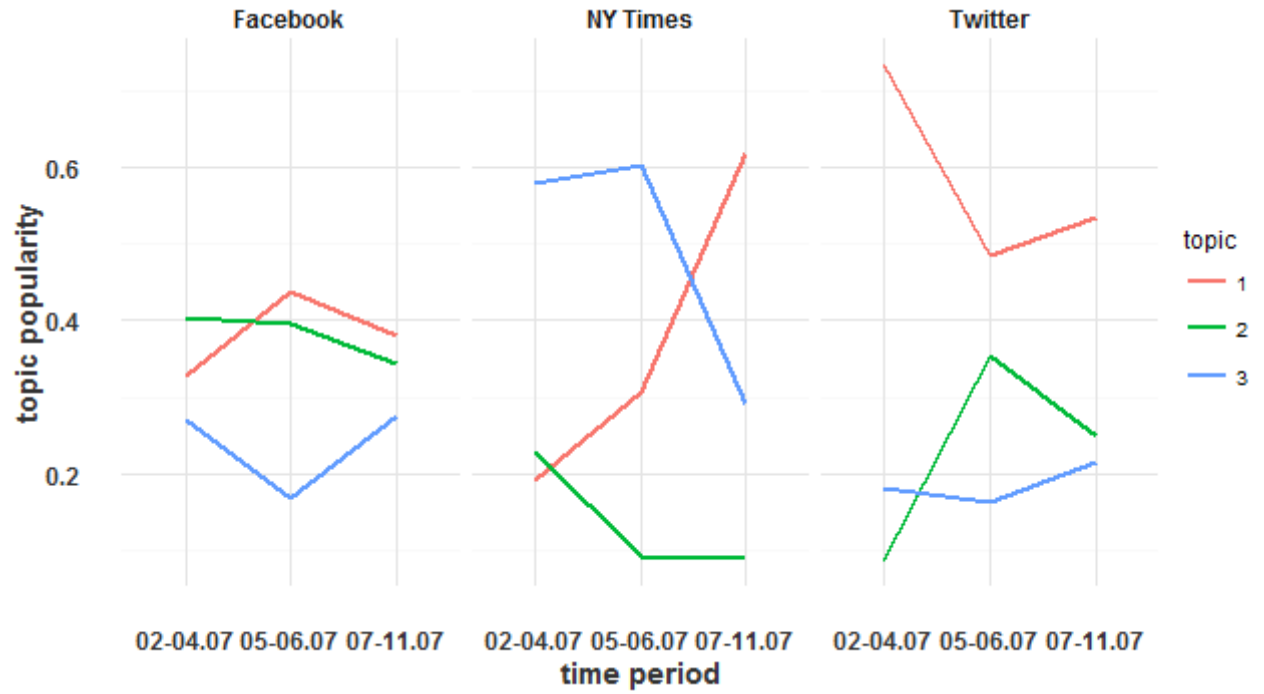


Similarity between Facebook comments and tweets



Similarity between tweets and NYT articles

Last, the interaction of Time and Media was analyzed. As expected, we found differences in the way topics fluctuated over time in the different sources, albeit in unexpected ways. NYT and tweets articles presented great variation, suggesting day-bound sensitivity to the developments. This was to be expected from Twitter, as it is famous for its immediacy. However, the immediacy of NYT articles was rather surprising, as they might have lagged behind due to the necessary investigation and editing for such kind of journalistic pieces. Unlike traditional paper-based NYT articles, this immediacy is now enabled by the publication online. Another unexpected finding was the relative stillness of Facebook posts over time. Since they are published at the minute, and nowadays mostly from mobile, we had thought they would present greater immediacy than NYT articles. We could hypothesize on this, but this would be best analyzed in further research. The plot below illustrates this interaction.



Topic fluctuations over time for the three content sources

6 Discussion

In this small-scale study, we analyzed the impact of a racism-related crisis in the American society online. This crisis started with the killing of two black citizens by policemen under dubious circumstances, which was followed by massive media attention and street demonstrations, and then continued to the fatal shooting of five policemen by a black militant (the crisis continued yet further after our analyses). The impact of this crisis was large, with a state funeral being organized for the killed policemen, and a presidential address warning of the direction of social tensions, and the need for greater empathy from all social sections involved.

We scraped the divided the social reaction to these events from three online sources, namely, the NYT online edition, public comments on the NYT Facebook page, and finally NYT-related tweets. The method was based on keywords highly relevant to this crisis, namely: ‘black’ OR ‘racism’ OR ‘police’ OR ‘dallas’ OR ‘alton’ OR ‘sterling’ OR ‘philando’ OR ‘castile’. We analyzed the Media factor and the Time factor separately, and more interestingly we looked at the interaction between these two factors.

As results, we found, first, that NYT articles were the most neutral, closely followed by Facebook posts. In contrast, tweets presented greater negativity overall. Next, we looked at topics within each time frame in each of the three sources. These topics differed across sources, even though there were also considerable overlaps. For instance, NYT articles and related tweets shared the content of their second topics, both of which revolved around ‘shooting.’ We went further to quantitatively measure any such overlaps or otherwise differences across sources. Cosine similarity—which ranges from 1, totally related, to 0, not related at all—confirmed our naked eye feeling. For instance, for the overlap between the aforementioned topics, there was a cosine similarity of .71. In contrast, a cosine of .03 came up for other comparisons, which also makes sense due to the intrinsic differences among these sources.

Last, the interaction between Time and Source was qualitatively analyzed by means of a plot, and we found that Twitter and NYT articles were most sensitive to live developments in the crisis, whereas Facebook comments lagged behing in this immediacy. All of these findings were discussed within a framework of qualitative big data analysis.

The data mass probed in these analyses could be described as medium-sized data in the big data field. This field is relatively recent, and the sucessful, seminal examples we count on tend to feature larger sizes of data. In particular, for time frames, it is rather uncommon to find such a tight scale as we excerpted. This fact complicates the drawing of assured conclusions from our findings, because we lack well-known precedents along these lines. While the social sciences have developed their tools for small samples, the tools of big data are currently designed for the larger amounts of data.

Additional materials

All materials are made public on the /RRisto GitHub page: https://goo.gl/bhRZ3K (https://goo.gl/bhRZ3K).

References

Ellison, R. (1952). *Invisible Man*. New York: Random House.

Feagin, J. R. (2004). Documenting the Costs of Slavery, Segregation, and Contemporary Racism: Why Reparations Are in Order for African *Americans Harvard BlackLetter Law Journal, 20*, 49-81.

Josephson, S., & Miller, J. S. (2015). Just State the Facts on Twitter: Eye Tracking Shows That Readers May Ignore Questions Posted by News Organizations On Twitter But Not on Facebook. *Visual Communication Quarterly, 22*(2), 94-105.

Oh, S., & Syn, S. Y. (2015). Motivations for sharing information and social support in social media: A comparative analysis of Facebook, Twitter, Delicious, YouTube, and Flickr. *Journal Of The Association For Information Science And Technology, 66*(10), 2045-2060.

Reilly, K., Kaufman, S., & Bodino, A. (2003). *Racism: A Global Reader*. London: M. E. Sharpe

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing And Management, 52*(1), 5-19.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal Of The American Society For Information Science And Technology, 62*(2), 406-418.

Wang P., He W., & Zhao J. (2014). A Tale of Three Social Networks: User Activity Comparisons across Facebook, Twitter, and Foursquare. *IEEE Internet Computing, 18*(2), 10-15.

---

1. A later update: On July 17, 2016—days after the current analysis—, the LMD crisis was extended with the killing of two policemen in the same Louisiana city where Alton Sterling had been killed.↩