Has the Black Lives Matter Movement Changed the Way the Media Portrays Police Shootings?

Anandi Gupta

Executive Summary

Media coverage of police shootings plays an important role in raising awareness and shaping public opinion about the issue. In this paper, I analyze a corpus of 2,615 New York Times articles on police shootings published between 2010-2022 to examine whether the media attention to this issue has increased since the birth of the Black Lives Matter movement in 2013. I also conduct a sentiment analysis to examine whether the language used to describe police shootings has shifted away from "neutral" or objective terms to more explicitly negative language. I find that media coverage of police shootings has increased since the start of the BLM movement, but attention has not been consistent during this period. Importantly, media attention towards the issue has decreased during 2022, during which over 1,000 people have been killed by the police, but less than half the number of articles about police shootings have been published as of November than were published in 2021. I additionally find no evidence that the language used to describe police shootings has become more negative over time. Limitations of the study and policy implications are discussed.

Background

Over the last decade, the Black Lives Matter ("BLM") movement has evolved dramatically from a twitter hashtag expressing outrage following the acquittal of George Zimmerman, the police officer who fatally shot Trayvon Martin, an unarmed Black teenager, in 2013 (Jackson et al., 2020), to a major social and political movement demanding a critical examination of systemic racism in the United States, particularly in the context of policing and the criminal justice system, and widespread reforms to increase police accountability. The movement reached its peak in 2020, when an estimated 15 to 26 million people in the United States participated in protests following the murder of George Floyd by police officer Derek Chauvin in Minnesota (Buchanan et al., 2021).

Media coverage is playing an increasingly important role in increasing visibility of the BLM movement as well as creating reliable documentation on the frequency of police shootings and information on the victims. Police departments are not required to publish data on the use of

lethal force, resulting in an undercounting of police shootings in official statistics such as those published by the FBI (VerBruggen, 2022). Thus, more comprehensive databases that attempt to track police shootings such as the Washington Post's *Fatal Force* database (Washington Post, n.d.) rely on news reports to supplement data from law enforcement websites.

In addition to the quantity of news coverage, the manner in which the media covers police shootings can shape the public's perception of these events as well as their "support or rejection of policies that might solve social ills such as racism and police brutality" (Jackson, 2020). Historically, media reporting of police violence has used passive language when talking about police violence, obscuring who is responsible, downplaying the aggressiveness of police officers, and branding victims as "suspects", as was seen, for example, in early media coverage of Breonna Taylor's killing (Hagle & Little, 2020; Hagle, 2020). Since the murder of George Floyd in 2020, there have been several appeals for the media to move away from using neutral and objective language like "officer-involved shootings" and in August, 2020 the Associated Press Stylebook updated their guidelines "avoid this vague jargon for shootings and other cases involving police" (Soderberg & Friedman, 2022).

Much of the existing literature examining media portrayals of police violence, protests, or the BLM movement more generally, is qualitative. However, there exists a small body of quantitative literature that utilizes text analysis methods to investigate the frequency of and language-use within media coverage on these topics. For example, in 2020, FiveThirtyEight published findings from a study analyzing closed captioning data of cable news broadcasts and headlines of online news articles showing that the usage of the phrase "Black Lives Matter" declined during the initial years of the Trump presidency, appearing less than half as frequently between 2017 and 2019 as it did from 2014 to 2016 (Mehta, 2020). More recently, the Huffington Post, in partnership with the Garrett Project, recently released a study analyzing the number of times the word "officer-involved" (or similar words such as "police-involved") appeared in news articles between the period 2000-2021. They found that the usage of the word "officer-involved" rose steadily in the 2000s and early 2010s, appearing in about 15% of all articles on police violence in 2013 (as compared to 5% in 2000). Further, they found that the usage of the phrase declined in 2020 (following the murder of George Floyd) but began to reappear more frequently in 2021, in which 8% of all articles on police violence used this language (Soderberg & Friedman, 2022).

In this project, I aim to build upon the findings of these studies by leveraging similar text analysis techniques to compare the frequency of newspaper articles covering police shootings, specifically, between 2010 and 2022. Given that the absolute number of police shootings has remained relatively stable (approximately 1000 per year) between 2015 and 2019 (Sullivan, 2019), changes in frequency would reflect changes in media attention to the issue rather than a change in the number of incidents that warrant reporting. I also examine the most frequently occurring tokens and tokens with highest importance in each time period. Finally, I conduct a sentiment analysis to examine whether the average sentiment or tone of these news articles has changed since the birth of the BLM movement and assess whether calls for the media to move away from "neutral objectivity" when portraying police violence have been successful.

Data

For my project, I analyze text from a corpus of over 2,500 newspaper articles classified under the term "Death and Injuries By Police" in the *Nexis Uni* database (LexisNexis, n.d.) (freely available to Georgetown students and faculty using a university login), which is consistent with the methodology used for identifying news articles related to police shootings in the Huffington Post study (Soderberg & Friedman, 2022). However, it is important to note that the use of this classification term to identify articles about police shootings may erroneously capture articles about the BLM movement more generally, or articles about protests condemning police brutality such as the massive protests that erupted nationwide following the murder of George Floyd, which may bias the results of the sentiment analysis and cause the average sentiment of the corpus to appear more positive than if I had been able to exclude these articles from my corpus or analyze each of those sets of articles separately.

To arrive at my corpus of articles, I then apply filters to include articles published after 2010, so as to include a few years of news articles prior to the start of the BLM movement, which will enable me to analyze how media coverage of police shootings has changed since its inception. I additionally filter on publication type to include only newspapers for consistency in length and to avoid news transcripts which, based on a manual review, occasionally refer to phrases such as "police shootings" sarcastically. Additionally, I apply filters for "location by publication" and "geography by document" to include only articles for the United States and for language to include articles in English only. However, upon manually examining the distribution of news sources within this selection, it appears that the make-up of news articles is not evenly

distributed across media organizations or ideologies. For example, the two sources that appear most frequently are the University Wire and The New York Times which together comprise almost 8000 articles, whereas Fox News is not included in the database (except for transcripts which I exclude). Given that I am most interested in investigating changes in frequency, token importance, and average sentiment of media coverage over time, any shifts in the make-up of news sources over the years would likely bias my results. This is particularly concerning given the consolidation of several news organizations during this time period (Center for Innovation and Sustainability in Local Media, 2018). Therefore, I filter on "source" to include articles only from the New York Times, as it represents the largest commercial newspaper in the corpus, reducing my corpus to 2,717 articles (as of November 5, 2022). Dropping duplicates and generic articles such as those titled "Corrections" results in a final corpus of 2,615 articles.[1] However, a key limitation of this approach of a using single newspaper is that other media sources such as non-partisan or more conservative media sources such as the Wall Street Journal or Fox News may follow dissimilar trends across time, which are not captured by my analysis. In an ideal world, I would have been able to analyze a more comprehensive corpus of articles with a balanced make-up of media organizations and partisan ideologies across time.

After arriving at my corpus of articles, I then develop a list of tokens to include in my token importance and sentiment analyses. To do this, I first determine the minimum number of documents in which a word must appear to be included as a feature in the analysis by examining the number of features corresponding to different minimum document frequency thresholds.

[1] Note, 15 articles were excluded due to a Unicode error when reading in the document into Python. However, manual examination of these articles indicates that the excluded articles are evenly distributed across the years and locations of incidents, and thus excluding them should not systematically bias my results.

.

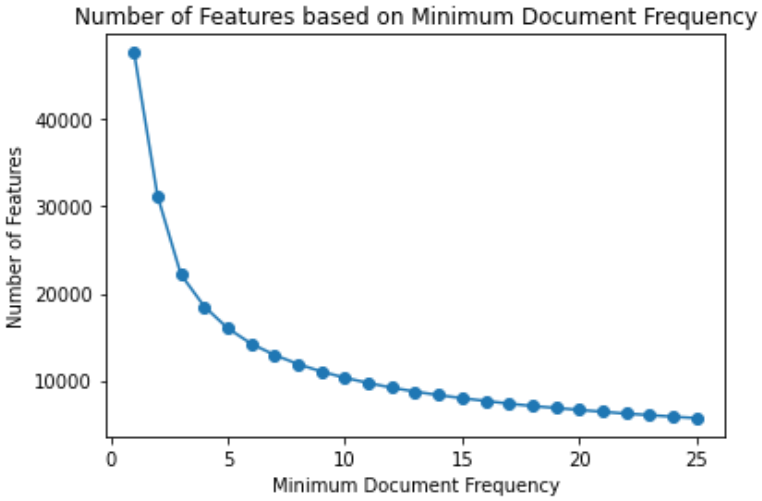

Figure 1: Selection of Minimum Document Frequency Threshold

In Figure 1 above, it appears that the number of features decline steadily as the minimum document frequency increases between 0 and 5 documents from ~48,000 features to ~16,000 features. However, as minimum document frequency increases beyond 5, the number of features decreases more slowly. Thus, I exclude tokens that appear in less than 5 documents from my analysis, as this threshold for minimum document frequency enables me to remove words that appear too infrequently to be meaningful for my analysis without losing too much information.

In addition to setting the minimum document frequency threshold, I apply several pre-processing steps outlined in the Implementation Appendix before conducting my analyses. Figure 2 below shows the 100 most commonly occurring tokens after applying these additional pre-processing steps.

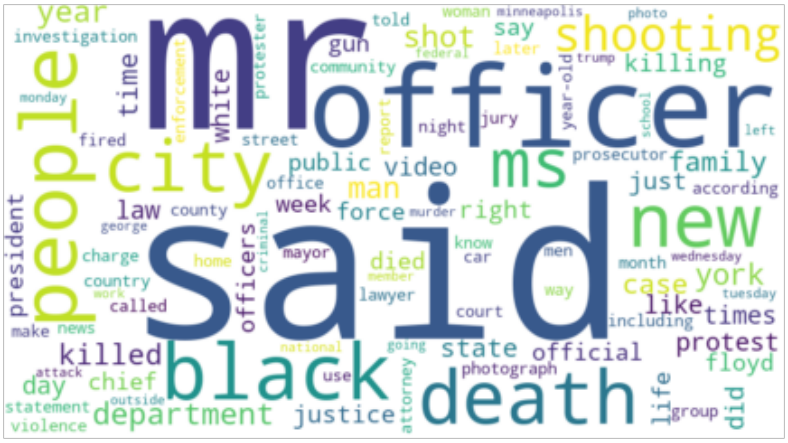

Figure 2: Top 100 Most Frequently Occurring Terms After Pre-Processing

Methodology

After conducting the pre-processing steps described in the Implementation Appendix and examining changes in the frequency of articles about police shootings across time, I use two text mining techniques to analyze my corpus of newspaper articles:

Calculation of Term Frequencies, Document Frequencies, and TF-IDF Weights

In order to measure token importance, I calculate term frequencies, document frequencies, and term-frequency-inverse document frequency (TF-IDF) weights. TF-IDF weighting is a common approach used in text analysis to identify topics that are important within a corpus. For example, Benhardus and Kalita (2013) used TF-IDF weighting to identify topics trending from Twitter streaming data. Similarly, Mee at al. (2021) used TF-IDF weights to analyze trends in tweets associated with Members of the British Parliament. TF-IDF weights represent the importance of a token relative to the corpus of documents and are calculated by combining term frequency (the number of times a token appears in a document) and document frequency (the number of documents in which the token appears), such that higher term frequency implies higher importance, whereas higher document frequency indicates lowers importance. Thus, words that appear frequently in a few documents will be given higher importance than words that occur in almost every document. Given that newspaper articles vary widely in length, I further apply normalization to my TF-IDF weights, so that the resulting weight is relative to the document length.

Sentiment Analysis (using dictionary methods)

In order to assess whether the average sentiment of news articles on police shootings has become more negative after the BLM movement and the George Floyd protests, I rely on dictionary methods to conduct a sentiment analysis for articles published before and after the start of the BLM movement, and following the George Floyd protests in 2020. Dictionary methods are widely used for assessing the "sentiment" or "tone" of political speeches, advertisements, and online activity of political users. For example, Haselmayer, et al. (2021) used sentiment analysis to examine gender differences in the level of negativity in parliamentary speeches in Australia.

I rely on the Valence Aware Dictionary and sEntiment Reasoner ("VADER"), a lexicon developed by Hutto and Gulbert (2014) and available open source under the MIT license, for my

primary sentiment analysis. VADER is increasingly being used in academic literature to conduct sentiment analyses of social media posts and other text data, particularly in the political sphere (e.g., Pinto & Murari, 2019). Although the VADER lexicon has been deemed a "gold-standard" sentiment lexicon as it has been empirically validated by humans, it is important to note that it is "specifically attuned to sentiments expressed in microblog-like contexts" and thus works best for shorter text content such as Twitter data or news headlines as compared to the longer news articles used in my analysis (Hutto & Gilbert, 2014). However, the developers note that VADER can be adapted for analyzing longer texts including articles, reports, and publications (Hutto, 2016).

I use "compound scores" calculated for each document by NLTK's Sentiment Intensity Analyzer based on the VADER lexicon to categorize documents as positive, negative, or neutral. Here, compound scores represent the sum of all sentiment scores assigned to the tokens in each document, which have been normalized between -1 and +1 such that -1 is extremely negative and +1 is extremely positive. Consistent with existing academic literature and the developers' recommendations, I categorize documents with a compound score >= 0.05 as positive, <= -0.05 as negative, and the remaining documents as neutral (Hutto, 2016).

Finally, I validate my results by using the AFINN sentiment lexicon (Nielsen, 2011), which is a list of terms that have been assigned valence scores ranging from -5 to +5, as an alternate dictionary and comparing my results.

Findings

Frequency of media coverage

From Figure 3 below, it is evident that the frequency of news articles related to police shootings has fluctuated drastically over the years, despite the number of shootings staying relative stable since data started being collected more systematically in 2015 (Sullivan, 2019). In 2010, only 3 articles were identified using my search criteria. From 2011-2013, I identified approximately 50 articles per year relating to police shootings. In 2014, the number of articles about police shootings increased to 202 articles, likely in response to the infamous shooting of Michael Brown in Ferguson in August 2014 (Demby, 2016). Media attention was sustained until 2017, after which the number of articles decreased steadily until the massive protests that erupted after the murder of George Floyd in 2020. In 2022, 122 articles about police shootings were published in the New York Times prior to November 5, 2022 (the date I pulled the articles),

which is less than half the number of articles about police shootings published in 2021, indicating that media attention has likely waned again.

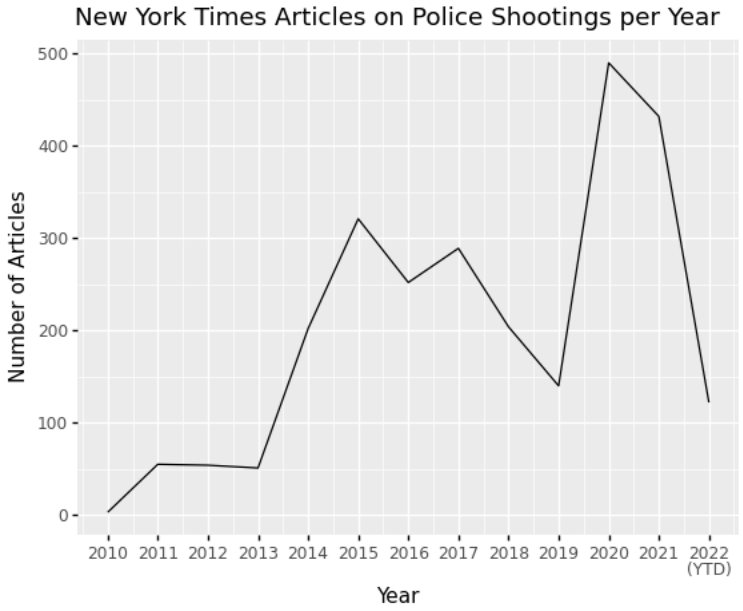

Figure 3: Number of Articles Related to Police Shootings Per Year

Term Frequencies, Document Frequencies, and TF-IDF Weighting

To assess changes in token importance between the pre-BLM (i.e. prior to the acquittal of the police officer who killed Trayvon Martin on August 13, 2013), pre-George Floyd (i.e., prior to the murder of George Floyd on May 20, 2020), and post-George Floyd periods, I computed term frequencies, document frequencies, and a TF-IDF matrix where each cell represents the TF-IDF weight for a token within a document. In Figure 4 below, I present the term frequencies for the top 10 most frequently occurring tokens for each of the three relevant time periods. As seen in the figure, the most frequently occurring terms were fairly consistent across the relevant time periods. However, words associated with race such as "black" began to appear more frequently after the start of the BLM movement in 2013, rising to the top 10 most frequently occurring terms in both the Pre-George Floyd and Post-George Floyd time periods. Interestingly, active words such as "shot" and "killed" appeared in the top 10 most frequently occurring terms in media coverage prior to the start of the BLM movement, whereas more "timid" or indirect words such as "said" began to be more frequently used after the start of the BLM movement. This is a particularly important result given Hagle and Little's thesis that the use of the passive voice can "shift the blame away from the police" (Hagle & Little, 2020).

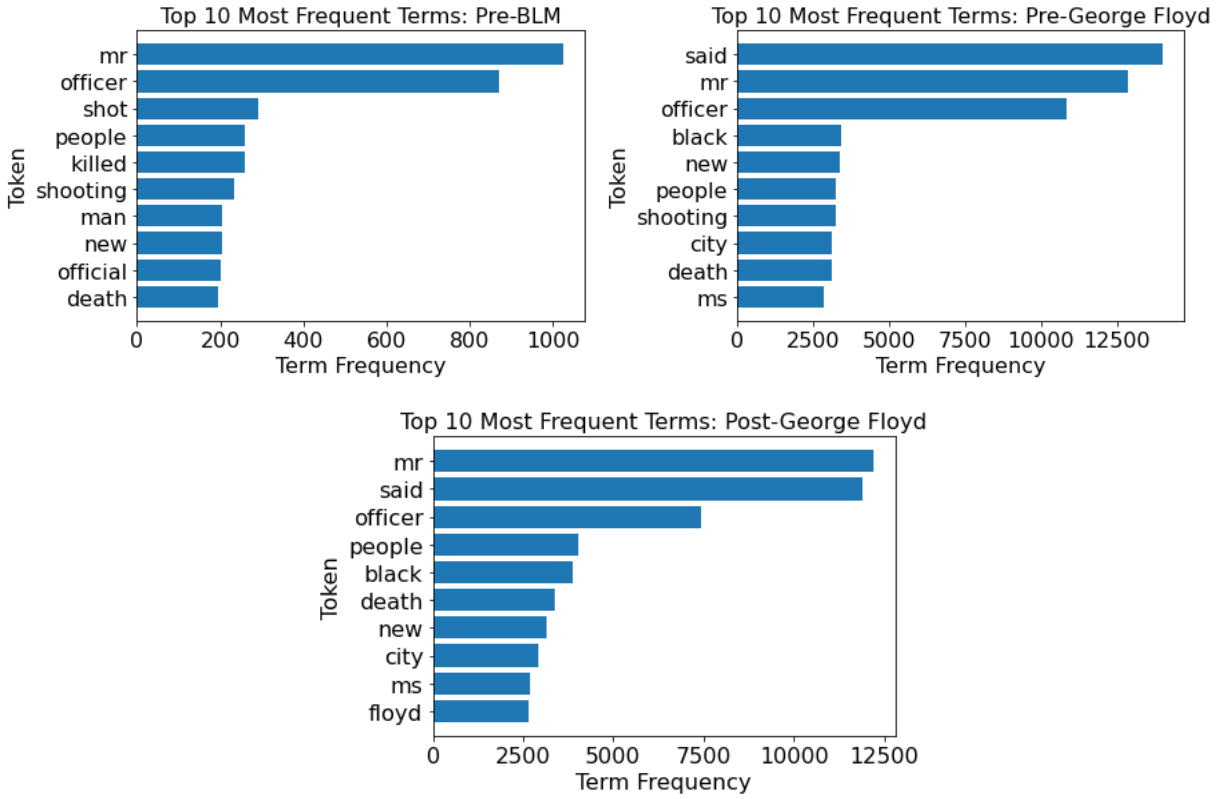

Figure 4: Top 10 Most Frequently Occurring Terms in Each of the Relevant Time Periods

Sentiment Analysis

Figure 5 below shows the proportion of news articles published during the pre-BLM, pre-George Floyd, and post-George Floyd periods that were categorized as positive, negative, or neutral, and Figure 6 shows the average compound sentiment score for articles by publication date.

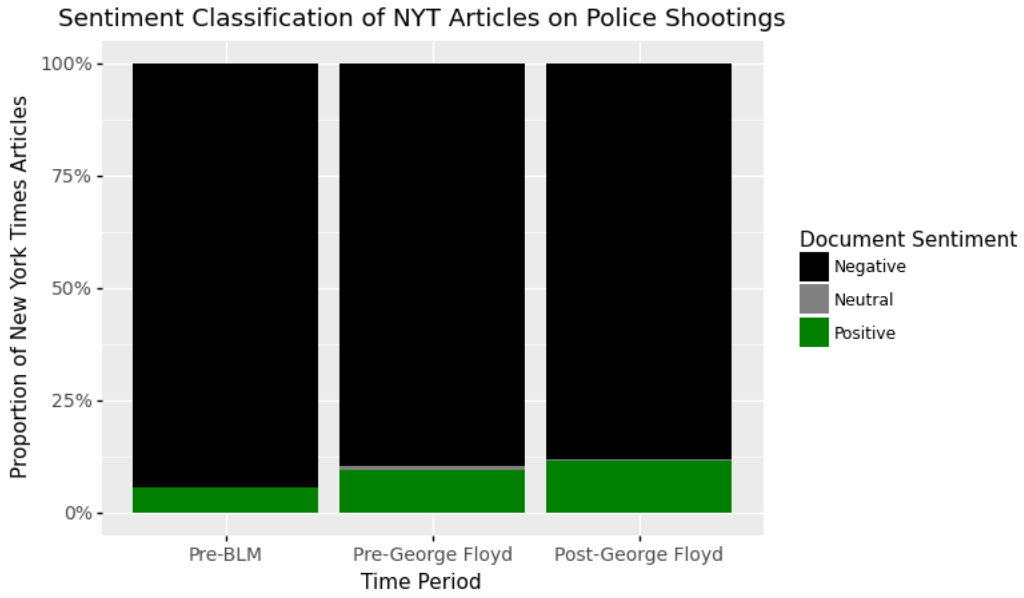

Figure 5: Proportion of News Articles Classified as Negative, Positive, or Neutral in Relevant Periods

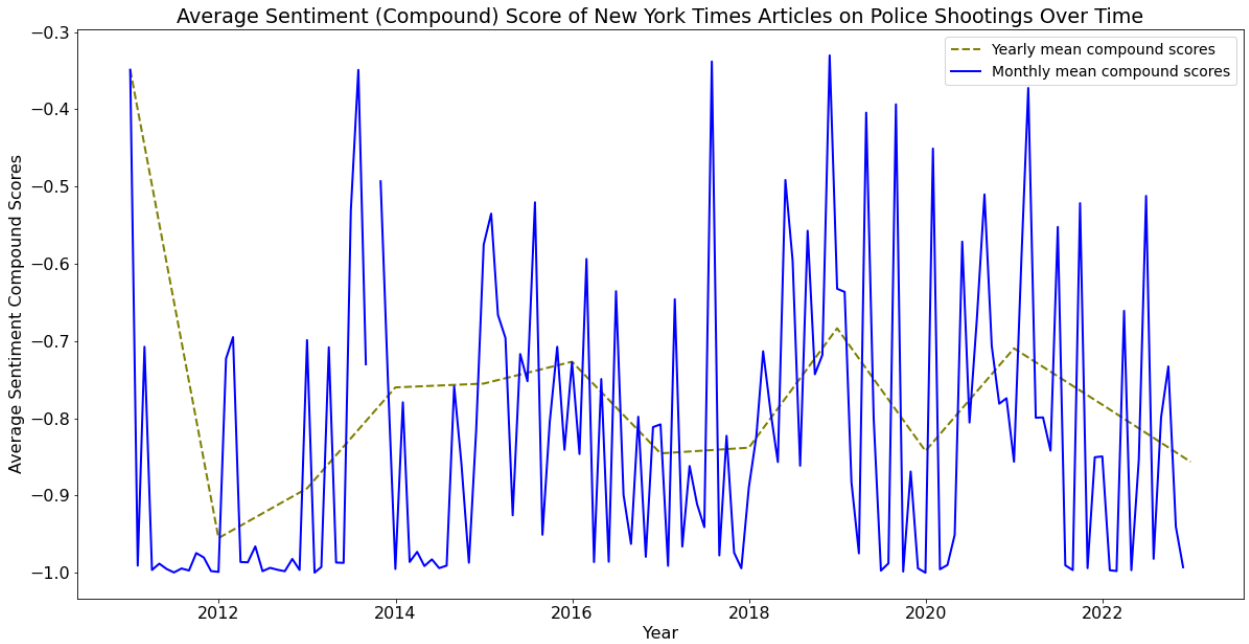

Figure 6: Average Sentiment of News Articles on Police Shootings Over Time

From these figures, it is evident that articles about police shootings in the New York Times were largely negative. However, it does not appear that the average sentiment of news articles on police shootings or the proportion of negative articles on this topic follow clear trends over time. For example, average sentiment was fairly steady between 2014 and 2016, but

decreased in 2017, and then spiked in 2019. Thus, my results suggest that there has not been a clear shift from more neutral language to more explicitly negative language since the BLM movement started in 2013. In fact, my analysis suggests that the corpus may have become less negative and more neutral over time, which contradicts my initial hypothesis. One explanation for these results could be that that the corpus includes articles about the BLM movement more generally or about protests against police brutality, both of which would have likely been portrayed positively in a liberal newspaper like the New York Times. Further, such articles would have been absent in the pre-BLM time period and would have likely become more commonplace after the George Floyd protests in 2020. Thus, even if the articles about police shootings exclusively have become more negative, we would expect to see the proportion of positive articles increasing in the post-George Floyd era, and average sentiment to be more neutral in this period due to the presence of both positive and negative articles. This would also explain why average sentiment was low in 2020, when there was widespread outrage against police brutality and language used to describe police shootings specifically was likely more negative, despite the relatively high proportion of positive articles in that year.

The results of the sentiment analysis run using the AFINN sentiment lexicon further validates my findings. Note, although similar, the annual trends from the AFINN analysis are not identical to the VADER analysis. However, such discrepancies can likely be attributed to the small size of the AFINN dictionary, which contains less than 10 percent of the tokens in my corpus. Consistent with my primary results, Figure 7 below shows that the average sentiment of articles in my corpus calculated using the AFINN lexicon was negative in all years. Further, average sentiment scores do not follow a clear trend or pattern over time, but it does appear that news articles in my corpus have become slightly less negative or more neutral since 2014.

Figure 7: AFINN Sentiment Analysis Results for Validation

Conclusion

Key Insights and Policy Implications

My research highlights the impact that the BLM movement has had on media coverage of police shootings in the U.S. Importantly, although the frequency of media coverage of police shootings has increased since the start of the BLM movement, which indirectly affects the knowledge and opinions held by the public about the issue, media attention has not been consistent during this time period. The declining media coverage of police shootings in 2022 is especially concerning, as it may suggest that police brutality has reached the "gradual decline of intense public interest" stage of the issue-attention cycle and may soon move into the "post-problem" stage where it is "replaced at the center of public concern" by other issues and is in "pro-longed limbo" (Downs, 1972). Further, despite the policy changes that were implemented in response to the widespread protests against police brutality in 2020, such as the adoption of body camera mandates in certain states/metropolitan areas (Wagner, 2021), almost 1000 people continue to be killed by police every year (Washington Post, n.d.). News organizations thus have an important role to play in slowing the decline in public interest by continuing to cover police shootings frequently and comprehensively, so that police brutality continues to be prioritized on the political agenda.

Further, news organizations must be more cognizant of the language they are using to portray police shootings. Despite calls for the media to move away from "neutral objectivity" when portraying police violence, the average sentiment of news articles has not consistently become more negative over time. Given the power of the media to influence public perception of issues relating to police brutality, it is imperative that the media begins to consciously use the active voice and explicitly negative/critical language so as to not "obscure who is responsible" (Hagle & Little, 2020) and hold the police accountable for their actions.

Limitations and Future Work

As previously discussed in the Data and Methods sections, it is important to consider the limitations associated with the data and dictionaries used for this analysis. First, the use of a single data source, the New York Times, makes it difficult to draw generalized conclusions about the media portrayal of shootings, especially as it is highly likely that non-partisan or conservative news sources would have followed dissimilar trends over time. Second, the choice of the New York Times, which is a liberal newspaper, may have underestimated any improvements in the language choices used in the portrayal of police shootings in response to the BLM movement for two reasons: 1) the number of articles in the Pre-BLM (baseline) period were relatively few and overwhelmingly negative, and 2) the New York Times likely used positive language in articles about the BLM movement or protests against police brutality, which may have been erroneously included in the corpus due to the tagging methodology used to identify relevant articles. Third, this analysis relies on existing dictionaries and lexicons which likely contain only a small portion of the tokens that appear in the corpus. Further, these lexicons were not developed for this specific domain, and thus may not be able to handle context-specific meanings of terms such as "officer-involved." Thus, I recommend that future research attempt to analyze similar research questions using a more comprehensive corpus of news articles that includes news sources with different ideologies and use customized dictionaries that are domain-specific.

Finally, there are no apparent ethical or privacy-related to issues associated with the data used in this project, because I relied exclusively on publicly available news articles. However, ethical issues may arise if the results of this analysis are misinterpreted or shared without sufficient context. For example, the New York Times may face reputational damage for depicting police shootings more positively in recent years if important contextual information

about the limitations associated with the tagging of news articles for the analysis is not communicated to the relevant audience.

Bibliography

Benhardus, J., & Kalita, J. (2013). Streaming trend detection in twitter. *International Journal of Web Based Communities*, *9*(1), 122-139.

Buchanan, L., Bui, Q., & Patel, J. K. (2021, October 25). *Black Lives Matter May Be the Largest Movement in U.S. History*. The New York Times. Retrieved October 5, 2022, from https://www.nytimes.com/interactive/2020/07/03/us/george-floyd-protests-crowd-size.html

Center for Innovation and Sustainability in Local Media. (2018, October 18). *Bigger and Bigger They Grow - Consolidation of Newspaper Ownership*. The Expanding News Desert. https://www.usnewsdeserts.com/reports/expanding-news-desert/loss-of-local-news/bigger-and-bigger-they-grow/

Demby, G. (2016, August 11). *The Butterfly Effects Of Ferguson*. NPR.org. https://www.npr.org/sections/codeswitch/2016/08/11/489494015/the-butterfly-effects-of-ferguson

Downs, A. (1972). Up and down with ecology: The issue-attention cycle. *The public*, *28*, 38-50.

Hagle, C. (2020, June 8). *Early media coverage of Breonna Taylor's killing branded her a "suspect" and sanitized police violence*. Media Matters for America. Retrieved October 5, 2022, from https://www.mediamatters.org/black-lives-matter/early-media-coverage-breonna-taylors-killing-branded-her-suspect-and-sanitized

Hagle, C., & Little, O. (2020, June 2). *Media coverage of police violence across the US sanitizes the state-sanctioned violence*. Media Matters for America. Retrieved October 5, 2022, from https://www.mediamatters.org/new-york-times/media-coverage-police-violence-across-us-sanitizes-state-sanctioned-violence

Haselmayer, M., Dingler, S. C., & Jenny, M. (2021). How Women Shape Negativity in Parliamentary Speeches—A Sentiment Analysis of Debates in the Austrian Parliament. *Parliamentary Affairs*.

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

Hutto, C. (2016). *GitHub - cjhutto/vaderSentiment: VADER Sentiment Analysis.* GitHub. https://github.com/cjhutto/vaderSentiment

Jackson, S. J., Bailey, M., & Welles, B. F. (2020, August 28). *Trayvon Martin and the Hashtag Campaign That Set the Stage for Black Lives Matter*. The MIT Press Reader. Retrieved October 5, 2022, from https://thereader.mitpress.mit.edu/trayvon-martin-hashtag-black-lives-matter-movement/

LexisNexis. (n.d.). "Death and Injuries By Police" articles from North America, 2010-2022. Nexis Uni | Academic Research Tool for Universities & Libraries. Retrieved November 5, 2022, from https://www.lexisnexis.com/en-us/professional/academic/nexis-uni.page

Mee, A., Homapour, E., Chiclana, F., & Engel, O. (2021). Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit. *Knowledge-Based Systems*, *228*, 107238.

Mehta, D. (2020, June 11). *National Media Coverage Of Black Lives Matter Had Fallen During The Trump Era — Until Now*. FiveThirtyEight. Retrieved October 5, 2022, from https://fivethirtyeight.com/features/national-media-coverage-of-black-lives-matter-had-fallen-during-the-trump-era-until-now/

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings* 93-98. 2011 May. http://arxiv.org/abs/1103.2903.

Pinto, J. P., & Murari, V. (2019). Real time sentiment analysis of political twitter data using machine learning approach. *International Research Journal of Engineering and Technology (IRJET)*, *6*(04), 4124-4129.

Soderberg, B., & Friedman, A. (2022, January 14). *Major Media Outlets Can't Stop Describing Police Violence As 'Officer-Involved' Incidents*. HuffPost. Retrieved October 5, 2022, from https://www.huffpost.com/entry/police-violence-officer-involved-analysis-lapd_n_61df310fe4b0a26702885448

Sullivan J. (2019) *Four years in a row police nationwide fatally shoot nearly 1,000 people.* The Washington Post. Retrieved from: https://www.washingtonpost.com/investigations/four-years-in-a-row-police-nationwide-fatally-shoot-nearly-1000-people/2019/02/07/0cb3b098-020f-11e9-9122-82e98f91ee6f_story.html

VerBruggen, R. (2022, March 9). *Fatal Police Shootings and Race: Evidence Review, Future Research*. Manhattan Institute. Retrieved October 5, 2022, from https://www.manhattan-institute.org/verbruggen-fatal-police-shootings

Wagner. (2021, April 30). Legislatures Require Police Body Camera Use Statewide. *National Conference of State Legislatures*. https://www.ncsl.org/research/civil-and-criminal-justice/legislatures-require-police-body-camera-use-statewide-magazine2021.aspx

Washington Post. (n.d.) *Fatal Force.* Retrieved November 5, 2022 from https://www.washingtonpost.com/graphics/investigations/police-shootings-database/

Implementation Appendix

Importing News Articles into Python

In order to import my data into a usable format in Python, I bulk downloaded the 2,615 articles identified in the *Nexis Uni* database (LexisNexis. (n.d.). as rich text files and used the "striprtf" package in Python to read in the files and parse the text from each file into a string. I then extracted the body and publication date for each file by splitting the text using the delimiters "Body," "Load-Date," and "End of Document" and saved this information for each file as a row in a Pandas dataframe.

Pre-Processing of Text

After applying tokenization and removing tokens that appear in less than 5 documents, as described in the Data section, I then applied basic pre-processing steps to my text such as removing stop words that occur in Scikit-Learn's built-in 'English' stop words list, tokens that contain only digits, and tokens containing only 1 character. I additionally tweaked the token pattern parameter using regular expressions to ensure that hyphenated words such as "officer-involved" are treated as a single token.

In order to better understand the features included in the corpus of documents and identify the additional pre-processing steps that are appropriate for my analysis, I then examined which features occur most frequently. In Table 1 below, I show the 20 most commonly occurring tokens, and their corresponding term frequencies. In Table 2, I show the 20 tokens that appear in the greatest number of documents, and their corresponding document frequencies. As seen in both tables, the two most commonly occurring tokens both in terms of term frequency and document frequency were "police" and "said," which occur in almost every document. Given that it is logical to expect that articles relating to police shootings would generally contain the word "police," including such tokens are not meaningful for my analysis. Therefore, I set a maximum document frequency parameter to exclude tokens that appear in over 95% of the documents in the corpus. Further, some of the most commonly occurring words such as officer and officers have the same meaning. Thus, I applied lemmatization to ensure that such words are considered a single token. Finally, I updated my list of stop words to include terms relating to the New York Times website such as "www," "com," "https," and "nytimes."

I use the pre-processed data to calculate term frequencies, document frequencies and TF-IDF weights, as well as for the AFINN sentiment analysis. However, for the VADER sentiment

analysis, I apply NLTK's Sentiment Intensity Analyzer to the raw text rather than the tokenized, pre-processed text, because VADER uses "word-order sensitive relationships between terms" such as punctuation, capitalization, degree modifiers, contrastive conjunctions, and preceding tri-grams, which are often lost in pre-processing and cannot be captured by a bag-of-words model (Hutto & Gilbert, 2014).

Table 1: Top 20 Most Frequently Occurring Tokens

Token	Term Frequency
said	27,307
police	26,665
mr	26,095
officers	11,859
officer	9,743
people	7,437
black	7,273
new	6,716
death	5,788
ms	5,654
city	5,155
shot	4,516
york	4,463
department	4,412
man	4,335
shooting	4,327
killed	4,178
times	3,844
family	3,348
did	3,293

Table 2: Top 20 Tokens that Appear in the Greatest Number of Documents

Token	Document Frequency
police	2,591
said	2,398
mr	2,185
death	1,999
new	1,968
officers	1,955
people	1,936
officer	1,772
times	1,748
york	1,706
killed	1,698
black	1,665
com	1,660
shot	1,659
man	1,645
nytimes	1,636
city	1,560
photograph	1,550
did	1,513
www	1,484