COP TOPICS: TOPIC MODELING-ASSISTED DISCOVERIES OF POLICE-RELATED THEMES IN AFRICAN-AMERICAN JOURNALISTIC TEXTS

A Thesis
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
MASTERS OF ARTS

---

by
Nicole Lemire Garlic
Diploma Date (August 2017)

Examining Committee Members:

Matthew Lombard, Ph.D., Advisory Chair, Klein College of Media and Communication
Nancy Morris, Ph.D., Klein College of Media and Communication
Peter Logan, Ph.D., College of Liberal Arts

ABSTRACT

The analysis of mainstream newspaper content has long been mined by communication scholars and researchers for insights into public opinion and perceptions. In recent years, scholars have been examining African-American authored periodicals to obtain similar insights. Hearkening back to the 1950s and 1960s civil rights movement in the United States, the highly-publicized killings of African-American men by police officers during the past several years have highlighted longstanding strained police-community relations. As part of its role as both a reflection of, and an advocate for, the African-American community, African-American journalistic texts contain a wealth of data about African-American public opinion about, and perceptions of, police.

In years past, media content analysts would manually sift through newspapers to divine interesting police-related themes and variables worthy of study. But, with the exponential growth of digitized texts, communication scholars are experimenting with computerized text analysis tools like topic modeling software to aid them in their content analyses. This thesis considers to what degree topic modeling software can be used at the exploratory stage of designing a content analysis study to aid in uncovering themes and variables worthy of further investigation.

Appendix A contains results of the manual exploratory content analysis. The list of topics generated by the topic modeling software may be found in Appendix B.

DEDICATION

To my faithful family and friends.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Throughout the history of the United States, there have been acute friction points between the police and members of the African-American community. Undoubtedly, some police have served the African-American community well, and not all relations between the two groups have been negative ones. Yet, stories of police brutality and the killing of African-Americans during protests, arrest, or while held in police custody, have left indelible wounds in the minds and perceptions of those who ingest them.

From the civil rights movement to today, the media has played a significant role in publicizing these police-related stories. On March 4, 1965, the story of unarmed Jimmy Lee Jackson who was shot by a state trooper after fleeing a civil rights march in Alabama was reported in the African-American newspaper, The Chicago Defender. The story painted a somber picture of "Hundreds of singing, swaying Negroes" giving "a martyr's funeral to . . . an unknown farm laborer who allegedly was shot down by a state trooper while trying to protect his mother during a racial demonstration" ("Hundreds file past bier of Jimmy Lee Jackson," p. 3). The story went on to highlight how the authorities refused to disclose the shooter's identity.[1] The expansive media attention given to the killings of unarmed African-Americans like Michael Brown, in Ferguson, Missouri, Philando Castile in Falcon Heights, Minnesota, Alton Sterling in Baton Rouge, Louisiana, Brendon Glenn, in Los Angeles, California, and Freddie Gray in Baltimore,

[1] State trooper James Fowler subsequently pled guilty to killing Jimmy Lee Jackson forty-five years after his death, in 2010 (Brown, 2010).

Maryland (Funke & Susman, 2016), are but a few recent examples of a sobering, and longstanding pattern of strained police relations.

Communication scholars have long demonstrated that stories unquestionably shape the public's understanding of traumatic events. Tracing themes that relate to police as they have been discussed by African-Americans in the public sphere through journalistic texts gives insight into the stories and concepts that shape understanding within the African-American community. In mainstream contexts, media scholars have demonstrated that newspapers reflect commonly-believed narratives (Lule, 2001; Wright, 1975). African-American newspapers have been described by historians as performing more of an advocacy function than mainstream newspapers (Dolan, Sonnett, & Johnson, 2009). Yet, there is no scholarship directly addressing 20th and 21st century African-American journalistic accounts of police.

Content analyses have been used by communication scholars to reveal narratives (or themes) embedded in journalistic accounts to encourage productive dialogue in the public sphere. As Mao and Richter (2014) explain, the "effectiveness of media content analysis is proven by the extent to which it helps uncover trends in media messages, changes in journalists' and other opinion formers' positions, and changes in community members' attitudes" (para. 1, Background of the Study). Moreover, Neuendorf (2002) explains (according to Mao & Richter, 2014) that tracing communication themes can help inform communication strategies that positively influence policy. In addition, print media informs the public about current events and frames public opinion (Mao & Richter, 2014).

The recent heightened conflicts between police and members of the African-American community call for a more nuanced understanding of how African-Americans conceptualize police. Uncovering themes in African-American newspapers can do just that. Although, in our modern times, members of the African-American community engage with these issues in social media, to locate themes originating from the civil rights movement and continuing throughout today, newspapers are a better source.

Given the longstanding nature of these police conflicts, a wide-ranging content analysis that extends from the civil rights movement and that encompasses newspapers from across the country would be ideal. But content analyses are burdensomely resource-intensive endeavors that require significant researcher time to design, conduct, and report the study. Moreover, to use resources effectively, content analysts must explore their chosen texts, early in the content analysis design phase, to locate themes in the texts that present interesting and provocative research questions.

Based on the suggestion of Jacobi, Atteveldt and Welbers (2016), who have employed topic modeling algorithms in their content analyses of journalistic texts, this thesis is a case study that tests whether using topic modeling software in the exploratory phase of content analysis design will aid researchers in locating themes (and/or attendant variables) for potential inclusion in research questions and, ultimately, the content analysis codebook. The case study proceeds by first conducting a manual exploratory analysis to uncover potential themes that could be studied in a future, full-scale content analysis. As will be illustrated below, through the process of interpreting the output of topic modeling software, I uncovered additional themes beyond those I found in the

manual exploratory analysis. Therefore, the results of this case study suggest that topic modeling is a useful software tool in the exploratory phase of content analysis design.

CHAPTER 2

LITERATURE REVIEW

Police Stories in African-American Contexts

There is a paucity of research in the communication field addressing police-related themes in African-American journalistic texts. Indeed, there is only a burgeoning body of scholarship addressing African-American periodicals generally (Gardner & Moody, 2015). Washburn's (2006) book, *The African American Newspaper: Voice of Freedom*, is an admitted exception. As Gutiérrez (2008) notes, Washburn's work is an exhaustive treatment that covers the 181-year history of African-American newspapers in the United States and details key issues covered by the papers through time, including the civil rights movement. But, the book is not a systematic content analysis of the newspapers.

In the past ten years, communication scholars have shown an interest in analyzing African-American newspapers but not in police-related areas. Bacon (2007), Burrowes (2011), Güven (2016), Helwig (2009), and Terry (2013) analyze African-American newspapers in the 19th century. There are some articles addressing 20th century African-American newspapers that note the presence of a range of non-police related issues (Carroll, 2011; Dolan et al., 2009; Everbacil, 2005; Hall, 2006; Lee, & Len-Rios, 2014; Qing & Armstrong, 2012; Ross, 2008; Theresa, 2015; Thornton & Tajima, 2014). Scholarship outside of the communication discipline, however, provides some insight into interactions between police officers and members of the African-American

community and the stories that are told about these interactions. In the Journal of Black Studies, O'Kelly (1980) published a thematic analysis of content from four large African-American newspapers, focusing specifically on editorials discussing what O'Kelly refers to as the black protest movement. Sampling 1,623 editorial pages, O'Kelly located two police-related themes: police assisting civil rights violators (like the Klu Klux Klan) who had murdered civil rights leaders, and post-rioting police brutality. Additionally, in his review of rioting and violent police-interactions represented in twentieth-century newspapers, novels, and films, Worgs (2006) locates the theme of police as representing the "White power structure" (p. 22), police corruption (referring to police who work alongside "politicians and/or the mafia" (p. 34)), and police harassment. He further points to incidents where African-American citizens enacted violence upon police or engaged in resistance to police officers' oppressive conduct.

More recent sociological scholarship highlights several police-related themes that emerge from historical and present-day interactions. These interactions include over-patrolling in urban ghettos to the point of a military-like occupation, the war on drugs, broken-window policing, and police officers as keepers of the social order (Steinmetz, Schaefer, & Henderson, 2016). These scholars also point to aggressive police searches and seizures, and investigatory stops, as recurring themes.

Other scholarship suggests that demographic factors such as gender and income status may influence African-American community members' characterizations of police encounters. According to psychologists Broman, Mavaddat, and Hsu (2000), the results of their survey of middle-class African-Americans indicated that males in this group have a higher incidence of perceiving police discrimination against them than do females. And,

survey results from Parker, Onyekwuluje, and Murty (1995) suggested that community members who live in poor, urban areas are more likely to have negative, involuntary interactions with police because police more routinely patrol blighted communities with vacant lots, abandoned buildings, and empty homes. In addition, these scholars found that "Blacks were more likely than Whites to be critical of the police; Blacks were more likely than Whites to report that the police had searched them without reason, used insulting language, and roughed them up unnecessarily" (p. 405). The scholars suggest that journalists' retelling of sensational stories of negative police interactions in the primarily African-American cities of Washington, DC and Atlanta, Georgia may also have influenced these reports.

Criminal justice scholar Tyler (2005), in examining how trust influences the degree to which members of majority and minority communities cooperate with the police, notes two types of trust that are noticeably low in African-American and other minority communities—institutional trust and motive-based trust. Institutional trust in the police refers to "beliefs about the degree to which the police are honest and care for the members of the communities they police" (p. 324) whereas motive-based trust "involves inferences about the motives and intentions of the police" (p. 325). Tyler explains that his survey research revealed that both types of trust were "strongly shaped" by community members' perception of the fairness of police policies and procedures. More specifically, community members' perception of fairness centered on the procedural components of police policy and practices, including the "neutrality of decision making, respectful and polite inter-personal treatment," and whether the police welcomed "input into decisions with whether police procedures were fairly applied" (p. 338). With this police-related

context in mind, I now provide some background on the content analysis and topic modeling software literature.

Content Analysis Methodologies

Content analysis is a central form of mass communication research, widespread in usage and utility in the media studies field (Lombard, 2002). Harold Lasswell is credited with conducting the first U.S.-based content analysis study of mass media in 1927 (MacNamara, 2005). As Thomas (1994) notes, Berelson (1952) crafted an early content analysis definition in the 1950s, describing it as "a research technique for the objective, systematic, and quantitative description of the manifest content of communication" (p. 689). By then, the use of content analysis by media scholars had skyrocketed and the technique continued to increase in popularity throughout the remainder of the 20th century.

But even before Lasswell's 1927 mass media study, others were engaged in what was then referred to as quantitative newspaper analysis (Krippendorf, 2004). As early as 1893, a researcher catalogued New York newspaper coverage, demonstrating that coverage of scandals, gossip, and sports had outpaced literary, scientific, and religious coverage. Large-scale newspaper content analyses like these were conducted throughout the early 20th century.

In more recent years, scholars have focused on concretely defining media content analysis, clarifying its methodological approaches, and making its practice more rigorous. Neuman (2014) defines content analysis as a "technique for examining the content or information and symbols contained in written documents or other communication media

(e.g., photographs, movies, song lyrics, advertisements)" (p. 49). He goes on to explain that, in conducting a content analysis, newspaper articles or other bodies of material are identified, and then a recording system is created for counting words or themes that appear in the media. He notes that content analyses contribute to social science scholarship by documenting "specific features in the content of a large amount of material that might otherwise go unnoticed" (p. 49).

Krippendorf (2004) and Thomas (1994) take issue with the notion that newspapers and other texts are containers that hold content ripe for picking by astute content analysts. As Thomas (1994) explains: "It would be simpleminded to claim that every person in a culture who sees an episode of a TV program will attend to, understand, or evaluate all characteristics of that episode in the same way" (p. 687). Rather, Thomas contends, content analyses explicate customs and beliefs collectively expressed in cultural artifacts like mass media texts. Krippendorf states it this way—mass media texts contain data from which analysts may *infer* what those texts mean to people, and what the information conveyed by the texts accomplishes.

Krippendorf (2004) adds that content analysts derive meaning from texts through making abductive inferences from the texts to answer particular research questions. In addition, because texts may only be understood in a context, Krippendorf reasons, a key component of the content analyst's work is to engage context-sensitive readings of the texts. Content analyses can be validly and reliably completed, in his view, through quantitative or qualitative methodologies.

In contrast, for Neuendorf (2010), another leading content analysis scholar, content analysis is "a summarizing, quantitative analysis of messages that relies on the

scientific method, including attention to objectivity/intersubjectivity, *a priori* design, reliability, validity, generalizability, replicability, and hypothesis testing" (p. 277). While she recognizes that media researchers employ both quantitative and qualitative methods when conducting content analyses, she argues that qualitative approaches to assessing media content are better understood as non-content analysis, humanistic approaches akin to rhetorical analysis or narratology, for example. She contends that true content analyses employ quantitative methodology modeled after scientific research. Shoemaker and Reese (1996) likewise acknowledge this qualitative/quantitative distinction, seeing the methodologies as reflective of two different scholarly traditions—behaviorist and humanist traditions. And, hearkening back to Berelson's early content analysis definition, they agree with Neuendorf (2010) that true content analysis is quantitative in nature.

Others argue that the qualitative/quantitative distinction has been overemphasized or, at least, that the types of methodologies may be employed together in some circumstances. For example, Gray and Densen (1998) argue for an "integrative" mixed-method approach that combines both. According to them, this integrative approach, rather than highlighting the divergence between the two types of research, "relies on the selection of techniques according to their suitability in tackling particular research questions" (pp. 419-20). This means that researchers who employ the integrative approach choose which methodological approaches best fit the research questions they ask. Yet others see qualitative and quantitative approaches as two complementary ends of a spectrum (McNamara, 2005).

Regardless of whether a quantitative or qualitative approach is employed, the preparatory stages (or development phase) of designing a content analysis is essentially

the same. It is in these early stages of study design that analysts must formulate their research questions and locate texts that can answer those questions (Krippendorf, 2004). Analysts must also define the units of analysis in those texts and obtain a sample of the texts. Finally, analysts must create coding categories (Krippendorf, 2004; McNamara, 2005).

But despite the similarity in process, qualitative and quantitative scholars approach these preparatory stages in distinct ways. Qualitative content analysts embark upon the development phase with an open mind, expecting to reform their research question and coding categories as their familiarity with the texts increases throughout the lifetime of the study (Krippendorf, 2004). Quantitative analysts, contrastively, see this stage through the lens of the scientific method (McNamara, 2005). For quantitative analysts, the preparatory work ensures objectivity and intersubjectivity by casting the decided-upon coding categories in stone before the actual coding—the substance of the study—takes place. Quantitative analysts refer to this beforehand-determination as *a priori* design.

Krippendorf (2004) and Neuendorf (2002) describe the coding category creation process as hermeneutic and iterative. In creating categorizations, content analysts draw upon their knowledge of scholarly literature to provide context for their interpretation of the texts. Neuendorf and Krippendorf suggest that, even for quantitative studies, content analysts delve into the pool of texts and conduct a qualitative-esque, preliminary reading of a representative sample of the texts. This is akin to the grounded theory approach, where analysts locate and identify issues/messages for future analysis by reviewing existing research in the field along with a subset of the texts that will be studied

(McNamara, 2005). Because crafting coding categories is an iterative and dynamic process that requires the researcher to familiarize him or herself with the range of categories present in a body of texts, I will refer to this portion of the content analysis design process as exploratory analysis.

Automated Content Analysis

Perhaps needless to say, manual content analysis is a labor-intensive process (Evans, 2014; Guo et al., 2016). While this is true for both qualitative and quantitative content analyses, this is particularly the case for quantitative analyses because quantitative ones are undertaken by multiple researchers and coders and tend to involve analysis of larger samples of data. They are designed in this fashion to allow for the demonstration of reliability (hence the generalizability) of the findings. Qualitative analyses, being from the humanist tradition, are completed by sole analysts who analyze smaller numbers of texts.

After the development phase of designing the content analysis, the analyst must actually embark upon the study. For quantitative content analysts, the study involves multiple activities that must be coordinated amongst the several researchers and coders. These activities include: (i) devising a coding scheme and drafting a code book to be used by multiple coders, (ii) training those coders, (iii) confirming whether variables collected are useful, (iv) determining validity and reliability, and (v) analyzing the data (Mao & Richter, 2014).

Understandably, given the time and resource-intensive nature of this type of study, media scholars have sought to automate quantitative content analyses and use

software to otherwise code media content. Indeed, Krippendorf (2004) argues that our modern revelations of communication, and the increasingly algorithmic world in which we live, require a more nuanced understanding and practice of content analysis research that fits our modern-day "large worlds of electronically available data" (p. xxi). He recognizes that text analysis software can be of assistance to content analysts, perhaps even participating "in content analysis as much as human analysts do" by carrying out the labor-intensive aspects of processing texts (p. xxi).

For large bodies of texts, content analyses often do not involve review of the entire corpus but only a sample thereof (Guo et al., 2016). Coding entire texts, referred to in the literature as a "census" (Krippendorf, 2004), is often cost and resource prohibitive. Nonetheless, even the process of coding a sample is time-consuming and resource-intensive. As Martindale and West (2002) quip: "The main problem with people is they are slow and sloppy in their work but want to be paid a lot of money" (p. 377). "Computers," they continue, "are fast, precise, and demand no wages" (p. 377).

Perceiving some promise in automated methods, communication scholars and scholars from other disciplines have increasingly gravitated toward automated content analysis software tools in recent years (Krippendorf, 2004; Lacy et al., 2015; O'Connor, Bamman, & Smith, 2015). Content analyses that utilize these tools have been referred to by communication scholars as "computer-aided text analysis" or "CATA" (Krippendorf, 2004; Neuendorf, 2011). However, the use of this term in not consistent within the discipline and, even more so, across disciplines who conduct content analyses with software tools. Some prefer the term computational text analysis (O'Connor et al., 2015), while others use "computer-assisted text analysis" (Guo et al., 2016). Yet others refer to

the automated process as "computer-assisted content analysis" (Chuang et al., 2014). Focusing on the math underlying this sort of software, DiMaggio (2015) and Lacy, Watson, Riffe, & Lovejoy (2015) refer to it as algorithmic text analysis. Finally, scholars who engage in qualitative, humanistic content analyses use "textual analysis" (Lacy et al., 2015). In this thesis, my use of "computer-aided text analysis" or "CATA" encompasses all of these various terms.

The types of CATA software utilized by communication researchers include: (a) dictionary-based software; (b) concordance-making software; (c) computational software programs that divvy texts into smaller parts (such as words or paragraphs) and perform numerical operations upon those parts; (d) software that distills two-pair word associations; (e) software programs that trace the use of a word over time (mimetic software); (f) statistically-derived programs that draw out co-occurring words; (g) software that displays connections between words as part of a physical network; and (h) clustering software that shows related words/concepts as grouped-together in a plot (Krippendorf, 2004; Leydesdorff & Nerghes, 2015; Neuendorf & Kane, 2010). This listing is not exhaustive, but illustrates the range of software tools available.

In the communication field, many have utilized dictionary-based software (Guo et al., 2016). The dictionary method is currently most widely used for computerized content analysis because it can automatically classify texts based on pre-existing keywords input by the researcher. The list of input words is the "dictionary." Often, these words are a pre-set list developed by an outside source (Krippendorf, 2004; Neuendorf, 2011).

All of the software options fit roughly into two categories—supervised and unsupervised methods (Guo et al., 2016; O'Connor et al., 2015; Van der Meer, 2016).

Supervised methods rely on human-derived and coded data to "train" a computer model that can then apply the codes to a larger number of documents (Brown, 2016; Guo et al., 2016). While the dictionary method is not typically referred to as a supervised method, it shares with that method the need for the researcher to input predetermined categories (Grimmer & Stewart, 2013; Guo et al., 2016). Unsupervised methods, in contrast, unleash software onto big data without first "training" the software to recognize user-specified words or patterns (Guo et al., 2016). This latter category of software claims to extract thematic data from large unstructured databases.

There are promises and pitfalls in automated content analysis for communication scholars (Krippendorf, 2004; McNamara, 2005; Neuendorf, 2011). As for promises, computers can statistically compute the frequency of patterns in texts more easily than humans (Martindale & West, 2002). Computers are better than humans at computational tasks and, at times, at standardization (Neuendorf, 2011). Simply put, computers are faster than humans at handling large volumes of data.

Some question whether the pitfalls of automated content analysis caution against the venture into technological territory, however (Krippendorf, 2004; Schmidt, 2012). Indeed, the number of challenges inherent in utilizing automated content analysis software is daunting. As noted above, supervised methods rely on the researcher to pre-determine categories. While this may be helpful for the researcher who has already devised a coding scheme, it is not useful when a researcher is looking to generate topics or themes from a large body of texts (Guo et al., 2016). Scholars question whether it is reasonable to ever expect computers to be able to generate themes. Humans understand words in context while computers do not (Krippendorf, 2004). More and more algorithms

are designed, with each passing year, to better approximate human understanding and comprehension. But only time will tell if software employing these algorithms will deliver on the textual analysis skills they promise.

And, even if a software program could extract themes, there are questions of validity. Did the software accurately and consistently categorize like materials (Jacobi et al., 2016)? Did the software fully take into account negation and ambiguity (Neuendorf, 2011)? These are not-so-easily-resolved validity questions the content analyst who utilizes CATA must answer.

Topic Modeling

Among the various CATA programs discussed in communication literature, topic modeling software is a recent addition. Topic modeling—also spelled as "topic modelling"—is a type of algorithmic text analysis originally developed by machine learning and computer scientists (Lacy et al., 2015; Leydesdorff & Nerghes, 2015). It has been characterized as an "unsupervised machine learning technique" (Hong & Davidson, 2010). While topic modeling software has existed for over a decade (Mohr & Bogdanov, 2013), it has gained increasing popularity in social science fields in the past several years (Guo et al., 2016; Jacobi et al., 2016).

Blei (2012) is a seminal study that brought topic modeling to the attention of many social science researchers. In this work, Blei, a computer scientist and machine learning scholar, demonstrated how a topic modeling algorithm that he developed could be used to trace "themes" in a large body of texts (corpus) over time. The corpus was 17,000 articles from the scholarly journal Science. Through use of the topic modeling

algorithm, this work discusses 100 "topics" that were found in the corpus. Topics approximate themes, as explained in more detail below.

Popular topic modeling algorithms implement *latent dirichlet allocation* (LDA), which is a probabilistic statistical model that assumes latent patterns of words in the texts exist throughout a corpus and calculates topics as probability distributions over all words in the corpus (Blei, 2012; Evans, 2014; Mohr & Bogdanov, 2013; Puschmann & Scheffler, 2016). In other words, these algorithms presume that each document within a corpus is a mixture of topics, and that the topics appear in multiple documents throughout the corpus (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009). The algorithm, then, treats the topic assignments within each document as a multinomial random variable. Finally, in addition to outputting the topics found throughout the corpus, the algorithm specifies, for each document, the topics found therein and what proportion of that document can be attributed to each of those topics. In short, "a typical topic model views each document as an unordered 'bag of words' which occurs with different frequencies. It then 'explains' the observed word frequencies in a given document in terms of a suitably weighted mixture of topical word frequencies where the weights indicate the different proportions of topics that appear in the document" (Guo et al., 2016, p. 336).

Evans (2014) succinctly describes how a basic-LDA topic model may lead a researcher to a theme:

> In the basic LDA model, any document can be described as a mixture of topics. So, for example, an LDA analysis of scientific abstracts might find one topic with the words "genetic embryo somatic dna" and another topic with the words "viral allograft antigen lupus." The analyst can then apply topic labels to indicate that one topic is focused on reproductive genetics and the other topic is focused on immunology. LDA estimates the probability that "viral"

> will be associated with "viral allograft antigen lupus" (immunology), the probability that the topic (immunology) will show up in any document, and the exact mixture of the resulting topics for each document in the corpus (e.g. 75% immunology, 25% reproductive genetics).

(p. 2). This means that once a topic modeling algorithm outputs a topic, the researcher interprets the words to determine the theme represented by those words. As part of this process, the researcher considers the proportion of the topic in the document being interpreted.

To be sure, content analysts do not uniformly refer to the categories they construct from topic modeling results as "themes." DiMaggio et al. (2013), for example, use "frame" and "voices" to describe their categories and Jacobi et al. (2016) use "issues." However, Jockers (2014)—a formidable literary scholar with programming skills who has done much to advance the propagation of topic modeling and other textual analysis software tools in the humanities—treats topic modeling-generated topics as indicative of "themes" (p. 137). Similarly, as noted by Mohr and Bogdanov (2013), Bonilla and Grimmer (2013) distill topics from newspaper stories into thematic categories. This thesis focuses on the exploratory phase of content analysis design, where conceptual decisions about which variables to include and how to define them are made (Neuendorf & Kane, 2010). Therefore, I find the term "theme" more appropriate than "frame" or "voice," and will use it here.

For content analyses, whether algorithmic text analysis is superior to pure human coding is an ongoing debate between scholars of several social science, computer, and humanist fields, including communication, machine learning, history, political science, and more (DiMaggio, 2015; Lacy et al., 2015). In many ways, this scholarly debate

mirrors the CATA debate amongst communication content analysts. Nonetheless, the topic modeling debate has some unique points worthy of mention here.

In the topic modeling debate, social scientists and computer scientists take inverted positions that reflect their assumptions about whether computers or humans are better analyzers of texts. Some social scientists insist that humans introduce error into the coding process by allowing their individual world views to influence how they interpret text (DiMaggio, 2015). These researchers see computer coding as a means for eradicating (or at least reducing) human bias in their quantitative content analyses (DiMaggio, 2015). In contrast, computer scientists who draft the algorithms that underlie the models see themselves as approximating the veritable ideal of human coding. As DiMaggio explains, they view human coding as the *sine qua non* of textual interpretation. Indeed, their algorithms are designed to mimic human neural processing.

Social scientists and computer scientists further differ in the areas of topic modeling that they would like to see improved. Social scientists are concerned with generalizability of a coding scheme that utilizes topic modeling, and with the reproducibility of results of a topic-modeling assisted study (Chuang et al., 2015). Computer scientists are, instead, concerned with increasing scale, that is, the ability to apply the results from an analysis to a larger set of data than from which the initial analysis was drawn (Chuang et al., 2014; Evans, 2014).

Like other CATA programs, topic modeling has its own specific promises and pitfalls (Grimmer & Stewart, 2013). One of the key claims scholars who utilize topic modeling make is that, as an inductive method, it has the potential to incite new areas of research and uncover formerly unrecognized trends within datasets (Jockers, 2014). In

other words, topic modeling has the potential to help unearth topics, the existence of which the researcher was previously unaware (Schmidt, 2012). This claim is rooted in the assumption that the latent space uncovered by the software is a semantically meaningful one (Chang et al., 2009). Other promises of topic modeling software focus on its ability to quickly categorize documents based on common words (Mao & Richter, 2014, citing Wimmer & Dominick, 2006), and its potential to overcome researcher selection bias, where researchers are "predisposed to pick one coding scheme over another when they manually examine a text corpus" (Chuang et al., 2014). The software may also provide researchers with a resource-effective means of obtaining a detailed overview of a corpus (Törnberg & Törnberg, 2016).

The pitfalls of topic modeling espoused in the literature are that, like any computerized coding software, it will not pick up tone, context, or subtleties (Chang et al., 2013; Grimmer & Stewart, 2013), and that topic modeling may produce different results in subsequent runs of the software with the same inputs (Chuang et al., 2014; Chuang et al., 2015; O'Connor et al., 2015). For quantitative studies, that topic modeling software does not return the same precise output of topics each time it is run creates the reproducibility problem. Jockers (2014) explains that this occurs because, in LDA topic modeling, the algorithm is designed to begin its computations at a different starting point in the corpus. Jockers dismisses this as a concern, however, noting that multiple runs of the same topic model will lead to similar results. Chuang et al. (2014) offer the alternative suggestion that content analysts use multiple LDA and non-LDA topic models and then select only those topics that can be found in the output of each model.

In addition, topic modeling is touted as unreliable because the manual content analysis standards endorsed by Krippendorf (2004) and Lombard et al. (2002) are inapplicable to it. These standards call for the use of (1) multiple human coders; (2) human inspection of individual distinctions; (3) inter-coder reliability measures; and (4) proper level of agreement amongst the coders. Chaung et al. (2014) argue that these manual content analysis standards cannot be satisfied by a sole researcher who interprets topic modeling software outputs without additional software tools that translate the standards into the automated environment. Other content analyst scholars contend that reliability is not a concern at all for computerized analyses because, with the computer as the sole coder, there is no need for the hallmark reliability test for content analyses—intercoder reliability (Guo et al., 2016; Krippendorf, 2004).

While topic modeling, with its focus on uncovering latent topics, appears to fit squarely in the unsupervised method of computer aided text analysis, there are supervised topic models as well. In supervised topic modeling, the corpus is manually coded. The themes found during the manual coding are then used to "train" software to code in the same manner as the manual coding process (Van de Meer, 2016, p. 956). In other words, the training process is designed to help the software replicate the manual coding. Once the software has been trained, it is then run on the entire corpus (Taghandiki, Zaeri, & Shirani, 2016). For unsupervised topic modeling, this training process is omitted.

Topic Modeling and the Exploratory Stage of Content Analysis

The above review of automated content analysis, and topic modeling, uses, promises, and pitfalls focused solely on the use of software tools to complete the "meat"

of a content analysis study, that is, the actual coding of texts. Hence the communication scholars' enumerated pitfalls must be seen in this light. It is only where the software is being used to complete the (typically) manual human coding in the study that the questions of reliability and replicability arise. This is true for both general reliability concerns, and inter-coder reliability. A replacement for inter-coder reliability is conceivably necessary only when the software replaces the coders.

It is understandable that the scholarship is directed towards the manual coding process. Arguably, the manual coding process is the area in which CATA would yield the most time and resource savings for content analysts. This is because the manual coding process is the most labor intensive portion of a content analysis.

Recently, however, communication scholars employing topic modeling algorithms have begun to consider its use during the exploratory stage of content analyses study design where the researcher crafts coding categories. Jacobi et al. (2016) argue that topic modeling may be particularly useful for "inductively showing which topics occur" which should, in turn, "help the researcher to create or improve the codebook by suggesting new codes and examples" (p. 102). And, Guo et al. (2016) have applied topic modeling software to social media tweets to conclude that the software elucidated a greater number of topics than had been uncovered in prior scholarly research. In short, as O'Connor et al., more simply state: topic modeling "can be used for exploratory analysis" (2015, p. 3).

Considering these scholars' recommendations, I consider whether topic modeling is a useful tool for generating coding categories or themes to be later categorized in a full content analysis. As explained above, the exploratory stage of content analysis design is a

hermeneutic and iterative process typically undertaken by the principal researcher. The construction of a theme by a researcher, even in preparation for a quantitative social science content analysis, necessarily involves qualitative methodology because the researcher must "group together words, phrases, or sentences and look for more abstract themes. Unless this were done, one could never get beyond the surface-level meaning of the text" (Martindale & West, 2002, p. 377).

To assess whether augmenting this process with topic modeling software would enhance it, I ask:

> RQ1: Does topic modeling unearth latent themes in African-American newspaper articles referencing police that I would not have uncovered through a manual exploratory content analysis?

Because African-American newspaper, police-related stories are understudied by communication scholars, it is ripe for exploration. There is no formidable body of prior content analyses or other communication scholarship from which themes may be drawn. But, as noted above, there is work by scholars in the Black studies and other fields that provide context for my interpretation of the texts. In addition, my own experiences as a member of the African-American community and basic understanding of potential issues present in police interactions can guide my uncovering and identification of themes. Engaging in this case study, I ask,

> RQ2: What potential themes could be included in a codebook created for a full content analysis of police-related discussions in African-American newspapers published between 1959 and 2015?

CHAPTER 3

METHOD

To compare the efficacy of topic modeling for elucidating themes not apparent from the manual method of theme elucidation, the method for this study was two-pronged. The first prong was a qualitative study of a representative sample of the police-related newspaper articles. For this study, the unit of analysis was each instance of a use of "police" within a newspaper article. The second prong employed topic modeling on the articles, using each newspaper article as an input into the topic modeling software. The software searched for "topics" by grouping co-located words throughout the entire set of articles (corpus). These topics are comparable to instances of a use of the term "police" within an article.

To be clear, the purpose of the two-pronged approach is not to determine if one method supersedes the other. Rather, consistent with Jacobi et al.'s (2016) assertion, the purpose is to determine if conducting the topic modeling analysis *in conjunction with* a manual analysis of a sample is a useful exercise. The topic modeling process will be useful if it aids the researcher in preparing the first draft of a codebook.

Compiling the Corpus

The corpus of newspaper articles was downloaded from the electronic library database Ethnic Newswatch in batches of 500 articles at a time. This database contained full-text, digitally-transcribed articles from African-American newspapers published

between January 1, 1959 and December 31, 2015.[2] This was a distinguishing feature of the Ethnic Newswatch database. Other library databases containing African-American newspapers did not include publications dated after the early 2000s and the articles were not digitally transcribed. Instead, the databases contained pdf copies of the articles that were not readily downloadable into .txt files that could be put through topic modeling software.

As for the criteria used to select the newspaper articles, I included in my search eight African-American newspapers that have been in operation since 1959 and continue to publish at the time of this study. In choosing the newspapers to include, I utilized a purposive sampling technique that considered readership statistics (Mao & Richter, 2014, citing MacNamara, 2006). The chosen newspapers were Philadelphia Tribune, New York Amsterdam News, Sentinel (Los Angeles), Miami Times, Sun Reporter (San Francisco), Call & Post (Cincinnati and Columbus), Chicago Defender, and Tri-State Defender (Memphis).

Prior to the 1960s, African-American newspapers as a whole touted much more significant circulation figures than in recent times (Washburn, 2006). As they have suffered the same fate as all newspapers in the increasingly digitized environment, readership has fallen significantly since its height in the 1960s. Nevertheless, a report by

[2] This data was compiled as part of a Digital Humanities scholar project during the 2015-16 academic year, in conjunction with Temple University's Digital Scholarship Center and HASTAC (Humanities, Arts, Science and Technology Alliance and Collaboratory). That project was a proof-of-concept exercise that involved downloading the corpus and performing a test topic modeling run on a subset of the documents. While this thesis utilizes the same data set, the pre-processing (discussed below), the *a priori* manual content analysis, and the full-scale topic modeling analyses is original research. This research has not been used to meet the requirements for another degree, and has not been published.

the 2008 Project for Excellence in Journalism State of the News Media Report indicates that several papers, including, the Amsterdam News saw an increase in readership (Project for Excellence in Journalism, 2008). In the 2010s, several of these newspapers had an audited six-month circulation rate (for a once-a-week cycle) of at least 5,000; New York Amsterdam News, Philadelphia Tribune, and Chicago Defender had circulation rates approximating 12,500, 19,250, and 5,500 as recently as 2014 ("African-American Newspaper Circulation Fact Sheet," 2014). Dolan et al. (2009), in their content analysis of African-American newspaper coverage of Hurricane Katrina likewise included Chicago Defender and Amsterdam News because of their prominence amongst existing papers.

In addition to considering readership statistics, I considered geographic diversity in choosing which papers to include in the corpus. The chosen newspapers range in location from the West Coast to the East, and include two Mid-Western cities and a Southern city. Finally, to ensure that no foreign bureaus were inadvertently included, I explicitly excluded non-United States cities and countries from the search.

Regarding the time period, I chose the 1959 through 2015 dates to encompass articles from both the Civil Rights Era, where there were frequent interactions with police and the African-American community, and the current Black Lives Matter movement that grew out of outrage against the killing of African-American men by police. This time frame also encompasses the 1970s and '80s community policing movement, which increased interactions by sending police into communities to develop trust after the combative years of civil unrest in the 1960s (Kappeler & Gaines, 2015). In addition, the library database did not include articles published prior to 1959.

In terms of inclusion criteria for the police-related articles, the only keyword used to search for relevant articles was "police," which resulted in 20,429 articles. While utilizing the single term "police" may have led to an over-inclusion of articles, such as those that mention police only in passing (Lacy et al., 2015), it was necessary to ensure that enough articles were obtained to more accurately test the degree to which unsupervised topic modeling can be of assistance to researchers. I also did not test the search term, as Lacy et al. (2015) suggest (citing Stryker, 2006), for concern of introducing selection bias.

The number of articles are displayed in the following list by source, and in descending order.

Philadelphia Tribune	(4467)
New York Amsterdam News	(3039)
Sentinel (Los Angeles)	(2403)
Sun Reporter (San Francisco)	(2403)
Miami Times	(2322)
Call & Post	(2003)
Tri-State Defender	(1936)
Chicago Defender	(1849)

Splicing

Before analyzing the articles, the first task was to splice them into individual .txt files. As noted, the articles were downloaded from the Ethnic Newswatch database in batches of 500 documents at a time. This was necessary because it would have been

rather time inefficient to download over 20,000 articles individually. Downloading the articles in batches, however, made it further necessary to portion the documents into individual files for both the manual coding process and for submission into the topic modeling software.

For the manual coding process, the articles were spliced through code written in R that selected out of the larger batches a document number that identified the article, the full text of the article, an html link to the article in the database, and the article's metadata. This metadata, which was mostly consistent across all sources in the database, included newspaper ethnicity, title of the article, publication (newspaper) title, volume, issue, publication year, pages, publisher, document type, and place of publication (location), among other metadata. For the topic modeling process, the articles were further spliced to remove all metadata and header information. Only the full text of the body of the article was included in each individual .txt file.

Manual Coding

To test whether topic modeling would lead me to uncover new themes that I likely would have not uncovered from conducting a fully manual content analysis, I first manually coded a subset of the newspapers, as if I were a lead researcher preparing a codebook. The subset consisted of a 153-article randomized sample. If this were a full manual content analysis, the review of the randomized sample would have eventually involved 2-3 researchers who would each read a subset of articles and jointly devise themes (Mao & Richter, 2014). Intercoder reliability tests would also be performed at this stage (Lombard, Snyder-Duch, & Campanella Bracken, 2002). However, while

intercoder reliability is the "gold standard" for content analyses, not all published content analyses report it. Lombard, Snyder-Duch, and Campanella Bracken have shown that, between the years of 1994 and 1998, only 69% of published content analyses reported some form of intercoder reliability, with most of those articles failing to specify the measure used to calculate it. Even more importantly, this thesis is engaging in an exploratory stage study only—not a full-scale content analysis.

I highlighted key words in the articles, and then formed categories from the groups of words, following Mao and Richter's (2014) inductive approach. As suggested above, this forming of categories is more of a qualitative process as it involved my human judgment (Törnberg & Törnberg, 2016). Moreover, as Lombard, Snyder-Duch, and Campanella Bracken explain, there are two types of content that are typically coded in content analysis studies—manifest (surface) content and latent content that lies beneath the surface. My analysis of the 153 articles includes both types of content. Like qualitative analysis, culling latent content from newspaper articles is necessarily a subjective process (Lombard, Snyder-Duch, & Campanella Bracken, 2002).

Based upon my reading of the articles, I created a list of themes (coding categories) that were mutually exclusive and that did not overlap (Mao & Richter, 2014, citing Lynch & Peer, 2002). I also noted, on the coding sheet, metadata about each article, including the title of the article, document type (e.g., news or editorial), year of publication, the newspaper name, and location where the article was published. As suggested above, if this were a full content analysis, I would have had independent coders confirm that the themes were exhaustive and comprehensive (Mao & Richter, 2014). But, for purposes of this study, where the research question is whether topic modeling

elucidates additional themes beyond those discovered through manual coding, my sole review of the article suffices.

Topic Modeling Process

Next, I prepared for the topic modeling process by pre-processing the entire corpus of 20,429 articles.[3] Pre-processing is often described as "cleaning" the corpus to prepare it for submission into topic modeling software (Guo et al., 2016). Cleaning may generally include "stemming" the corpus, removing "stopwords," and removing all punctuation, special characters, spaces, and numbers. Using code written in R, I "stemmed" the words in the corpus, which means that I reduced each word in the corpus to its base stem, or root, form (Guo et al., 2016). Stopwords are words that contain little thematic value and are routinely excluded from the topic modeling process. I first eliminated standard stopwords from the corpus, relying on those provided by the topicmodels program (Guo et al., 2016; Jockers, 2014). The stopwords on this list were commonly used words like "a," "an," and "the." However, after reviewing the initial results, and following Jockers's (2014) suggestion, I created a customized stopword list in order to exclude several groups of words specific to this corpus. I excluded the names of states within the United States, city names, and proper names of individuals. Excluding these additional categories led to more generally-applicable thematic topics.

---

[3] The pre-processing stage also includes tokenization of the corpus. Tokenization is the process of converting the words and sentences found in a corpus into a sequence of tokens (Galvis Carreño & Winbladh, 2009). "A token is generally a word, but could be a paragraph, a sentence, a syllable, or a phoneme" (p. 584). In this study, the token is a word. The topic modeling software automatically separated this corpus into 6,131,913 tokens.

Topic modeling software was then run on the entire corpus. I ultimately utilized the user-friendly, GUI version of MALLET—the web-based, "TopicModelingTool." MALLET, an acronym for "MAchine Learning for LanguagE Toolkit," is a topic modeling software program designed by University of Massachusetts computer scientists (Leydesdorff & Nerghes, 2015). Consistent with Evans (2014), I considered several LDA-based topic model software options before making my final choice. My initial preference was for an R-based implementation for several reasons. For one, R provides advanced visualization tools that aid in topic interpretation (Graham, Milligan & Weingart, 2013; Jockers, 2014), such as a wordcloud generator (Fellows, 2015). In addition, the code used to implement topic modeling in R is reproducible (Graham, Milligan & Weingart, 2013). However, in testing the "topicmodels" R package, and the "mallet" R package, I experienced significant lag in processing time. Another reason for choosing the R implementation would be that the researcher may customize more topic modeling parameters than in TopicModelingTool, such as the "seed" parameter, which is a random integer the software uses to determine where in the corpus to begin its computations (Awati, 2015).[4] But, as Jockers (2014) suggests, the most critical parameters to customize are the number of topics, "burn-in" (repeated runs of the program help average out variations, and "burn in" is the number of runs), and the

[4] There are additional features in the R implementation of MALLET that may prove helpful in conducting a full-scale content analysis such as testing topic significance by assessing the Dirichlet parameter for each topic to determine what proportion of the corpus has been assigned to a particular topic (Evans, 2014). Because this thesis focuses on the exploratory stage of preparing for a manual content analysis, and the exploratory stage is not the most time intensive portion of a content analysis study, it is questionable whether spending the time to assess topic significance at this early stage would be beneficial.

number of times that the software will pour over the data (iterations) (pp. 146-147). All but the "burn-in" parameter can be customized in the TopicModelingTool just as in "topicmodels" or "mallet" in R.

There is another, more all-encompassing reason for choosing TopicModelingTool over the R implementations—its easy-to-produce and examine results (Handel, 2014). Running TopicModelingTool requires pointing the mouse to the pre-processing input file of .txt documents, the location of the customized stopwords list, and a few keystrokes indicating the number of topics and iterations. With this customized information, the software produces two output files that neatly display the topics and documents in linked html files. In my testing, this output format simplified the process of reviewing articles connected to each topic generated by the software. There is another R package designed to create topic browsers, "stmCorrViz," but that package must be deployed in connection with another topic modeling package such as the Structural Topic Modeling package "stm" (Roberts, Stewart, & Tingley, 2017; Coppola, Roberts, Stewart, & Tingley, 2016). TopicModelingTool is turnkey.

For this thesis, the number of iterations chosen was 400—a number in the range recommended by Jockers (2014). I excluded from topic generation any words that had less than a .05% probability of being present throughout the corpus (Guo et al., 2016). The number of words within each topic that were displayed for my review was 30.

In choosing the number of topics to be generated, I explored topics lists of several sizes—30, 50, and 300—and ultimately settled on 50 topics. As Törnberg and Törnberg (2016) note, choosing the best number of topics is more of an art than a science, hence I qualitatively evaluated the output of the varying numbers of topics in reaching my

decision. The 30-topic topics were generally broad and a single theme was not readily apparent. For example, one topic read "black white african racial peopl race men racism polit histori racist mani panther riot color male power nation brutal countri women group struggl poor equal hispan societi social class history."[5] This topic could be interpreted as including the themes of "racism," "militant groups," or "race-based socio-economic class."

In contrast, some of the 300-topic topics were so fine-tuned that the relation to police disappeared. Topic # 13 of the 300 topic model results is an example: "problem need issu address solut deal mani concern solv peopl caus lack help number look believ real done resolv attent communiti blame face talk system understand won someth discuss agre." Some of the 300-topic topics were usable, such as "violenc communiti violent citi peopl mani neighborhood peac end anti need prevent crime involv increas youth effort take problem life anoth commit children togeth senseless rise approach high lead plagu," which could be interpreted as "youth violence prevention." But because a potential 300-topic list would be too extensive for a manual content analysis codebook, this larger set of topics was rejected. Moreover, this is consistent with Mimno, Wallach, Talley, Leenders, and McCallum's (2011) finding that more finely grained topics produced by generating a larger number of them tend to be poor in quality.

Next, I interpreted the topics and reduced them to themes (Jacobi et al., 2016; Jockers, 2014). Following Evans' (2014) mixed-method approach to interpreting topics, I assigned a one-concept label (theme) to each topic. I then excluded topics that combined

[5] Note that the topics will appear to be missing letters at the end of words because, as discussed above, the corpus was stemmed during the pre-processing stage.

multiple themes into one topic or, conversely, contained a portion of a theme that spanned multiple topics (Evans, 2014). Additionally, I excluded those topics that appeared nonsensical. Moreover, I performed internal validation by reviewing the articles most indicative of each topic to confirm that my interpretation of each topic was correct (Jacobi et al., 2016). I completed this task by utilizing the TopicModelingTool's html topic browser to locate the articles listed marked by the software as containing the specified topics. Reviewing at least three, and up to five, articles that were highly-rated per topic, this validation involved review of 150-180 articles apart from those reviewed in the manual exploratory content analysis.

Finally, I compared the themes that I derived from the randomized 153 sample manually coded sample to these topic modeling software results to determine if any additional narratives were uncovered by the topic modeling software.

CHAPTER 4

RESULTS

Exploratory Manual Content Analysis Results

The initial 153-article randomized sample yielded 17 themes, with several articles containing multiple themes. The themes found, in order of prevalence, were:

- Crime/Accident/Event Reporting (40 Articles (8 Domestic Violence Specific))
- Police Violence Against Citizen(s) (40 Articles)
- Police-Community Relations (27 Articles)
- Police Accountability Measures (18 Articles)
- Police Protecting/Not Protecting Citizens in African-American Communities or Investigating Crimes Against Them (14 Articles)
- Police Officer Bragging/Memorial (11 Articles)
- African-American Police Officers in the Workplace (11 Articles)
- Treatment of African-Americans Within Criminal Justice System, Including but Not Limited to Police (11 Articles)
- Resistance by Citizens (11 Articles)
- Discriminatory Policing/Racial Profiling/Police Harassment (10 Articles)
- Offending Police Officer's Treatment by Criminal Justice System (9 Articles)
- Citizen Violence Against Police (8 Articles)
- Police Plea to Public (2 Articles)
- Police Corruption (2 Articles)
- Police Officers Assisting Groups Hostile to African-Americans (1 Article)
- Police Training (1 Article)

- Police Competence/Incompetence (1 Article)

The articles selected ranged in dates from 1963 through 2014, and hailed from every newspaper in the corpus.

In the manual coding process, there were 10 articles that were excluded from the analysis. One article was a creative writing piece submitted by a high school student, another article made no mention of police. Two more articles referred to "police" in a manner not contemplated by the study—as "Police," the musical group, and the "internet police." Another article discussed police at airports, with no mention of African-American civilian interactions. Two articles related to immigrants, without mentioning African ones, and how they might be less inclined to report crime. Three others were excluded because they related to police office construction, EPA emission waivers, and a possible government shutdown in Washington, DC, respectively.

Crime/Accident/Event Reporting

As indicated above, the first most commonly-found theme of crime/accident/event reporting discussed police as experts in the context of reporting on crime, or utilized the police as an information source. For example, in a 1971 Sun Reporter article, the police were cited as an information source:

> The couple had an argument with unidentified neighbors which must have found them taking opposite sides.
>
> The argument resumed within in the couple's home, and of course it degenerated into that state of name calling which is a sign that logical expressions have ran their course.
>
> Mr. Ray, according to police, began to assert his masculinity in a most vulgar primitive manner, with his fists.

As with this newspaper article, several that contained themes in the category of "crime/accident/event reporting" relayed details about domestic violence-based crimes. Those articles that discussed police as experts on crime did not address a single crime or incident, but cited to police accounts of crime statistics or general trends. A 1997 New York Amsterdam News article, for example, quotes a state assemblyman as stating: "Law enforcement has identified areas where the gangs are, in some cases they're loosely organized, and in other cases, the young people are mimicking the crimes being committed by gang members." Other articles that contained this theme were not about criminal activity, but were about community events or car accidents.

Police Violence Against Citizen(s)

Tying with crime/accident/event reporting, the police violence theme encompassed violent conduct perpetrated by White and African-American police officers. One article explained:

> The St. Petersburg Times conducted a computer analysis of arrest records and determined that African-American men under the age of 35 are twice as likely to be involved in hostile police encounters than White men in the same age group.
>
> No analysis needed.
>
> Almost any African-American male over the age of 10 could have told them that with alarming accuracy and from lamentable experience.
>
> The newspaper also said the race of the arresting officers apparently played no role in the incidents of alleged abuse. In fact, analysts say African-American officers may be slightly more prone to use force on African-American suspects than White officers.

Police-Community Relations

The next most prevalent theme found in the articles was police-community relations. The articles in which this theme was found highlighted both positive and negative relations with the community. Youth programs in which police participated or sponsored were mentioned. Relief drives and other demonstrations of police as positive, contributing members of the community were also included. In addition, there was discussion of town halls and the general concept of community policing. As an example, one article discussing the Rodney King verdict acquitting several White police officers who had severely beaten King stated: "In a press statement, the Sentinels, an organization of Black Cincinnati police officers, said, the verdict 'has made it very difficult for the Black community to trust the system and police officers.' Echoing the sentiments of the National Black Police Association, the Sentinels believe a national strategy should be developed to educate the public on how to organize and fight police abuse."

Police Accountability Measures

The articles that discussed the theme of police accountability ranged in addressing the lack of accountability and the need for it, to critically analyzing attempts at accountability. One article discussed a new home rule charter that would allow the City of Philadelphia to create a Police Advisory Board. Another article noted how respected civil rights organizations, like Amnesty International, recognize the difficulty in holding police accountable for their use of violence: "last year's Amnesty report added another influential institutional voice into the mix of community activists and civil rights groups that have long complained of the difficulties of challenging police misconduct."

Police Protecting/Not Protecting Citizens in African-American Communities or Investigating Crimes Against Them

Articles addressing whether police were "protecting/not protecting citizens in African-American communities or investigating crimes against them" were found in a significant portion of the sample. One article told how police forces had failed to monitor or fully investigate criminal activity in Philadelphia's housing projects. Another blamed the police for failing to investigate a crime, reporting a murder victim's brother's lament: "She was strangled just last spring, and Jerome Hall believes that the police let the trail of the killer run cold and are preparing to declare the investigation inactive because his sister, Patricia Hall, 33, was a Black drug user."

Police Officer Bragging/Memorial

Eleven articles "bragged" on African-American police officers, characterizing them as strong community leaders and role models. Several of these were in the context of memorials to those who had died. For example, a 2003 Sentinel article reported: "A public memorial was scheduled today in Anaheim for slain Marine Corps 1st Sgt. Edward Smith, a black reserve Anaheim police officer who was mortally wounded April 4 in Iraq." The article quoted the slain officer's wife as saying that he "would do anything to help somebody" and that he was "just the best man I've ever known." The article went on to note that the officer had received the "Top Cop" award when he graduated the police academy, and that he had told his fellow police officers that he would carry his SWAT hat with him to Baghdad.

African-American Police Officers in the Workplace

There were also eleven articles addressing African-American police officers in the workplace. These articles discussed challenges that these officers face, as well as the need to increase the proportion of officers of color in departments. One 2008 article, for example, praised then-Newark, New Jersey Mayor, Cory Booker, for swearing in an "especially diverse mix of 66 new officers," comprised of "13 African American males, 23 Latino males and 17 white males. Six African American females and seven Latino females." Another 2012 article in the New York Amsterdam News discussed a discrimination suit filed by an African-American graduate from the police academy who was not hired by the New Jersey State Police, and quoted the then-New Jersey Attorney General as saying, "The NAACP has a seat at the table as we continue to review and revise ways to attract the most qualified candidates of all backgrounds to be part of the State Police." As an example of African-American police officers being discriminated against in the workplace, a 2002 Miami Times article noted that African-American officers are often not promoted and that they formed an organization to challenge the City of Miami's promotion practices.

Treatment of African-Americans Within Criminal Justice System, Including but Not Limited to Police

The next most frequent set of themes found in the articles was treatment of African-Americans within criminal justice system, including but not limited to police. This theme appeared in eleven articles. The articles containing this theme addressed how African-Americans who had been wrongfully arrested fared in court, for example.

Resistance by Citizens

Also appearing in eleven articles was the theme resistance by citizens. This theme was found in articles addressing protests, vigils, and community outrage at police conduct. Some of these resistive events were non-violent and some violent. For example, a 2001 Chicago Defender article reported that the NAACP publically called for a judge to exclude uniformed police officers from the courtroom during a criminal trial against four African-American defendants who were alleged to have murdered a police officer. According to the NAACP, the officers' presence as spectators to the proceedings compromised the fairness of the trial. As another example, a 1998 New York Amsterdam News article reported that, in the suburban town of Ossining, New York, following the killing of "a 24-year-old Black man by three white cops . . . angry Black residents participated in a Town Hall meeting organized by the Police Community Relations Council to take questions." The article further noted that the residents "marched from a local church to the site of the meeting, chanting, 'No justice, no peace.'."

In terms of violent resistance, a poignant 1992 article recounts, first-hand, the rioting that followed the announcement of the Rodney King verdict in Los Angeles:

> Suddenly I heard repeated thumps like a battering ram knocking up against a resistant object. I looked up and catty-corner across the street, a buff African-American with (it seemed like) two dozen others were trying to break down the door of the A&W Liquor Store at 43rd and Central with a long metal bar. The door and large bay windows of the store were completely encased with wrought iron bars interlaced with more wrought iron that reminded one of the pattern of jig saw puzzle. Notwithstanding the bars, the assault on the door continued. The cruisers slowed their vehicles and watched, some pulling to the curb to await the victory over the passageway. BAM, BAM went the bar, the battering continued. I looked for the police, but none were in sight or forthcoming, as it soon became obvious. I thought about the owners, relatively young Chinese couple and their relatives who helped run the store. I thought how they had tried so hard to be pleasant, sing-songing their good morning and have a

> nice day. If these vandals get in, their business is gone. BAM, BAM, the assault on the door wasn't working out too well.
>
> For one moment, one brief moment, I thought I should do something to try to stop these people, to try to stop this ridiculous onslaught, when I heard the crash of an object going through the window, leaving a gaping hole with jagged glass all around it. I immediately reconsidered my thought as, perhaps, noble but completely, impractical. These people were in no mood to listen to reason.

After describing more violent acts, the article concluded, "It was a sick, sad, sorry experience."

Discriminatory Policing/Racial Profiling/Police Harassment

This theme was found in articles addressing disproportionate enforcement of laws against African-Americans, as well as profiling or harassment based on race. A 1974 Sun Reporter article discussed the arrest of two African-American men for the shooting of a cab driver in Oakland, California. The two men, one a salesman and the other a store clerk, were allegedly victims of discriminatory policing, according to a "spokesman at the local Muslim Temple" who described the arrest as "harassment." In a more recent, 2014, article, it was noted that a convicted "cop-killer," who was spared from death row due to his mental capacity, had never claimed in all of his legal appeals that "he was the victim of any misconduct by the prosecution or the police."

Offending Police Officer's Treatment by Criminal Justice System

Nine of the articles contained a theme discussing an (allegedly) offending police officer's treatment by the criminal justice system. The officers who brutalized Rodney King are one example. Another example of this theme is an article discussing an officer, Robert Ralson, who lied about a self-inflicted wound, claiming that two African-American men shot him. Thankfully, the article noted, "based on numerous

inconsistencies, investigators quickly decided that Ralston had shot himself and lied about it." However, the article criticized the District Attorney for refusing to file criminal charges against Ralston.

Citizen Violence Against Police

Another theme of note is the citizen violence against police theme. Killings and assaults against police officers were included in this theme. For example, one article discussed a sentence being commuted for a "convicted cop-killer." Another article argued that an assault weapons ban is needed to protect citizens and police from criminals wielding these massively destructive weapons. This article gave the vivid example of an officer who was murdered by such a weapon.

Remaining Themes

The last few themes were found in only a handful of articles (Police Plea to Public, Police Corruption, Police Officers Assisting Groups Hostile to African-Americans, and Police Training). A few articles included a police plea to the public to help locate a perpetrator, to support legislative action, or to heed holiday drinking and gun fire warnings. A few other articles addressed police corruption, such as a former Miami police officer who embezzled funds from the Miami Community Police Benevolent Association while serving as the organization's president. Lastly, one article suggested that police officers had been assisting groups hostile to African-Americans, namely, the Klu Klux Klan. The complete results of the manual content analysis are included in Appendix A.

Topic Modeling Results

Turning to the topic modeling process, the results strikingly illustrate and confirm that individual articles may contain multiple themes. This is noteworthy because, in reviewing the results, some articles were highly rated in more than one topic. An article that summarized significant events in African-American history over 100 years is a prime example of this.[6] Other, less comprehensive-in-scope-articles also contained multiple themes. Moreover, because the unit of input in the TopicModelingTool is a single article, the software will not indicate which words the software pulled from each article but the topic browser will point to the article as a whole.

Among the 50 topics, there were several new themes discovered. The complete list of topics may be found in Appendix B. These topics led me to discover the following new themes:

- Media Representation of Issues Involving Police and the African-American Community
- Police Officers Working with Schools
- Police Relationships with Famous Athletes
- Police Targeting Black Consciousness Movements and Leaders
- Pro-Police Legislation that Disfavors African-Americans/Minorities
- Police Investigating and Enforcing Domestic Violence and Sexual Assault Laws
- Police as Resources (Need More Police/Effect of Layoffs)
- Police Strategy in Drug War
- Police Assistance During Times of Emergency

[6] For this reason, this article was highly-rated in several topics.

- Police Treatment of Family of Accused/Victim
- Police Interactions with Perceived Terrorists, Domestically and in African Countries
- Police "Cleaning Up" Inner City Neighborhoods

As an example of this latter theme, one article discussed the police forcing homeless persons out of the city to improve the city's appearance.

Interestingly, some new themes were discovered in passing, that is, these themes were not the focus of the topic but were otherwise present in one of the articles connected to the topic. An example is the "Police as Mediators" theme. The topic was "black peopl polit african parti panther power organ govern struggl state movement unit histori leader free right dr world liber countri system social human oppress freedom cultur self war call" (Topic 9). As noted above, this topic represents the theme "Police Targeting Black Consciousness Movements and Leaders."

New subthemes were also discovered. For example, Topic 49 pointed to articles discussing the theme "Police Mishandling/Planting of Evidence." This is a subtheme of the larger theme found in the manual analysis—"police corruption." Topic 45 read "offic file alleg claim accord lawsuit sheriff report suit complaint attorney deputi lewi charg comment incid time action letter damag case week deni settlement order howev receiv refus harass violat." The articles associated with this theme discussed activities, short of seeking criminal prosecution, of holding police officers accountable for their untoward conduct. Examples included seeking to prevent a chief of police who was proven discriminatory from obtaining a new job, warning neighbors about officers' threatening conduct, and filing civil law suits. I, therefore, interpreted this topic as "Citizens Using

Extra-Criminal Judicial Means to Hold Police Accountable," which is a subtheme of the larger theme of "Police Accountability."

Review of several topics, and their connected articles, also made clear that the "Police Violence Against Citizens" topic needed to be broken down into two subthemes: "Police Violence Against Individual Citizens" and "Police Violence Against Groups of Demonstrators." As found in the exploratory manual analysis, a number of articles addressed police violence against individuals, almost exclusively against Black men when that violence resulted in death. Compare this to Topic 6 ("march peopl protest street ralli crowd demonstr day polic peac event organ group call support gather mani activist hall stand citi want show front youth demand action saturday held parad"), which focused on violent police conduct during large-scale protests. Other topics revealed that "Police Violence Against Citizens" should be further broken down into "Police Violence Against Individuals While Being Arrested," and "Police Violence Against Individuals Being Interrogated/Held in Custody." The former included articles discussing excessive force during arrest while the latter focused on torture during interrogation.

To be clear, "Police Violence Against Groups of Demonstrators," which recalls images of water hoses being sprayed on protesters in the South during the civil rights movement, is distinguishable from the theme "Resistance by Citizens." This latter theme includes resistance that may not have devolved into violence, as illustrated by the following quote from an article associated with Topic 6: "Unions and New York City students demanded a redress to issues such as layoffs and pension cuts and tuition increases and educational cuts, respectively. Other activists protested inner-city disparities and called on Mayor Michael Bloomberg to stop strangling city schools and

Police Commissioner Ray Kelly to check out-of-control police and reel in violent police tactics."

Another subtheme found was based on the degree of violence. Topic 26 read "shoot kill shot death fire famili polic murder dead die year fatal unarm bullet old life men wit incid involv tragedi happen justic time victim killer tragic gun anoth occur." Review of its associated articles made clear that there is a distinction between police violence that leads to death (fatal) and police violence that is non-fatal.

Yet another subtheme was found, this one related to the "Discriminatory Policing" theme. That theme refers generally to articles discussing stop-and-frisk policies that were used to harass African-Americans, "driving while Black" traffic stops, and the like. Topic 16 revealed, in passing, a subtheme of African-Americans who called the police to report a crime but were mistaken by the police, instead, as the perpetrator.

Furthermore, Topic 24 made clear that the "Treatment of African-Americans within Criminal Justice System, including but not Limited to Police" topic needed to account for a particular case. That topic centered in on articles discussing Mumia Abu Jamal, an African-American journalist in Philadelphia who was (he says, wrongfully) accused and convicted of killing a police officer. There has been a longstanding movement in Philadelphia to secure his release. This topic suggests that a significant number of articles will relate to his case, so it would be useful in a full content analysis to be able to track how many articles within the larger theme are specific to his plight.

In addition, one topic led me to recognize a theme I had not noticed while conducting the *a priori* manual content analysis but, in retrospect, I now see was included therein. Topic 30 read "elect vote presid polit mayor candid state support campaign

democrat voter republican issu senat parti bush term leader endors african race governor poll public politician repres primari win seat posit." At first glance, this topic appears to encompass elections but upon review of the articles with the highest number of words, I discovered the theme of "Police Governance," that is, who is in charge of police organizations and who sets their policies. Articles connected to Topic 30 questioned whether police leadership were favorable to African-American interests, regardless of the leader's racial or ethnic identity. Several of the 153 articles discussed police governance in this way as well, but I did not pick up on that theme during my initial analysis.

While these new themes and subthemes were discovered, many of the topics produced themes that confirmed those found through the manual coding process. The following themes were confirmed: "Police Relations with Community," "Discriminatory Policing/Racial Profiling/Police Harassment," "Police Protecting/Not Protecting Citizens in African-American Communities or Investigating Crimes Against Them," "Crime/Accident/Event Reporting," "Police Officer Bragging/Memorial," "Resistance By Citizens," and "Offending Police Officer's Treatment by Criminal Justice System." Review of the topics and associated articles also confirmed, as suggested above, that "Police Accountability" is a viable topic despite there being only 1 article containing this theme in the manual coding dataset. Topic 22 is one of the topics that led me to this conclusion.

Noticeably, some of the topics produced non-police related themes from the corpus. Topic 18 states "news report inform media press call camera public authoraffili video post use releas time radio show week confer phone newspap garner wright record stori publish photo surveil statement writer accord." Upon review of the highest-rated

articles in this topic, it became clear that the topic includes articles that slightly reference police. What connects the articles is that each involve the media and the use of cameras, but the articles do not speak to how police relate to the media or camera use. For example, one article discussed discriminatory policing at a media (press) convention in the 1960s while another discussed a political cartoonists depiction of a police shooting.

Several of the topics also highlighted that the corpus contained articles discussing "police" in ways not useful for the research study. As an example, Topic 7 included articles referring to "fashion police." Topic 5 contained a highly-rated article discussing how, in the international scene, the United States believes it has the right to "police the world."

Also noteworthy is that review of the topic-associated articles exposed troubling misspellings in the corpus. As explained above, the Ethnic Newswatch database is fully digitized, in contrast to most of the newspaper databases that contain scanned copies of print articles. This means that the pre-internet era articles incorporated into the database were transformed from their original print form to computerized text, most likely through OCR (optical character recognition) software. While OCR is a useful tool, it is not a foolproof one and some errors can appear in the digitized documents. By way of example, one of the articles connected to Topic 33 states "In a July 2005 press conference, Gricar's daughter Lara called not knowing what happened to her *fattier* 'the hardest thing I've ever had to go through'." Here, "fattier" should have been rendered as "father."

Moreover, some of the errors could have been present in the original print articles. An article in Topic 8 likely contains an error of this kind. One of the topic's articles

stated: “CORE is best known for the wave of ‘freedom rides’ which began in May, 1961. These latter demonstrations eventually led to a firm anti-discrimination *police* in interstate transportation.” Here, the context strongly suggests that “police” should have been “policy.”

Finally, five topics produced by TopicModelingTool were nonsensical and review of the attendant highly-rated articles did not produce any clear theme. For example, Topic 40 contained too many references to proper names to render a useful theme: “memphi state tri counti week defend citi jr henri ford montgomeri tennesse year shelbi turner club willi marshal director dr publish time wallac morri person citizen hay high name night.”[7] And, Topic 45 confusingly stated “year day week beach water month station food price town close sever take hour sinc end orlean mani start visit roll face pay haitian old holiday local hotel anoth restaur.” These and other nonsensical topics were rejected.

As noted above, the 50 topics are displayed in Appendix B, along with the themes that I interpreted from them. The chart, further, lists comparable themes derived from the manual analysis in the right column. Nonsensical topics are marked as “[rejected].”

Police-Related Themes

Based on my synthesis of both the manual and exploratory topic modeling analyses, I would include the following themes in a codebook prepared for a full content analysis of police-related themes in African-American journalistic texts. The unit of analysis would be each instance of a use of “police” within a newspaper article.

1. Crime/Accident/Event Reporting:

[7] While many proper names were included in the stopword list, all possible names were not included and the stemming process affected the functionality of the stopword list. The interaction between the stemming process and stopword exclusion is discussed in more detail in the Discussion chapter.

(1) domestic violence crime;
(2) non-domestic violence crime;
(3) accident; or
(4) community event.

2. Police Violence Against Citizen(s):

   (1) against individual citizen
      a. fatal police violence against individuals while being arrested;
      b. non-fatal police violence against individuals while being arrested;
      c. fatal police violence against individuals being interrogated/held in custody;
      d. non-fatal police violence against individuals being interrogated/held in custody; or

   (2) against group of demonstrators
      a. fatal police violence;
      b. non-fatal police violence.

3. Police-Community Relations

4. Police Accountability:

   (1) citizen-staffed review boards;
   (2) governmental review boards; or
   (3) citizens using extra-criminal judicial means to hold police accountable.

5. Police Protecting/Not Protecting Citizens in African-American Communities or Investigating Crimes Against Them

6. Police Officer Bragging/Memorial

7. African-American Police Officers in the Workplace:

(1) challenges facing African-American police officers;
(2) challenges facing African-American applicants; or
(3) appeals for increased diversity in hiring.

8. Treatment of African-Americans within Criminal Justice System, including but not Limited to Police:

(1) Mumia Abu Jamal case; or
(2) not Mumia Abu Jamal case.

9. Resistance by Citizens

10. Discriminatory Policing/Racial Profiling/Police Harassment

(1) Victim mistaken as perpetrator; or
(2) Victim not mistaken as perpetrator

11. Offending Police Officer's Treatment by Criminal Justice System

12. Citizen Violence Against Police

13. Police Plea to Public

14. Police Corruption:

(1) police mishandling/planting of evidence;
(2) torture while in custody;
(3) embezzlement and other financial crimes; or
(4) rape.

15. Police Officers Assisting Groups Hostile to African-Americans

16. Police Training

17. Police Competence/Incompetence

18. Media Representation of Issues Involving Police and African-American Community

19. Police Officers Working with Schools

20. Police Relationships with Famous Athletes

21. Police Targeting Black Consciousness Movements and Leaders

22. Pro-Police Legislation That Disfavors African-Americans/Minorities

23. Police Investigating and Enforcing Domestic Violence And Sexual Assault Laws

24. Police as Resources (Need More Police/Effect of Layoffs)

25. Police Strategy in Drug War

26. Police Assistance During Times of Emergency

27. Police Treatment of Family of Accused/Victim

28. Police Interactions with Perceived Terrorists, Domestically and In African Countries

29. Police as Mediators

30. Police “Cleaning Up” Inner City Neighborhoods

Of these themes, eleven would not have been discovered without the aid of topic modeling. And, as discussed in more detail above, topic modeling helped locate subthemes and clarify existing themes. It also confirmed those themes discovered by the exploratory manual content analysis.

CHAPTER 5

DISCUSSION

Police-Related Themes

Through the manual exploratory content analysis and interpretation of the topics generated by the Topic Modeling Toolkit, a range of police-related themes were revealed that are worthy of potential future inclusion in full-scale content analysis studies. Many of these themes mirror those discussed in scholarship addressing relations between police officers and African-American community members, such as the themes of police brutality and discriminatory policing. Even though these themes are indicated in scholarly literature, they have not been the subject of content analysis studies. Nor have the other themes that are absent from the literature, such as police treatment of the family of an accused person or victim, or police interactions with perceived terrorists, been the subject of content analyses. This study suggests that police-related themes in African-American journalistic texts are ripe for further study.

Incorporating Topic Modeling in the Exploratory Stage

The results of this case study are consistent with Jacobi et al.'s (2016) assertion that topic modeling software is useful for uncovering themes not discovered through manual exploratory content analysis. As illustrated above, the topic modeling software elucidated twelve new themes beyond those discovered manually. As an individual researcher, there were other benefits to the use of topic modeling in this study as well. For one, the use of topic modeling confirmed themes already found in the manual analysis and helped illuminate new subthemes. This is useful for individual researchers

because it may reduce, if not eliminate, manpower time in refining themes set for inclusion in a codebook.

It is important to further note that some of the new themes and subthemes were "in passing" theme discoveries. These discoveries cannot be directly attributed to the topic modeling algorithm because they were found while reviewing articles pegged to topics altogether different from the theme I discovered. During the manual content analysis, I reviewed 153 articles and, through interpreting the topics, I reviewed an additional 150+ articles. The sheer process of reviewing these additional articles accounts for some of the new themes and subthemes.

But, the topic modeling topics themselves sparked categorizations that I had not considered and that were not indicated to me in the relevant literature. In addition, where a simple review of the words in the the topic itself did not suggest a new theme or subtheme, reviewing the highest-rated articles associated with the topic led me to consider whether there was a latent relationship between the articles and, in several instances, there was. Without those articles having been grouped together by the topic modeling algorithm, I would have not have considered or discovered the latent theme. Review of the topic modeling results also highlighted new potential search terms. These results are consistent with Törnberg and Törnberg's (2016) presumption that topic modeling provides a detailed overview of a corpus.

One of the more time-consuming aspects of using topic modeling software is the preprocessing stage—splicing, cleaning, tokenizing, and performing other preparatory tasks. Even though the newspaper texts in this study were already digitized through an OCR process, the preprocessing of these texts nonetheless took significant researcher

time to complete. Moreover, as a novice researcher who is still learning computational textual analysis tools, the learning curve was steep. Beyond reading (and comprehending) topic modeling scholarship written by computer scientists as well as social scientists and humanists, I took online courses in R through DataCamp and Coursera to better familiarize myself with the R interface in order to tailor pre-written code to the preprocessing of this corpus. This online training was in addition to the training and assistance that I received through Temple's very capable Digital Scholarship Center.

In addition, because R is an open-source software program, it is not as well-documented as some comparable commercial products (Jockers, 2014). This meant that, when I encountered a glitch or coding conundrum, I needed to seek outside assistance to resolve the challenge, either by consulting with an expert or drilling down through layers of online help forums. These outside sources always produced a workable solution, but each consultation ratchets up the time spent on merely preparing the corpus to be input into the software. Surely, the R skills and formidable knowledge of topic modeling scholarship that I developed will suit me well in future endeavors. But the significant amount of time spent, at least in part, undercuts the expediency value of topic modeling as a tool in the exploratory stage of content analysis design.

Further, the results of this study suggest that any stemming of the corpus should be done after the stopwords have been removed. In the study, proper names were included in the stopwords lists per Jockers (2014) recommendation that removing those names leads to more meaningful terms being present in the list of highly rated topic words. With TopicModelingTool, however, one cannot avoid stemming before removing the stopwords. Therefore, several of the names were not actually removed because the

stopworded-names did not match the redacted forms in the stemmed corpus. A workaround would be to use R to remove the stopwords and then stem the corpus.

In addition, stemming the corpus made the interpretive process more time consuming because I could not simply read the article text in the TopicModelingTool topic browser and take advantage of that key user-friendly feature of the software. The browser presents whatever input it is given. So, if a stemmed version of the corpus is inputted by the researcher, that is what will appear in the topic browser. A stemmed (and other preprocessed) article is difficult to comprehend. As an example, see the following excerpt of preprocessed text from an article associated with Topic 50:

> william shackleford oakdal ave welcom guest home
> gwendolyn o hay oakdal ave shackleford made obnoxi ms
> hay becam belliger shackleford sister ms hay becam involv
> argument the argument turn physic ms hay shackleford belt
> sister

Although the full forms of some word stems are apparent, *e.g.,* "involv" for "involve," to reasonably make sense of the entire article, I had to return to the original unstemmed text housed in a separate location on my computer. Without running another round of topics on a non-stemmed version of the corpus, it is not possible to say whether stemming should be avoided.

Altogether then, when the process of interpreting the themes is viewed as a whole, the topics provided a guided tour through the corpus—directing me to both poignant, previously undiscovered themes and subthemes, and perspectives for refining search terms and concepts. In so doing, it admittedly led me to some unrelated junk on the way, but not so much as to frustrate discovery of the meaningful topics. And, the value of confirming topics previously discovered in the manual exploratory content analysis

should not be underestimated as it may reduce time spent in refining the themes and variables for inclusion in a codebook.

Limitations

In terms of the themes suggested for future content analyses, a limitation of this study is that it did not incorporate metadata into the topic modeling analysis, which could have potentially lead to a better understanding of the topics (O'Connor et al., 2015). Metadata could have brought additional meaning to the data by, for example, comparing in what region the articles associated with a given topic were published, the articles' time periods, or the publication in which they appeared (Puschmann & Scheffler, 2016). However, in an exploratory study like this one, time considerations weighed against further analyzing the topic modeling data. Researchers conducting a full-scale topic-modeling assisted study should consider incorporating metadata into the analysis and interpretation of the topic modeling results. Finally, this thesis does not speak to whether topic modeling could be successfully employed in lieu of a manual content analysis, nor whether it should be utilized in the context of a full-scale content analysis study.

Conclusion

This thesis contributes to the existing literature by demonstrating that police-related themes in African-American newspapers are ripe for full content analysis study by communication scholars. Interest in these publications is growing, and that trend should continue. Methodologically, this thesis confirms Jacobi et al's (2016) assertion that topic modeling is useful in the exploratory stages of content analysis study design in

elucidating latent themes. Use of topic modeling software is not without costs, however, and novice researchers should consider the time needed to familiarize themselves with computerized text analysis software tools and R before embarking on such a project. Future research may consider whether other forms of CATA software will assist researchers in the exploratory content analysis stage. Text regression, which may categorize texts based on a set of documents trained by the researcher, is one possibility (O'Connor et al., 2015).

REFERENCES

African-American Newspaper Circulation Fact Sheet. (2015, April 27). Retrieved from http://www.journalism.org/2015/04/29/african-american-media-fact-sheet-2015/pj_2015-04-29_sotnm_african-american-media-01/

Awati, K. (2015, Sept. 29). A gentle introduction to topic modeling using R [Blog post]. Retrieved from https://eight2late.wordpress.com/2015/09/29/a-gentle-introduction-to-topic-modeling-using-r/

Bacon, J. (2007). "Acting as Freemen": Rhetoric, Race, and Reform in the Debate over Colonization in Freedom's Journal, 1827-1828. *Quarterly Journal of Speech*, 93(1), 58-83. doi:10.1080/00335630701326860

Berelson, B. (1952). *Content analysis in communication research*. New York: Hafner.

Black Press Research Collective (2017). *Archives & Online Resources*. Retrieved from http://blackpressresearchcollective.org

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993-1022.

Blei, D. M. (2012). Topic modeling and digital humanities. *Journal of Digital Humanities, 2*(1). Retrieved from http://journalofdigitalhumanities.org/2- 1/topic-modeling-and-digital-humanities-by-david-m-blei/

Bonilla, T., & Grimmer, J. (2013). Elevated threat levels and decreased expectations: How democracy handles terrorist threats. *Poetics, 41*(6), 650-669.

Brown, R. (2010, Nov. 15). 45 years later, an apology and 6 months. *The New York Times*. Retrieved from http://www.nytimes.com/2010/11/16/us/16fowler.html

Brown, S. (2016, Jan. 26). Tips for computational text analysis. Retrieved from http://matrix.berkeley.edu/research/tips-computational-text-analysis

Broman, C.L., Mavaddat, R. & Hsu, S. (2000). The experience and consequences of perceived racial discrimination: A Study of African Americans. *Journal of Black Psychology, 26*(2), pp. 165-180, http://dx.doi.org/10.1177%2F0095798400026002003

Burrowes, C.P. (2011). Caught in the crosswinds of the Atlantic. *Journalism History*, *37*(3), 130-141.

Carroll, B. (2011). This is IT! *Journalism History, 37*(3), 151-162.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). *Reading Tea Leaves: How Humans Interpret Topic Models.* Paper presented at the Neural Information Processing Systems (NIPS) Conference. Retrieved from https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models

Chuang, J., Roberts, M.E., Stewart, B.M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J. (2014). Computer-assisted content analysis: Topic models for exploring multiple subjective interpretations. Retrieved from http://scholar.princeton.edu/sites/default/files/bstewart/files/nipshpml2014.pdf

Chuang, J., Roberts, M.E., Stewart, B.M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J. (2015). TopicCheck: Interactive alignment for assessing topic model stability. Paper presented at the Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, Denver, Colorado, May 31 – June 5, 2015 (pp. 175–184).

Coppola, A., Roberts, M., Stewart, B., Tingley, D. (2016). stmCorrViz: A Tool for Structural Topic Model Visualizations. Retrieved from https://cran.r-project.org/web/packages/stmCorrViz/index.html

Dolan, M. K., Sonnett, J. H., & Johnson, K. A. (2009). Katrina coverage in Black newspapers critical of government, mainstream media. *Newspaper Research Journal*, *30*(1), 34-42.

Evans, M. (2014). A computational approach to qualitative analysis in large textual datasets. *PLOS One, 9*(2), 1-10, http://dx.doi.org/10.1371/journal.pone.0087908

Everbacli, T. (2005). Breaking baseball barriers: The 1953 -1954 Negro League and expansion of women's public roles. *American Journalism*, *22*(1), 13-33.

Fellows, I. (2015). Package 'wordcloud'. Retrieved from https://cran.r.project.org/web/packages/wordcloud/wordcloud.pdf

Funke, D. & Susman, T. (2016, July 12). From Ferguson to Baton Rouge: Death of black men and women at the hands of police. *Los Angeles Times*. Retrieved from http://www.latimes.com/nation/la-na-police-deaths-20160707-snap-htmlstory.html

Galvis Carreño, L.V. & Winbladh, K. (2013). Analysis of user comments: An approach for software requirements evolution. *2013 35th International Conference on Software Engineering (ICSE)*, 582-591, http://dx.doi.org/10.1109/ICSE.2013.6606604

Garnder, E. & Moody, J. (2015). Introduction: Black periodical studies, 106 American periodicals. *American Periodicals, 25*(2), 105-111.

Graham, S., Milligan, I., & Weingart, S. (2013). Advanced topic modeling with R. *The Historian's Macroscope* (working title under contract with Imperial College Press). Retrieved from http://themacroscope.org

Gray, J.H. & Densten, I.L. (1998). Integrating quantitative and qualitative analysis using latent and manifest variables. *Quality & Quantity*, 32, 419–431.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297, http://dx. doi.org/10.1093/pan/mps028

Gutiérrez, F.F. (2008). The African American newspaper: Voice of freedom. *Journalism and Mass Communication Quarterly, 85*(2), 438-439.

Güven, E. (2016). The image and the perception of the Turk in Freedom's Journal. *Journalism History, 41*(4), 191-199.

Guo, L., Vargo, C.J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2) 332–359. http://dx.doi.org/10.1177/1077699016639231

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In M. Lapata & H. T. Ng (Eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08* (pp. 363–371). Morristown, NJ, USA: Association for Computational Linguistics. http://dx.doi.org/10.3115/1613715.1613763

Hall, J. (2006). Aligning darkness with conspiracy theory: The discursive effects of African American interest in Gary Webb's "Dark Alliance." *Howard Journal of Communications, 17*(3), 205-222.

Handel, J.M. (2014, Aug. 20) MALLET GUI. Tools for Algorithmic Text Analysis. Retrieved from https://confluence.cornell.edu/display/map6pub/Mallet+GUI

Helwig, T. (2009). Black and White print: Cross-racial strategies of class solidarity in Mechanics' Free Press and Freedom's Journal. *American Periodicals, 19*(2), 117-135.

Hong, L. & Davison, D. (2010). Empirical study of topic modeling in Twitter. *1st Workshop on Social Media Analytics (SOMA '10), July 25, 2010*. Washington, DC, USA. Retrieved from http://snap.stanford.edu/soma2010/papers/soma2010_12.pdf

"Hundreds file past bier of Jimmy Lee Jackson." (1965, March 4), *Chicago Defender (Daily Edition)*. Retrieved from http://search.proquest.com.libproxy.temple.edu/docview/494144335?pq-origsite=summon&accountid=14270

Jacobi, C., van Atteveldt, W. & Welbers, K. (2016) Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism, 4*(1), 89-106, http://dx.doi.org/10.1080/21670811.2015.1093271

Jockers, M.L. (2014). *Text Analysis with R for Students of Literature*. Switzerland: Springer.

Kappeler, V.E. & Gaines, L.K. (2015). *Community Policing*. Taylor & Francis. Retrieved from <http://www.myilibrary.com.libproxy.temple.edu?ID=732625>

Lacy, S., Watson, B.R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly*, 92(4), 791-811. http://dx.doi.org/ 10.1177/1077699015607338

Lee, H., & Len-Ríos, M. E. (2014). Defining obesity: Second-level agenda setting attributes in Black newspapers and general audience newspapers. *Journal of Health Communication, 19*(10), 1116-1129, http:dx.doi.org/10.1080/10810730.2013.864729

Leydesdorff, L. & Nerghes, A. (2008). Co-word maps and topic modeling: A comparison using small and medium-sized corpora (n < 1000). *Journal of the Association for Information Science and Technology, 68*(4), 1024-1035.

Lombard, M., Snyder-Duch, J., & Campanella Bracken, C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4), 587-604. http://dx.doi.org/10.1093/hcr/28.4.587

MacNamara, J. (2005). Media content analysis: Its uses; benefits and best practice methodology. *Asia Pacific Public Relations Journal, 6*(1), 1–34.

Mao, Y. & Richter, S. (2014). Content analysis: Canadian newspaper coverage of homelessness. *SAGE Research Methods Cases*, http://dx.doi.org/10.4135/978144627305014526829

Martindale, C. & West, A.N. (2002). Quantitative hermeneutics: Inferring the meaning of narratives from trends in their content. In M. Louwerse, W.V. Peer (Eds.), *Thematics: Interdisciplinary Studies* (pp. 377-396). Amsterdam: John Benjamins Publishing Co.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011*

Conference on Empirical Methods in Natural Language Processing, pages 262–272, Edinburgh, Scotland, UK, July 27–31, 2011.

Mohr, J.W. and Bogdanov, P. (2013). Special issue title: Topic models and the cultural sciences. *Poetics, 41*(6). Harlow: Pearson Education Ltd.

Neuman, W.L. (2014). *Social Research Methods: Qualitative and Quantitative Approaches*. Edinburgh Gate: Pearson Education Limited.

Neuendorf, K. A. (2011). Content analysis—A methodological primer for gender research, *Sex Roles, 64,* 276–289, http://dx.doi.org/10.1007/s11199-010-9893-0

Neuendorf, K. A., & Kane, C. L. (2010). The content analysis guidebook online. Retrieved from http://academic.csuohio.edu/ kneuendorf/content

O'Kelly, C.G. (1980). Black newspapers and the Black protest movement, 1946-1972. *Phylon: The Atlanta University of Race and Culture, 41*(4), 313-324. Retrieved from http://www.jstor.org/stable/274856

Parker, K.D., Onyekwuluje, A.B., & Murty, K.S. (1995). African Americans' attitudes toward the local police: A multivariate analysis. *Journal of Black Studies, 25*(3), 396-409. Retrieved from http://www.jstor.org/stable/2784645

Project for Excellence in Journalism. (2008). *The State of the News Media 2008: An Annual Report on American Journalism*. Retrieved from http://www.stateofthemedia.org/2008/ethnic-intro/black-press/

Puschmann, C. & Scheffler, T. (2016). Topic modeling for media and communication research: A short primer. *Alexander von Humboldt, Institut für Internet und Gesellschaft, HIIG Discussion Paper Series*, Discussion Paper No. 2016-05. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2836478

Roberts, M., Stewart, B. & Tingley, D. (2017). 'stm': Estimation of the Structural Topic Model. Retrieved from https://cran.r-project.org/web/packages/stm/index.html

Ross, F. J. (2008). The Cleveland Call and Post and the election of Carl B. Stokes. *Journalism History, 33*(4), 215-223.

Schmidt, B.M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1). Retrieved from http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/

Shoemaker, P. J., & Reese, S. D. (2014). *Mediating the message in the 21st Century: A Media Sociology Perspective* (2nd ed.). New York: Routledge Taylor & Francis Group.

Shoemaker, P. J., & Reese, S. D. (1996). *Mediating the message: Theories of influences on mass media content* (2nd ed.). White Plains: Longman.

Steinmetz, K.F., Schaefer, B.P., & Henderson, H. (2016). Wicked overseers: American policing and colonialism. *Sociology of Race and Ethnicity, 3*(1), 68–81. http://dx.doi.org/ 10.1177/2332649216665639

Taghandiki, K. Zaeri, A. & Shirani, A. (2016). A supervised approach for automatic web documents topic extraction using well-known web design features. *International Journal of Modern Education and Computer Science*, *8*(11), 20-27. Retrieved from https://search-proquest-com.libproxy.temple.edu/docview/1884174211?accountid=14270

Teresa, C. (2015). "We needed a Booker T. Washington … and certainly a Jack Johnson": The Black Press, Johnson, and issues of representation, 1909–1915. *American Journalism*, *32*(1), 23-40. http:dx.doi.org/10.1080/08821127.2015.999539

Terry, T. C. (2013). Pacific appeal campaigns for Black Man's role in Civil War. *Newspaper Research Journal*, *34*(3), 22-35.

Thomas, S. (1994). Artifactual study in the analysis of culture. *Communication Research, 21*(6), 683-697. http://dx.doi.org/10.1177/009365094021006002

Thornton, M., & Tajima, A. (2014). A "model" minority: Japanese Americans as references and role models in Black newspapers, 2000–2010. *Communication & Critical/Cultural Studies, 11*(2), 139-157. http://dx.doi.org/10.1080/14791420.2014.902086

Törnberg, A. & Törnberg, P. (2016). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse & Society, 27*(4), 401–422. http://dx.doi.org/ 10.1177/0957926516634546

Tyler, T. R. (2005). Policing in black and white: Ethnic group differences in trust and confidence in the police. *Police Quarterly, 8*(3), 322-342. http://dx.doi.org/10.1177/1098611104271105

Qian, W., & Armstrong, C. L. (2012). Black newspapers focus more on community affairs stories. *Newspaper Research Journal*, *33*(4), 78-90.

Washburn, P. (2006). *The African American Newspaper: Voice of Freedom*. Evanston: Northwestern University Press.

Worgs, D.C. (2006). “Beware of the frustrated . . .”: The fantasy and reality of African American violent revolt. *Journal of Black Studies, 37*(1), 20-45. Retrieved from http://www.jstor.org/stable/40034371

Van de Meer, T. (2016). Automated content analysis and crisis communication research. *Public Relations Review, 42*, 952-961, http://dx.doi.org/10.1016/j.pubrev.2016.09.001