Noname manuscript No.
(will be inserted by the editor)

Text as Issue

Measuring Issues Preferences among Minority Groups through Ethnic Newspapers

Jae Yeon Kim

Received: date / Accepted: date

Abstract Survey research has been central to studying racial and ethnic politics in the US. However, most of these surveys were developed in the 1990s and 2000s, so they are not useful if researchers are interested in historical questions. The text-as-data approach provides a solution to this problem by turning ethnic newspaper articles into data. In this study, I present a mixed-method framework that combines a case selection strategy, content analysis, and text classification to utilize a large collection of ethnic newspaper articles for descriptive inference. As a demonstration of this framework, I apply machine learning techniques to 78,383 articles from Asian American and African American newspapers from the 1960s through the 1980s. Content analysis assesses data quality by measuring what and how human coders label the training data. Text classification demonstrates that Asian American newspapers issued linked progress articles by 110% more than African American newspapers did. By contrast, African American newspapers produced linked hurt articles by 133% more than Asian American newspapers did. The gap between the two groups widened up to 10 times when the training data were measured by the minimum rather than the maximum threshold.

Keywords Computatioanl text analysis · Content analysis · Machine learning · Racial and ethnic politics · Political communication

1 Introduction

Racial and ethnic politics in American politics has evolved based on the development of new surveys. The American National Election Studies (ANES) consists of high-quality panel data that go back to 1948 and have been central

Jae Yeon Kim
210 Barrows Hall 1950, Berkeley, CA 94720
Tel.: +1-510-646-5183
E-mail: jaeyeonkim@berkeley.edu

in investigating public opinion in American politics campbell1980american, zaller1992nature, bartels1999panel. However, when it comes to studying the politics of ethnoracial minority groups, the data have clear limitations because these groups take up only a very small portion of ANES data conway2004politics. Other prominent panel data on public opinion, such as the General Social Survey (GSS), are no exception. Racial and ethnic politics researchers have tried to overcome this data limitation by creating new surveys. To name a few, the National Black Election Study[1] was developed in 1984, the Latino National Political Survey[2] in 1989, the National Surveys of Latinos[3] in 2002, the Pilot National Asian American Political Survey[4] in 2000 and its official version[5] in 2008, and the Comparative Post-Election Survey[6] in 2008. This new stream of data has enabled a large number of research to be conducted in African American, Latino, and Asian American politics gurin1990hope, tate1993protest, dawson1994behind, fraga2011latinos, wong2011asian, mcclain2018can. Nevertheless, most of these surveys are short lived and thus not comparable to the ANES or GSS in terms of longevity. More importantly, these surveys were mostly developed in the 1990s and 2000s, so they are not useful if researchers are interested in historical questions, such as the political origins and development of minority political movements. For instance, the 1960s and 1970s were the pinnacles of minority political activism in the US. Yet, because of data limitation, the investigation of this critical period has been left to anecdotal examples munoz1989youth, $wei_a sian_1 993, joseph 2006 black, maeda 2012 rethinking, ishizuka 2016 serve, linder 2018 text.$

The text-as-data approach provides a solution to this long-standing problem in the study of racial and ethnic politics in the US. This approach expands the data infrastructure in racial and ethnic politics by turning ethnic newspaper articles into data. Ethnic newspapers have been an essential part of mobilization networks for ethnoracial minority groups. Because mainstream media did not cover minority issues, minority activists founded ethnic newspapers to develop their own political agendas and discuss their unique political issues le1992asian, dawson1994black, rodriguez1999making, dawson2001black, kannegaard2008press, harris2010barbershops. Therefore, ethnic newspapers are invaluable historical resources to fill gaps in the quantitative analysis of US racial and ethnic politics—they trace prevalent issues and how these tendencies varied between ethnoracial minority groups over time. Traditionally, content analysis was the main means to analyze newspaper articles; researchers manually collected, read, and interpreted these documents. Large-scale text analysis was impossible until modern computational techniques, such as web

[1] For more information, see https://www.icpsr.umich.edu/icpsrweb/ICPSR/series/163.

[2] For more information, see https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/6841.

[3] For more information, see https://www.pewresearch.org/topics/national-survey-of-latinos/.

[4] For more information, see https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/3832.

[5] For more information, see https://naasurvey.com/data/.

[6] For more information, see https://cmpsurvey.org/.

scraping, natural language processing, and machine learning, automated the labor-intensive data collection and analysis process [**?**,**?**].

Nevertheless, more is not always better. The advantage of large-scale computational text analysis over traditional content analysis is scale. A large sample has some nice features for hypothesis testing because it reduces the size of standard errors and establishes the law of large numbers (or the central limit theorem). However, because newspaper articles are collected through non-probability sampling, the large size of the data may also increase the degree of bias in them [685-688]meng2018statistical. For instance, not all ethnic newspapers made their records public, not to mention digitized. Because these decisions would have been influenced by the financial status of such newspapers, among other factors, the collection of these newspaper articles would have non-random missing data and would likely not be representative.

Furthermore, measuring issues that appear in ethnic newspaper articles is difficult. These issues could be conceptualized and measured at a high (meta issues) or low level (specific issues). Meta issues or broad issue frames are useful in text classification, a supervised machine learning technique. Reducing the number of classes that need to be classified increases the number of observations for each class, which is the volume of training data. Furthermore, a small number of classes simplify the coding scheme for human coders and lower their cognitive loads, thus enhancing the reliability of the training data mikhaylov2012coder. Nevertheless, these meta issues are conceptually valid only if they are closely associated with specific issues they are supposed to represent. More importantly, different thresholds could be applied to label training data. The minimum threshold indicates that at least one human coder must agree with the coding decision. The maximum threshold indicates that all human coders must agree with the coding decision. These different thresholds may influence machine learning outcomes by providing different qualities of training data. The maximum threshold produces more reliable training data because it is based on strictly consistent coding decisions among human coders.

In response to these methodological and conceptual challenges, I present a mixed-method framework that combines a case selection strategy, content analysis, and text classification. This framework helps utilize a large collection of ethnic newspaper articles for descriptive inference. The key idea is that computational text analysis techniques build upon and never replace human decisions [268]grimmer2013text. Technical guidelines exist for the automated part of the text analysis process, such as preprocessing documents, training algorithms, and evaluating their performance. Yet scholars have rarely engaged with how data collection and measurement decisions—humans in the machine learning loop—affect machine learning outcomes mikhaylov2012coder, gitelman2013raw, geiger2020garbage. The framework addresses this problem by structuring data collection and demonstrating the sensitivity of machine learning outcomes to measurement decisions.

As a demonstration of this framework, I apply machine learning techniques to 78,383 articles from Asian American and African American newspapers from the 1960s through the 1980s. I intentionally selected Asian American

and African American newspapers based on the West Coast because Asian Americans and African Americans in the region shared strong social and political networks during the period under investigation. The case selection strategy reduces alternative explanations. Meta issues among ethnoracial minority groups in the US could be divided into two categories: providing collective gains (linked progress issue) and preventing collective losses (linked hurt issue). Content analysis assesses data quality by measuring what and how human coders label the training data. Text classification demonstrates that Asian American newspapers issued linked progress articles by 110% more than African American newspapers did. By contrast, African American newspapers produced linked hurt articles by 133% more than Asian American newspapers did. The gap between the two groups widened up to 10 times when the training data were measured by the minimum rather than the maximum threshold.

This study makes several contributions. Substantially, the findings deepen the understanding of why building interracial coalitions has been difficult in the US [**?**,**?**]. Historians of US race relations have pointed out preference misalignment as a hurdle to the formation of interracial coalitions brilliant2010color, kurashige2010shifting. The study presents the first systematic evidence of the magnitude of this problem by using a large-scale computational text analysis. Although the evidence is purely descriptive, it provides new theoretical insights into the dynamics of interracial coalition building in the US. Methodologically, the study demonstrates why content analysis, ensuring the quality of training data, is vital in machine learning applications. Using less reliable data leads to not only less accurate predictions but also more extreme interpretations. Machine learning has a strong potential to expand data infrastructure in political science by making the collection and analysis of large-scale data efficient. This feature is especially attractive for a data-hungry field, such as racial and ethnic politics. However, in using this method for knowledge accumulation, acknowledging its limitation is equally crucial. This powerful method reaches its potential only if the quality of training data is not compromised. Examining data quality is essential in using the text-as-data approach for political science research to reassure the credibility of prediction results and interpretations.

2 Section title

Text with citations [**?**] and [**?**].

2.1 Subsection title

as required. Don't forget to give each section and subsection a unique label (see Sect. 2).

Paragraph headings Use paragraph headings as needed.

$$a^2 + b^2 = c^2 \tag{1}$$

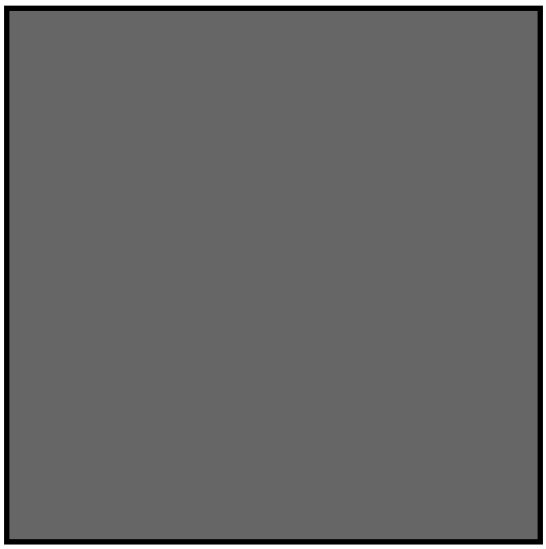

Fig. 1 Please write your figure caption here

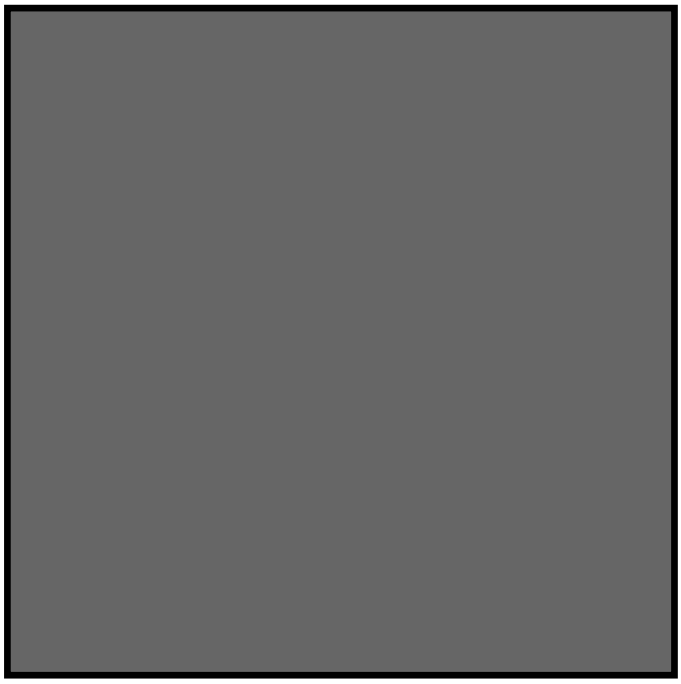

Fig. 2 Please write your figure caption here

Table 1 Please write your table caption here

first	second	third
number	number	number
number	number	number

Acknowledgements I thank Taeku Lee, Eric Schickler, Paul Pierson, Irene Bloemraad, Hakeem Jefferson, Jonathan Simon, Laura Stoker, Ruth Collier, Joel Middleton, Andrew McCall, Max Goplerud, and Christopher Stout for their comments. I am also grateful to Angela Yip, Brenna Uyeda, Gregory Eng, and Jenny Feng for their research assistance. This paper received the Don T. Nakanishi Award for Distinguished Scholarship and Service in Asian Pacific American Politics from the Western Political Science Association (2020).