Chapter 8

Big Data, Text Mining, and News Content: Where is the Big Data?

Debora Cheney
University at Albany, USA

ABSTRACT

Mining the natural language text of news content has great potential for academic researchers seeking to understand and visualize patterns and relationships buried within everyday news coverage and content. Mining news can help researchers across many disciplines understand the impact of news, biases in news coverage, and language usage. It can also help them detect unknown patterns in news coverage. However, researchers must understand the challenges of using news text for text-mining-based research. Many challenges are inherent in the news form, including the complexity of the news environment; changing patterns of news consumption and distribution; growing use of social media; and the use of visual and audio information. Additional challenges relate to determining if the news content is available in a digital format, access and license restriction on use of the news text, and how complete and completely searchable the news text really is. This chapter explores these challenges and the impact they may have on how researchers gain access to news text, and methodologies used.

INTRODUCTION

Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns. In the case of text mining, however, the data sources are documents collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections. (Feldman & Sanger, 2007, p. 1)

DOI: 10.4018/978-1-4666-8580-2.ch008
Because of its daily, and, now minute-by-minute record of events, news text is an excellent reflection of news readers and where their interests lie—culturally, socially, and politically. Mining news text has become an important methodology and focus of contemporary research methods and data analytics in academic research (Housewright, Schonfeld, & Wulfson, 2013, p. 42). It is being used for a wide range of research, including an analysis of the impact Hong Kong finance news wire stories have on stock prices (Li et al., 2014); analysis of slave advertisements in the Richmond Dispatch from 1860-1865 (Nelson, n.d.); country comparisons of historical events referred to in news coverage from 1990-2010 (Au Yeung & Jatowt, 2011); analysis of hate blogs following the Obama presidential election (Sela, Kuflik, & Mesch, 2012); and ideological differences between local and international press coverage of Kenyan elections and post-election crises (Pollak, Coesemans, Daelemans, & Lavrac, 2011).

These examples illustrate the potential variety of news sources that may be used for news text mining, the potential for analysis across many decades, the possibility of international scope, and the wide variety of potential research questions. News text is used by academic researchers in many disciplines—history, political science, journalism, linguistics, rhetoric and communications studies, the humanities, business, marketing and advertising—to extract and visualize patterns and relationships within news text.

Because of the sheer scope of big data and text mining, it is possible to lose sight of the underlying characteristics of the news text data itself and the challenges they present, yet they are important. Some text mining projects may be more viable than others because of the nature and structure of the news corpus. Text mining social media, for example, may be appropriate for understanding how groups of individuals responded to an election, candidate or issue (Mohammad, Zhu, Kiritchenko, & Martin, 2015). Other research, for example, related to Civil War newspaper content, will be affected by how and whether the appropriate news text data and the necessary years are available in a digital form for text mining. It is important for researchers to understand that many challenges exist and to decide how they will be resolved in order to ensure that the research question and outcomes are articulated appropriately within the framework and to the extent allowed by the news text data.

Researchers engaged in news text mining research will experience significant challenges in two broad areas. One, challenges related to the form and function of the news, which pertain to the different types of news reporting and intended news readership, the changing patterns of news consumption and distribution, the growing use of social media, and the use of visual and audio news forms. Second, challenges related to accessing and identifying appropriate news sources for text mining research, which pertains to gaining access to the news text data in compliance with copyright and licensing law, and ensuring the needed news text data is complete, fully searchable, and available for the years or decades required of the research question.

THE FORM AND FUNCTION OF THE NEWS: CHALLENGES

In reality, there is no such thing as “the news.” Today’s news consists of many different types of news reporting with different purposes and intended readership, and an increasing array of methods used to attract readers. Although traditionally delivered on newsprint, today’s news
Related Content

**Anthropocentric Poetry: Arts-Based Research for a New Era**
Lucinda McKnight (2019). *Educational Research in the Age of Anthropocene (pp. 234-258).*
[www.igi-global.com/chapter/anthropocentric-poetry/212479?camid=4v1a](www.igi-global.com/chapter/anthropocentric-poetry/212479?camid=4v1a)

**Sustainability: An Overview of the Triple Bottom Line and Sustainability Implementation**
Maria Salome Correia (2019). *International Journal of Strategic Engineering (pp. 29-38).*
[www.igi-global.com/article/sustainability/219322?camid=4v1a](www.igi-global.com/article/sustainability/219322?camid=4v1a)

**A Literature Review on Alkali Silica Reactivity of Concrete: Consequences and Challenges**
[www.igi-global.com/article/a-literature-review-on-alkali-silica-reactivity-of-concrete/204390?camid=4v1a](www.igi-global.com/article/a-literature-review-on-alkali-silica-reactivity-of-concrete/204390?camid=4v1a)

**Grounded Theory and Market Research**
[www.igi-global.com/chapter/grounded-theory-and-market-research/117406?camid=4v1a](www.igi-global.com/chapter/grounded-theory-and-market-research/117406?camid=4v1a)