Text Data Mining Beyond the Open Data Paradigm: Perspectives at the intersection of Intellectual Property and Ethics

Megan Senseney
University of Illinois at Urbana-Champaign, USA.
mfsense2@illinois.edu

Eleanor Dickson Koehl
University of Illinois at Urbana-Champaign, USA.
dicksone@illinois.edu

ABSTRACT

This poster highlights outcomes from an IMLS-funded National Forum project on text data mining with content that is subject to use conditions due to intellectual property rights. It argues that developing strong frameworks for conducting text mining with IP-limited data is an urgent priority for supporting responsible, sustainable research in the twenty-first century.

KEYWORDS

Text data mining, intellectual property, ethics.

INTRODUCTION

Copyright law and resource licensing complicate research with text data (Brook, Murray-Rust, & Oppenheim, 2014). Within the context of text data mining, researchers often use, or wish to use, web-based content, news media, scholarly journal articles, or large collections of digitized books. To work with these data, scholars must interpret the terms of use for publicly available content, negotiate with content providers for access through formal licensing, and operate within an ambiguous fair use framework for material that are in copyright. As authors in their own right, scholars must also make a series of complicated choices regarding the terms under which they publish and make available their own scholarship in an environment where it may also be treated as data for analysis. The existing legal and socio-technical landscape gives rise to ethically complicated situations: researchers want to use text data but lack clarity on which uses are permissible; authors want to mine journal content but may not engage in publishing practices that make their own work mineable; universities want to benefit from the use of altmetrics but, in doing so, risk compromising and commodifying scholarly production. Navigating the contours of intellectual property for text data mining raises a host of ethical concerns for scholars, both as producers and consumers. This poster articulates the ethical dimensions of text data mining in the United States with content that is anything but open.

BACKGROUND

Addressing the complicated set of intellectual property issues surrounding text data mining requires a socio-technical perspective that draws on expertise from a range of stakeholders. To map the landscape and facilitate collaborative action, a team of scholars from the University of Illinois hosted an IMLS-funded National Forum in April 2018.[1] Twenty-five leading figures selected from among researchers who use TDM techniques, librarians, legal experts, content providers, and representatives of scholarly and professional societies converged in Chicago to establish a shared understanding of issues encountered across the stakeholder groups, assess strategies and recommendations for action, and make commitments for future work. While the focus of the event was on legal and logistical strategies for working with these data, attendees were quick to assert that intellectual property issues are deeply entwined with research ethics. This poster discusses the ethical implications that emerged from the National Forum project and argues that developing strong frameworks for conducting text mining with IP-limited data is an urgent priority for conducting responsible, sustainable scholarship in the twenty-first century.

DATA AND METHOD

In advance of the National Forum, project members undertook a two-part research initiative: a literature review and a set of semi-structured interviews with participating stakeholders. Potential stakeholders were identified through the literature review and subsequent snowball sampling. Each participant agreed to prepare a forum statement and an analysis of Strengths, Weaknesses, Opportunities, and Threats (SWOT) prior to the event. Interview protocols were designed to assist participants in preparing their materials, while also providing an opportunity to speak extemporaneously and confidentially with the project team during the early phase of the project. Upon completion of all interviews, the project team reviewed notes and interview transcripts for prominent themes and then coded each interview using a set of 26 thematic codes divided into six categories. Using the codebook, the team conducted a conventional qualitative content analysis of the transcribed interviews to identify key topics

[1] Project information along with a copy of the pre-forum discussion paper and participants' statements and SWOT analyses are available at https://publish.illinois.edu/limitedaccess-tdm/

and establish cross-cutting themes and tensions identified by participants from across different stakeholder communities (Hsieh & Shannon, 2005). The findings reported below combine initial analysis with data drawn from a further round of selective coding on initial interview transcripts, SWOT analyses, and forum notes for ethics as a sensitizing concept.

FINDINGS

Participants raised ethical concerns at two levels within the academic ecosystem: the individual and the institutional. At the individual level, several participants remarked on scholars' research practices in the face of ambiguity. One participant proposed a black hat, gray hat, and white hat model for understanding users' assumptions about access and use where "gray hat" practice operates "in a space where legal and ethical compliance is uncertain." Another participant remarked that in the face of uncertainty, "some researchers do not admit to TDM or are unwilling to share their projects because of fear of being sued, leading to difficulty in reproducibility." At the institutional level, participants emphasized the need for university administrators to pay greater attention to library licensing, data governance, and intellectual privacy. Participants were at odds over whether model licenses for TDM were beneficial in providing much-needed clarity to allowable data access and use or detrimental in their tendency to restrict activities that constitute fair use. Reversing a conversation about TDM practice where researchers are situated as data consumers, one participant also advocated for more attention to understanding the role of researchers as data producers, both in terms of the research they produce and the data that are systematically collected as a byproduct of their scholarly activity in the form of altmetrics. From this perspective, universities bear an ethical obligation to develop data governance policies that better protect academic freedom in the face of digital surveillance.

DISCUSSION

Researchers who wish to utilize text data mining methods experience a chilling effect on the scholarship when faced with legal and ethical ambiguity. For those who aren't deterred entirely, maintaining the quality and validity of a research project is an uphill battle as scholars are faced with compromising on corpus creation, utilizing black box tools that introduce new levels of technical opacity, and communicating research with underspecified "snippets" and derivative data. To conduct TDM confidently and ethically, several scholars who participated in the National Forum expressed interest in compiling a best practice guide or convening "the text-analytic equivalent of an Institutional Review Board for human subjects research, where it becomes clear when a research project has followed all appropriate guidelines." At the level of local policy, university administrators must re-examine the ways they license content and how they implement data governance policies in light of the text data mining practices of scholars and the vendors who wish to profit from scholarly production. Administrators are responsible for ensuring that academic institutions meet their ethical obligations to protect the privacy and intellectual freedom of individual scholars and to serve as responsible stewards of the scholarly record. The university library is well placed to advocate for policy reform along these lines, as it is bound by a code of professional ethics to protect privacy and confidentiality, navigate the balance between the interests of information users and rights holders, and refrain from advancing private interests at the expense of the institution and its members (American Library Association, 2017).

CONCLUSION

Working with content that falls outside the open data paradigm introduces significant challenges both legal and ethical. In discussing contemporary research practice, nearly every participant at the national forum described text data mining as vital for navigating scholarly literature at scale, recovering undiscovered public knowledge, and formulating new research questions. Yet in the absence of clear institutional policies or disciplinary best practices, scholars are left to operate in a legal and ethical gray zone that stymies groundbreaking research. The current climate hampers research activity and undermines scholarly communication. While research into strategies and recommendations for library administrators is ongoing, this poster aims to foster further discussions within the ASIS&T community about the role libraries ought to play in fostering rather than inhibiting text data mining research through institutional policies, better licensing strategies, and research services that improve research with text data, from corpus creation and analysis to publication and redistribution.

REFERENCES

American Library Association. (2017, May 19). Professional ethics. Retrieved from http://www.ala.org/tools/ethics

Brook, M., Murray-Rust, P., & Oppenheim, C. (2014). The social, political and legal aspects of Text and Data Mining (TDM). *D-Lib Magazine*, *20*(11/12).

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, 15(9), 1277-1288.