

Chapter 11

Text Data and Mining Ethics



Abstract Before leaping to the critical legal and ethical issues related to text mining, it is vital to comprehend (i) the importance of data management for text mining, (ii) the lifecycle of research data, (iii) data management plan that strategizes the various data security, legal, and ethical constraints, (iv) data citation, and (v) data sharing. This chapter covers all the above-stated concepts in addition to legal and ethical issues related to text mining (such as copyright, licenses, fair use, creative commons, digital management rights), algorithm confounding, and social media research. It further presents text mining licensing conditions by selected prominent publishers and a “do’s and dont’s” list to help library professionals conduct text mining efficiently.

11.1 Text Data Management

In order to conduct research responsibly, it is vital to comprehend the role of data management to reuse, share, and store textual data. Data management is the everyday handling task of research data starting from the active phase of a project to its completion. It includes activities such as planning, documenting, and formatting of data and practices that support access, long-term preservation, and data usage. When a research project has been completed, the data could be used to answer additional questions not considered during the original project and to extend the study over time to perform longitudinal comparisons. It becomes very difficult if you have not effectively managed your data. Also, many publishers, research institutions, and funding agencies ask how the researchers will perform the task of data management during and after the completion of a research project by demanding more transparency in research projects and data management processes.

11.1.1 *Plan*

Data management plan (DMP) is a formal document that explains how the data will be generated, collected, used, managed, described, preserved, and stored for future access for a research project. It summarizes the data management strategies that will be implemented both during and after completing a research project. A DMP should include information on:

- Data description (metadata) and formats
- Handling of data (during and after the project ends)
- Data collection, processing, and generation
- If the data will be shared in open-access
- Data security and other legal or ethical constraints
- How the data will be curated and preserved

The funding agencies require grantees to submit DMP because they provide assurances that the data will be available for use for a very long time. This is important for two primary reasons:

1. It supports transparency and openness.
2. It provides greater returns on public investments in research by making the data discoverable, accessible, and reusable.

There are two open-access tools, DMPTool¹ and DMPonline,² among others for DMP, which researchers and information professionals can use to create DMPs. These tools provide guidance and templates for creating DMP in compliance with institutional and different international funder requirements.

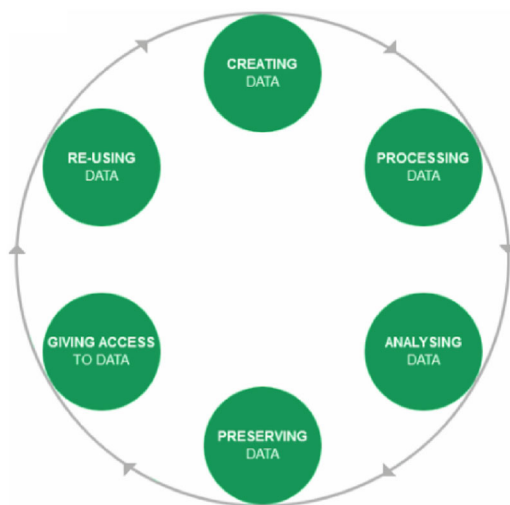
11.1.2 *Lifecycle*

A research data lifecycle represents various activities related to data occurring throughout a research process. When performing any study which deals with data, it is essential to think about various aspects related to the management of data processed and produced in the process. Research lifecycle models help from planning to archiving of the project for proper data management. Each stage produces a specific data product in the data lifecycle that requires a range of considerations, responsibilities, and activities. There are numerous data lifecycle models from simple to complex, each with a different focus or perspective. Some of the selected lifecycle models are illustrated in Figs. 11.1, 11.2 and 11.3. These plans maybe a formal document required for submission for a grant or woven into

¹ <https://dmptool.org/>.

² <https://dmponline.dcc.ac.uk/>.

Fig. 11.1 UK Data Archive Research Data Lifecycle
 (©2020 University of Essex—reprinted with the permission of UK Data Archive. <https://ukdataservice.ac.uk/manage-data/lifecycle.aspx>, Accessed 17 August 2020)



lab protocol, or maybe more informal. Regardless of whether a DMP is required or not for a research project, taking time to put it together will be beneficial down the line and save a lot of time and frustration later. In a typical research data lifecycle (Fig. 11.4), data undergo the subsequent stages of processing, analyzing, preserving, accessing, and reusing, which are identified as:

- *Stage 1: Discovery and Planning* phase helps to identify the data type and format even before commencing the data collection process. This stage may involve collecting new data, combining existing data with the new data, or analyzing the existing data. Any sensitive data such as name, phone number, location, etc., should be removed and handled ethically. Further, metadata standards, the type of documentation generated, and identification of potential users of the data should be considered and finalized in this phase. Finally, it is essential to determine the potential cost of data management, which involves formatting, documenting, storing, cleaning and anonymizing, and archiving.

For information professionals, this stage offers an opportunity to help the researchers to develop a thorough data management plan for their research project. Librarians can assist researchers with the integration of datasets from different repositories and provide hands-on training on locating data resources and tools for managing, manipulating, and viewing datasets. They can assist in data wrangling, interpreting codebooks and other documentation, and troubleshooting problems with a dataset. This is where information professionals can be of service, leveraging their knowledge of organizational systems and preservation methods by asking questions and offering suggestions, which help to ensure that the research project begins and moves forward smoothly.

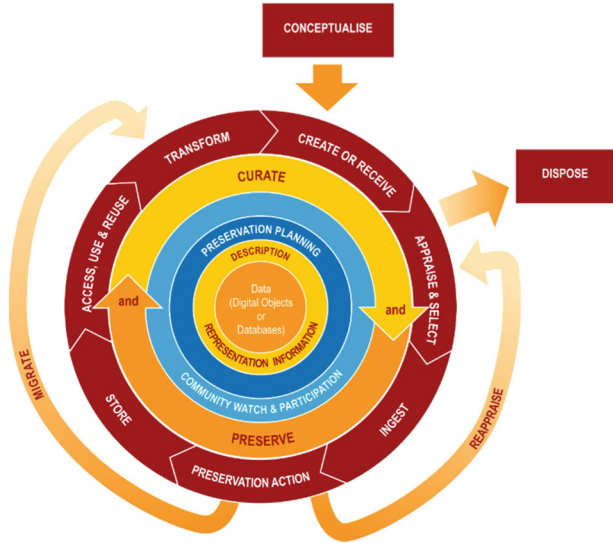


Fig. 11.2 Digital Curation Centre (DCC) Lifecycle Model (UK) (©2020 Digital Curation Centre—reprinted with the permission of Digital Curation Centre. <https://www.dcc.ac.uk/guidance/curation-lifecycle-model>. Accessed 18 Aug 2020)

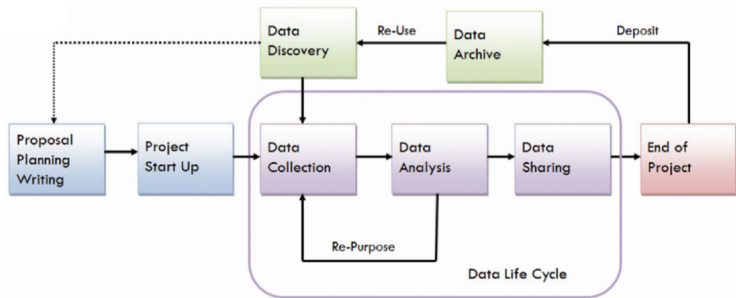


Fig. 11.3 University of Virginia Library Research Data Lifecycle (©2020 The Data Management Consulting Group, University of Virginia Library—reprinted with the permission of University of Virginia Library. <https://data.library.virginia.edu/data-management/>, August 17, 2020)

- *Stage 2: Initial Data Collection* phase ensures the implementation of the best data management practices which include backup and storage strategies, file organization, deciding on file organization schemes, and quality assurance protocols, including naming conventions and file versioning policies. Managing the data in the above way will make this phase effortless and decrease duplicates

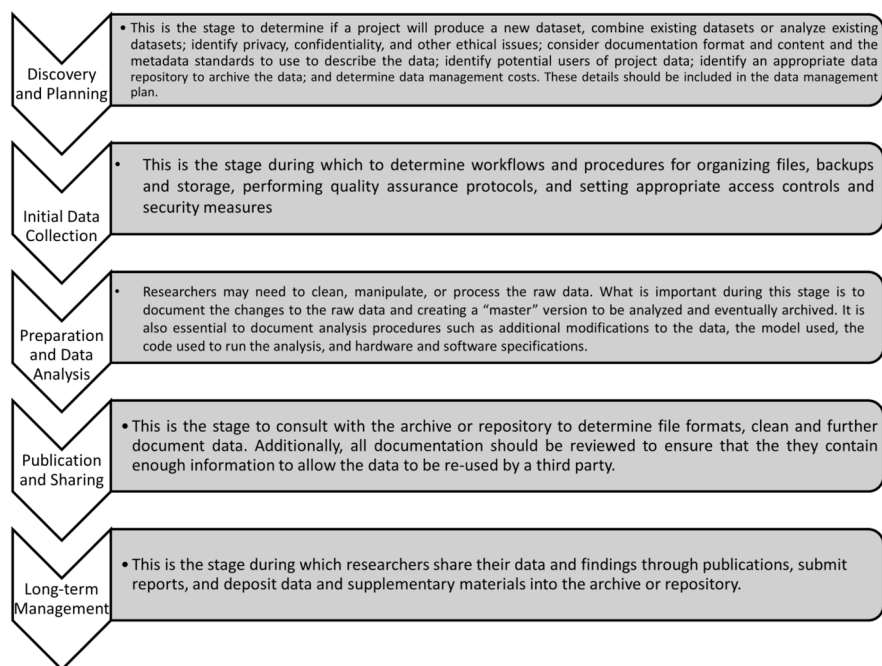


Fig. 11.4 Data management across the research lifecycle

and out-of-sync versions of the data. Proper backup and storage approaches will defend against data loss. The draft secures backup and storage protocols and quality assurance protocols. Further, one should consider data security and access controls during this phase and take appropriate steps to ensure that data is safely stored and accessible only to authorized individuals.

For information professionals, this stage offers an opportunity to advise the researchers on how the data should be organized. These include conventions for file naming, persistent unique identifiers (i.e., digital object identifiers or DOI), and versioning control. Information professionals can work with researchers to provide strategies for preventing errors from entering data and individual consultations or group training sessions on activities for both monitoring and maintenance of the quality of the data during the research project. Further, they can help the researchers to make sure that (a) the data is protected against accidental data loss by backing them up regularly; (b) appropriate data access controls are in place, mainly if researchers are dealing with sensitive data; (c) only authorized research personnel have access to the data; (d) the schedule and the responsibility for data backup are included in DMP; and (e) the data can be found and understood by others by helping the researchers with data description activities.

- *Stage 3: Preparation and Data Analysis* phase helps in cleaning, manipulating, processing, performing statistical analysis, and visualization of the data. A master version of the processed data should be analyzed and eventually be archived. The final version of the master version of processed data should be stored in read-only mode so that it cannot be inadvertently altered. All the modifications performed to the data, the model used to analyze the data, the code used to run the analysis, the acronyms and units of measurement used, and the hardware and software specifications to conduct the research should be documented meticulously for validity and reproducibility of the research process. In the case of a big data project, enough data storage, computing power, and bandwidth should be considered for processing and analysis.

For information professionals, this stage offers an opportunity to guide the researchers on how best to describe data and files. This could include teaching them how to write readme style metadata, identifying an existing metadata schema, and creating a data dictionary. They can assist by (a) offering referrals to research computing and support services; (b) providing instruction on data processing and analysis, specifically guidance on using computing language, such as R and Python, and statistical environments like Stata, SPSS, and SAS; and (c) providing assistance in data visualization and representation.

- *Stage 4: Publication and Sharing* phase helps in the preparation of data files and other research materials necessary for interpretation and usage of data in the future.
- *Stage 5: Long-Term Management* phase dictates that data should be stored and made available for sharing in a trustworthy data repository open to the public. Mendeley Data,³ Registry of Research Data Repositories,⁴ DataCite,⁵ and FAIRsharing⁶ are examples of such data repositories. The selection of an appropriate data archival repository may also depend on data type and research discipline.

One should consult with information professionals or data repository staff for guidance while preparing the data for sharing. The specific needs and requirements of the repository, the ability to reuse the data, and proper standard metadata should be applied before submitting the data in a repository. Information professionals can assist the researchers in providing data citation and persistent DOIs for their datasets.

All research projects do not necessarily have to go through all the stages of the data lifecycle or go through all the stages in a particular order. A data lifecycle is a valuable tool for understanding the typical path of research data. Researchers may contact information professionals for their help in (a) locating potentially useful

³ <https://data.mendeley.com/>.

⁴ <https://www.re3data.org/>.

⁵ <https://datacite.org/>.

⁶ <https://fairsharing.org/>.

datasets created and preserved by others, (b) determining if the data submitted by others are usable, and (c) determining if datasets submitted by others are appropriate for reuse. Information professionals can provide training to researchers in areas such as searching data archives, understanding other's metadata, and identifying and using tools for secondary data analysis.

11.1.3 Citation

Data management helps to provide a standardized method for secondary users to cite data and can also be used by data producers to cite their own data as standalone research products. In 2014, a group called **FORCE11**⁷ issued a joint declaration that enumerated “eight data citation principles viz. *importance, credit and attribution, evidence, unique identification, access, persistence, specificity and verifiability*, and *interoperability and flexibility* that cover the purpose, function, and attributes of citations and recognize the dual necessity of creating citation practices that are both human and machine understandable” [1]. Furthermore, the Registry of Research Data Repositories⁸ helps researchers, funding organizations, libraries, and publishers to check whether a data repository chosen by them supports the creation of unique data citations that embody the joint declaration of data citation principles.

Why Cite Data?

By providing data citation to your work, you make it easier for others to identify and acknowledge your data. Data citation promotes the reproduction of research results, tracks the use and impact of data, and gives a structure that recognizes and rewards the data creator.

Citation Elements

DataCite,⁹ an international standard body founded in 2009, works with data repositories to assign persistent identifiers such as digital object identifiers (DOIs) to data. DOIs (i) are effective methods of data citation, discovery, and access, (ii) ensure that data can be discovered online regardless of where they are located, and (iii) provide methods for identification that is machine actionable as well as provide stable access to the data resource using a persistent identifier. It recommends the following minimum citation elements:

- *Creator (Publication Year). Title. Publisher. Identifier.*
- Two additional properties (Version and Resource Type) may also be added: *Creator (Publication Year). Title. Version. Publisher. Resource Type. Identifier.*

⁷ <https://www.force11.org/datacitationprinciples>.

⁸ <https://www.re3data.org/>.

⁹ <https://datacite.org/cite-your-data.html>.

11.1.4 *Sharing*

The primary key players in data sharing are data creators/producers (by making the data available), secondary users, and data repositories (by enhancing the discovery and reuse of data by creating a formal data citation). It helps in transparency, openness, and maximum funders' return on investment. It helps the research community to:

- Reinforce open scientific inquiry
- Support the replication and verification of original results
- Promote new research
- Test alternative or new methods
- Encourage collaboration and multiple perspectives
- Provide necessary teaching resources
- Avoid efforts in duplicate data collection
- Protect against fraudulent or faulty data
- Enhance overall impact and visibility of research projects
- Preserve data for future use

Some of the significant challenges associated with data sharing are:

- It takes time and efforts to make the data shareable.
- Perceived risks from the loss of control of data.
- Data containing sensitive or confidential information need to be de-identified which may affect the secondary usability of secondary analysis.
- Ownership of the data may be unclear or problematic.
- Lack of incentives for sharing data.
- Lack of experience and knowledge in data management.

11.1.5 *Need of Data Management for Text Mining*

Even though data is generated exponentially, only a small amount of data is known, let alone published and accessible practically. The accessibility and reuse strategies of textual data by humans and machines can differ, especially in the case of text mining. Textual data is generally transformed into structured data with the help of NLP, statistical processing of data, advanced pattern recognition, clustering techniques, machine learning, etc., for further analysis. Therefore, textual data is needed to be provided in the *right* format with the *right* metadata for the machine to understand it and process it further.

Data creation is both a time-consuming and expensive process that includes steps such as data collection and curation, metadata addition, annotation, maintenance, preservation, and legal clearance. In many scientific processes, data can be additionally reused for secondary value services which were not thought of initially when the project was started. Hence, appropriate tools and mechanisms (technical,

scientific, legal, social, and organizational) are required to allow efficient access, sharing, reusing, and re-purposing of textual data.

Thus, there are numerous reasons why data management is important, which include:

- Discovery and interpretation of data by other researchers
- Sustaining the value of data by enabling others to verify and build upon the published results
- Facilitating long-term preservation and access to datasets

11.1.6 Benefits of Data Management for Text Mining

1. *For researchers/data provider:* A complete and efficient DMP allows researchers to discover, document, secure, maintain, preserve, and deploy powerful computational facilities; allows interoperability and lawful sharing; provides recognition in the form of ownership, citation, and publicity; provides an added value by allowing reuse and new modes of re-purposing of data; and allows new collaboration. Thus, data management helps researchers to:

- Do better research
- Optimize the use of data during research
- Collaborate with other researchers
- Verify or refine the published results
- Reduce scientific fraud
- Promote the development of new research questions
- Provide resources for training new researchers
- Discourage unintended redundancy
- Ensure that data is preserved for future researchers to discover, interpret, and reuse
- Sustain the value of data

2. *For data users:* Anyone—researchers, private companies, the general public, scientists—can be a data user and can benefit by (i) accessing a substantial amount of data, tools, and technologies; (ii) quickly identifying and accessing the data; (iii) having a permanent persistence record of data in a repository for future use and preservation; (iv) having clear licensing and terms of use; and (v) creating new collaborations.

11.1.7 Ethical and Legal Rules Related to Text Data

Advancements in storage capacity and computing power lead to the exponential growth of textual data, where data collection progressed from paper-based to

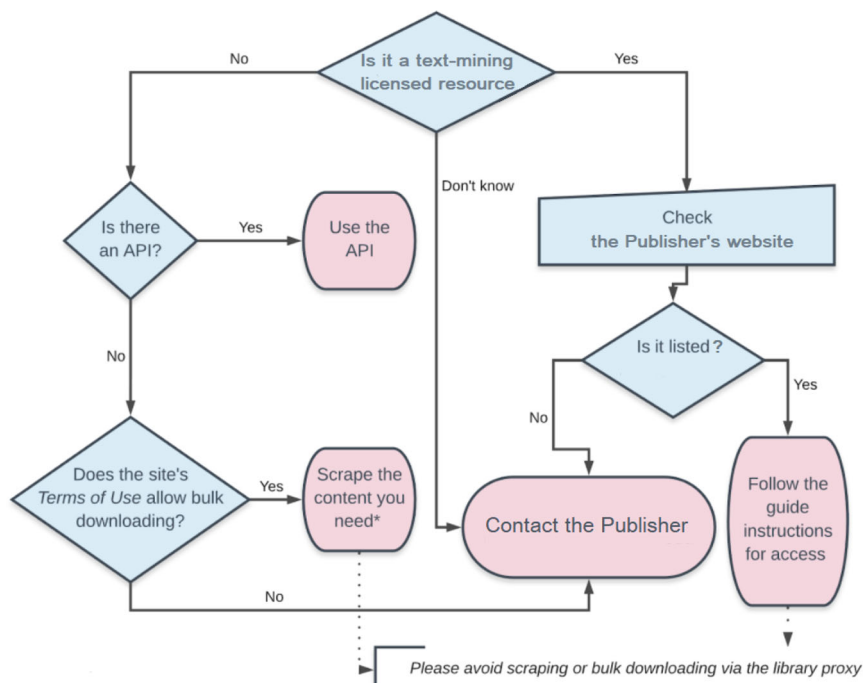


Fig. 11.5 Ethical framework for text mining of databases

digital records. This generation of data leads to new techniques and technologies to transform data to useful knowledge and information. Thus, large databases present some critical ethical challenges that needed to be considered during text mining (Fig. 11.5). Some of the prominent rules to make the data discoverable, interoperable, and reusable not only to humans but also to machines are summarized below.

11.1.7.1 FAIR Data Principles

In order to reuse data, one should comply with FAIR (findable, accessible, interoperable, and reusable) principles, which were first proposed by Wilkinson et al. [3] in 2016 and were adopted by the EU.¹⁰ These principles emphasize the capability of machines to automatically find and use the data, where in this context data is not only data in the conventional sense but algorithms, tools, and workflow. Figure 11.6 summarizes the FAIR principles for research data and should be complied by the researchers and librarians while performing text mining.

¹⁰ <https://ec.europa.eu/research/participants/docs>.

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

Fig. 11.6 FAIR guiding principles for research data (©2016 Springer Nature, all rights reserved—reprinted under Creative Commons CC BY license, published in Wilkinson et al. [3])

11.1.7.2 Creative Commons

Creative Commons(CC)¹¹ is a US-based nonprofit organization that is the leader in developing legal tools for sharing creative works over the Internet since the 2000s and is the far most widespread open content licensing model. DataCite¹² recommends using the CC0 license for the data. CC helps to achieve (i) robust legal code, (ii) human-readable summary that is understandable at a glance, (iii) machine-readable layer of code that can help make resources interoperable, and (iv) both copyright and database laws (CC version 4.0). Creative Commons [4] classifies CC licenses into six major categories:

1. **CC BY (Attribution):** It allows others to distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original work. It is the most accommodating license offered and allows maximum dissemination and use of the licensed work.
2. **CC BY-SA (Attribution-ShareAlike):** It lets others remix, tweak, and build upon your work even for commercial purposes, as long as they credit you and license their new creations under identical terms. This license is often compared with *copyleft* free and open-source software licenses.
3. **CC BY-ND (Attribution-NoDerivs):** It allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit to you.

¹¹ <https://creativecommons.org/>.

¹² <https://datacite.org/cite-your-data.html>.

Additionally, it does not permit adaptations of the work, which may lead to significant problems with the combination of different contents.

4. **CC BY-NC (Attribution-NonCommercial)**: It allows others to remix, tweak, and build upon your work non-commercially. Further, their new works must also be acknowledged by you and be non-commercial but do not have to license the derivative works on the same terms.
5. **CC BY BY-NC-SA (Attribution-NonCommercial-ShareAlike)**: It lets others remix, tweak, and build upon your work non-commercially, as long as they credit you and license their new work under identical terms.
6. **CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)**: It is the most restrictive of all the CC licenses and only allows others to download your work and share them as long as they credit you, but they cannot change the work in any way or use it commercially.
7. **CC0 (No Copyright: Public Domain)**: It lets anyone copy, modify, distribute, and perform the work, even for commercial purposes, all without asking permission.

11.1.7.3 Digital Rights Management

Digital assets are easy to copy, mix, and share and can lead to the dematerialization of the copyrighted materials through digitalization. Finck and Moscon referred *data right management (DRM)* as “software and hardware that defines, protects and manages rules for accessing and using digital content (text, sounds, videos, etc.) by leveraging the exclusive rights recognized by copyright and neighboring rights” [5]. They denoted it “as an additional layer of paracopyright constraints or technological self-protection that annihilate legally recognized protections such as exceptions and limitations in the EU and fair use under US copyright law. It is designed to give maximum control over digital content in self-enforcing conditions and terms of access and use, for instance, publishing and selling of electronic books, digital movies, digital music, interactive games, computer software, and other material in digital format” [5]. Thus, in the context of text mining, DRM follows copyright rules and exceptions (such as fair use that involves search engines, commentary, parody, criticism, research, news reporting, library archiving, teaching, scholarship, etc.) but for digital content specifically.

11.2 Social Media Ethics

It is very difficult to apply standard ethical practices such as *informed consent* and *anonymity* to social media mining. Researchers must develop ethical guidance when using social media data. Some of the critical areas of concern within social media research are:

1. **Is the data public or private?**: As social media platforms post data in public spaces, the users argue that the data is public. Depending upon the platforms, the users (including researchers) often agree to a set of *terms and conditions* stated by third parties. However, many people do not know that the users do not always

read those *terms and conditions*. Therefore, *whether the data is public or private depends upon the context of the data*. For instance, a discussion on Twitter or a post in an open debate can be considered a public data, whereas in the case where people are posting within a private or closed Facebook group, where one has to either become a member or have a password, that data might be considered as a private data.

2. **Do we have to acquire informed consent from the users of social media platforms?:** In most social media platforms, the users are usually not aware that they are being observed by the researchers or their data is being used by them. Therefore, informed consent is not present in the design of research itself as it would have been in traditional research methodology. Again, whether or not to seek informed consent depends upon certain situations, such as if you are working with data that can be considered as private that might also contain some sensitive data (such as users talking about things like illegal activities, marital status, financial status, a disease they are fighting with, etc., that might bring risk to them if that data is exposed to new contexts or new audience), then you need to seek consent before reusing the social media data. However, usually, with very large datasets, it is difficult to get consent from every user, so you have to consider how you would be using the data. Will you just aggregate the data together, analyze it, and present statistical results, or will you be citing individual tweets/units of data? Another important aspect is whether you are working with children or people who are considered as vulnerable adults, in which case again you have to consider the question of informed consent very seriously.
3. **Is there a need to anonymize the data?:** Anonymizing data is again a very problematic ethical concern with social media research than it is with traditional forms of data collection. It again depends on the way you are using the data where you might not give the name of the people when reusing their data in papers/presentations. However, if you use the unit of data presented on the platform in its original wording, then it can quite often be used to trace back that person's profile or to the owner of the social media data. In order to overcome that problem, you need to pay attention to the *terms and conditions* of the platform where you are accessing the data from because all the social media platforms have different sets of *terms and conditions*, and they change regularly as well. For instance, some platforms stipulate that you cannot reword units of data if you are going to use them in your research, and you have to use them word for word that makes it even more difficult to protect anonymity of the users. Moreover, it becomes much more vital to protect the anonymity of the users if those users had an expectation of privacy when they were posting the data or if the data is considered as sensitive or placed them under any risk or harm when exposed in a new context or if you are working with data that was created by children or vulnerable adults.

So, as researchers, it is your responsibility to make sure that when you are using the data of social media users in your research, (i) you do not place those users under any risk of harm, (ii) you make sure that you are protecting your participants

from the risk of harm, and (iii) you think about the nature of the data itself and the context from where you took the data. If you are only analyzing the data but not reporting back the data in your output word for word, then that is perhaps less risky. However, if you want to provide quotes and cite the data in its original format, then that is potentially riskier to your participants.

11.2.1 Framework for Ethical Research with Social Media Data

Figure 11.7 summarizes the discussion from Sect. 11.2 and presents a framework in the form of a flowchart to help the researchers to conduct research using social media data ethically. In addition to that, you might also need to consider your role as a researcher and whether you are also participating in the conversation online. Is there a blurring boundary between researchers and participants? Are you taking part in the conversation, and if so, are you doing so openly as a researcher who is seeking data, or are you doing so as another social media user with interest in the discussion taking place? In that case, you really need to think about the ethics of contributing to that conversation and, in some cases, perhaps guide the conversation for your own purposes as a researcher and whether or not to disclose your primary purpose for participating in that group.

Conversation about social media ethics is ongoing because technology is constantly changing, and social media platforms are continuously changing, as are their *terms and conditions*. So, we can hope that new ethical frameworks would emerge and refine with time as technology and its uses change. The framework itself should be considered as guidance but not as prescriptive because each social media context is unique. The ultimate responsibility lies with the researchers and their ethics committee to ensure that the approach they are taking is ethical.

11.3 Ethical and Legal Issues Related to Text Mining

“Law and technology have a complex relationship. Technology shapes legal development, while it is also shaped by law” [5]. Library professionals need to understand the ethical and legal issues related to text mining (Table 11.1). Many LIS job profiles, including university librarian, data librarian, scholarly communication librarian, subject librarian, and research librarian, require the candidate to have a good ethical and legal knowledge of text mining. Even though text mining has enhanced the overall functioning of libraries, there are legal and ethical barriers related to it.

The ethical and legal aspects related to text mining are much more complicated than can be anticipated. Therefore, it is hard to recognize the instances where text

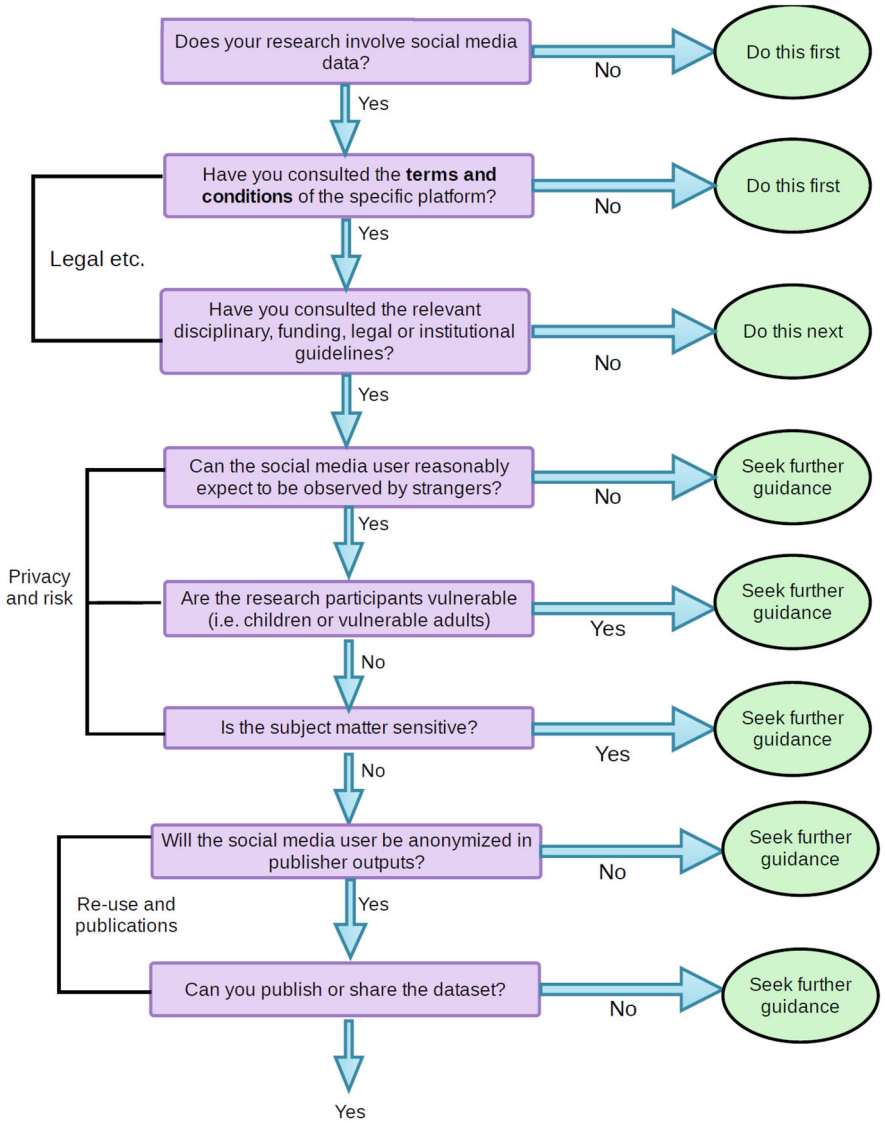


Fig. 11.7 Ethical framework for social media data (adapted from Townsend [6])

Table 11.1 Do's and don'ts of using text data for mining by librarians (adapted from McNeice et al. [2])

Do's	Don't's
Establish if you will use or mine data that contains sensitive data	Only think of data protection issues when you usually start the process of text mining
Assign a data protection officer if text mining is one of your organization's core activities, or if your organization does text mining regularly	Collect textual data and just assume it does not concern with personal, private, and sensitive data
<i>Impact assessment (IA)</i> : establish what data you will use for what purposes, and who will have access to the data within and outside your organization, and whether your use of textual data brings any legal or ethical risks	Store and retain all data just because it may be helpful in the future
Check whether you have the legal grounds to collect and use the textual data	Randomly transfer or provide access to the data to third parties
<i>Privacy by design</i> : based on your impact assessment (IA), design your whole text mining project in a way that guarantees that you can safely and adequately use the textual data	Reuse textual data from one project to another, without making sure it is compatible with data protection rules, even though you had made sure that the use in the first project is compatible
Look into sector-specific regulation or self-regulation and codes of conduct within your domain, which may provide you more guidance and certainty on what kind of analyses and techniques you can employ on the textual data	Share any textual data with the public, without proper consultation
Anonymize data so you are not dealing with any personal, private, or sensitive data. Note that if you pseudonymize personal, private, or sensitive data, one can still be able to identify the anonymized data if additional information is used	Make decisions affecting the data subjects solely on automated processing of their personal data—it should be prohibited
	Ignore the request by the owner/publisher/funder to access, rectify, or erase data
	Transfer textual data to others without permission

mining becomes unlawful. In the absence of apparent legal exceptions such as fair use, intellectual property rights are most likely to be present for the content created by others. Additionally, it is essential to be cautious in respecting the privacy and personal data of any data subjects. It is always vital to assess the potential legal issues before commencing a text mining project and try to minimize or avoid these issues in your project's design. As all the countries have different laws and regulations related to text mining, this chapter will be restricted to present the text mining rules and regulations for the European Union (EU). Thus, when conducting

Table 11.2 Difference between copyright, neighboring rights, and database rights (©2017, all rights reserved—reprinted under Creative Commons CC-BY license, published in McNeice et al. [2])

Copyright	Neighboring rights	Database rights
Protects the original and creative expressions of the authors	Protects performers (for instance, actors or musicians) and producers of performances or recordings thereof	Protects the investments of producers in creating those databases
It lasts for 70 years after the death of the author (according to EU copyright framework). Historical sources may be out of copyright	It lasts for 50 years after first publications or 70 years in the case of phonograms in the EU	It lasts for 15 years after publication in the EU. If a database is substantially modified, it starts again from the day of the modified version
Examples: Books, websites, research papers, newspaper articles, films, lyrics, musical compositions, original databases, and collections	Examples: Sound recordings, films, broadcasts, fixation of live performance	Examples: Relational databases, SQL databases, tables on a website, playlists on Spotify

text mining research, the two most common legal aspects a data miner should keep in mind are:

1. *Intellectual property rights*, especially copyrights, neighboring rights, and data rights (Table 11.2). It is essential to look at whether the content which you intend to mine is attached to any intellectual property rights or not. If it does, you might need to seek permission for the same from the rights holder. Text mining is most probably to be affected by neighboring rights or copyright, in comparison to data mining as *data* and *facts* are not creative expressions and are free from copyright, except for copyright associated with the collection of the data. Countries like the UK and France let you use the content for text mining without any permission from rights holders but are limited for non-commercial and scientific research purposes. “In many European countries, other text mining exceptions may also exist if you use the content for:
 - (a) Private and non-commercial purposes;
 - (b) Non-commercial research or teaching purposes in general; and
 - (c) Temporary copies are necessary to enable lawful use of work” [2].
2. *Data protection rules*, more specifically the *General Data Protection Regulation (GDPR)*¹³ in the EU which replaced the Data Protection Directives in 2018 as a primary law across EU nations to protect personal data. The GDPR rules for data privacy and protection include:

¹³ <https://gdpr-info.eu/>.

- The requirement of consent by subjects for data processing
- Protecting the privacy by performing anonymization of the collected data
- Producing notifications for a data breach
- Transferring the data across borders safely
- Appointing a data protection officer to oversee GDPR compliance

The following are some of the vital aspects covered in the EU's data protection rules:

- (a) *Personal data*: Personal data is related to an identifiable or identified living person and includes any data that can directly or indirectly let you identify the individual. In the EU specifically, anything from collecting personal data to removing personal data has to comply with European data protection rules. It is essential to get the consent of the data subjects and the consent must be unambiguous, informed, and properly registered. Personal data can be used for research purposes (scientific and historical) in the EU.
- (b) *Privacy*: McNeice et al. explained in their FutureTDM report that when “accessing research data made available by other organizations, it is important that mining activities do not inadvertently disclose confidential information or breach the privacy of research subjects. Although the primary responsibility for the ethical collection, storage, and access to research data sits with the research owner, it may be possible to filter data in ways that can reveal confidential or identifying details” [2]. They found that “this is why some data owners require researchers to make an application to use their data or may license its use via a formal agreement or Creative Commons (CC) license” [2].
- (c) *Data minimization vs. maximization*: Data maximization is the process of making *big data* text mining so valuable by gathering and using as much data as possible, whereas data minimization allows one to (i) collect personal data only for explicit, specified, and legitimate purposes, (ii) further use the data which is compatible with the purpose it was collected (purpose limitation), and (iii) use only adequate, relevant, and limited data which is necessary for text mining purposes. Hence, as summarized by the library guide of the University of Queensland that “even if the license permits text mining, some approaches to text mining are considered poor etiquette due to the inconvenience they can cause to data providers. For example, bulk scraping or non-rate-limited programmatic querying via APIs can place a significant burden on data providers’ servers, causing slow response times or even downtime for other users” [7]. The guide suggested that the “best practice is to check the requirements of the data provider and comply with their preferences regarding data mining activities” [7].
- (d) *Sensitive data*: Sensitive data is data related to political opinions, ethnic or racial origin, philosophical or religious beliefs or trade union membership, health data, biometric data, genetic data, and data related to a person’s sexual orientation or sex life. European data protection law is stringent for sensitive data and generally prohibits its usage unless you have legal grounds.

- (e) *Anonymization of data*: For any text mining project, it is essential to anonymize any personal, sensitive, and private data so that it cannot be reidentified.

11.3.1 Copyright

The library guide of the University of Queensland summarizes how “the process of mining is conducted (e.g., whether the material is copied, reformatted or digitized) without permission could be considered copyright infringement. The ability to data mine relies heavily on technologies that are considered *copy-reliant*, where copies must be made of the data in order for it to be analyzed” [7]. It mentioned that “currently, the Copyright Act 1968 makes no specific exemption for text or data mining. Limited text mining might be covered by the fair dealing exceptions. However, if an entire dataset needed to be copied, this would clearly exceed a *reasonable portion* of the work. While copyright does not apply to raw data or factual information, it does cover the arrangement of data within a database or the *expression* of data” [7].










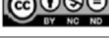
The Database Directive gives copyright exceptions for personal use, teaching, and scientific research. Thus, the most problematic issue with text mining with library databases, according to Ducato and Strowel, is the “broadness of concept of reproduction (for copyright) and extraction (for the database right). When a text mining tool runs the analysis, it copies all or part of the work and transfers all or substantial part of the content of a database to another medium or adapts or translates the content (such as the conversion of PDF to another format). These operations that are essential in the text mining process fall under copyright or database rights where user cannot perform text mining without the authorization of right holder” [8].

11.3.2 License Conditions

The library guide of the University of Queensland explains how “data providers will have their own specific standards and procedures that you must follow to use the data they provide legally. You must ensure from the outset of your project that the activities you intend to perform during the course of your text mining and the subsequent publication of your research results comply with any licensing *terms and conditions*” [7]. The guide further elucidates how “many data providers license their data to be mined for research purposes only and either prohibit or require special negotiation for text mining with potential commercial applications” [7].

Data miners should ensure that they follow the terms of use of the data while carrying out a text mining project (Table 11.3). Thus, they should look out for different types of Creative Commons categories (refer to Sect. 11.1.7.2) and the restrictions associated with the data in addition to intellectual property rights and data protection rules. Appendix C summarizes text mining licensing conditions for some of the selected prominent publishers.

Table 11.3 Comparison between different license categories

Copyright	Creative commons	Public domain
All rights reserved	Some rights reserved	No rights reserved
		
	<div><div>most free</div><div>      </div><div>least free</div></div>	
NA*		NA*
All <i>original</i> work is protected under Copyright as designated by the owner	Any work that owners have chosen to be designated as Creative Commons	Work published before 1923, work of dead owners, or when owners designate the work as public domain
Work <i>cannot</i> be adopted, copied, published, or used without the owner's permission	Work <i>may</i> be used without permission but as per the rules (above) set by the owner	Work <i>can</i> be adopted, copied, published or used without the owner's permission

*NA stands for *not applicable*

11.3.3 *Algorithmic Confounding/Biasness*

Mass-scale digitization of data, artificial intelligence (AI), and machine learning algorithms led to the automation of simple and complex decision-making processes. Automated data analysis tools, including algorithms used to process text mining, also present ethical challenges. An algorithm can be defined as a set of instructions for how a machine should achieve a particular task. They are the products that involve humans and machine learning. There is long overdue of standards and enforcement of accountability, fairness,¹⁴ and transparency for algorithms in text and data mining procedures. Nowadays, many decisions are governed by algorithms as the datasets are too large to be processed by humans. Thus, humans have become dependent on algorithms to make all kinds of recommendations and decisions for them. They are used to perform the following tasks in the process of text mining and machine learning:

¹⁴ Fairness is a nonmathematical, human determination grounded in shared ethical beliefs.

- Ranking information
- Curating information
- Classifying information
- Predicting information
- Clustering information and many other sub-tasks covered in the previous chapters

It may appear that algorithms are neutral and result in unbiased decisions as “they take in objective points of reference and provide a standard outcome” [9]. In reality, we may know what is going inside and coming outside the algorithm, but there is currently no available system of external auditing present that can assess what happens to the data during the process.¹⁵ They are *opinions embedded in mathematics*.¹⁶ As opinions, they are also different where some algorithms such as *word to vector* prefer a specific group of words over another and present incorrect results. “Thus, when an algorithm’s output results in unfairness, we can refer to it as bias. Algorithmic accountability is the process of assigning responsibility for harm when algorithmic decision-making results in discriminatory and inequitable outcomes” [9]. Algorithmic accountability helps to determine whether to use a system and whether human judgment is needed to check decisions to counter the biasness (Fig. 11.8).

It is essential to evaluate bias and make the algorithms as accountable as humans making the same decision [11]. In terms of accountability, it is easier to evaluate a non-AI software that applies a simple set of rules, where the process is transparent and can be checked or challenged by those who disagree with the rules, compared to AI software, where most algorithms are opaque and mathematically abstract so that humans cannot understand the rules and cannot challenge them. For instance, a support vector machine (SVM) uses thousand-dimensional space in relation to make thousand-dimensions separators. SVM is understandable in only two or three dimensions but not more. The opaqueness of algorithms can be checked for biasness. Therefore, it is not a problem but can generate unintended bias due to the nature of the data used to run the algorithm. In 2019, Booker [12] proposed the US Algorithmic Accountability Act to force companies to assess the impact of algorithms that make sensitive decisions and correct any found biases. In the same year, the EU released *Ethics Guidelines for Trustworthy AI*¹⁷ outlining seven governance principles: (i) human agency and oversight, (ii) technical robustness and safety, (iii) privacy and data governance, (iv) transparency, (v) diversity, nondiscrimination, and fairness, (vi) societal and environmental well-being, and (vii) accountability. These principles enable inclusion, equal access and treatment, and inclusive design processes in the AI’s lifecycle.

¹⁵ Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

¹⁶ O’Neil, Cathy. 2016. *Weapons of Math Destruction*. Crown.

¹⁷ <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.

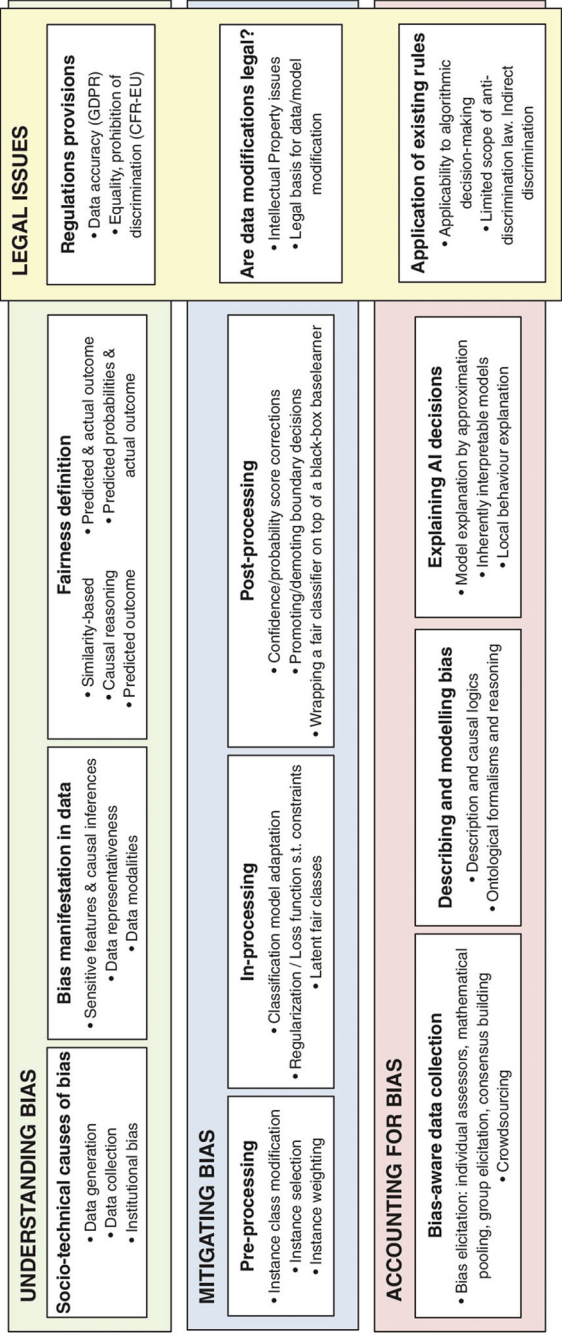


Fig. 11.8 Summary of different issues and solutions related to biasness (©2020 Wiley, all rights reserved—reprinted under Creative Commons CC BY 4.0 license, published in Ntoutseti et al. [10])

There many different ways in which biasness can be created in the algorithm:

- While creating the algorithm by its creator.
- The output of machine learning algorithm often reflects the patterns of input, so if any unexpected results come, it is worth investigating if they are accurate or have inherent bias.
- Insufficient or over-representation of training data.
- The way the algorithm is applied.
- The data on which the algorithm is applied.

11.3.3.1 Types of Biases

1. **Response or Activity Bias:** It occurs in the content generated by humans such as tweets, reviews, posts, etc. As only a small proportion of people within particular demographic groups and geographic areas contribute to opinion, content, and preferences that are unlikely to reflect the opinions of the whole population, this can lead to response or activity bias.
2. **Selection Bias Due to Feedback Loops:** It occurs when a model itself influences the generation of data that is used to train it, such as a ranking system or recommender system. Machine learning models have built-in feedback loops, where the generated data is fed back into the system as training data for the model and is influenced by the user's responses. The user's responses are used to generate label examples by tracking their views, clicks, and scrolls. This bias tends to favor models that generated the data when evaluated using held-out samples from this data and homogenizes user's behavior over time.
3. **Bias Due to System Drift:** It occurs when the model or algorithm changes how users interact with the system over time, for instance, the failure of Google Flu Trends for influenza-like illness based on search data to forecast the expected number of flu cases for a season [13]. It is of two types:
 - (a) **Concept Drift:** It occurs when the target or concept being learned is changed, such as the definition of fraud changes in fraud prediction system.
 - (b) **Model Drift:** It occurs when there is a change in the user interaction model, such as adding new modes of user interaction like share or like buttons or the addition of search assist feature.

When there is drift in a static model, its performance can degrade over time. Also, if the data from different periods are used to train a model, then the model may not perform well.

4. **Omitted Variable Bias:** It occurs when the vital attributes that influence the outcome are missing and happens when data generation relies on human input, or the process that records the data does not have access to the attributes due to privacy concerns.
5. **Societal Bias:** It occurs when humans produce content on the web and social media, such as gender or race stereotypes.

11.3.3.2 Ways to Mitigate Biases

Though algorithms are deployed to correct humans' biasness, they usually either codify the existing biases or create new ones. They need to be consistently monitored and adjusted with time as they can outdate rapidly. Some of the ways biases can be mitigated are by:

- Understanding how the data was generated
- Performing comprehensive exploratory data analysis (EDA) techniques
- Training the algorithm with sufficient training data which is representative of all the classes
- Using tools that identify bias in models and algorithms such as FairML,¹⁸ IBM AI Fairness 360,¹⁹ Accenture's "Teach and Test" Methodology,²⁰ Google's What-If Tool,²¹ and Microsoft's Fairlearn²²
- Using R packages for interpretability (such as `iml`,²³ `lime`^{24,25}) and fairness (such as `fairness`²⁶);
- Making the data, process, and outcome open, thus making it *transparent* and helping us to judge "whether it is fair and whether its outcomes are reliable where different contexts may call for different levels of transparency (EU directive 2016/680 General Data Protection Regulation states the *right to an explanation* about the output of an algorithm)" [9]
- Re-purposing the data and algorithms but keeping in mind the standards that were set and ethical in one setting may create biasness in a new application, therefore creating algorithms and standards that can be adapted from one application to another
- Following the "set of standards proposed by the Association for Computing Machinery US Public Policy Council²⁷" [9] and applying them at every stage in the algorithm creation process
- Enforcing accountability in policies during "auditing in pre-and post-processing as well as standardized assessment" [9] as critically algorithms do not make mistakes, but humans do

¹⁸ <https://dspace.mit.edu/handle/1721.1/108212>.

¹⁹ <https://aif360.mybluemix.net/>.

²⁰ <https://newsroom.accenture.com/news/accenture-launches-new-artificial-intelligence-testing-services.htm>.

²¹ <https://cloud.google.com/ai-platform/prediction/docs/using-what-if-tool>.

²² <https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/>.

²³ <https://cran.r-project.org/web/packages/iml/index.html>.

²⁴ <https://www.rdocumentation.org/packages/lime/versions/0.5.1>.

²⁵ <https://cran.r-project.org/web/packages/lime/lime.pdf>.

²⁶ <https://cran.r-project.org/web/packages/fairness/fairness.pdf>.

²⁷ https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

11.3.3.3 Language-Based Models and Biasness

In recent studies, it has been identified that there are different variety of biases in language-based models (such as sentiment analysis and word embedding) that might disseminate social biases against certain groups of sociodemographic factors (such as race, gender, geography). The output of these language-based models is primarily dependent on the annotated datasets and is sensitive to social bias created by humans. An algorithm that uses both text and metadata to learn is likely to be highly biased as metadata consists of the author's nationality, discipline, etc., when compared to an algorithm with text-only data. Even with text-only data, algorithms will still learn bias due to the language problems generated by second-order effects for text-based machine learning. Additionally, when using chatbots to provide real-time recommendations, the “dialogue of chatbot can be modelled with available metadata to adjust the features of the replier in terms of gender, age, and mood” [14]. Biases in NLP models and tools can emerge from the following sources:

- Amplifying particular views on social media such as “retweeting” the minority opinions and introducing bias
- Creating user bias by putting particular keywords
- Bias emerging from algorithmic decision-making
- Bias by “auto-complete function of search engines” [15]
- “advertisements based on search terms and image search results” [15]
- Biasness created in human-authored data for training the algorithm

Algorithms can magnify and perpetuate the biases if not removed from the confounding factors but can potentially mitigate biases effectively if appropriately designed. Language-based bias can be mitigated by:

- Performing gender-neutral word embedding
- “tagging the data points to preserve the gender of the source” [15]
- Using R packages for interpretability (such as DALEX²⁸) and fairness (such as fairness²⁹)

Appendix C

Text Data and Mining Licensing Conditions

Table C.1 summarizes text mining licensing conditions of some selected prominent publishers.

²⁸ <https://cran.r-project.org/web/packages/DALEX/index.html>.

²⁹ <https://cran.r-project.org/web/packages/fairness/fairness.pdf>.

Table C.1 Selected prominent publishers and their TDM licensing conditions (McNeice et al. (2017) FutureTDM: Reducing Barriers and Increasing Uptake of Text and Data Mining for Research Environments using a Collaborative Knowledge and Open Information Approach. https://project.futuretdm.eu/wp-content/uploads/2017/07/FutureTDM_D5.3-FutureTDM-practitioner-guidelines.pdf. Accessed 5 Nov 2020)

	Elsevier	Wiley	Springer	Emerald
Where is the licence?	https://www.elsevier.com/tdm/userlicense/1.0/	http://olabout.wiley.com/WileyCDA/Section/id-826542.html	http://www.springer.com/tdm	https://www.emeraldinsight.com/page/tdm
Do I need to check other licences or documents?	NO The TDM Agreement supersedes any and all prior and contemporaneous agreements	UNCLEAR The click-through TDM agreements say that it supersedes all other prior and contemporaneous agreements, but all that it is superseded by any separate TDM agreement	YES The TDM clause may not have been included in existing SpringerLink subscription agreements but can be added by existing subscribers	NO Not mentioned in policy
Does this license affect my use of open access content?	NO Individual OA licenses supersede anything the contrary in the TDM Agreement	NO If more permissive licenses of OA apply, you may use content in accordance with article-level restrictions	NO Springer content is usually allowed without restrictions	NO Not mentioned in policy
Is TDM permitted?	YES	YES	YES	YES
Can I carry out TDM for any purposes?	NO You may not extract, develop or use the dataset for any direct commercial activity or indirect commercial activity related to specific projects; direct or indirect commercial purposes require prior written consent from Wiley	NO You may only text and data mine Wiley content for non-commercial scholarly research purposes	NO You may only access content for non-commercial research purposes	NO TDM rights are granted purely for internal non-commercial research purposes
Do I need to tell anyone what I am doing with TDM?	MAYBE You must provide TDM output and any related content to Elsevier on request	NO Not mentioned in policy	NO Not mentioned in policy	NO Not mentioned in policy
Are my TDM activities monitored?	YES You are required to use an API key; Elsevier maintains information about you which may be used in aggregate, and may be used to promote Elsevier offers to you	YES You are required to use an API key	NO No authentication is required when retrieving SpringerLink content for TDM	NO No authentication is required for CrossRef's TDM API

Note: Table C.1 is colored coded as follows:

Ideal for TDM activities	Close to ideal for TDM activities	Some negative implications for TDM activities	Very negative implications for TDM activities
--------------------------	-----------------------------------	---	---

	Elsevier	Wiley	Springer	Emerald
Can I access any content I like?	NO You are licensed to solely to access content made available via the CrossRef API	NO You may only access content made available via APIs	YES You may use all subscribed content	YES Emerald suggests using CrossRef's TDM service to identify and access content
Can I access content any way I like?	NO You are licensed to use a set of proprietary APIs to access data; you may not use any automated programs to search or scrape any Elsevier web site or application	NO You must access content using a Wiley-approved API, and you may not bypass the API; you may not use any automated programs to search or scrape any Elsevier web site or application	YES You are encouraged but not required to download content directly from the SpringerLink platform; friendly DOI-based URLs are provided, tools and methods are suggested, and no API key is required	YES You are encouraged to use CrossRef's TDM services, but not forbidden from accessing content in other ways
Are there limits on how much content I can access, or how quickly?	UNCLEAR No details are provided about rate limiting through Elsevier's API	SOMETIMES You must abide by any rate-limiting which may be conveyed from time to time	VOLUNTARY You are asked to be considerate and limit your download speed to a reasonable rate	SOME There are no hard limits on the number of items that may be downloaded, but you may be blocked if your downloading constitutes unfair usage
Do I need to ask or inform anyone before carrying out TDM?	NO Not mentioned in policy	NO Not mentioned in policy	NO Not mentioned in policy	ADVISED You are advised to inform Emerald you wish to mine their site to avoid being blocked due to unfair usage
Are there limitations on the types of TDM analysis I can perform?	YES You are licensed to extract semantic entities for the purpose of recognition and classification of relations and classifications between them	NO You are licensed to carry out computational analysis including but not limited to identification of entities, structures, and relationships	NO None mentioned in policy	SOME License includes specific definitions of TDM activities and outputs; these are broad but you must not perform systematic or substantive extracting of content
Are there restrictions on how I can store and share datasets I am using for TDM?	NO None mentioned in policy	YES You may load and technically format content on your servers for use for specific TDM projects; you may not otherwise create any form of central repository for Wiley content, or any product or service that could potentially substitute any existing Wiley services	NO None mentioned in policy	SOME You may load and technically format XML content on your server, PC or laptop to enable access and use of content for allowed TDM purposes; you may not make results of TDM outputs available on any externally facing server or website

Note: Table C.1 is colored coded as **Ideal for TDM activities** **Close to ideal for TDM activities** **Some negative implications for TDM activities** **Very restrictive for TDM activities**

	Elsevier	Wiley	Springer	Emerald
Are there restrictions on how I can share new knowledge I generate as a result of TDM?	YES Results may be used by you and your company or institution, but may not be used with existing Elsevier products; a specific proprietary notice must be used when sharing results externally	YES You may communicate TDM outputs as part of original non-commercial research, including in articles about that research	NO None mentioned in policy	YES There are no restrictions on where and how you can publish your research results, but you may not make results of TDM outputs available on any externally facing server or website
Am I required to share the outputs of my TDM research?	YES You must provide TDM output and any related content to Elsevier on request to ensure compliance with their agreement	NO None mentioned in policy	NO None mentioned in policy	NO None mentioned in policy
Can I support my results with experts from the content I have mined?	YES Limited to the dependent text of a maximum length of 200 characters surrounding the semantic matched, or metadata, must include a DOI link to the original material	YES Limited to query- permitted under national copyright laws; must include a DOI link to the original material	YES Limited to quotations of up to 200 characters, 20 words, or maximum of 200 characters; must include a DOI link to the original material	YES You can use snippets up to a maximum of 200 characters, provided these are referenced as you would reference a copyrighted work; you must contact Emerald if larger extracts are exceptionally required
Can I retain datasets for verifiability and reproducibility of my results?	NO You may not substantially retain the dataset; all Elsevier content stored for TDM must be permanently deleted on termination of the agreement	NO You must delete all Elsevier content downloaded for TDM project, or on termination of the agreement with Wiley	UNCLEAR Not mentioned in policy	NO You may not substantially retain content; all copies of Emerald content that have been locally loaded for TDM must be destroyed on termination or expiry of this license
Do I have other responsibilities or obligations?	YES You are responsible for complying with data protection and relevant privacy laws when using or processing personal data	YES You must implement and maintain data security measures to protect Wiley content in line with international industry standards; you are responsible for complying with data protection and relevant privacy laws when using or processing personal data	NO None mentioned in policy	NO None mentioned in policy

Note: Table C.1 is colored coded as

Ideal for TDM activities	Close to ideal for TDM activities	Some negative implications for TDM activities	Very restrictive for TDM activities
--------------------------	-----------------------------------	---	-------------------------------------

References

1. Martone M (ed) (2014) Data Citation Synthesis Group: joint declaration of data citation principles, San Diego, CA: FORCE11. <https://doi.org/10.25490/a97f-egy>
2. McNeice K, Caspers M, Gavrilidou M (2017) FutureTDM: reducing barriers and increasing uptake of text and data mining for research environments using a collaborative knowledge and open information approach. https://project.futuretdm.eu/wp-content/uploads/2017/07/FutureTDM_D5.3-FutureTDM-practitioner-guidelines.pdf. Accessed 5 Nov 2020
3. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18>
4. Creative Commons (2020) About the licenses. <https://creativecommons.org/licenses/>. Accessed 6 Nov 2020
5. Finck M, Moscon V (2019) Copyright law on blockchains: between new forms of rights administration and digital rights management 2.0. *IIC* 50:77–108. <https://doi.org/10.1007/s40319-018-00776-8>
6. Townsend L (2017) Social media research & ethics. SAGE research methods [streaming video]. SAGE, London. <https://doi.org/10.4135/9781526413642>. Accessed 26 Feb 2021
7. Berends F (2020) Library guides: text mining & text analysis: considerations - ethics, copyright, licencing, etiquette. <https://guides.library.uq.edu.au/research-techniques/text-mining-analysis/considerations>. Accessed 6 Nov 2020
8. Ducato R, Strowel A (2019) Limitations to text and data mining and consumer empowerment: making the case for a right to “Machine Legibility.” *IIC* 50:649–684. <https://doi.org/10.1007/s40319-019-00833-w>
9. Caplan R, Donovan J, Hanson L, Matthews J (2018) Algorithmic accountability: a primer, data & society. https://datasociety.net/wp-content/uploads/2019/09/DandS_Algorithmic_Accountability.pdf. Accessed 8 Nov 2020
10. Ntoutsis E, Fafalios P, Gadiraju U, Iosifidis V, Nejdil W, Vidal M-E, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E, Kompatsiaris I, Kinder-Kurlanda K, Wagner C, Karimi F, Fernandez M, Alani H, Berendt B, Kruegel T, Heinze C, Broelemann K, Kasneci G, Tiropanis T, Staab S (2020) Bias in data-driven artificial intelligence systems—an introductory survey. *WIREs Data Min Knowl Discovery* 10:e1356. <https://doi.org/10.1002/widm.1356>
11. Lepri B, Oliver N, Letouze E, Pentland A, Vinck P (2018) Fair, transparent, and accountable algorithmic decision-making processes. *Philos Technol* 31:611–627. <https://doi.org/10.1007/s13347-017-0279-x>
12. Booker C (2019) Booker, Wyden, Clarke introduce bill requiring companies to target bias in corporate algorithms. <https://www.booker.senate.gov/news/press/booker-wyden-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms>. Accessed 12 Nov 2020
13. Butler D (2013) When Google got flu wrong. *Nat News* 494:155. <https://doi.org/10.1038/494155a>
14. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, Gigante A, Valencia A, Rementeria MJ, Chadha AS, Mavridis N (2020) Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digit Med* 3:1–11. <https://doi.org/10.1038/s41746-020-0288-5>
15. Diaz M, Johnson I, Lazar A et al (2018) Addressing age-related bias in sentiment analysis. In: Proceedings of the 2018 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, pp 1–14

Additional Resources

1. Barocas S, Hardt M, Narayanan A (2019) Fairness and Machine Learning. <http://www.fairmlbook.org>. Accessed 17 June 2021
2. D'Ignazio C, Klein LF (2020) Data feminism. The MIT Press, United States. <https://data-feminism.mitpress.mit.edu/>. Accessed 17 June 2021
3. Noble SU (2018) Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press
4. <https://uc-r.github.io/lime>
5. <https://www.kdnuggets.com/2019/09/python-libraries-interpretable-machine-learning.html>
6. <https://onthebooks.lib.unc.edu/>
7. <https://dhdebates.gc.cuny.edu/read/untitled/section/557c453b-4abb-48ce-8c38-a77e24d3f0bd>