Ethics of Data Mining

My Post:

"…when we are dealing with data, [that] data represent[s] peoples' lives." (North, 2023).
Data mining ethics can be easily understood through the lens of consumer privacy. This type of ethical concern is heavily tied to corporations' collection, dissemination, and use of consumer purchase activity or general web activity. These same concerns can be seen in data mining activities occurring within LIS systems for collection of data on patron activities as well. When data mining is being used in this way, to collect user data, it is important to keep individuals' identities and private information safe. While I agree with the viewpoint, it does feel a bit like the cats are already out of the bag at this point with how advertisements are able to track us through the web nowadays.
Another aspect of ethics comes up when looking at data mining regarding research practices in an academic setting. My last discussion looked at how text and data mining might be useful for research services in academic libraries. When using published articles for text mining consideration for copyright and publication policies may come up. According to Senseney et al "texts that comprise datasets for analysis are frequently protected by copyright or other intellectual property rights (2018)." I think further issues could arise if students or faculty are required to make their research data available once their research is published, as is the case with some grants and institutional mandates. Of course, all of this depends on the literature being used.

References

North, M. (2023). *Data mining for the masses: With implementations in RapidMiner and R (4th ed).* MyEducator.

Senseney, M., Dickson, E., Namachchivaya, B., & Ludäscher, B. (2018). Data mining research with in-copyright and use-limited text datasets: Preliminary findings from a systematic literature review and stakeholder interviews. International Journal of Digital Curation, 13(1), 183-194. https://doi.org/10.2218/ijdc.v13i1.620

Responses to responses of my post

Student 3:

*responses thoughtfully to my post and asks:

I also found your example of a data mining policy for academically published articles interesting, but I wonder how this is enforced? Is there a policy that requires text-data mining to report what datasets were used?

My Response:

Hi Student 3,

It's interesting to hear that signing in via one of the larger sites is less secure. While I understand that it means trackers are more likely to follow you from one site to the next/use information from both sites, I was under the impression that maybe it was more secure because of the affiliation with the larger company? (This is probably just me being less tech literate then I should be.)

In terms of dataset reporting it's less TDM specific and more a trend/standard that seems to be being suggested & sometimes implemented in RDMS circles. As I understand it, the idea is to make datasets available so that others can confirm findings and perform further analysis.

Response to Student 1s Discussion Post: Regarding student perspectives of Data Mining in Higher Education Libraries:

Hey Student 1,

I liked the idea of having student representation on the protection boards in academic libraries. This should facilitate conversation between the librarians doing the data mining and students whose information is being used. Some questions came up for me though. The students on the board may be particularly sensitive to the subject of data privacy or possibly have a greater interest in data security than the average student which could prompt them to seek out such a position. If this is the case, do these individuals really encapsulate what the average student thinks? This also becomes a question for the surveys, are those answering actually a good representation of the student body? This seems like such a tricky situation due to how much student demographics may shift from one year to the next.

Additional Discussion to Student 1s post — My Response to Student 2s Response where in they talked about wanting to understand how prevalent data leeks are at Higher Education Institutions

> I had to go looking when I read your last sentence and I'm a bit sorry I did. These are a bit old, but still crazy. 30,000 records and 10,000 records with compromising student information leaked through hacks just from these two instances that you mentioned. It would be useful to know where the breaches happened (which department/databases) but these articles were quite short. Also I could understand why the schools wouldn't want that information to get out.
>
> Perez-Hernandez, D. (2014). Data breach at Iowa State U. exposes nearly 30,000 social security numbers.
>
> Fischman, J. (2008). Harvard security breach exposes sensitive student data. Wired Campus, The Chronicle of Higher Education.

Response to Student 2s Discussion Post: Regarding how the National Endowment of the Humanities & UC Berkeley Library Taskforce works to ensure the legal and ethical compliance of TDM Projects

Hi Student 2,

The discussion of unintended consequences is important when talking about most technologies, but especially so when regarding individuals' personal data. As you mentioned the potential for individuals to be inadvertently identified is a major concern, particularly regarding minority or disadvantaged groups. This seems to be the intersection where biases can be exploited for monetary gain or to further push power dynamics in one group's favor.

The thing that really gets me with these discussions is how lopsided ethics discussions appear to be. I used to work for a large company; my take away from that environment was that ethical considerations only seemed to matter if the actions being taken made the company look bad. If the unethical choices were potentially disguised enough from public view or could be manipulated away through minor actions by the company while creating more revenue, that was the path that was typically taken.

Response to Student 3s Discussion Post: On how minority populations specific needs are often overlooks and the need for better privacy policies & disclosures

Student 3,

This is a very well researched post with some interesting examples I've never thought about before. The example of privacy policies as informed consent is particularly thought-provoking. I agree with Student 1's comment about how these policies can be both confusing and somewhat useless if the majority of users don't actually read them. More than that, if users are actually supposed to read through these policies, realistically the companies should make them much more accessible. The average American reads at a 7th-8th grade level, while most of these policies are full of legal jargon.

Instinctually I want to push back a little bit on the idea that this issue is a shared responsibility between users and creators. While some of these services can be viewed as optional, quite a few of them are not. For example, any apps that are needed for work, many apps related to finance (banking), online health portals or even government websites. I found it crazy that I had to agree to terms of use on an ID site that had to be used to gain tax documents from the IRS the other year. Even taking something like dating apps or social media into account, deciding not to use these services in some circumstances is the equivalent of isolating oneself if the community they live in is lacking 3rd places. As you mentioned there are ways to limit the amount of risk as a consumer, but companies need to do better.