Readings Week 1

Text Mining Terminology: Choose 1 or more

1. What exactly is text mining?
   1. From the Ferreira-Mello et al. (2019) article, text-mining is described as the process to extract high-quality information from unstructured data using techniques such as natural language processing, text classification and clustering. It is also described as the process of extracting interesting and non-trivial information and knowledge from unstructured data. The “unstructured” portion of this definition means that the data has not been sorted, categorized, quantified in any way to make it usable for analysis. The data is more meant to be usable by humans, such as online posts, chat messages, and essays. Text mining has applications in finance, business, medicine, biology, among other fields. When applied to education, text mining is typically called “educational text mining”.
2. How is it defined or described in your readings?
3. What alternative terms or phrases are used? How would you explain it to someone with no background in text mining?
4. What are some other new terms, words, concepts that you have come across in the resources that were unfamiliar to you, or that you had come across before but feel you have a better understanding of after this week?

Methods

1. What text mining methods or techniques are commonly used in analyzing text and for what purpose?
2. How has text mining be applied to educational contexts, or in other fields that might be relevant to education?
3. How might text mining be applied in your own professional context?
   1. From the Ferreira-Mello et al. (2019) and Fesler et al. (2019) articles, there were several applications of text mining that I realized have already been applied to my own professional practice as a school psychologist. I recall that a few years ago a new version of an academic assessment that school psychologists often give, the Weschler Individual Achievement Test, Fourth Edition (WIAT-4) was released and included an online automatic essay scoring component. There was a lot of discussion and mistrust about the system in our field which was not helped by the fact the publisher Pearson did not really explain how the system worked. I realize now that they are likely using a text mining/natural language processing to evaluate essays. I see text mining being useful for several other clinical applications in school psychology. For example, a practice that we often use in the field is called “testing the limits” and it is done as a qualitative approach to see why a child struggled to answer a particular type of question during testing or better understand how they approach problem solving in general. We usually use questions that they have already seen before to do this; however, if we had automatic question generation, then we could better evaluate them using questions that are at a similar difficulty level to the ones that they have already seen.
4. How has/could text mining be applied to address systemic issues or persistent problems in Education?

Affordances

1. What are some of the advantages of text mining over more traditional approaches to analyzing text as data?
2. What are some of the challenges and limitations of text mining in comparison to traditional qualitative analysis?
   1. While text mining often allows researchers to analyze a larger corpus of data than would be feasible using human-mediated techniques, text mining often still requires qualitative review of data to ensure that ratings completed using text mining are valid. For example, Fesler et al. (2019) points out that when using unsupervised dictionary-based text analysis requires comparison of the dictionary to hand coding to ensure that the dictionary being use accurately captures the valence of the text it is being applied to. By using a confusion matrix, you can compare the computer generated codes compared to your hand codes to see if the computer accurately classified your hand codes and vice versa. Similarly, supervised machine learning requires qualitative coding or labeling of data to use the approach. In my self-selected reading, Martí‐Parreño et al. (2016) used qualitative procedures to provide context to the topics identified using text mining. A limitation of the approach is that when you derive unigrams or bigrams from text, you lose the context of the original document that it was in. By qualitatively reviewing articles from which topics were identified using text-mining, Martí‐Parreño et al. (2016) were able to get more context for their data which facilitated their ability to identify how themes were represented across articles in their systematic review.
3. What ethical issues should be considered when mining text for educational purposes or in general?

Text-Based Data Sources

1. What sources of data are commonly used in text mining, particularly in educational contexts?
2. Are some data sources more suitable or appropriate for text mining purposes than others?
3. What data sources in education are publicly available for analysis?
4. What sources of data are you interested in potentially exploring for an independent analysis or final course project?
   1. There are a few different data sources that I would be interested in using for an independent analysis. I am interested in doing a systematic review of research on multitiered systems of supports (MTSS). This would involve taking analyzing manuscripts (articles, book chapters, dissertations) and identifying themes of trends in the literature. Since a goal with this project would be to classify the range of interventions and implementation methods represented in the literature on MTSS. My advisor and I have been qualitatively coding these manuscripts for their content, so it would be interesting to see how well text mining does at classifying documents and if they match our human codes. There are a few ways to approach this. If I use an unsupervised approach, we would be able to see how well the algorithm does at identifying certain themes or ideas in the literature and potentially uncover new themes that were not identified through human coding. Alternatively, and likely the approach I would take, is using a supervised machine learning method in which I train an algorithm on a subset of pre-coded data and see how well the algorithm does at classifying documents compared to human codes using a confusion matrix.

The other data source that I am interested in is Reddit data. Specifically, last semester I began an independent project in which I was looking at educator perspectives and sentiments about MTSS as expressed on the r/Teachers subreddit. The goal with this project would be to identify topics teachers naturally bring up on r/Teachers related to MTSS and to see if they identify any consistent barriers to MTSS implementation. This would involve using lexical-based methods such as using a dictionary to conduct a sentiment analysis as well as topic modeling to identify topics that are most frequently discussed.

APA Citation (note: this can be easily [retrieved via Google Scholar](https://moodle-courses2425.wolfware.ncsu.edu/mod/forum/discuss.php?d=265065#0))

Martí‐Parreño, J., Méndez‐Ibáñez, E., & Alonso‐Arroyo, A. (2016). The use of gamification in education: a bibliometric and text mining analysis. *Journal of Computer Assisted Learning*, *32*(6), 663-676. <https://doi.org/10.1111/jcal.12161>

What was the **purpose** of your article?

This article was written to provide a review of current literature on gamification and educational games used in education. The authors used bibliometric analysis to identify the distribution of topics over time, social network analysis to identify clusters of researchers and institutions conducting research in this area, and text mining to understand the topics being discussed by articles.

How was Text Mining **defined** and/or characterized?

Text mining is defined as a method used to “…overcome the limitations of data mining to handle huge amounts of unstructured textual documents (Martí-Parreño et al., 2016) .”

What **data source(s)** were analyzed or discussed?

Data were 139 journal articles published between 2010 to 2014 sourced from ISI Web of Science. Key search terms included “game-based learning”, “serious games”, and “gamification”. The authors then reviewed the documents by reading titles and abstracts and selected those that were most relevant to the study.

How, if at all, did your article touch upon the **application(s)** of text mining to "understand and improve learning and the contexts in which learning occurs?”

The article sought to improve learning and its contexts by understanding how research has evolved on gamification and game-based learning. Text-mining in particular was used to identify the topics addressed by research over time. They specifically looked at unigrams and found that terms such as “effectiveness”, “attitudes”, “engagement”, “cognitive”, and “assessment” were most frequently represented in the literature, and that “assessment” and “engagement” increased in frequency between 2010 and 2014.

Did your selection address any **ethical or legal considerations** of text mining?

Ethical and/or legal issues were addressed in this article.