Title: Text Mining Applications to Support Health Library Practice: A Case Study on Marijuana Legalization Twitter Analytics

Structured Abstract

Background: Twitter is rich in data for text and data analytics research, with the ability to capture trends.

Objectives: This study examines Canadian tweets on marijuana legalization and terminology use. Presented as a case study, Twitter analytics will demonstrate the varied applications of how this kind of research method may be used to inform library practice.

Methods: Twitter API was used to extract a subset of tweets using seven relevant hashtags. Using open source programming tools, the sampled tweets were analyzed between September to November 2018, identifying themes, frequently used terms, sentiment, and co-occurring hashtags.

Results: More than 1,176,000 tweets were collected. The most popular hashtag co-occurrence, two hashtags appearing together, was #cannabis and #CdnPoli. There was a high variance in the sentiment analysis of all collected tweets but most scores had neutral sentiment.

Discussion: The case study presents text-mining applications relevant to help make informed decisions in library practice through service analysis, quality analysis, and collection analysis.

Conclusions: Findings from sentiment analysis may determine usage patterns from users. There are several ways in which libraries may use text mining to make

evidence-informed decisions such as examining all possible terminologies used by the public to help inform comprehensive evidence synthesis projects and build taxonomies for digital libraries and repositories.

Keywords: Social media, Twitter, Text mining, Evaluation, Sentiment analysis, Taxonomies

Key Messages:

- Twitter is rich in data for text mining research to identify themes, sentiment, and textual analysis.
- Using marijuana legalization as a case study, findings shed light on the variety, frequency, sentiment, and co-occurrence of terms used in the Canadian Twittersphere.
- Text mining can be used to inform library practices, services, and operations such as finding terminologies to build the taxonomy of a digital library or repository.
- Library management may use sentiment analysis as a method to extract user comments related to library facilities and services.

Background

Text mining is a relatively new research method adopted by social science researchers (Ignatow & Mihalcea, 2018). As a subset of data mining, text mining involves using software and programming to retrieve large volumes of data and turning unstructured text into a structured format in order to extract meaningful insights. Twitter, a microblogging social media platform, is rich in data for text and data analytics research. With the ability to capture trends, knowledge dissemination, and gather other types of information (e.g. user information, network of followers), researchers are able to evaluate the sentiment (e.g. positive, neutral, and negative could be considered a subset of all possible sentiment that could be analyzed) behind the messages and content for themes (Cavazos-Rehg et al., 2015; Dai & Hao, 2017; Lamy et al., 2016). Previous studies have made use of Twitter as a source of information and evidence for conducting sentiment analysis on topics such as the Olympics and the Oscars (Thelwall et al., 2011) and examining political polarization on social media (Gruzd & Roy, 2014). In the health sciences, text mining techniques helped inform several projects such as using Twitter as an effective outbreak surveillance tool during the Ebola outbreak (Odlum & Yoon, 2015), and as a medium for potentially sharing valid and invalid health information concerning antibiotics (Scanfeld et al., 2010). Marijuana has also been explored and some American studies found that the majority of tweets related to marijuana or cannabis edibles were positive (Cavazos-Rehg et al., 2015; Lamy et al., 2016).

There are notable benefits for academic libraries to incorporate data mining and big data-related technologies to improve services (Sandhu, 2015). Katsurai and Joo (2021) evaluated trends of data mining methods in library and information science (LIS) and they found big data, machine learning, text mining, information retrieval, and dimension reduction to be the most popular methods. Information retrieval (IR) systems process search queries and retrieve documents through a matching algorithm to the end user who made the request (Sharma & Patel, 2013). As biomedical literature continues to grow exponentially (Fodeh & Zeng, 2016), some scholars have improved the performance of document retrieval in health sciences by using a probabilistic IR model that incorporates a passage retrieval system (Sarrouti & Ouatik El Alaoui, 2017). Passage retrieval extends the IR method by identifying relevant passages or pieces of texts instead of whole documents and this can be achieved with higher probability when the query matches passages that already exist in the system (Sarrouti & Ouatik El Alaoui, 2017). Text mining extends the IR model and focuses on information extraction rather than information retrieval (Hersh, 2020). Hersh (2020) posits that even though IR and data extraction systems are distinct, IR systems play a critical role “for aiding processes like information extraction and text mining” (p. 11).

Text mining became a popular data mining method in LIS, particularly during the period 2009-2013 (Katsurai & Joo, 2021). Some examples include using Twitter data to understand existing library practices or support strategic planning. To understand patron engagement with the library, one study analyzed retweets and assessed how followers engaged on Twitter and they found that most tweets were related to “institutional boosterism” (Stewart & Walker, 2018). In other words, library hashtags did not correspond to library-related content but were associated with slogans, mascots, and university events instead. This enabled users who were affiliated with the university to discover and retweet content, rather than library followers. Sewell (2013) evaluated Twitter followers of a comparable library to determine the usefulness of the platform as a means of communicating with the target population in their own library. Similarly in another study, the authors relied on analytics to help improve patron outreach on Twitter (Al-Daihani & Abrahams, 2016). Public libraries also used text mining techniques and found that Facebook posts including community events or awards/photos generated more user likes and shares, and these findings helped staff make informed decisions for improving engagement via social media (Joo et al., 2020).

There are practical text mining applications to literature searching. Quality systematic reviews and other evidence synthesis projects require support from information professionals with searching expertise. The Cochrane Handbook recommends researchers to consult with librarians or information specialists for the search process, especially authors of non-Cochrane reviews (Higgins et al., 2021). There is also a growing body of evidence to suggest that librarians improve the quality of the search when they are involved in some capacity (Aamodt et al., 2019; Meert et al., 2016; Schellinger et al., 2021). In order to develop comprehensive searches, librarians employ

several strategies for recognizing relevant free-text terms and controlled vocabulary terms. Text mining is a lesser known but creative way to identify terminologies on a larger scale (Milward & Singh, 2012), where librarians can find tweets on a particular topic (e.g. marijuana), and discover other terms related to the topic, such as weed, Ganja, or hash. Librarians can also create a corpus by exporting references from databases, such as MEDLINE and CINAHL, and use free software (i.e. Voyant and R) to analyze the data for term frequency and correlating terms (McGowan, 2021). McGowan (2021) argued, “text mining tools help librarians improve search precision, increase search sensitivity, and translate search strategies across multiple research databases.” Despite these promising examples, there have been few studies that investigated data mining applications in library systems (Huancheng et al., 2019). In addition, librarians and other information professionals have been slow to adapt these strategies.

Objectives

The objective of this study is to present a methodological article, illustrated by a case study, to demonstrate the varied applications of how this kind of research method can be used to inform library practices, services, and operations. The case study examines tweets on marijuana and the use of terminologies associated with it on Twitter in Canada during the time period from September to November 2018. The findings shed light on the nature, variety, frequency, sentiment, and the co-occurrence of terms used in relation to marijuana and legalization in the Canadian Twittersphere.

Methods

Marijuana as a Case Study

A key aspect of text mining is to gain a deeper terminological and conceptual insight into how Twitter is used as a mode of communication and transmission of competing and conflicting opinions, perspectives, and information. There is very little research that explores the vocabulary and variety of terminology used in relation to marijuana on social media platforms, particularly in a Canadian context. With the legalization of marijuana use in Canada on October 17, 2018 (Government of Canada, 2018), it offered a timely opportunity to examine how social media environments such as Twitter provide a platform for the movement of data, information, opinions, ideas, sentiments, and facts on such a contested topic. This kind of analysis provides evidence-based perspective of the ways in which the vocabulary and conversations on marijuana legalization are shaping up.

Marijuana legalization is a contested issue and few countries in the world have legalized marijuana for recreational and medical purposes. Some countries include Uruguay, Portugal, Jamaica, and a handful of American states such as California, Colorado, and Washington (Kalvapalle, 2017). There is a wealth of literature exploring public perceptions concerning marijuana legalization with individuals who are in support of or are opposed to this policy. Based on studies from the United States and Canada, pro-legalization interview respondents remark on the economic benefits such as increased tax revenue and reduced criminal justice involvement/costs (McGinty et al., 2016; Osborne & Fogel, 2017). Anti-legislation respondents argue that legal reform may potentially increase marijuana use among users and nonusers, and develop adverse health effects such as marijuana-impaired driving (McGinty et al., 2016). An effective way of capturing public opinion on marijuana use and marijuana legislation is through social media platforms, such as Twitter, where messages by the public are made publicly available.

To our knowledge, little research has explored terminology use, conversations, and public perceptions on marijuana legalization on Twitter in a Canadian context.

Building on previous research with open data, a research team member (KL) with programming and computing science background used Twitter API (Application Programming Interface) to extract Twitter data (tweets) for the time period September 1st to November 30th 2018. Social media tends to be more active during broadcasts and around important events and sensitive topics, therefore, this date range was chosen to provide a one-month window before and after the official marijuana legalization date of October 17, 2018, which optimized the chances of retrieving relevant tweets on the topic of interest. The data collected contains elements such as the tweets' timestamps, screen names, and tweet texts. We gathered tweets using a Compute Canada server under Dr. Geoffrey Rockwell at the University of Alberta using a combination of Twarc (https://github.com/DocNow/twarc), for scraping tweets from Twitter, and crontab (https://crontab.guru/), used to run Twarc in the background. Twarc is a command line tool and Python library for archiving tweets as JSON data. Crontab is a UNIX command designed to create a list of commands to be executed by the operating system at a specified time (i.e. 12-hour intervals in this case). Since the project required a great deal of technical expertise, this paper only provides a high-level overview of what the process entailed. We recommend interested readers to consult the following resources for step-by-step instructions on how to apply these tools for scraping Twitter data (Custodio, 2021; RapidAPI Staff, 2021; Twitter Developer Platform, 2022; UNLV University Libraries, n.d.), or connecting with IT professionals in their respective organizations.

Initially, the research team scraped tweets if they met the following criteria: a) tweets originating from Canada by using Twitter's geolocation information, and b) containing hashtags relevant for the purposes of the project such as #medicalcannabis. However, the first condition had to be removed since the geolocation was only available if Twitter users enabled the geolocation feature, which greatly limited the amount of Twitter data captured. In order to collect a representative sample of how marijuana was discussed on Twitter, the research team conducted several keyword searches, including cannabis, marijuana legalization, and Cannabis Act, to determine appropriate hashtags to use, based on popularity. We carefully selected hashtags related to marijuana that focused on the Canadian context such as Bills C-45 (the Cannabis Act) and C-46. Bill C-46

amended the Transportation Provisions of the Criminal Code by “strengthen[ing] existing drug-impaired driving laws” that related to the Cannabis Act (Government of Canada, n.d.). Alternatively, using hashtags such as #cannabis or #weed would have been too broad and unlikely to retrieve tweets in the Canadian context or about marijuana legalization. In the end, the search process resulted in the following hashtags to be included:

- #C45
- #C46
- #CannabisAct
- #CannabisCommunity
- #CdnPoli
- #LegalizeIt
- #PromiseKept

By scraping these particular hashtags we aim to narrow the scope to the Canadian context. However, it is still possible that there may be tweets that do not originate from Canada, but given the Canadian focused hashtags used, the context of the tweets will still be related to marijuana and Canada. For each tweet, we were concerned with the following elements: Twitter handle, screen name, tweet, replies, timestamp, retweet counts, and likes.

We made use of Python code to sort, analyze, and categorize the tweets, identify high frequency terms and themes within tweets, and conduct co-occurrence analyses. We utilized Voyant (https://voyant-tools.org/), an open access, web-based and visualization system, to conduct further text visual analysis of tweets. Preliminary analysis of tweets focused on high frequency terms used in relation to various hashtags and the contents of tweets, variety of terms used to refer to marijuana, and the terms and themes that co-occurred with marijuana and legalization.

The project team was also interested in the sentiment expressed by tweets on marijuana legalization. Sentiment analysis is a method to contextualize and quantify subjective information by assigning a quantitative value on the type of emotion expressed in the text: positive, neutral, or negative. We used the Natural Language Toolkit source code (NLTK, 2022), an open source software application, to build into the library in order to determine the sentiment of each tweet. On the toolkit website, they list the types of words that correspond to negative sentiment (e.g. can’t, aren’t, didn’t) and words that are associated with positive sentiment (e.g. absolutely, amazingly, incredibly, really). Sentiment scores range from +1 (positive) to -1 (negative), and a score of zero is neutral sentiment (neither positive nor negative).

Co-occurrence analysis evaluates how often two hashtags appear together in one post or one tweet. It may also apply to text strings instead of hashtags. This type of analysis adds value in its ability to visualize the text corpus and identify relationships between the entities, keywords, and their juxtaposition. It also provides a different type of insight as to how Twitter users conceptualized, understood, and discussed the legalization of marijuana in Canada.

Results

Within the three-month period (September to November 2018), more than 1,176,000 tweets were collected. We compiled hashtag frequency after removing retweets to eliminate an overrepresentation of the same hashtags. The most frequently used hashtag was #CdnPoli (n=107,389), followed by #CannabisAct (n=965), #C45 (n=422), #LegalizeIt (n=328), #CannabisCommunity (n=41), #C46 (n=32), and #PromiseKept (n=20). The most popular hashtag co-occurrence, two hashtags occurring together, was #cannabis and #CdnPoli with a total of 587 co-occurrences. In Table 1, #Canada was commonly used in conjunction with cannabis-related keywords and this strong correlation points to the prominence of this topic in the weeks leading up to and following the marijuana legalization date. The UN hashtag also appeared regularly alongside #C45, which stands for the United Nations.

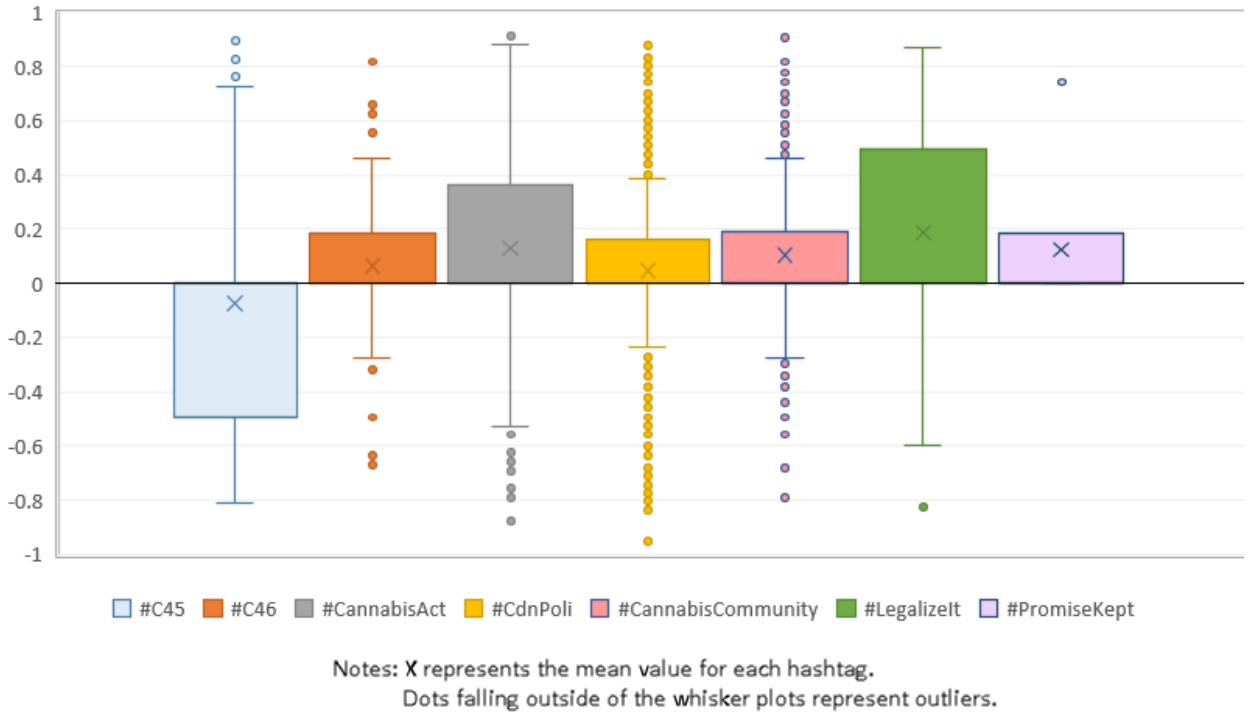
Table 1 – Frequency of Hashtag Co-occurrence

Hashtags	# of co-occurrences
#cannabis #CdnPoli	587
#CannabisAct #cannabis	199
#CdnPoli #grassroots	167
#Canada #CannabisCommunity	126
#Canada #LegalizeIt	93
#C45 #UN	87

The total number of sentiment scores obtained from each hashtag varied due to the fact that there were some hashtags used more frequently than others (e.g. #C45 had 422 sentiment scores compared to #C46, which received 32 sentiment scores). Figure 1 portrays the sentiment scores visualized in box plots. The graph highlights the data range with whiskers (vertical lines). Dots falling outside of the whiskers represent outliers and the X marked in each box is the mean value. Box plots normally have horizontal lines intersecting the box to differentiate between the first quartile and third quartile ranges, and the horizontal line represents the median value. However, the box plots presented are all missing these data points because the median scores for all six hashtags are zero. This means that the first quartile and median values are so

compressed that they fall where the line is zero. Most of the median and mode values were also zero, which indicate that most tweets were neutral and were not overly positive or negative in sentiment. The only noticeable differences are reflected in the hashtags #C45 and #LegalizeIt, where their box plots fall below the x-axis (in the case of #C45) and above the x-axis (#LegalizeIt). Therefore, #C45 has implied negative sentiment while #LegalizeIt leans more strongly towards positive sentiment.

Figure 1 – Sentiment Analysis of Hashtags Visualized in Box Plots



Notes: X represents the mean value for each hashtag.
Dots falling outside of the whisker plots represent outliers.

Table 2 highlights the proportion of tweets for each sentiment category. The colour scheme in the table provides a visual representation of the categories that had more tweets. The darker the colours, the higher the percentages of tweets that fell in that particular category. Most of the darker colours fell under the Neutral category, which confirms that most of the tweets had neutral sentiment or a sentiment score of zero.

Table 2 – Proportion of Tweets that Fall into Each Sentiment Category

	Positive	Neutral	Negative
#C45	19%	47%	34%
#C46	25%	59%	16%
#CannabisAct	35%	56%	9%
#CdnPoli	31%	47%	22%
#CannabisCommunity	27%	66%	7%
#LegalizeIt	47%	45%	8%
#PromiseKept	17%	83%	0%

Three hashtags, #CannabisCommunity, #LegalizeIt, and #PromiseKept, were used more broadly so they were not all restricted to the Canadian context. For instance, many tweets were from America or they were not related to marijuana legalization at all. Due to the vagueness and ambiguity of these hashtags, there were other reasons why users used them. For example, politicians or businesses used #PromiseKept in their tweets to promote the introduction of new legislation for small business tax cuts or marketing campaigns. While we made efforts to filter these tweets so that they were more "Canadian focused," it became too difficult to automate a system in order to screen out irrelevant tweets. As a result, we conducted further text analysis by excluding tweets with these hashtags since they were deemed to be irrelevant, and included tweets pertaining to marijuana legalization in Canada (i.e. only using hashtags directly related to Canada marijuana legalization - #C45, #C46, and #CannabisAct). Taking the tweets with the remaining three hashtags, we created a word cloud using a corpus that included 209,626 total words and 13,689 unique word forms. After some data cleaning to remove strange symbols and RT (i.e. Retweets), they were uploaded into Voyant to create a word cloud with 45 items (Figure 2). The most frequent words that Voyant detected were *cannabis* (n=5470), *cannabisact* (n=3570), *canada* (n=2576), *govcanhealth* (n=1158), and *legalization* (n=1121). The larger the words in the word cloud, the more frequently they were mentioned in the corpus. In addition to the top five most frequent words listed, other prominent terms include *c45, legal, marijuana, new,* and *jodieemery.* Jodie Emery is a cannabis rights activist and politician (https://jodieemery.ca/), which provides some insight as to why Twitter users commonly mentioned her on Twitter during that timeframe.

Figure 2 – Word Cloud of Tweets from Marijuana Legalization in Canada

Discussion

The findings of this study offer interesting insights on how Twitter users discussed marijuana in the months leading up to and following marijuana legalization in Canada through hashtag frequency counts, hashtag co-occurrence, sentiment analysis, and text analysis. Frequent word counts in the semantic analysis are useful for libraries to identify popular words to support decision-making and strategic planning (Al-Daihani & Abrahams, 2016). For example, researchers in one study exploring academic libraries' tweets identified the most frequent bi-gram (two-word sequence such as "special collections") and tri-gram terms (three-word sequence such as "save the date") (Al-Daihani & Abrahams, 2016), which offer compelling information on what library user groups are discussing. Hashtag co-occurrences helped show relationships between words that may be difficult to capture through other means. Based on the number of pairings identified from hashtag co-occurrence, there are reasonable explanations on why users commonly used the hashtags together, without the need to analyze the tweet for context. For example, #cannabis and #CdnPoli were two hashtags that had the highest frequency of co-occurrence, since it was a timely and newsworthy topic readily discussed in mainstream media. As the data was captured approximately one month prior to and after the marijuana legalization date, this study clearly shows that the topic was on top of mind of many people including those who are active on social media. One interesting hashtag pair, #C45 and #UN, required more explanation since there was no apparent immediate connection on why Twitter users were talking about Bill C-45 and the United Nations. Upon further exploration, we observed that the popularity of #UN stems from the fact that cannabis legalization violates three United Nations drug-control treaties: the 1961 Single Convention on Narcotic Drugs, the 1971 Convention on Psychotropic Substances, and the 1988 Convention against Illicit Traffic in Narcotic Drugs and Psychotropic Substances (Cain, 2018; Macnab, 2018). Several decades

ago, Canada pledged to ban multiple illicit drugs, including marijuana, and legalizing marijuana would have been a direct contradiction to the drug treaties. Therefore, the UN treaties became a heated debate on social media. The value of analyzing hashtag co-occurrence networks is evidenced in its ability to determine the "relative prominence of individual hashtags" (Wang et al., 2016, p. 854) and how they relate to one another. The UN connection, for instance, reveals a unique perspective of how Twitter users discussed marijuana legalization, which was discovered through hashtag co-occurrence.

Text and sentiment analysis are particularly useful methods for social media analytics as they provide an overview of trending discussions and topics that are emerging among social media users. Our sentiment analysis of marijuana tweets suggests that users did not have strong positive or negative feelings on the subject matter even though it was a highly controversial topic and heavily debated in parliament for decades prior to passing in legislation (Gligorijević, 2020). There are many text mining applications to health sciences research such as providing policy makers with discourse directly from the public on how they talk about controversial topics and share information. However, the focus of this paper will be on data mining approaches in the library setting.

Text Mining Applications to Library Practice

Text mining covers a diverse range of analytics including text and sentiment analysis, and text visualization. These strategies are among the very useful techniques and research methods that many LIS professionals and social scientists use. In the following, we provide an overview of some of the related research that provide ideas and approaches to apply these techniques in various contexts.

Big data and data mining projects are gaining traction in many industries. Libraries use data-driven assessment in a number of areas including "financial expenditures, such as serials, interlibrary loan, and online purchases, to other areas of library service" (Travis & Ramirez, 2020, p. 34). While several data mining techniques exist, one interesting classification approach for libraries involves a framework that includes four quadrants: service analysis, usage analysis, quality analysis, and collection analysis (Siguenza-Guzman et al., 2015). One of the ways to assess library service is determining the relationship between library instruction and how library workshops may increase information literacy skills and student learning. This may be achieved by analyzing course generated data and student grades (Travis & Ramirez, 2020). In another study, researchers used Google Analytics' Event Measurement to improve user experience on the library website (Vecchione et al., 2016). Event Measurement (formerly called event tracking) is a feature that identifies how users navigate to and from certain web pages, and this type of analytics provides insights on users' movements through a particular website (Google Analytics, n.d.). Vecchione and colleagues (2016) found that users preferred to use only a limited number of tools and links on the library website, which was a strong indication that the site needed to be revised to optimize the user experience. Text mining is a subset of data mining and the Twitter marijuana case study illustrates how this methodology may be applied to support

librarians' regular practice. The primary benefit of using tweets or other social media platforms as the text corpus over other text formats (e.g. blog posts, newspapers, scholarly articles), is the ability to spot trends, commentary, and public opinion from the community on a large scale. By applying the same classification framework, there are multiple ways in which a Twitter text mining methodology may be used for analysis in libraries; we will focus on three of the four quadrants.

Service Analysis

Academics conduct evidence synthesis projects on a regular basis and library professionals have the searching expertise to provide research support in this area by way of using information retrieval systems. Several publications demonstrated the feasibility of pulling references or identifying words and subject headings from database records as the text corpus to speed up the search process (CADTH, 2018; McGowan, 2021). This study presents the possibilities of identifying additional terms that may increase search sensitivity through Twitter. For instance, if a researcher were interested in conducting a comprehensive literature review or grey literature search on marijuana legalization, researchers may not have considered searching for the name Jodie Emery. Analytical tools such as Voyant help to analyze and visualize the text corpus in a word cloud where the more prominent words become immediately noticeable. Even though many of the prominent terms in this particular word cloud do not convey novel information, visualizing text in this manner may still add value to text mining projects. In another project evaluating the frequency of Twitter posts related to health technology, the co-authors created a word cloud with the following dominant words: 'care,' 'new,' 'mental,' and 'fitness' (Lee et al., 2019). The popular words they found did not appear original or novel, similar to our case study, but reviewing the entire visualization holistically may offer other useful explanations or observations. Word clouds also do not require a lengthy amount of time to develop, thus providing a good rationale to include them in text mining projects. Figure 2 took mere minutes to create in Voyant. While text mining tools may enhance the search process for librarians and other information professionals, this method alone is not sufficient to identify all relevant terms that would describe the same concept (O'Keefe et al., 2022; Thomas et al., 2011).

Quality Analysis

Libraries may assess the quality of their services by surveying patrons in the traditional sense (e.g. questionnaires) but this can be a labour intensive process. Text mining, including sentiment analysis, is an alternative source to extract user data based on their posts on social media platforms. Additionally, sentiment analysis is a data mining method readily used in LIS work (Katsurai & Joo, 2021). The sentiment analysis in the marijuana case study did not have overly strong positive or negative sentiment, which showed that users did not feel strongly either way about marijuana legalization in Canada. Expanding on previous library Twitter studies concerning patron outreach (Al-Daihani & Abrahams, 2016), libraries are able to analyze tweets about the library or pose questions on Twitter to find out how users feel about their library by asking, "Tell us what you think about our library," "Where is your favorite place to study," or "What can we do to serve you better?" In a study looking at library social media engagement, one public library in California mentioned that social media acted like "a thermometer on

what people know, want to know, and don't know about their services" (Wardell & Kelly, 2022, p. 102), which helps shape library spaces, programming, and services. In addition to using content analysis for identifying themes, sentiment analysis may determine how users feel about library programming or initiatives. Negative tweets may be analyzed further to identify potential changes for the library to act upon. Content analysis and sentiment analysis are cost effective ways to learn users' perceptions about the library without conducting surveys or questionnaires.

Collection Analysis

The 45 items presented in the word cloud presents an interesting perspective on the terminologies used by the public on Twitter, providing a snapshot of the most frequently used and popular terms. By excluding the obvious terms, *cannabis* and *cannabisact*, users are also describing cannabis with words such as *cbd, cannabidiol,* and *pot.* This type of finding may extend knowledge on how marijuana terminologies are evolving and emerging in a subset of the population via the Twittersphere. Using a user-centred approach, we can identify terminologies to include in taxonomies. This may subsequently increase discovery in databases, digital libraries, or repositories. Some evidence suggests that existing taxonomies developed by other organizations are not as reliable for supporting the creation of new controlled vocabularies and taxonomies for digital libraries. While creating the taxonomy for key terms in a Mathematics Education digital repository, the authors found that they were unable to use taxonomies previously created by UNESCO and ERIC due to their limitations (Gómez & Cañadas, 2013). They also evaluated the thesaurus in a mathematics education database and found that it came up short since they did not apply specifically to mathematics education (Gómez & Cañadas, 2013). While they developed their own taxonomy by using a curricular approach for describing items and consulting international experts, consulting social media and message board posts on mathematics education could have provided other sources of terms to supplement the terminologies in the repository. The variety of terms found in analyzed tweets may be used to enhance and diversify knowledge organization systems on certain topics by providing different access points in digital libraries, databases, and digital repositories.

Text mining strategies are not without their limitations and challenges. Due to the nature of the work on how the corpus is collected, a substantial amount of programming knowledge is required. One study interviewed library staff about big data in academic libraries and they expressed concerns about using advanced technologies for analytical work including analytics tools, visualizations, and data curation (Hamad et al., 2022). Therefore, the challenges with staff competency using text mining methodologies cannot be overlooked. If librarians and other information professionals do not have a computing science or programming background, this can be a steep learning curve but not impossible. Furthermore, the availability of numerous web-based tools and tutorials provide an opportunity for library and information professionals to self-learn and develop data analysis and visualization techniques in order to be able to apply those techniques to inform and improve their current library practices, services, and operations. Examining tweets can only go so far; we did not analyze emojis, which can have a vital impact on the meaning and sentiment behind messages, such as with sarcasm. The sentiment analysis demonstrated in this case study did not identify irony or cases where users might have used negative or positive words sarcastically. Since sentiment

analysis does not interpret emojis or humour in the English language, there is a potential margin of error on how the sentiment scores were calculated.

Conclusions

While there is literature to indicate the importance of combining both information retrieval and text mining capabilities to information systems (Liddy, 2000), text mining in LIS is not commonly practiced. The marijuana legalization Twitter project presents a case study on the practical applications of using text mining and visualization as a methodology to inform and improve library practices, services, and operations. This methodology has several practical applications from analyzing library services, assessing quality, to developing taxonomies, and information literacy instruction. While this study provided possibilities on how Twitter analytics may be used in theory, the analytical methods reported in this study could be utilized in future projects where new digital libraries, archives, or repositories are to be developed. Other text mining applications in library services contexts may also be explored, for instance, gauging the sentiment of how patrons feel about academic library services during the COVID-19 pandemic. Meaningful insights may be gathered to improve communications, help in evidence synthesis projects, and service delivery.

References

Aamodt, M., Huurdeman, H., & Strømme, H. (2019). Librarian co-authored systematic reviews are associated with lower risk of bias compared to systematic reviews with acknowledgement of librarians or no participation by librarians. In *Evidence Based Library & Information Practice, 14*(4), 103–127. https://doi.org/10.18438/eblip29601

Al-Daihani, S. M., & Abrahams, A. (2016). A text mining analysis of academic libraries' tweets. *Journal of Academic Librarianship, 42*(2), 135–143. https://doi.org/10.1016/j.acalib.2015.12.014

CADTH. (2018). Text mining opportunities: White paper [White paper]. Ottawa. https://www.cadth.ca/sites/default/files/pdf/methods/2018-05/MG0013_CADTH_Text-Mining_Opportunites_Final.pdf

Cain, P. (2018). *Marijuana ban will stay in UN treaties — for now*. https://globalnews.ca/news/4739793/marijuana-ban-un-treaty/

Cavazos-Rehg, P. A., Krauss, M., Fisher, S. L., Salyer, P., Grucza, R. A., & Bierut, L. J. (2015). Twitter chatter about marijuana. *Journal of Adolescent Health*, *56*(2), 139–145. https://doi.org/10.1016/j.jadohealth.2014.10.270

Custodio, A. (2021). *How to check if a Cron Job has run (Crontab Log).* https://www.inmotionhosting.com/support/edu/cpanel/did-cron-job-run/

Dai, H., & Hao, J. (2017). Mining social media data on marijuana use for Post Traumatic Stress Disorder. *Computers in Human Behavior*, 70, 282–290.

https://doi.org/10.1016/j.chb.2016.12.064
Fodeh, S., & Zeng, Q. (2016). Mining big data in biomedicine and health care. *Journal of Biomedical Informatics*, 63, 400–403. https://doi.org/10.1016/j.jbi.2016.09.014
Gligorijević, N. (2020). Legalization of cannabis in Canada. *CRIMEN: Casopis Za Krivicne Nauke*, 2020(1), 79–86.
Gómez, P., & Cañadas, M. C. (2013). Development of a taxonomy for key terms in mathematics education and its use in a digital repository. In *Library Philosophy & Practice* (pp. 1–9). https://digitalcommons.unl.edu/libphilprac/903/
Google Analytics. (n.d.). *Event Measurement.* https://developers.google.com/analytics/devguides/collection/analyticsjs/events
Government of Canada. (2018). *Taking stock of progress: Cannabis legalization and regulation in Canada.* https://www.canada.ca/en/health-canada/programs/engaging-cannabis-legalization-regulation-canada-taking-stock-progress/document.html
Government of Canada. (n.d.). *Legislative Background: reforms to the Transportation Provisions of the Criminal Code (Bill C-46)* https://www.justice.gc.ca/eng/cj-jp/sidl-rlcfa/c46/p3.html
Gruzd, A., & Roy, J. (2014). Investigating political polarization on twitter: A Canadian perspective. *Policy and Internet*, *6*(1), 28–45. https://doi.org/10.1002/1944-2866.POI354
Hamad, F., Fakhuri, H., & Abdel Jabbar, S. (2022). Big data opportunities and challenges for analytics strategies in Jordanian academic libraries. *New Review of Academic Librarianship*, *28*(1), 37–60. https://doi.org/10.1080/13614533.2020.1764071
Hersh, W. R. (2020). *Information retrieval : a biomedical and health perspective* (4th ed.). Springer.
Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. (2021). *The Cochrane handbook for systematic reviews of interventions version 6.2*. https://training.cochrane.org/handbook/current/chapter-04
Huancheng, L., Tingting, W., & Rocha, Á. (2019). An analysis of research trends on data mining in Chinese academic libraries. *Journal of Grid Computing: From Grids to Cloud Federations*, *17*(3), 591–601. https://doi.org/10.1007/s10723-018-9461-3
Ignatow, G., & Mihalcea, R. (2018). *An introduction to text mining: research design, data collection, and analysis.* SAGE Publications, Inc.
Joo, S., Lu, K., & Lee, T. (2020). Analysis of content topics, user engagement and library factors in public library social media based on text mining. *Online Information Review, 44*(1), 258–277. https://doi.org/10.1108/OIR-11-2018-0345
Kalvapalle, R. (2017). *Weed around the world: what legal marijuana looks like in other countries*. https://globalnews.ca/news/3378603/marijuana-laws-around-the-world/
Katsurai, M., & Joo, S. (2021). Adoption of data mining methods in the discipline of Library and Information Science. *Journal of Library and Information Studies*, *19*(1), 1–17. https://doi.org/10.6182/jlis.202106_19(1).001
Lamy, F. R., Daniulaityte, R., Carlson, R. G., Sheth, A., Nahhas, R. W., Martins, S. S., & Boyer, E. W. (2016). “Those edibles hit hard”: Exploration of Twitter data on cannabis edibles in the U.S. *Drug and Alcohol Dependence*, 164, 64–70. https://doi.org/10.1016/j.drugalcdep.2016.04.029
Lee, J., Kim, J., Hong, Y. J., Piao, M., Byun, A., Song, H., & Lee, H. S. (2019). Health information technology trends in social media: Using Twitter data. *Healthcare Informatics Research*, *25*(2), 99–105. https://doi.org/10.4258/hir.2019.25.2.99
Liddy, E. (2000). Text mining. *Bulletin of the American Society for Information Science &*

Technology, 27(1), 13–14. https://doi.org/10.1002/bult.184
Macnab, A. (2018). *International law experts say legal cannabis defies Canada's UN treaty obligations*. https://www.canadianlawyermag.com/news/general/international-law-experts-say-legal-cannabis-defies-canadas-un-treaty-obligations/275627
McGinty, E. E., Samples, H., Bandara, S. N., Saloner, B., Bachhuber, M. A., & Barry, C. L. (2016). The emerging public discourse on state legalization of marijuana for recreational use in the US: Analysis of news media coverage, 2010-2014. *Preventive Medicine*, 90, 114–120. https://doi.org/10.1016/j.ypmed.2016.06.040
McGowan, B. S. (2021). Using text mining tools to inform search term generation: An introduction for librarians. *Portal: Libraries & the Academy*, *21*(3),603–618. https://doi.org/10.1353/pla.2021.0032
Meert, D., Torabi, N., & Costella, J. (2016). Impact of librarians on reporting of the literature searching component of pediatric systematic reviews. *Journal of the Medical Library Association*, *104*(4), 267–277. https://doi.org/10.3163/1536-5050.104.4.004
Milward, D., & Singh, G. (2012). Clarifying the social media blur. In *Information Outlook, 16*(20), 10–13. https://www.linguamatics.com/publications/clarifying-social-media-blur
NLTK. (2022). *Source code for nltk.sentiment.vader*. https://www.nltk.org/_modules/nltk/sentiment/vader.html
Odlum, M., & Yoon, S. (2015). What can we learn about the Ebola outbreak from tweets? *American Journal of Infection Control, 43*(6), 563–571. https://doi.org/10.1016/j.ajic.2015.02.023
O'Keefe, H., Rankin, J., Wallace, S. A., & Beyer, F. (2022). Investigation of text-mining methodologies to aid the construction of search strategies in systematic reviews of diagnostic test accuracy-a case study. *Research Synthesis Methods*. July 31, 1–20 https://doi.org/10.1002/jrsm.1593
Osborne, G. B., & Fogel, C. (2017). Perspectives on cannabis legalization among Canadian recreational users. *Contemporary Drug Problems, 44*(1), 12–31.
RapidAPI Staff. (2021). *How to use the Twitter API in 4 easy steps [Tutorial]* https://rapidapi.com/blog/how-to-use-the-twitter-api/
Sandhu, G. (2015). Re-envisioning library and information services in the wake of emerging trends and technologies. *2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services, ETTLIS 2015 - Proceedings*, 153–160. https://doi.org/10.1109/ETTLIS.2015.7048190
Sarrouti, M., & Ouatik El Alaoui, S. (2017). A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering. *Journal of Biomedical Informatics*, 68, 96–103. https://doi.org/10.1016/j.jbi.2017.03.001
Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control, 38*(3), 182–188. https://doi.org/10.1016/j.ajic.2009.11.004
Schellinger, J., Sewell, K., Bloss, J. E., Ebron, T., & Forbes, C. (2021). The effect of librarian involvement on the quality of systematic reviews in dental medicine. *PLoS ONE*, *16*(9), e0256833. https://doi.org/10.1371/journal.pone.0256833
Sewell, R. R. (2013). Who is following us? Data mining a library's Twitter followers. *Library Hi Tech, 31*(1), 160–170. https://doi.org/10.1108/07378831311303994
Sharma, M., & Patel, R. (2013). A survey on information retrieval models, techniques and applications. *International Journal of Emerging Technology and Advanced*

Engineering, 3(11), 542–545.
Siguenza-Guzman, L., Saquicela, V., Avila-Ordóñez, E., Cattrysse, D., & Vandewalle, J. (2015). Literature review of data mining applications in academic libraries. *Journal of Academic Librarianship*, *41*(4), 499–510. https://doi.org/10.1016/j.acalib.2015.06.007
Stewart, B., & Walker, J. (2018). Build it and they will come? Patron engagement via Twitter at historically Black College and University Libraries. *Journal of Academic Librarianship, 44*(1), 118–124. https://doi.org/10.1016/j.acalib.2017.09.016
Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science & Technology, 62*(2), 406–418. https://doi.org/10.1002/asi.21462
Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods, 2*(1), 1–14. https://doi.org/10.1002/jrsm.27
Travis, T. A., & Ramirez, C. (2020). Big data and academic libraries: The quest for informed decision-making. *Portal: Libraries & the Academy*, *20*(1), 33–47. https://doi.org/10.1353/pla.2020.0003
Twitter Developer Platform. (2022). *Getting started: About the Twitter API.* https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api
UNLV University Libraries. (n.d.). *Twitter data collection using Twarc*. https://www.library.unlv.edu/sites/default/files/inline-images/fy9jEgQjTudzio3zy68UnbyPBnNJBBJp1kmnPlVR5KaFiv2QTq.pdf
Vecchione, A., Brown, D., Allen, E., & Baschnagel, A. (2016). Tracking user behavior with Google Analytics Events on an academic library web site. *Journal of Web Librarianship*, *10*(3), 161–175. https://doi.org/10.1080/19322909.2016.1175330
Wang, R., Liu, W., & Gao, S. (2016). Hashtags and information virality in networked social movement: Examining hashtag co-occurrence patterns. *Online Information Review*, *40*(7), 850–866. https://doi.org/10.1108/OIR-12-2015-0378
Wardell, J. & Kelly, K. (2022). Doing more with a DM: A survey on library social media engagement. *Evidence Based Library and Information Practice, 17*(3). https://doi.org/10.18438/eblip30141