


## Research and Applications

# A platform for connecting social media data to domain-specific topics using large language models: an application to student mental health

Leonard Ruocco, PhD<sup>\*.1,2</sup>, Yuqian Zhuang, MDS<sup>1,2</sup>, Raymond Ng, PhD<sup>1,3</sup>,  
Richard J. Munthali, PhD<sup>2</sup>, Kristen L. Hudec, PhD<sup>2</sup>, Angel Y. Wang , MPhil<sup>2</sup>,  
Melissa Vereschagin, BSc<sup>2</sup>, Daniel V. Vigo, Lic Psych, MD, PhD<sup>2,4</sup>

<sup>1</sup>Data Science Institute, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada, <sup>2</sup>Department of Psychiatry, University of British Columbia, Vancouver, British Columbia V6T 2A1, Canada, <sup>3</sup>Department of Computer Science, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada, <sup>4</sup>Department of Global Health and Social Medicine, Harvard University, Boston, MA 02115, United States

\*Corresponding author: Leonard Ruocco, PhD, Data Science Institute, University of British Columbia, 6339 Stores Road, Vancouver, British Columbia V6T 1Z4, Canada (leon.ruocco@ubc.ca)

## Abstract

**Objectives:** To design a novel artificial intelligence-based software platform that allows users to analyze text data by identifying various coherent topics and parts of the data related to a specific research theme-of-interest (TOI).

**Materials and Methods:** Our platform uses state-of-the-art unsupervised natural language processing methods, building on top of a large language model, to analyze social media text data. At the center of the platform's functionality is BERTopic, which clusters social media posts, forming collections of words representing distinct topics. A key feature of our platform is its ability to identify whole sentences corresponding to topic words, vastly improving the platform's ability to perform downstream similarity operations with respect to a user-defined TOI.

**Results:** Two case studies on mental health among university students are performed to demonstrate the utility of the platform, focusing on signals within social media (Reddit) data related to depression and their connection to various emergent themes within the data.

**Discussion and Conclusion:** Our platform provides researchers with a readily available and inexpensive tool to parse large quantities of unstructured, noisy data into coherent themes, as well as identifying portions of the data related to the research TOI. While the development process for the platform was focused on mental health themes, we believe it to be generalizable to other domains of research as well.

## Lay Summary

We present a novel artificial intelligence-platform that allows researchers to study large, unstructured and incoherent bodies of text using state-of-the-art natural language processing (NLP) tools. Our platform uses unsupervised NLP methods to structure the text as well as modern large-language models to understand whole sentences as opposed to individual words. With this platform, researchers can investigate a chosen theme-of-interest (TOI), identifying portions of the text related to their specific theme as well as other topics and themes that are correlated with their TOI. Mental health in the student population is a common research interest and we demonstrate the functionality of our platform through 2 case studies, in which we identify themes related to depression within text from student social media Reddit data. We also report on secondary topics of discussion correlated with the TOI, which offer insights into the context behind the detected depression-related themes.

**Key words:** natural language processing; artificial intelligence; mental health; topic modeling; social media.

## Introduction

In an increasingly digitized world, online social media sites have become spaces where many people choose to share ideas, voice concerns, and express their emotions. This provides a valuable source of textual data that can be studied for the purposes of healthcare research<sup>1</sup> and of better understanding the mental health needs within certain populations.<sup>2–5</sup> Identifying predominant themes within shared online content is, therefore, increasingly important. In particular, universities may benefit from greater awareness around emergent mental health topics within the social media space, as young adults enrolled in university may be especially at-risk. Young adults have a high prevalence

of mental health and substance use disorders<sup>6,7</sup> and a low propensity for seeking traditional mental health resources,<sup>8,9</sup> opting often for social media sites as a source of peer-to-peer support. As such, timely review of social media data is crucial, yet manual methods of coding are not feasible, given the sheer magnitude of data and the staggering amount of time that would be required for human review. Methods that provide a faster and more easily adaptable mechanism for exploring relevant research topics within social media sites—and other large text data—are therefore needed.

We have developed a software platform that utilizes state-of-the-art unsupervised natural language processing (NLP)

Received: November 1, 2023; Revised: December 20, 2023; Editorial Decision: December 29, 2023; Accepted: January 5, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the American Medical Informatics Association.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

methods as well as large-language models (LLMs) to investigate research topics/themes of interest.<sup>1</sup> Research has shown that identifying mental health-related signals on social media using machine learning<sup>10–12</sup> and NLP<sup>13,14</sup> is possible. More recently, deep learning methods have been used to great effect<sup>15–18</sup>; however, these studies have mostly been restricted to supervised learning tasks involving prediction and classification. On the other hand, the unsupervised domain of NLP offers the opportunity to study large volumes of unstructured and unlabeled text data by identifying patterns and structures, though this form of analysis has seen far less attention within the field of mental health informatics. Our platform incorporates these methods and thus has the ability to analyze large quantities of text data and to identify signals related to a specific theme-of-interest (TOI) by comparing the primary input text data to a secondary source of domain-specific text data. This approach provides the researcher with greater flexibility, as the TOI is chosen by the researcher and can be formed from any collection of documents that they deem appropriate to describe their TOI.

As a primary input data source into the platform, social media data captures a wealth of information but can be extremely noisy, and extracting coherent signals from the data is challenging. Our approach is to cluster the data into various coherent themes (eg, topics of discussion). Specifically, we use a popular unsupervised text processing method known as topic modeling (TM), which uses statistical algorithms to uncover latent semantic structure in the form of topics—distributions of correlated words that capture the salient themes in the data.<sup>19–22</sup>

Some recent studies have applied TM to social media data within the context of mental health, providing researchers with a valuable tool for understanding the experiences of those living with certain conditions as well as helping researchers in developing mental health interventions.<sup>23</sup> Others have studied online discourse using TM to better understand psychosocial stressors during the coronavirus disease 2019 (COVID-19) pandemic.<sup>24</sup> However, these studies have mostly been restricted to either certain mental health conditions or specific public health events and utilized TM methods such as Latent Dirichlet allocation, which are outperformed in the modern era of LLMs. A recent study implemented LLM-based methods to successfully detect depression-related signals in social media data, then used TM to validate the model performance,<sup>25</sup> highlighting the promising applications of these methods.

Our platform incorporates recent advances in powerful pre-trained LLMs based on the transformer architecture of neural networks.<sup>26</sup> These methods have been extended to the sentence-level where the corresponding embeddings are able to capture the semantic meaning of large sentence structures.<sup>27</sup> We leverage these models to find signals within the data pertaining to our TOI. This is done by embedding both the primary input text data—and the secondary input data describing our TOI—in a high-dimensional vector space<sup>28,29</sup> and then comparing them mathematically.

The LLM-based embedding of both the primary and secondary text data sources is a key step. The vector space, to which the input text is projected onto, contains a rich degree of semantic context for each of the embeddings provided by the highly complex LLM models which are trained on vast corpora of text data. Furthermore, through our word-to-sentence mapping functionality, our platform makes a key

innovative step and creates embeddings of whole sentences from both the primary and secondary data sources. This produces embedding vectors with much more informational content, with regards to natural-language understanding, since a sentence conveys far greater meaning than its unstructured collection of constituent words.<sup>30,31,27</sup>

Another key functionality of our platform is its ability to structure and organize the data, at several stages of the analysis, such that TOIs can be related to various other concepts and themes within the data, offering potential insight into which topics are highly correlated with the TOI. In conjunction with our word-to-sentence mapping functionality, these related topics can be understood at the sentence-level, providing far more context behind the discussion themes and their potential relationship to the TOI.

Motivated to better understand the mental health needs of Canadian university students, we include 2 case studies to demonstrate how the platform can be used. Specifically, we analyze dynamic trends across time as well as location-based comparisons of social media site data with depression as our TOI. In doing so, we can identify clear mental health-related signals and their potential relationship to various other themes emergent in the data. While our case studies and selected TOI relate to mental health, these examples are chosen to demonstrate the application of the platform, but the platform is not limited to this research domain.

The article is structured as follows: we begin by introducing the platform, outlining the various stages of analysis and their functionality. We then apply the platform to 2 mental health case studies with social media site data and discuss the corresponding results. Finally, we briefly discuss other areas of potential application for the platform.

## Methods

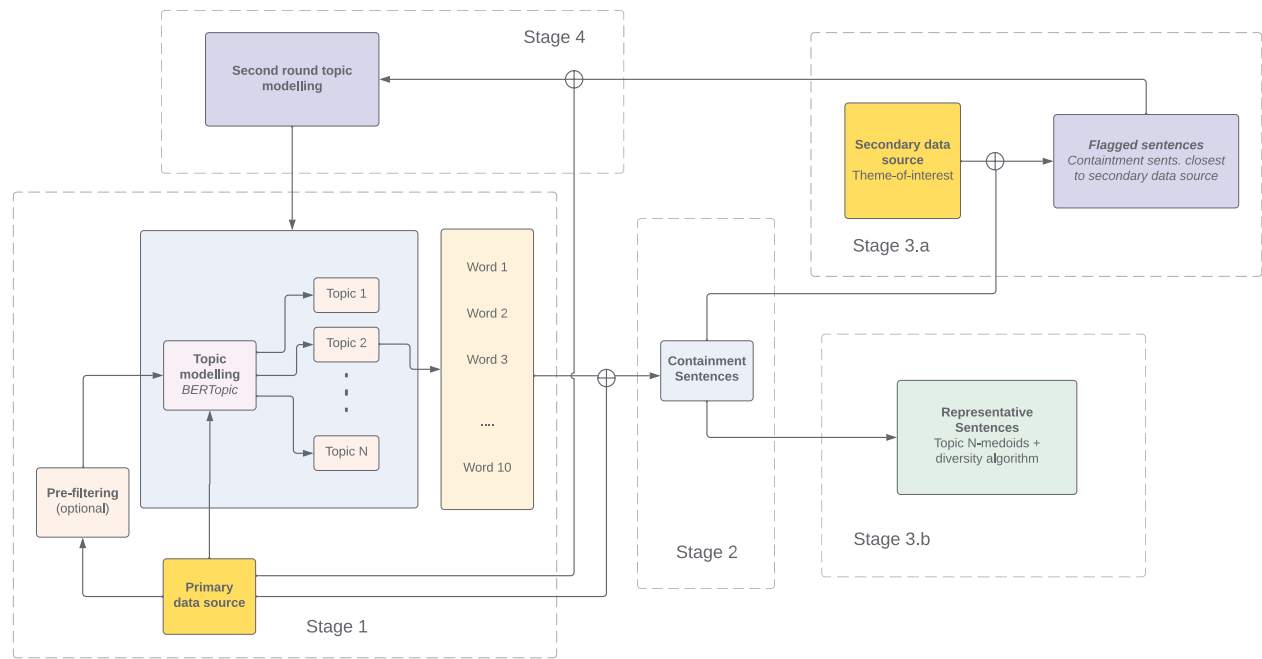
We present a platform, written in the Python programming language, which accepts various text-based data inputs and analyzes them using a combination of pre-trained and unsupervised NLP models. In [Figure 1](#), we outline the data processing and analysis workflow of our platform in stages 1-4.

### Data sources

The platform accepts 2 input sources (both indicated in yellow): the primary input text data intended to be analyzed and a secondary source of input text data to which the primary data is compared. It is the secondary input data source that. The primary input data can be large and potentially very “noisy” due to numerous topics of discussion with varying degrees of prominence. The secondary input data source forms the basis for a TOI and, for example, may contain chapters from canonical textbooks from the field of study or news articles covering aspects of the theme.

The primary input text data sources for our case studies are social media sites (ie, Reddit) specific to 2 universities: referred to as University A and University B, located in western and eastern Canada, respectively. The data were extracted for the period of interest (February 09, 2020 to August 15, 2022) using the Pushshift and Praw APIs. Specifically, the corpus was formed from the text bodies of all social media posts including comments, replies, and excluding titles or any metadata.

For our secondary input data source, we sought reference publications containing a rigorous description of the



**Figure 1.** Visual schematic of platform including text processing and analysis stages.

depressive syndrome that also avoided excessively technical language, reflecting a layperson’s description more closely. After careful consideration, we chose a guide by the Centre for Addiction and Mental Health (CAMH), a leading Canadian knowledge translation center, intended for people living with depression, their families, and anyone wanting to understand the basics of this illness and its treatment and management. Specifically, we used the chapter “Understanding Depression” from “Depression: An informative guide.”<sup>32</sup>

*Optional pre-filtering step:* An optional pre-filtering step is available at the beginning of stage 1 to generate a subset of the raw primary input data related to a given theme (not to be confused with the TOI, which can be a completely distinct theme). The corpus is filtered by a cosine-similarity threshold between embeddings formed from a user-generated list of words pertaining to the pre-filtering theme, and embeddings of words spanning the entire corpus. The word embeddings for this optional stage are produced using the GloVe pre-trained model<sup>33</sup> due to its scalability given the potential large size of the primary input data, and the large computational resources required to embed the entire dataset. As an example of this functionality, we apply pre-filtering in case study II of the Results section. Our pre-filtering theme pertains to “housing,” with a corresponding list of words chosen related to housing, and the TOI remains the “depression” theme.

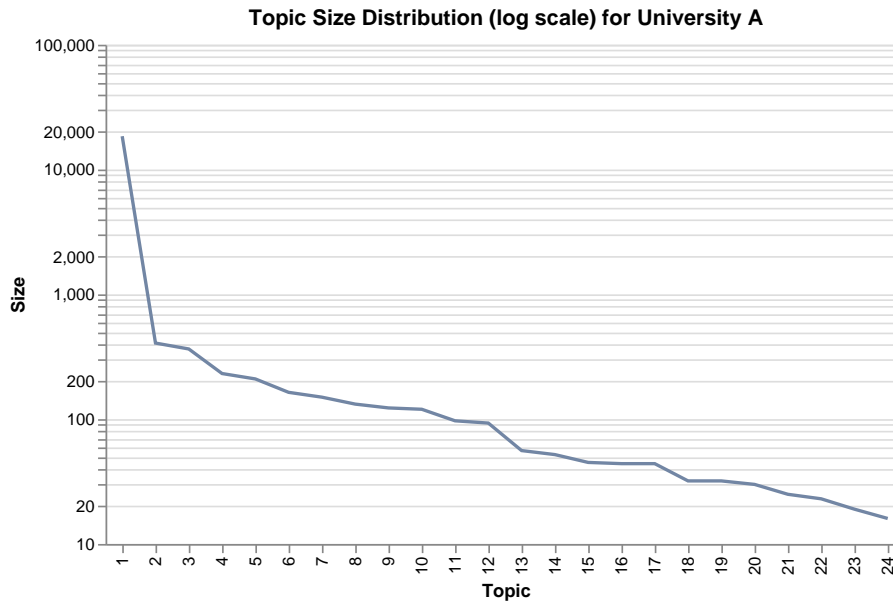
*Stage 1:* The first stage is designed to structure the primary input text data by performing TM, thereby clustering the data into various coherent topics of discussion. For this, we use BERTopic, a TM technique based on the BERT LLM,<sup>34</sup> that leverages transformer neural networks at the sentence embedding level. The corresponding clusters found by the model are highly coherent collections of words, forming the basis for each topic. Table S1 shows an example of 9 topics identified by BERTopic on University A 2020 Reddit data, which is discussed later. Each topic is characterized by 10 topic words. In general, a word may be one of the topic descriptions for multiple topics.

*Stage 2:* Although BERTopic outputs topic descriptors as bags of words, this method is well-known to lack context,<sup>28,35</sup> and we seek to incorporate more contextual information with sentences. In order to utilize sentence-level embeddings for the subsequent stages of the analysis, we find all “containment sentences” for each topic word. A containment sentence is defined as any sentence—drawn from the subset of documents for each topic—that contains a given topic word. This produces a one-to-many mapping of words to sentences for each topic.

Next, we use pre-trained LLMs to embed the text-based sentences. For this, we use SBERT: a BERT-based transformer that produces semantically meaningful sentence embeddings using siamese and triplet network structures.<sup>27</sup> The topics are output in ascending order according to their prevalence. Figure 2 provides a plot of the topic size—the number of paragraphs assigned to each topic—for each topic number. As each individual paragraph can only be assigned to one topic, this provides an indication as to how prevalent each topic is within the corpus. As shown in Figure 2, the first topic (Topic 1) is an order of magnitude bigger than all other topics and is very general in content. The first topic often lacks specificity and is not relevant to the TOI; the next stage of the analysis is designed to identify relevant topics and sentences.

*Stage 3:* The next stage is split into 2 parallel stages to analyze the “containment sentences” produced by stage 2. Stage 3a accepts as inputs both the containment sentences as well as the secondary input data source and begins the process of identifying a TOI—depression, in our cases—within the primary data source.

The sentences within the secondary input source document are parsed by period delimiters and fed into the SBERT transformer to produce embeddings. We apply the LexRank summarization algorithm<sup>36</sup> to the document in order to create a single embedding vector representation of the document to be compared to the embedded containment sentences. Using an eigenvector centrality method, LexRank identifies the



**Figure 2.** Topic prevalence for University A Reddit data using BERTopic. Topic size corresponds to the number of individual paragraphs assigned to a topic.

sentences within the document that best summarize the document as a whole. The centroid of the corresponding sentence embeddings forms the vector to which containment sentences are compared. We refer to this vector as the depression (DEP) anchor.

With both the primary and secondary data sources cast in the same embedded space, we are able to compare them geometrically. For this, we use the cosine similarity measure given by

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|},$$

where  $\mathbf{u}$  and  $\mathbf{v}$  are 2 vectors in the embedded space to be compared. The cosine similarity score is bounded between 0 and 1, with lower values indicating low similarity and higher values, closer to 1, indicating high similarity between vectors. This allows us to identify sentences from the primary data source that are closest, or most similar, to the DEP-anchor. We define these as “flagged sentences.”

For our anchor  $\mathbf{v}_{\text{dep}}$ , example flagged sentences from the University A Reddit data are shown in [Table 1](#). Flagged sentences  $\mathbf{u}_{\text{flag}}$  are defined as sentences with  $\cos(\mathbf{u}_{\text{flag}}, \mathbf{v}_{\text{dep}}) > \lambda_c$ , where  $\lambda_c$  is a hyperparameter representing some arbitrary cutoff value chosen by the user. Flagged sentences, with cosine similarity to the anchor greater than  $\lambda_c$ , are sorted in descending order of similarity. Therefore, in our case, the highest scoring sentences have the strongest semantic overlap with concepts surrounding depression.

In [Table 1](#), we see a strong relationship between the flagged sentences and mental health-related concepts, such as those being queried. This can be attributed to the quality of the secondary input source document and highlights the importance of carefully choosing such inputs. In our case, the platform is successfully identifying sentences with strong relationships to depression concepts due to the high specificity and relevance of the secondary input source. The corresponding cosine similarity scores between each sentence and the mental health anchor are also presented and indicate that, for

**Table 1.** Example flagged sentences

Flagged sentences ( $\mathbf{u}_{\text{flag}}$ )	$\cos(\mathbf{u}_{\text{flag}}, \mathbf{v}_{\text{dep}})$
“...recently my mental well being has hit an all time low”	0.57
“my circumstances are frustrating... I cant talk to my family and friends about my mental health...”	0.51
“...i just start to feel down and event start to feel uncomfortable physically”	0.40

Subset of flagged sentences for selected topics. Sentences have been lightly redacted to maintain anonymity.

this particular measure, scores approaching  $\cos(\mathbf{u}_{\text{flag}}, \mathbf{v}_{\text{dep}}) \approx 0.5$  have a strong correspondence to depression. For our purposes, we found a cutoff  $\lambda_c = 0.2$  approximately delineated relevant sentences. The cutoff is freely chosen by the user, and this particular value was chosen to illustrate the process for these case studies. Notice that only the first 2 flagged sentences explicitly contain the words “mental health.” The other sentences are nuanced and implicit with respect to mental health. This is the benefit of using a large language model to capture as much contextual information as possible.

In parallel to stage 3a flagged sentences operation, stage 3b runs to capture the semantic context of each topic at the sentence level. This is done by identifying “representative sentences,” which are a subset of the containment sentences that best represent a given topic. Representative sentences are calculated as the medoid sentences of a topic, where the  $n$ th-medoid is defined as the sentence with vector  $n$ th-closest to the centroid of the topic. The top- $N$  representative sentences are therefore the  $N$ -medoids ranked in descending order of proximity to the centroid. The representative sentences average over the semantic meaning of each embedding dimension, intended to capture the overall best representation of the semantic concepts within a given topic. As an example, [Table 2](#) shows representative sentences for the “COVID” topic in 2020 and the “COVID” topic in 2022, which are

**Table 2.** Case study I comparing University A across 2 time periods

Representative sentences for University A related to COVID	
January-June 2020	<p>“i most certainly believe that there has been many coronavirus cases. . .but this is the first official one i heard”</p> <p>“if there is a confirmed case of covid 19 in the. . .community, what is the. . .protocol for sharing this information?”</p>
July-December 2022	<p>“none of us are happy about wearing masks but we do it to protect others”</p> <p>“leave people alone about masks and let s not go back to endorsing putting in crazy restrictions that do more harm than good”</p>
Flagged sentences for University A related to COVID	
January-June 2020	<p>“i know we re not supposed to panic and we have to stay at home and all will be well, but i can t shake the feeling of being incredibly worried for my parents and grandparents health”</p> <p>“just because they claim ed that no one at university is reported to have it doesnt mean that people dont have it, why would we wait for people to start showing symptoms?”</p>
July-December 2022	<p>“... a veritable champion of alienation and fragmentation, feels obliged to keep us masked up without the slightest recognition of the psycho social ramifications”</p> <p>“do you feel exhausted from the mask mandate?”</p>
Second round TM (with word-sentence mapping) 2020-2022	
Topic 1.1	<p>“all these people talking about loneliness, anxiety, and difficulty making friends are reminding me of my first year”</p> <p>“this class in particular has caused me so much stress that i can barely function some days, i panic when i just think about it”</p>
Topic 3.1	<p>“any free counseling or psychiatrists for. . .students or. . .people to talk to?”</p> <p>“where can i start to look for counselling therapy for mental health issues that s covered with our health insurance?”</p>

case study I topics A2 and B7 respectively from the full list of topics in Table S1. While the differences between the 2 versions will be discussed later, it suffices to see that representative sentences are more informative than just the topic words.

In order to avoid all representative sentences clustering around one centroid, we diversify the sentences to capture more themes within a topic. We refer to this as the diversity heuristic, which first calculates and collects the medoid sentence of all sentences, then removes a certain user-defined percentage of remaining sentences closest to this medoid. The algorithm then iteratively recalculates a new medoid from the remaining sentences and removes its closest sentences until the number of sentences collected reaches a user-defined number. Additionally, we provide 2 options to select representative sentences: “local” and “global.” Local representative sentences are medoids of all sentences containing a topic word. Global representative sentences are medoids of all sentences within a topic. Each topic word has its own representative sentences while all topic words within a topic share the same global representative sentences.

*Stage 4:* The final stage leverages an optional second round of TM, this time focused on the original paragraphs in the

corpus from which the flagged sentences originated. The purpose is to allow the researcher to potentially find additional themes that may be correlated with the flagged sentences, thereby providing additional context. The words produced by BERTopic are again mapped to sentences from within the input paragraphs in much the same way as is done in stage 2. It is important to note that the current algorithm makes no attempt to establish a concrete causal relationship between these secondary topics and flagged sentences but simply identifies further emergent themes in the flagged sentences, which may be of interest to the researcher. Furthermore, the researcher should be aware that this functionality will depend on the hyperparameters chosen in previous stages of the platform, particularly the flagged sentences cutoff parameter in stage 3a and the pre-filtering option, because BERTopic’s clustering algorithms require a certain amount of data to reach convergence. Therefore, if the number of flagged sentences for a given topic is too low, the second round of TM may not be available.

## Results

We now present 2 case studies wherein we compare results across 2 time periods as well as 2 locations (University A and B). In these 2 cases, we demonstrate the platform’s ability to differentiate correlated themes across time and geographical locations.

### Case study I

We begin by comparing University A Reddit data across 2 6-month intervals between the first half of 2020 and the latter half of 2022. This roughly corresponded to the beginning of the COVID-19 pandemic and 2 years after that. A full list of topics from this case study appears in Section SA where various coherent topics of discussion are evident. For our example, we focus on a COVID-related topic, which appears highly ranked in 2020 (Topic 2) and has a lower ranking in 2022 (Topic 7). The COVID-related topic was a more prominent topic of discussion in 2020 as opposed to 2022, as indicated by its higher ranking in topic ordering for that period.

To better understand the themes behind each topic, we utilize the portion of the platform that accesses the representative sentences from each topic. In Table 2, we present a subset of the representative sentences for both topics, which reveal the broader themes surrounding the COVID-19 topic each year. In 2020, there is evidently a heightened concern around individual safety, whereas in 2022, the focus has shifted to mask compliance.

We also identify flagged sentences for each topic—ranked by their proximity to the depression anchor  $v_{dep}$ . In Table 2, we present a subset of the flagged sentences for each topic. In 2020, the flagged sentences appear to present themes of anxiety and concern about personal health and well-being, including that of relatives. However, in 2022, the themes have shifted toward frustration around the use of masks, which is in accordance with the representative sentences.

As an illustration of stage 4, Table 2 provides results from a second round of TM on University A data with a word-sentence mapping. BERTopic requires a substantial amount of input data in order to perform its clustering algorithm successfully. Therefore, the second round of TM in Stage 4 will only yield results for topics that produce enough flagged sentences. As such, we have extended the study period

accordingly across the full time period (2020-2022). Given that the input paragraphs in this case were the flagged sentences for our depression anchor, we expect depression themes to be present in the second-round TM. This is indeed the case, as can be seen in Table 2. However, it is also evident that secondary themes emerge from the analysis, such as lack of social interaction and stress due to studying (Topic 1.1). Topic 3.1 appears to focus more on acquiring counseling services, which one might expect to be correlated with a discussion around depression.

### Case study II

In this example, we compare Reddit discussion forum data across 2020-2022 from University A to University B. Here, we also utilize the optional pre-filtering step to filter the data according to a “housing” related theme. We pre-filter posts with a list of user-selected housing-related words, including: Rental, Lease, Tenant, Landlord, Rent, Inspection, Eviction, Move-in, Move-out, Sublet, Occupancy, Apartment, Condo, House, Bedroom, Residence, Roommate, etc.

Figure 3 presents a bipartite graph output of the TM stage (see Supplementary Material for a full list of corresponding topics). On the left side, each node corresponds to a topic from University A data in ascending order of abundance. Similarly, University B topics are shown on the right side. Lines connecting topics indicate a cosine-similarity between topic centroid vectors above a variable user-input threshold. In Figure 3, 4 topics are linked using a threshold set at  $\lambda_c = 0.75$ . We will focus on the example of connected topics A4 and B5, which relate to meal-plan arrangements as they pertain to student housing. This can be seen more clearly by inspecting the representative sentences for these topics in Table 3. The similarity of those sentences justifies a line in

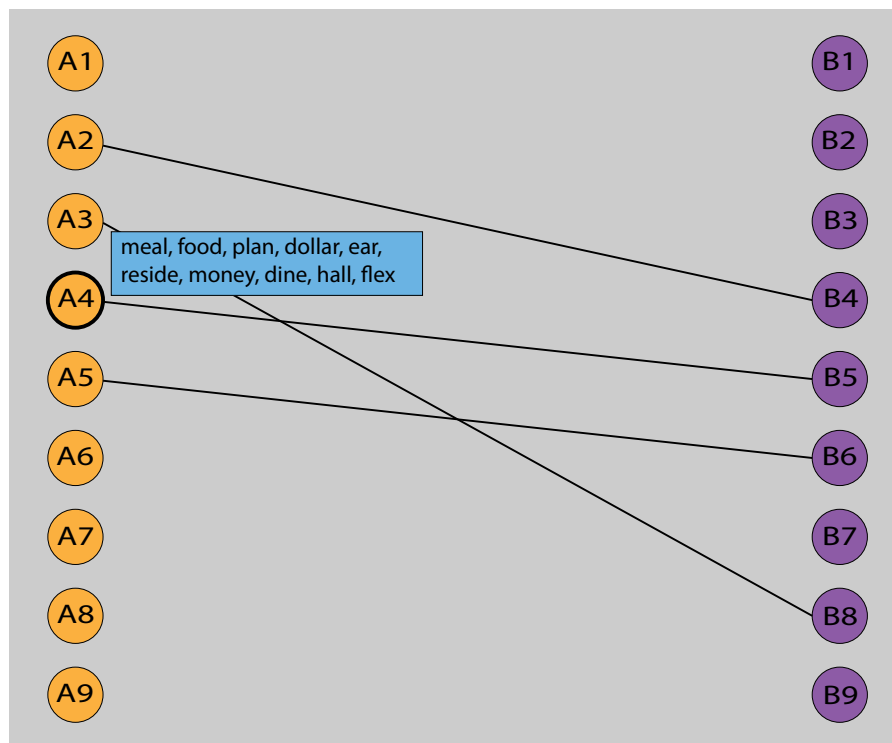
the bipartite graph linking the 2 topics across universities. As expected, if a lower threshold is used, more lines connect the nodes from the 2 universities. For very low thresholds, unconnected nodes correspond to topics unique or specific to that university.

Next, we inspect flagged sentences for the housing filtered data with a cosine-similarity threshold set to  $\lambda_c = 0.2$ . In Table 3, we present the top-2 flagged sentences for topics A4 and B5. The flagged sentences express concepts not only related to depression but also simultaneously related to the pre-filtering “housing” theme.

It is important to note: certain topics, despite having high prevalence in the data, might not yield any flagged sentences despite a sufficient cutoff value. For example, in the “housing” pre-filtered data of Figure 3, topics A3 and B8 describe students’ concerns regarding internet access while in on-campus residence but do not yield any flagged sentences. This is most likely due to the pragmatic nature of the topic and, therefore, does not coincide with discussions around depression. Similarly, topics A4 and B5 did not produce enough flagged sentences for a second round of TM.

### Discussion

The rapid and continued increase of social media use and of mental health-related concerns signals an important opportunity for exploration. A well-designed platform capable of facilitating an investigation of social media site data could offer valuable insights for healthcare research and beyond. We have demonstrated the functional strengths and potential data insights offered by our platform, illustrated through 2 case studies pertaining to mental health-related signals in social media data.



**Figure 3.** Bipartite-graph representation of topic modeling output comparing 2 input primary data sources: University A (yellow) and B (purple). Topics are listed in descending order of prevalence A1-A9, B1-B9. Connecting lines indicate a cosine similarity score between topics that exceed the user input cutoff value  $\lambda_c$ .

**Table 3.** Case study II comparing University A and B across 2 locations

Representative sentences for universities across full time-period 2020-2022 (with housing filter)	
University A Topic A4	“as students wishing to live in first year residence we are being forced to purchase a meal plan which is actually more expensive than simply using cash at the meal plan locations” “while looking through the meal plans, i came across a residence overhead cost”
University B Topic B5	“so i m going into first year and i plan on living on residence, and wanted to know how meal plans work?” “also is meal plan mandatory like other unis?”
Flagged sentences for universities across full time-period 2020-2022 (with housing filter)	
University A Topic A4	“i don t know if it s a byproduct of the first year residence food situation or school burnout or what but lately i haven t been feeling like eating much of anything even if i m hungry”
University B Topic B5	“the reason i ask is because i have struggled with an eating disorder for a while now and am concerned that without being able to have food i am more comfortable with, i just won t eat at all”

The main utility of the platform lies in its ability to identify signals—pertaining to a specified TOI—within large volumes of highly noisy text data. For our purposes, “noisy” refers to a largely incoherent body of text containing many topics of discussion with varying degrees of prominence. The text data can also contain colloquial elements, incomplete sentences, grammatical errors, and non-English language-based symbols, as is often the case with social media data. However, we recommend the user performs a certain degree of text pre-processing of the primary data source plus the secondary data source, as needed. For example, this might involve removing certain stopwords, emoticons, and URLs from the data. With reasonable preparation, our platform is able to perform multiple stages of analysis in the presence of noise and to extract coherent topics of discussion as well as identify signals in (parts of) the data pertaining to the TOI.

A key functionality and strength of our platform is its ability to identify whole sentences corresponding to topic-words, vastly improving the platform’s ability to perform downstream similarity operations given the increased semantic content of sentences. The platform then identifies representative sentences, and more importantly, users can flag sentences within topics that are “relevant” to a given TOI. Given that both representative sentences and flagged sentences derive from topics provided by BERTopic, the user is able to infer a degree of correlation between results. For example, students may describe their stress in multiple ways, and this can, to some degree, be inferred from the correlations between flagged sentences and other emergent themes in the data. Through word-to-sentence mapping in the case study examples, we gain significantly greater insight into the context around each topic as well as identifying clear signals in the data related to our TOI: depression. Additionally, in case study I, we show how a second round of TM performed on the flagged sentences can reveal TOI-related secondary themes in the data. Notably, in our use of the platform, we have discovered that a secondary data source providing a more informal and less clinical description of the TOI performs better when studying social phenomena relating to

personal experience. Similar considerations may be relevant for investigations outside mental health topics.

While the platform development was motivated by mental health research, we believe it to be potentially generalizable to other research domains, especially given the ability to handle large, noisy datasets. The primary input data source is not limited to social media data and could be a myriad of data sources, such as other internet content or large-scale survey responses. In fact, both the primary input data source, as well as the secondary data source, can theoretically be derived from any domain, provided certain criteria are met; in which case, we believe the platform to be potentially useful in performing similar data analysis to those we have presented in this paper. Examples of other research domains include sociology, public opinion research, and journalism, where researchers may choose to explore public perception of specific concerns (eg, COVID-19 mandates, cost-of-living, housing affordability) among the general public or specific populations.

## Conclusion

We have presented a platform for connecting social media data to a domain-specific research TOI using LLMs and unsupervised NLP methods. Our platform is able to take a large, unstructured dataset and identify coherent, structural themes within the data through TM and identify signals in the data related to the research TOI. This is enabled by the use of LLMs, which contain a rich contextual understanding of text, and is further augmented by our platform’s ability to parse and analyze text at the sentence-level.

Finally, our platform provides some graphical outputs for the user, such as a bipartite graph feature that allows one to compare between 2 lists of topics, whether they were generated from data across different time periods, and/or data collected from different locations. Even though we use mental health analysis throughout the paper as an application, we believe that this platform has broader applications.

## Author contributions

L.R. and Y.Z. developed the platform. L.R., Y.Z., and R.N. designed the platform. Y.Z. collected and preprocessed the data used for case studies. R.N. and D.V.V. formulated the project. All authors helped to analyze results. L.R. was the primary manuscript author, while K.L.H., R.N., and A.Y.W. contributed significant revisions. All authors contributed to manuscript preparation and revisions.

## Supplementary material

[Supplementary material](#) is available at *JAMIA Open* online.

## Funding

This work was supported by Health Canada’s Substance Use and Addictions Program [Arrangement #: 1920-HQ-000069].

## Conflict of interest

The authors have no competing interests to declare.

## Ethics statement

This work exclusively utilized non-identifiable, aggregated information in the public domain for which there is no expectation of privacy and, therefore, did not require Research Ethics Board review as outlined in the Canadian Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans—TCPS 2 (2022). For detailed guidance, the following document is available: Does Research Using Social Media Platforms Require Research Ethics Board Review? ([ethics.gc.ca/eng/reb-ccer\\_social-sociaux.html](https://ethics.gc.ca/eng/reb-ccer_social-sociaux.html)). Nevertheless, data security measures were implemented to safeguard the data and prevent uncontrolled access to the aggregated dataset.

## Data availability

The data underlying this article were extracted from publicly available content on Reddit and may be accessed at [www.reddit.com/r/\[University\]](http://www.reddit.com/r/[University]) (eg, [www.reddit.com/r/uvic](http://www.reddit.com/r/uvic)).

## References

- Hirschberg J, Manning CD. Advances in natural language processing. *Science*. 2015;349(6245):261-266.
- Rissola EA, Losada DE, Crestani F. A survey of computational methods for online mental state assessment on social media. *ACM Trans Comput Healthc*. 2021;2(2):1-31.
- Skaik R, Inkpen D. Using social media for mental health surveillance: a review. *ACM Comput Surv*. 2020;53(6):1.
- Zhang T, Schoene AM, Ji S, et al. Natural language processing applied to mental illness detection: a narrative review. *NPJ Digit Med*. 2022;5(1):46.
- Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digit Med*. 2020;3:43.
- Auerbach RP, Alonso J, Axinn WG, et al. Mental disorders among college students in the World Health Organization World Mental Health surveys. *Psychol Med*. 2016;46(14):2955-2970.
- Auerbach RP, Mortier P, Bruffaerts R, et al.; WHO WMH-ICS Collaborators. WHO World Mental Health surveys international college student project: prevalence and distribution of mental disorders. *J Abnorm Psychol*. 2018;127(7):623-638.
- Ebert DD, Mortier P, Kaehlke F, et al.; WHO World Mental Health-International College Student Initiative collaborators. Barriers of mental health treatment utilization among first-year college students: first cross-national results from the WHO World Mental Health international college student initiative. *Int J Methods Psychiatr Res*. 2019;28(2):e1782.
- Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. *BMC Psychiatry*. 2010;10:113.
- Saleem S, Prasad R, Vitaladevuni S, et al. Automatic detection of psychological distress indicators and severity assessment from online forum posts. In: *Proceedings of COLING, Mumbai, India. The COLING 2012 Organizing Committee*; 2012:2375-2388.
- De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. *ICWSM*. 2021;7(1):128-137.
- De Choudhury M. Anorexia on Tumblr: a characterization study. In: *Proceedings of the 5th International Conference on Digital Health*, New York, NY, USA. Association for Computing Machinery; 2015:43-50.
- Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. In: *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Baltimore, MD, USA. Association for Computational Linguistics; 2014:51-60.
- Coppersmith G, Dredze M, Harman C, Hollingshead K. From ADHD to SAD: analyzing the language of mental health on twitter through self-reported diagnoses. In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, CO, USA. Association for Computational Linguistics; 2015:1-10.
- Mago V, Liyanage C, Garg M, et al. Augmenting reddit posts to determine wellness dimensions impacting mental health. In: *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Toronto, Canada. Association for Computational Linguistics; 2023:306-312.
- Shah FM, Ahmed F, Joy SKS, et al. Early depression detection from social network using deep learning techniques. In: *IEEE Region 10 Symposium (TENSYP)*, Dhaka, Bangladesh. IEEE; 2020:823-826.
- Kim J, Lee J, Park E, et al. A deep learning model for detecting mental illness from user content on social media. *Sci Rep*. 2020;10(1):11846.
- Kanaan R, Haidar B, Kilany R. Detecting mental disorders through social media content. In: *IEEE 3rd International Multidisciplinary Conference on Engineering Technology (IMCET)*, Beirut, Lebanon. IEEE; 2021:23-28.
- Blei DM, Ng AY, Mi J. Latent dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022.
- Blei DM. Probabilistic topic models. *Commun ACM*. 2012;55(4):77-84.
- Blei DM, Lawrence C, Dunson D. Probabilistic topic models. *IEEE Signal Process Mag*. 2010;27(6):55-65.
- Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv, arXiv:2203.05794, 2022, preprint: not peer reviewed.
- Timakum T, Xie Q, Lee S. Identifying mental health discussion topic in social media community: subreddit of bipolar disorder analysis. *Front Res Metr Anal*. 2023;8:1243407.
- Leung YT, Khalvati F. Exploring COVID-19-related stressors: topic modeling study. *J Med Internet Res*. 2022;24(7):e37142.
- Dhankar A, Katz A. Tracking pregnant women's mental health through social media: an analysis of reddit posts. *JAMIA Open*. 2023;6(4):o0ad094.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017;6000-6010.
- Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics; 2019:3982-3992.
- Mikolov T, Chen K, Corrado GS, et al. Efficient estimation of word representations in vector space. In: *International Conference on Learning Representations*, Scottsdale, AZ, USA. ICLR 2013. Workshop Track Proceedings; 2013.
- Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*. 2013;2:3111-3119.
- Conneau, A Kiela, D Schwenk, H, et al. Supervised learning of universal sentence representations from natural language inference data. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics; 2017:670-680.
- Cer, D Yang, Y Kong, SY, et al. Universal sentence encoder for English. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium. Association for Computational Linguistics; 2018:169-174.
- Bartha C, Thomson C, Parker C, et al. *Depression: An Informative Guide*. Revised ed. CAMH; 2013.



33. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar. Association for Computational Linguistics; 2014:1532–1543.
34. Lee K, Devlin J, Chang MW, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*, Minneapolis, MN, USA. Association for Computational Linguistics; 2019:4171–4186.
35. Basile P, Rossiello G, Semeraro G. Centroidbased text summarization through compositionality of word embeddings. In: *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, Valencia, Spain. Association for Computational Linguistics; 2015:12–21.
36. Erkan G, Radev DR. LexRank: graph-based lexical centrality as salience in text summarization. *JAIR*. 2004;22:457-479.