Understanding College Students COVID-19 Response: A Structural Topic Model Approach using Reddit Data

Term paper

Jacy Li and Chendi Zhao

2020-12-17

Table of Contents

SET UP

```
setwd("C:/Users/赵晨笛/Desktop/DATA 727/all data")
data<-read.csv("pooleddata.csv")
```

Abstract

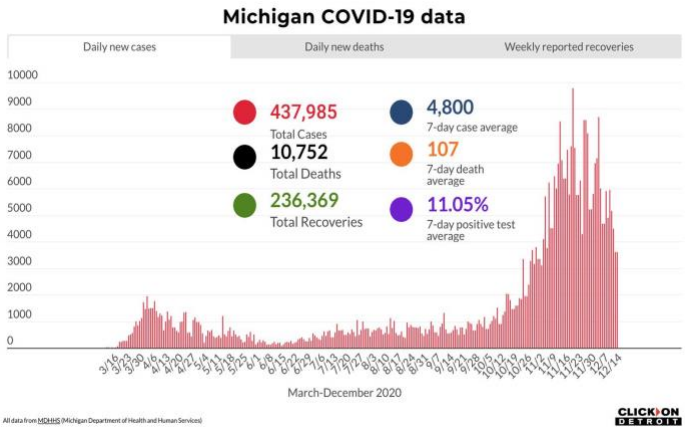
This study focuses on college students' responses to COVID-19 emergency management and the variability across time and groups in Michigan. Using Reddit data, we collect organic submissions in text formats from corresponding subreddits of the ten largest colleges and universities in Michigan for the period of March 1st, 2020, to October 31st, 2020. Taking a structure topic modeling approach, we derive COVID-19 related topics and present the student group's social media landscape. An analysis of topic prevalence and topic content indicates unmet needs on course housekeeping and needs beyond educational requirements. Our study gives school administrators validity to understand student concerns without validating privacy and sheds light on the implications of school emergency management plans on student responses.

Keywords: COVID-19, higher education, concern, response

1.Introduction

2020 has marked a worldwide public health crisis. The State of Michigan has seen surges of COVID-19 cases and deaths since March (Figure 1). As schools adopt better teaching modes to contain the virus, the education system is scrambling to maintain education quality in such a disruptive world. In an emergency event of such kind that persists, it becomes increasingly important to have reliable, up-to-date emergency management assessments. The ability to accurately capture student needs and attitudes are critical to delivering COVID-19 response plans.

Figure 1. Daily new cases in MI (Bartkowiak 2020)



Despite previous studies, where analysis deals with responses to options from pre-established categories, this study contributes to the literature of COVID-19 student experience by analyzing organic, self-reported data from social media. Besides, most surveys introduce hypothetical bias as subjects' responses are measured under hypothetical scenarios (Murphy et al., 2005). Reddit, the self-proclaimed "front page of the Internet" with its unique feature of subreddits, offers timely, continuous public data at almost no cost, allowing the study of smaller geographies and rarer subpopulations.

Therefore, our study aims to understand Michigan college students' response to the COVID-19 pandemic collected from subreddits of universities in the state of Michigan. The rest of the paper is structured as follows: summary of related work can be found in Section2.the methodology to explore data is presented in Section 3, the experiments and results in Section 4. Conclusion and final recommendations are presented in Section 5.

2.Related Work

In the United States, researchers paid particular attention to the most impacted students, medical students (and trainees), and have explored the COVID-19 impacts on their career perceptions (Byrnes et al., 2020) and responses to distance learning strategies (Chen et al., 2020; Gallagher & Schleyer, 2020). For the general student body, Aucejo et al. (2020) administered surveys to approximately 1500 students in a public university. Their study shows that the school shutdowns caused by COVID-19 could exacerbate existing achievement gaps resulting from heterogeneous economic and health burdens among students. Similarly, Chetty et al. (2020) found that students from indigent ZIP codes perform even poorly on an online math program.

However, previous studies have several drawbacks. One is that most studies focused mainly on the most impacted students, medical students (and trainees), instead of the student group as a whole. Also, social media data was rarely used for analysis. As active users on social media platforms, students' opinions can be captured quickly and at a low cost. While there are more studies trying to analyze students' views and behaviors via social media (Mulrennan and Colt, 2020; Li and Leung, 2020), we would be able to have a better understanding of the topic and provide more effective means to address students' concerns and difficulties.

This study explores college students' responses to COVID-19 emergency management and the variability across time and teaching modes (remote vs. hybrid) in Michigan. With datasets scripted from Reddit, we analyzed the content and prevalence of topics using a structural topic modeling (STM) approach. Our research questions are as follows:

RQ1: What topics have been widely discussed by college students in the state of Michigan during the coronavirus pandemic?

RQ2: What were the differences over time?

RQ3: What were the discrepancies between universities with different teaching modes?

3.Data Source and Methods

3.1 Data Collection

Our study scraped data from Reddit (http://www.reddit.com) with Python Reddit API. We selected Reddit for the following reasons: Firstly, Reddit is a highly active platform with more than 138,000 active subreddits, ranking as the 7th most-visited website in the U.S as of December 2020 (*Top Sites in United States* , 2020). Also, most of the subreddits are open to the public, allowing researchers to have access to large amounts of data at a low cost. Lastly, in terms of the current study, given that college students are an active group on social media, we can collect their opinion and thoughts easily on Reddit.

For data collection, we selected ten subreddits of universities in the state of Michigan. According to Wikipedia, these universities had more than 20,000 enrollment in fall 2016, 82.4% of all college students in Michigan (*List of Colleges and Universities in Michigan* , 2020). Table 1 shows the basic information about the selected universities and their subreddits. Data of Macomb Community College and Baker College was missing due to non-existent subreddit and non-existent posts, respectively. In total, 12478 submissions were collected from March 1st to October 31st, 2020.

Table 1. Basic Information about the selected universities and subreddits

School	Subreddit	Enrollment(Fall 2016)	Subreddit Member	Posts
Michigan State University	r/msu	50340	20100	4930
University of Michigan	r/uofm	44718	29000	6048
Wayne State University	r/waynestate	27238	1900	333
Central Michigan University	r/centralmich	25986	1700	81
Grand Valley State University	r/GVSU	25460	3000	601
Western Michigan University	r/WMU	23227	2100	247

Macomb Community College	NA	21734	NA	NA
Eastern Michigan University	r/emu	2246	1100	99
Baker College	r/Baker College	21210	17	NA
Oakland University	r/oaklanduniversity	20012	810	139
Total		281171	59727	12478

3.2 Model Specification

Topic modeling is an introductory text mining technique widely used in analyzing text from social media. Topics are formed in clusters of words based on topic-word proportions and topic-word distributions. Therefore, a topic is always presented by its prevalence and its content (vocabulary). In this study, a more advanced topic model, Structural Topic Model (STM), was employed for analysis. Compared with the primary topic modeling, STM innovates on statistical topic models that allow the inclusion of covariates (Roberts et al., 2014; Roberts et al., 2019). In other words, STM generates top topics and calculates the topical proportion while taking the covariates into account. Further, the correlation coefficient between the topics and covariates can be calculated as well. Hence, STM allows further investigation of how the topic's prevalence changes over time and how the topic's content changes under the influence of specific covariates, providing a more comprehensive, targeted, and dynamic analysis of the text data. STM has been widely used in political sciences (Fang et al., 2016; Duan et al., 2016).

The current study has three covariates: the submission time of the post (time), diagnosis ratio of the county where each university sits (ratio), and teaching mode of the university in the Fall 2020 semester (mode). Table 2 shows the basic information of the covariates. Time was measured as a continuous variable and incorporated into the model as UTC form; the Ratio was dichotomized as "below" and "above" by comparing each university ratio with the benchmark - average diagnosis ratio of Michigan; Mode is also dichotomized as "hybrid" and "remote" according to each university's policy. The hybrid teaching model allows students to take online or offline courses, while the remote mode only allows online teaching. In total, half of the eight universities have a diagnosis ratio above the average ratio (n=1320) in Michigan, and half of them have a diagnosis ratio below the average ratio (n=11158); Only two universities employed remote teaching mode (n=5029) while other universities used hybrid mode (n=7449).

Table 2. Basic Information of the Covariates

School	Posts	Rotio(MI avg= 3.21%)	Mode
Michigan State University	4930	Below(2.79%)	Remote
University of Michigan	6048	Below(2.63%)	Remote
Wayne State University	333	Above(3.39%)	Hybrid
Central Michigan University	81	Below(3.20%)	Remote
Grand Valley State University	601	Above(4.27%)	Hybrid
Western Michigan University	247	Above(3.27%)	Hybrid
Eastern Michigan University	99	Below(2.63%)	Remote
Oakland University	139	Above(3.37%)	Hybrid

We hypothesized that time and ratio could affect how much a specific topic was discussed, and mode could affect what was discussed within a specific topic. Therefore, we took time and ratio as the prevalence covariate and mode during the covid pandemic as the content covariate for further analysis. There were two models in total and we specified the number of topics we wanted as K=20. With Model 1, including only the two prevalence covariates, topics were identified and categorized into three groups based on the commonality of their themes. Then, we analyzed the prevalence of topics over time, both individually and in groups. Model 2 includes both the prevalence covariates and the content covariates, which allowed us to examine the content covariate effect on particular topics. With the two models, we expect to see variance among and within topics generated from the data.

```
processed <- textProcessor(data$selftext, metadata = data)
out <- prepDocuments(processed$documents,
                     processed$vocab,
                     processed$meta,
                     lower.thresh = 15)
docs <- out$documents
vocab <- out$vocab
meta <-out$meta

M1<-stm(documents = out$documents,
                        vocab =out$vocab,
                        K = 20,
                        prevalence =~ratio+s(created_utc),
                        max.em.its = 50, data = out$meta,
                        init.type = "Spectral")
M2<-stm(documents = out$documents,
                        vocab =out$vocab,
                        K = 20,
                        prevalence =~ratio+s(created_utc),
                        content = ~mode,
                        max.em.its = 50, data = out$meta,
                        init.type = "Spectral")
```

4.Results and Discussion

```
out$meta$ID <- seq.int(nrow(out$meta))
theta <- data.frame(M1$theta)
theta$ID <- seq.int(nrow(theta))
all <- merge(out$meta[,-c(1:2)], theta)

all %>% select(ID, created_utc, starts_with('X'))
all_1 <- gather(data = all, key = topic, value = theta, starts_with('X'
), factor_key=TRUE)
```

4.1 Topic Summary

Table 3 presents the results from Model 1 which includes only the three prevalence covariates. The first column are the labels summarized based on the keywords in the topics. The researchers each came up with their own lists and discussed the annotation to reach a consensus. The second column is top keywords in the outputs from the model. The third column is the average proportion of the topics in the reference time. Three topics (T7, T15, T20) were removed from the analysis since we believe that they were all related to internal management of the subreddits, such as removing posts.

```
labelTopics(M1, 1:20)
por<-all_1 %>%
  group_by(topic) %>%
  summarize(avg_theta = mean(theta))
```

Table 3. Topic Summary

#	Topic Labels	Topwords	Average Proportion
T1	Election	think, delete, can, curious, vote, view, regist	3.0%
T2	School reopen	campus, stay, home, back, move ,come, arbor	3.0%
T3	Social life-action	year, college, life, people, friend,make, graduate	3.1%
T4	Outside-covid	open, get, build, campus, around, walk, close	3.0%
T5	Social life-covid	people, case, parties, like, covid, social, virus	3.0%
T6	Strike	student, university, support, work, intern, education, concern	2.9%
T8	Admission	school, major, program, transfer, application, student, engineer	10.0%

T9	Social life-place	interest, club, join, group, game, team, play	4.1%
T10	Course communication	anyone, email, know, got, say, else, start	7.5%
T11	School covid test	will, student, test, health, campus, covid-19, communities	2.4%
T12	Financials -funding	still, will, pay, aid, financial, student, money	4.2%
T13	Mental health	time, feel, sad,stress, day, now, dont, like	4.9%
T14	Course - technology	find, can, use, help, student, websit, access	5.2%
T16	Course - housekeeping	class, grade, exam, professor, lecture, test, student	5.2%
T17	Financials - (lost) job	get, work,lost,hour	4.7%
T18	Housing	live, hous, dorm, apartment, room, year,roommate	6.4%
T19	Course - mode and registration	class, take, semester, course, online, credit, fall	11.3%

Following the same topic annotation method, we then assigned each topic into groups (Table 4) for further analysis. We identified three groups:

*Student Academic-related Attitudes: Students' opinions toward school academic response.
*Student Actions/Behaviors
*Time-specific Topics: certain topics that emerge at specific periods of time.

The last group includes election, strike, and university admission, which were important events in society or regular topics every year that brought the discussion to students. Thus, we did not take this group into further analysis and discussion. Except topics in this last group, other topics are all concerns about students' daily life and school life in the context of COVID-19. We summarized the main theme of each topic and divided them into two groups: attitude and behaviors. The attitude group consists of six topics about students' perception of pandemic and the resulting unprecedented shift in teaching patterns. The latter group includes eight broader topics ranging from course, social life, housing, financial to personal mental health.

Table 4. Topic Groups

#	Themes	Topics
Group 1	Student Academic-related Attitudes	T2,T10,T11,T14,T16,T19

Group 2	Student Actions/Behaviors	T3,T4,T5,T9,T12,T13,T17,T18
Group 3	Time-specific Topics	T1,T6,T8

4.2 Topic Prevalence Analysis

```
prep_M1 <- estimateEffect(1:20 ~ratio+s(created_utc),
                          M1,
                          meta = out$meta,
                          uncertainty = "Global")
```

Based on Model 1, we plotted each topic's prevalence by groups, and the results are shown in Figure 2 to Figure 4. Group 1 includes topics related to students' attitudes toward school response. As seen in Figure 2, school mode and registration (T19) was discussed the most by students, especially during the winter semester and summer vacation. Another topic related to the school issue, course communication (T10), is also widely discussed, and a decreased proportion is observed after September. Starting summer vacation, there were relatively small fluctuations in school reopen, covid test, and course technology topics (T2, T11, T14), but course-housekeeping (T16) was climbing up in the fall semester.

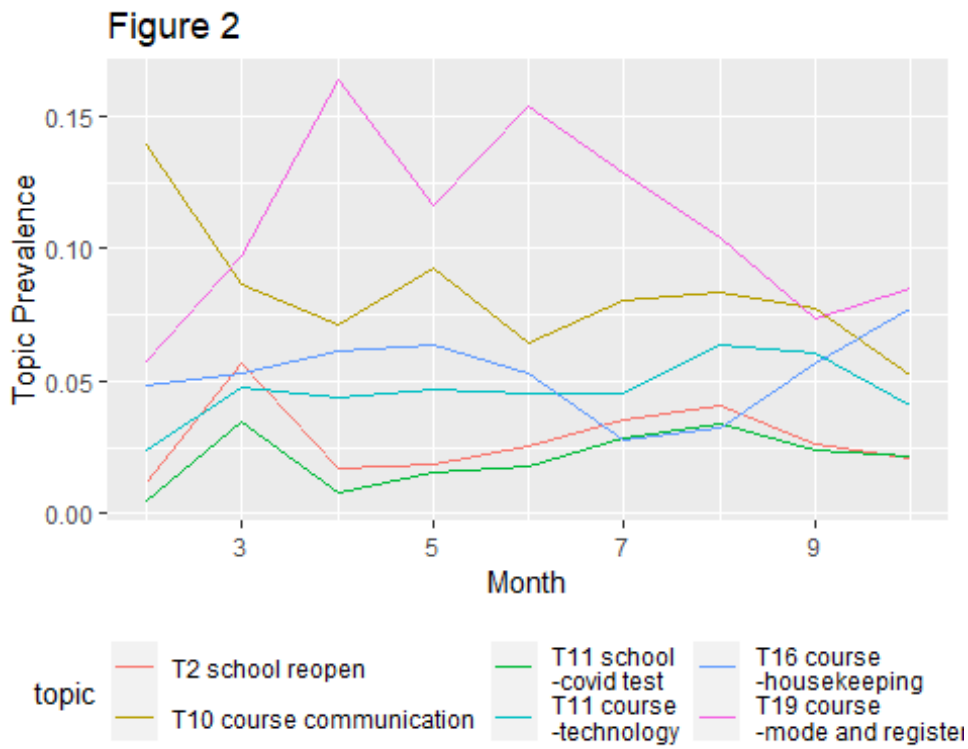
We previously hypothesized that there should be a decreased prevalence for course-related topics irrespective of time dynamic, but this was not the fact for all of the topics. Course mode and registration has the highest prevalence in Group1. There are two peaks for discussion surrounding the topic in April and June, followed by a sharp decline. This trend conforms with our hypothesis as most schools announced their COVID-19 response plans in April and June, which likely has spurred the discussion. However, discussions on communication and technology remain relevantly stable over the passage of time.

Moreover, we observed changes in topic prevalence over time. Starting from the Fall 2020 semester (August or September), fewer discussions surrounding the covid test, school reopen, course technology, and course communication were observed. We argue that the slight drop in prevalence indicates student needs were satisfied to some extent, suggesting a successful school response. This is particularly true for coursework-related topics, as technology and communication topics are intuitively expected to rise as the semester starts. On the contrary, discussion on course housekeeping tripled since the start of the semester, which reveals issues for professors and university leaders to improve.

```
group1<-all_1 %>%
  filter(topic==c("X2","X10","X11","X14","X16","X19"))%>%
  group_by(month, topic) %>%
  summarize(total = n(),
            med_theta = median(theta),
            avg_theta = mean(theta))
```

```
## `summarise()` regrouping output by 'month' (override with `.groups`
argument)

group1%>%
  ggplot(aes(x=month,y=avg_theta,color=topic)) +
  geom_line() +
  labs(x = 'Month', y = 'Topic Prevalence')+
  scale_x_continuous(breaks=c(3,5,7,9))+
  theme(legend.position = "bottom")+
  scale_color_discrete(labels=c("T2 school reopen",
                                "T10 course communication",
                                      "T11 school\n-covid test",
                                      "T11 course\n-technology",
                                      "T16 course\n-housekeeping",
                                     "T19 course\n-mode and registeration"))
+
  ggtitle('Figure 2')
```



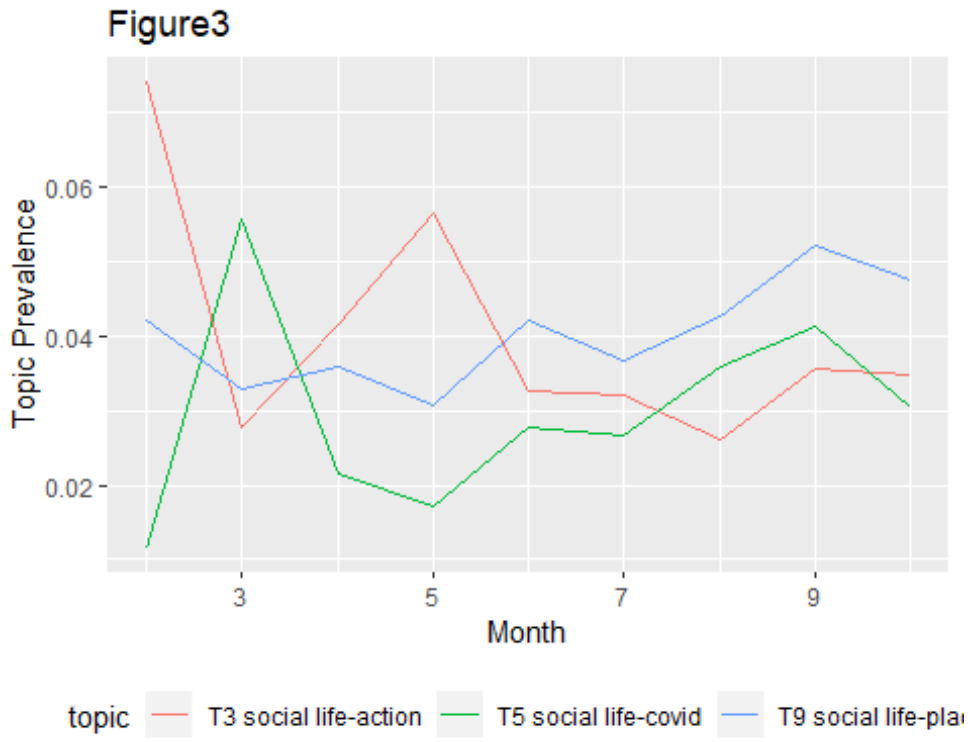
Topics in Group 2 were plotted separately. Figure 3 includes social life-related topics, and Figure 4 includes other topics in Group 2. It could be concluded from Figure 3 that there is a large fluctuation in social life-action and social life-covid (T3 and T5), especially during the winter semester and summer vacation. Compared with these two topics, social life-place (T9) showed a small fluctuation but is gradually increasing over time and reached its peak in September.

Discussion on covid impacts on social life peaked in March, following a peak about covid impacts on behavior in May. This trend is consistent with the student covid narration: after covid first drew public attention in March, Governor Whitmer issued an executive order 2020-21, ordering people to stay home stay safe; after schools recessed in late April, or early May, students shifted from coursework to think about social life in a covid world. In fact, such concerns on social life remain unresolved when schools reopen in fall - discussions on where (social life-place) and how (social life-action) climbed up, and the where questions reached a peak almost equivalent to social life-action's historical peak in May. There is a shift of focus from how to where, however.

```
group2_1<-all_1 %>%
  filter(topic==c("X3","X5","X9"))%>%
  group_by(month, topic) %>%
  summarize(total = n(),
            med_theta = median(theta),
            avg_theta = mean(theta))
```

```
## `summarise()` regrouping output by 'month' (override with `.groups` argument)
```

```
group2_1%>%
  ggplot(aes(x=month,y=avg_theta,color=topic)) +
  geom_line() +
  labs(x = 'Month', y = 'Topic Prevalence') +
  scale_x_continuous(breaks=c(3,5,7,9))+
  theme(legend.position = "bottom")+
  scale_color_discrete(labels=c("T3 social life-action",
                                "T5 social life-covid",
                                      "T9 social life-place"))+
  ggtitle('Figure3')
```



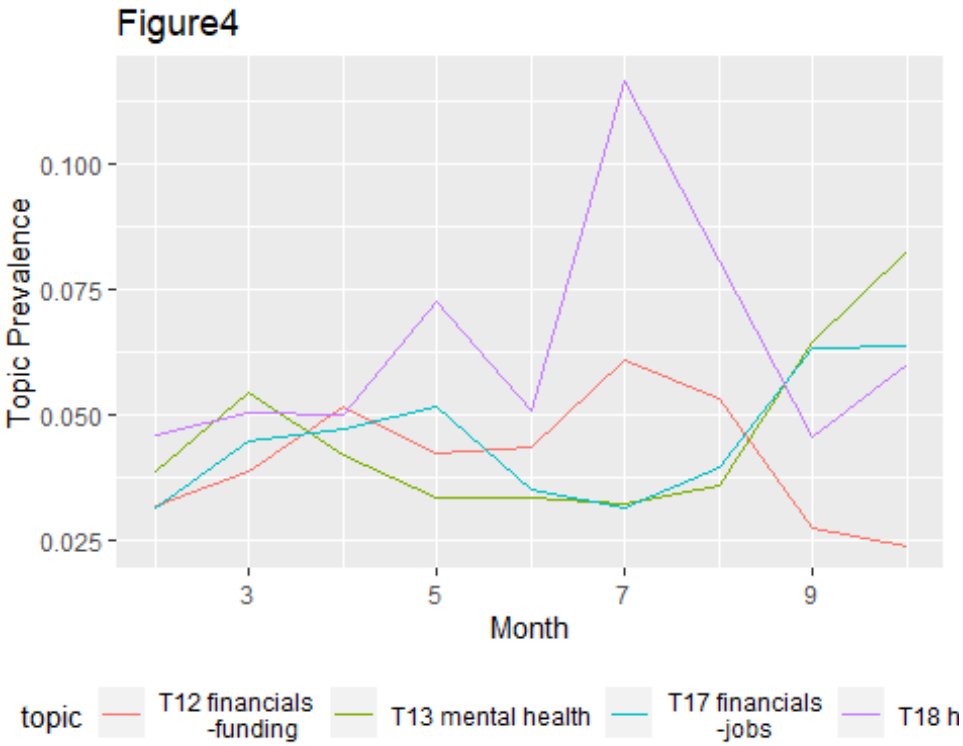
In Figure 4, housing (T18) was discussed the most during summer. Two topics about finance showed different trends: Funding (T12) peaked in the middle of summer and decreased rapidly after that. On the other hand, loss of job (T17) hit rock bottom simultaneously but increased rapidly, especially during the fall semester. Though discussion on lost jobs and funding were equally prevalent before June, starting June, the complementary effect of lost jobs and funding indicates an effective school and/or governmental response. On the contrary, mental health (T13) concerns became increasingly important and remained unsolved.

```
group2_2<-all_1 %>%
  filter(topic==c("X12","X13","X17","X18"))%>%
  group_by(month, topic) %>%
  summarize(total = n(),
            med_theta = median(theta),
            avg_theta = mean(theta))

## `summarise()` regrouping output by 'month' (override with `.groups`
argument)

group2_2%>%
  ggplot(aes(x=month,y=avg_theta,color=topic)) +
  geom_line() +
  labs(x = 'Month', y = 'Topic Prevalence') +
  scale_x_continuous(breaks=c(3,5,7,9))+
  theme(legend.position = "bottom")+
  scale_color_discrete(labels=c("T12 financials\n        -funding",
```

```
                                "T13 mental health",
                               "T17 financials\n        -jobs",
                                "T18 housing"))+
  ggtitle('Figure4')
```



4.3 Topic Content Analysis

```
labelTopics(M2, 1:20)

prep_M2<- estimateEffect(1:20 ~ratio+s(created_utc)+mode,
                         M2,
                         meta = out$meta,
                         uncertainty = "Global")
summary(prep_M2)
```

Schools tried to maintain individual health and campus health during the pandemic by switching to online classes partly or completely. Overall, there were two teaching modes: remote and hybrid. Under remote mode, classes were taken completely online, while under the hybrid mode, some classes were still offline, and students could choose to take classes online or in person. These two teaching modes might lead to different school experiences and lifestyles among students, affecting their concerns, attitudes, and behaviors. Therefore, to further understand how topical content varied under the influence of teaching mode, we fit Model 2 with prevalence variables and teaching mode as the content variable.

Topics were all affected by teaching modes, except topics about school reopen, social life-place, course-communication, and school covid test (T2, T9, T10, T11). As shown in Figure 5, topics about outside-covid, financial-funding, course-technology, course-mode, and registration (T4, T5, T12, T14, T19) were more likely to be discussed by students under remote teaching mode. On the other hand, topics related to social life-action, mental health, course-housekeeping, financial-jobs, and housing (T3, T13, T16-18) were more likely to be discussed by students under hybrid teaching modes.

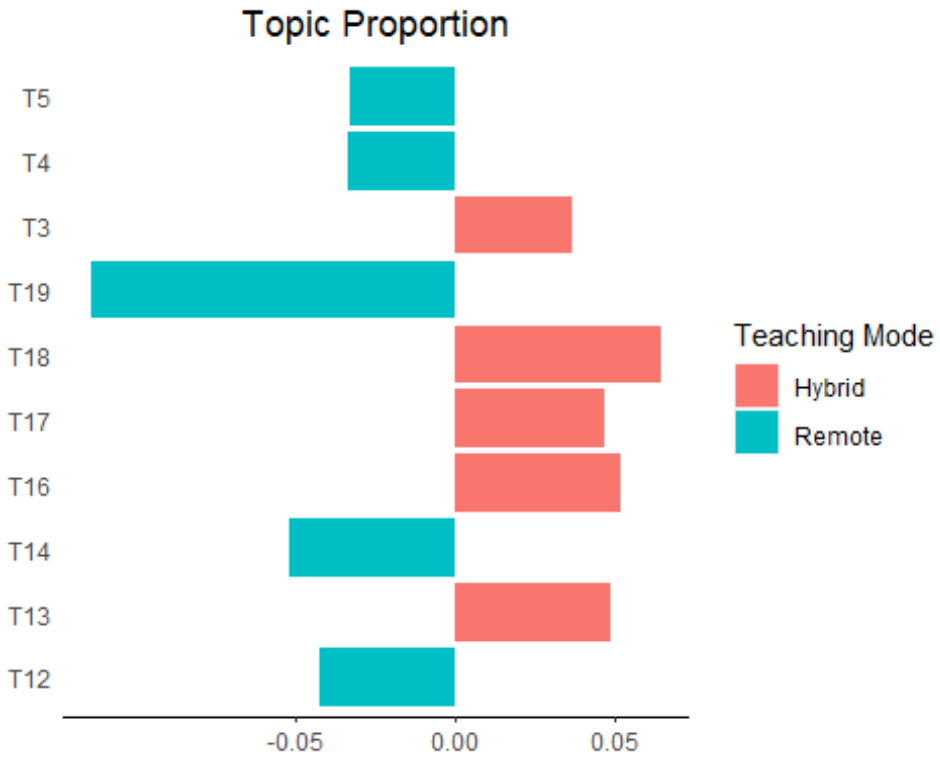
Considering the topical proportion and content, we can sum up the following points: Firstly, all students are concerned about the school's policy and general social life under the COVID-19 pandemic. Compared with topics in the remote group, topics in the hybrid group were more diverse. To be specific, in the remote group, student discussions were restricted to topics directly related to the university's educational responsibility. Students talked more about course technology issues, course mode and registration, and financial funding. In the hybrid group, topics about social life, campus housekeeping, housing, mental health, and job issues were more widely discussed to touch upon the social responsibility in higher education. It is also necessary to mention that words such as sad and pressure were in a large proportion within the mental health topic, indicating a negative sentiment.

```
por_2<-all_1 %>%
  group_by(topic) %>%
  summarize(avg_theta = mean(theta))%>%
  .[-c(1,2,6,7,8,9,10,11,15,20),]

## `summarise()` ungrouping output (override with `.groups` argument)

por_2$topic<-gsub("X","T",por_2$topic)
a<-ifelse(por_2$topic=="T4"|por_2$topic=="T5"|por_2$topic=="T12"|por_2$
topic=="T14"|por_2$topic=="T19",-1,1)
por_2$mode<-ifelse(por_2$topic=="T4"|por_2$topic=="T5"|por_2$topic=="T1
2"|por_2$topic=="T14"|por_2$topic=="T19","Remote","Hybrid")
por_2$avf_theta_2<-por_2$avg_theta*a
ggplot(data = por_2,aes(x = topic, y =avf_theta_2,fill = as.factor(mode
))) +
  geom_bar(stat = 'identity') +
```

```
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.ticks.y = element_blank(),
        axis.line.x = element_line(),
        axis.title = element_blank()) +
  labs(title  = 'Topic Proportion',fill='Teaching Mode')+
  scale_y_continuous(breaks=c(-.05,0,.05))+
  theme(plot.title = element_text(hjust = 0.5))+
  coord_flip()
```



5. Conclusion and Implication

In conclusion, our study derived a variety of topics discussed by Michigan college students from March 1st to October 31st, 2020. By summarizing and categorizing them into different groups, we further analyzed each topic's prevalence and observed trends and fluctuations over time and within each semester. Furthermore, we added the mode as a covariate and found a difference in students' topics with different teaching modes. As we have investigated potential explanations for our results, we would like to address limitations, implications, and future study suggestions.

There are a few limitations to our study. First, data collected only from eight universities' subreddits may limit the study findings' generalizability. Besides, massive available corpus in Reddit does not assure coverage of the population,

which likely introduces bias for demographic factors (Hargittai, 2018). For instance, users who post in a school subreddit may not necessarily be current students. Previous research on online teaching found math courses are particularly difficult for K-12 students (Kuhfeld et al. 2020). However, our research did not verify the finding. As we hypothesized, there should be at least one math topic among the top 20 topics generated by Model 1. Our lack of representativeness may contribute to this missing math topic. It might merely be due to the differences in learning capabilities between K12 and higher education students.

Second, Reddit offers little user-level information. The only available data is users' display names, and researchers could only make inferences on user-specific information based on user subreddit subscriptions and from the contents of their posts. Given that, we could not include some potential covariates, such as gender, major, and grade, which may reduce credibility without controlling basic demographic information. Future studies could focus on other social media or combine available data sources to expand the sample's quantity, better understanding the question, and increasing credibility.

Lastly, ethical concerns of researches using data gleaned from social networking sites have drawn the attention of researchers (Zimmer, 2010; Williams et al. 2017; Fiesler and Proferes, 2018). Ethical norms require a more reflexive approach that puts user privacy and safety center stage. While our study employs strategy for privacy and anonymity by using aggregated data, it was not feasible to obtain direct consent from users. In other words, the providers of our data were unaware of this research.

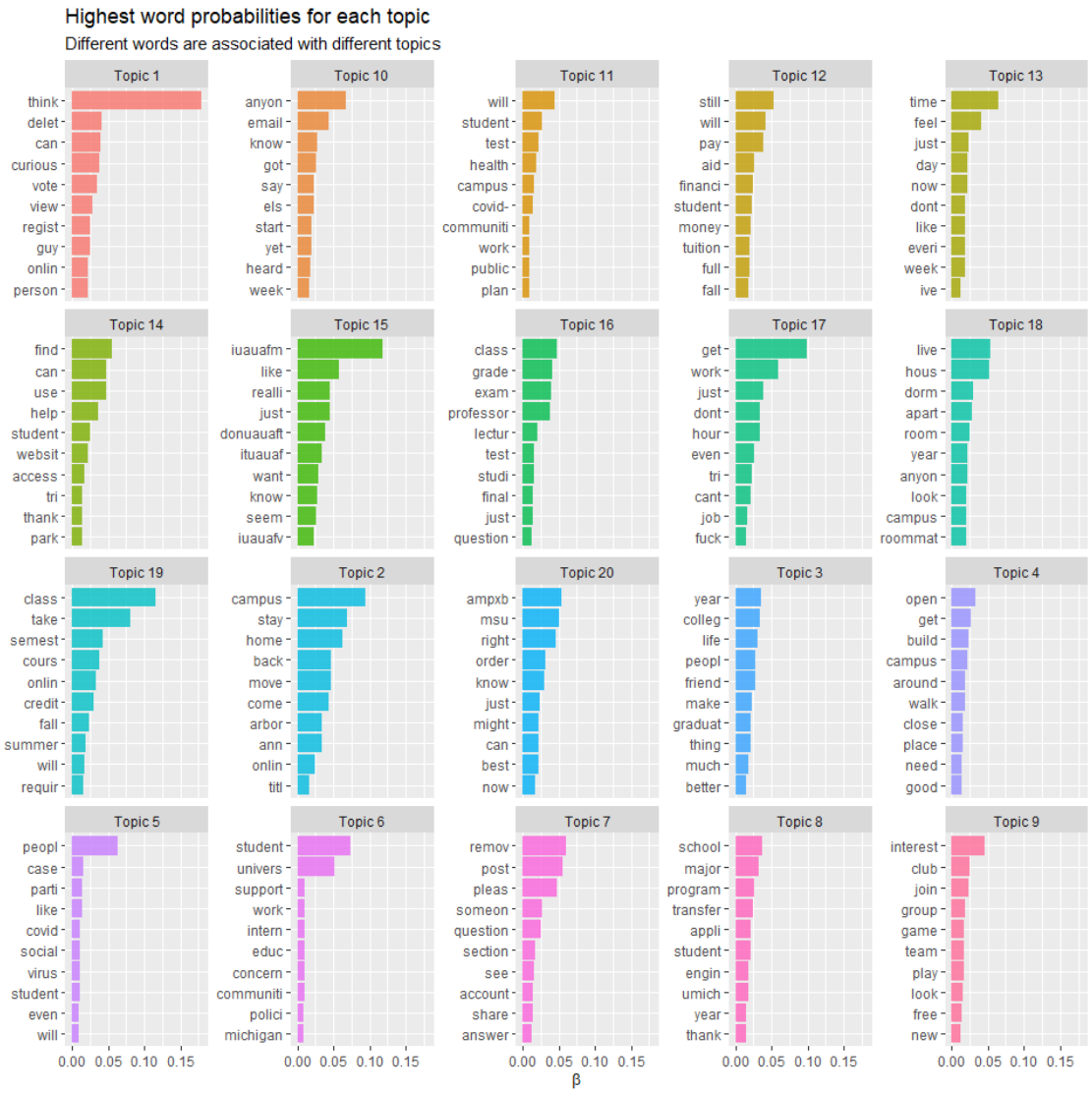
Building upon previous studies recommending interactive virtual classes as a coping strategy for distant learning (Chen et al., 2020), our findings identified the top concerns among students that the university leaders and professors should consider. Course mode and registration, course communication, and housing issues were among the highest student discussions regardless of time. In addition, a group-level analysis shows that students' needs on course housekeeping and where to seek social life have been increasing and remained unsolved since the start of the fall semester.

Our findings on topic prevalence and topic content analysis indicate that students, especially students receiving hybrid teaching, expressed needs beyond educational requirements. Although financial needs have been satisfied to some extent, we observed a growing trend in mental health and social life discussions. Universities should focus on these aspects and provide actionable suggestions to students as well.

Appendix

```
M1_a<-tidy(M1)
M1_a %>%
```

```
  group_by(topic) %>%
  top_n(10, beta) %>%
 ungroup() %>%
    mutate(topic = paste0("Topic ", topic),
           term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = as.factor(topic))) +
  geom_col(alpha = 0.8, show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free_y") +
  coord_flip() +
  scale_x_reordered() +
  labs(x = NULL, y = expression(beta),
       title = "Highest word probabilities for each topic",
       subtitle = "Different words are associated with different topics
")
```




```
group1<-all_1 %>%
  group_by(month, topic) %>%
```

```
  summarize(total = n(),
            med_theta = median(theta),
            avg_theta = mean(theta))

## `summarise()` regrouping output by 'month' (override with `.groups`
argument)

group1%>%
  ggplot(aes(x=month,y=avg_theta,color=topic)) +
  geom_line() +
  labs(x = 'Month', y = 'Topic Prevalence')+
  scale_x_continuous(breaks=c(3,5,7,9))+
  facet_wrap(.~topic,scale="fixed")+
  ggtitle('Figure 2')
```