Topic Modeling on Social Media to Understand the Shift of Interest in Different Types of Traveling

Khoa Dang, Youngmin Kim & Harshit Raj

December 13, 2023

Abstract

We were all provided a life-changing opportunity to experience different or new concepts due to the appearance of one of the worst pandemics, COVID-19, in traveling for example. Travel comes in different forms. In the project, our goal is to observe the impact of COVID-19 on different types of travel. By conducting topic modeling using BERTopic on Reddit data containing over 700,000 posts from different subreddits, we extracted different topics which were then grouped into 4 types of travel (cruise, camping & hiking, backpacking, road trip) based on the clustering and our evaluation. After that, we utilized sentiment analysis to enrich the findings even further. As a result, we managed to capture some evidence that suggest interesting trends and changes in people's interests online between 2018 and September 2023, such as the increase in demand for more extravagant trips on cruises.

1 Introduction

COVID-19 has made drastic changes in human behavior, even though humans seem to be creatures of habit [DSP21]. A lot of norms, or perspectives, rapidly became outdated and gradually replaced by new trends and concepts. According to a report by McKinsey and Company in 2021 [SLR21], it indicates that The pandemic sped up trends in working from home, online shopping, and automation when about 25 percent more workers than originally thought might have to consider changing jobs. Our concerns about mental health have finally become international spotlight due to the increase in the rate of depression, anxiety, or mental disorders [oH21].

Tourism serves as a significant revenue-generating source for many countries around the world. People love to travel for different reasons, from relaxation to exploration of new cultures and knowledge. Without an exception, it is expected that traveling and tourism fall into the victims of the pandemic. The economy of multiple countries was disrupted due to the restrictions of health guidelines to tackle the pandemic, and people have raised different concerns and interests about traveling before, during, and after the COVID-19 pandemic because they were forced to step out of their comfort zones to experience new things. The goal of our research is to study how people have adapted to the potentially new normal of traveling by analyzing their conversations on social media.

Travel could vary in many dimensions, destinations, types, and so on. People tend to visit new countries or cities, so the destination can be determined by their travel history. Moreover, travel destination varies too much to be the focus of this research. On the other hand, people tend to do the same type of travel repeatedly (i.e. People who like camping, do camping repeatedly). Travel types could vary from simple backpacking to trips on cruises. By further examining the contents of the traveling-related discussions to gain an overview of the people's perspectives, we will then focus on whether and how people have changed their references to different types of travel.

Research Question: What are the interests and concerns of people in the variation of travel types before, during, and after the pandemic? Do they change?

2 Related Research

This research investigates the potential of leveraging social media data to identify and explain shifts in tourism interest attributed to external events. In the contemporary digital era, the exploration of online social networks has become integral to understanding diverse facets of societal behavior, and the field of transportation planning and travel demand modeling has embraced the richness of big data from social media platforms. Previous work by Rashidi et al. [RAM+17] undertook a comprehensive review of literature from 2007 to 2015, examining the utility of social media data in travel behavior modeling and transportation planning. The study not only surveyed travel prediction experts but also delved into the integration of social media data into advanced travel prediction models. Such groundwork is foundational to understanding the opportunities and challenges associated with utilizing social media data in the domain of tourism.

Addressing the challenge of collecting high-resolution data on individual travel behavior, Hasnat and Hasan [HH18] explored the potential of location-based social media data, particularly Twitter, in identifying tourists and analyzing spatial patterns in their destination choices. By applying a data mining framework that included heuristic classifiers and spatial clustering techniques, the authors distinguished between resident and tourist users. Their findings underscored the unprecedented scale at which social media, particularly Twitter, can provide insights into tourist travel behavior. This work highlights the relevance of social media data in comprehending the dynamic choices of tourists, a population with rapidly changing travel preferences, and underscores its significance for transportation and city planners seeking innovative approaches to data collection and analysis.

In tandem with the broader exploration of social media's impact, the research done by Park, Ok, and Chae [SBPC16], emphasizes the burgeoning significance of social media analytics in unraveling the intricate dynamics of the cruise tourism industry. With a dedicated focus on Twitter as a robust data source, the study employs meticulous data collection and diverse analytical techniques, including word frequency analysis, content analysis, and network mapping. The findings reveal prevalent themes and patterns in cruise-related tweets, identifying major topics such as travel destinations, cruise companies, and the emotional sentiments expressed by users. As social media analytics continues to evolve, studies underscore the untapped potential of social media as a pivotal platform for businesses and organizations in the hospitality and tourism sector, urging strategic utilization to enhance marketing, promotion, and customer engagement practices.

3 Data

3.1 Data Description

The data was collected from Reddit, a widely popular online community where people can discuss a variety of topics with different levels of expertise freely. Each potential topic can be categorized by a subreddit so that the consistency of interest in the users' posts remains intact, which helps reduce the noise of irrelevant discussions within a topic. Reddit is a good and inexpensive source for collecting high-quality data (4). By utilizing the characteristics of this social media platform, it reduces the complexity of the data-gathering process of traveling-related content for our research. In this case, we have gathered the data of users' posts from 19 subreddits discussing traveling. The posts can be a question asking for advice, or a review and recommendation of a user's experience and trip to certain places.

Subreddits: Travel, backpacking, CampingandHikin, solotravel, roadtrip,AirBnB, Cruise, onebag, vagabond, TravelHacks, Flights, travelpartners, Shoestring, adventures, travelphotos, wanderlust, hitchhiking, couchsurfing, and longtermtravel. (Descriptions of subreddits in 8. Appendix)

These are listed in a comprehensive list of travel subreddits provided by a moderator of /travel subreddit[1], one of the most well-known subreddits for traveling content with over 8 million followers and

---

[1]https://www.reddit.com/r/travel/comments/1100hca/the_definitive_list_of_travel_subreddits_to_help/

Topics	2018-2019	2020-2021	2022-2023.9	Total
AirBnB	9701 (3.67%)	9,919 (5.31%)	11,443 (4.53%)	31,063(4.41%)
CampingandHiking	21,335 (8.06%)	12,904 (6.91%)	10,053 (3.98%)	44,292(6.29%)
Cruise	8,435 (3.19%)	6,102 (3.27%)	15,927 (6.31%)	30,464(4.33%)
Flights	3,367 (1.27%)	4,906 (2.63%)	9,380 (3.72%)	17,653(2.51%)
Shoestring	4,388 (1.66%)	2,618 (1.40%)	3,531 (1.40%)	10,537(1.50%)
TravelHacks	3,780 (1.43%)	4,283 (2.29%)	11,832 (4.69%)	19,895(2.83%)
adventures	3,053 (1.15%)	3,997 (2.14%)	2,898 (1.15%)	9,948(1.41%)
backpacking	16,659 (6.29%)	17,827 (9.55%)	18,039 (7.15%)	52,525(7.46%)
couchsurfing	912 (0.34%)	810 (0.43%)	908 (0.36%)	2,630(0.37%)
hitchhiking	955 (0.36%)	632 (0.34%)	1,154 (0.46%)	2,741(0.39%)
longtermtravel	474 (0.18%)	155 (0.08%)	171 (0.07%)	800(0.11%)
onebag	5,837 (2.21%)	5,318 (2.85%)	10,842 (4.30%)	21,997(3.13%)
roadtrip	10,487 (3.96%)	13,429 (7.19%)	12,979 (5.14%)	36,895(5.24%)
solotravel	18,029 (6.18%)	9,396 (5.03%)	11,406 (4.52%)	38,831(5.52%)
travel	141,989 (53.65%)	79,842 (42.77%)	116,862 (46.31%)	338,639(48.13%)
travelpartners	5,958 (2.25%)	2,536 (1.36%)	3,059 (1.21%)	11,553(1.64%)
travelphotos	2,325 (0.88%)	2,237 (1.20%)	2,891 (1.15%)	7,452(1.06%)
vagabond	5,929 (2.24%)	7,706 (4.13%)	7,072 (2.80%)	20,707(2.94%)
wanderlust	1,050 (0.40%)	2,058 (1.10%)	1,883 (0.75%)	4,991(0.71%)
All	**264,663 (100%)** (11,028/month)	**186,675 (100%)** (7,778/month)	**252,330 (100%)** (12,016/month)	**703,668 (100%)** (10,198/month)

Table 1: Number of posts of 19 Travel related Subreddits

approximately 8000 active users every day. After manually inspecting the list, we decided to choose the mentioned subreddits to align with the focus of our research based on their popularity, the quality of their posts, and interaction rates within their communities. The total number of posts is over 700,000 from January 2018 to September 2023.

On the other hand, we divided the proposed timeline into 3 groups to support the focus of our research. Pre-COVID period is before Jan 2020 because the first cases of COVID-19 happened at the end of 2019 and reached the US in January 2020. During-COVID period is between Jan 2020 and December 2021. Even though a lot of countries started to reopen for travel in November 2021; however, due to the complications of the Omicron outbreak, countries were on a high alert until January 2022. Post-COVID period starts from Jan 2022. Although there were some strict restrictions and health policies, multiple countries already reopened for tourism and the number of flights started to recover from the pandemic significantly according to the report by the International Trading Association[Adm22].
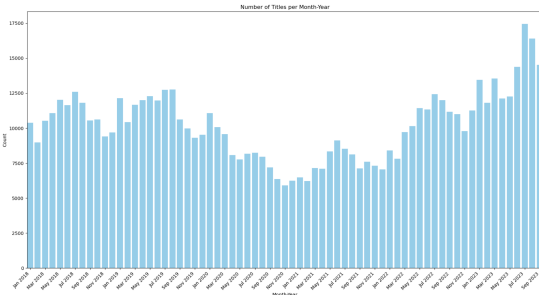
3.2 Explanatory Data Analysis

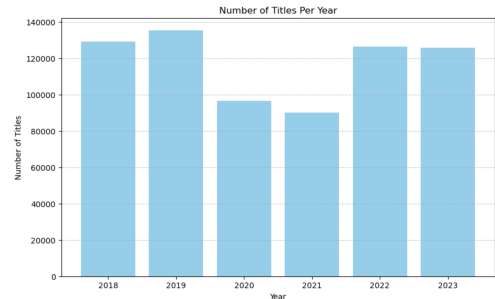Figure 1: Number of Titles per Month-Year

Figure 2: Number of Posts Per Year

Figure 1 portray the monthly distribution of posts which reveals an evident dip in the number of posts from 2020 to 2021. The quantitative analysis underscores this trend, with the average number of posts dropping from 11,028 in the preceding period of 2018 - 2019 to 7,778 in 2020 - 2021(Table 1), reflecting a notable decrease of approximately 29.5%. The stark decline can be seen in the Figure 1, where the average decrease of approximately 3,248 posts. However, the graph exhibits a increase in the period from 2022 - 2023, with the average number of posts rebounding to 12,016 per month. The number of posts yearly(Figure 2) also supports this observation where there is a decline in the years 2020 and 2021 and then a surge in the years 2022 and 2023.

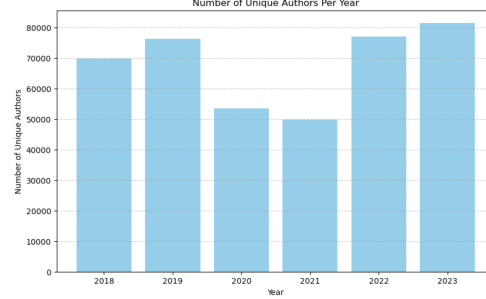|	Text Length
count	703668.0
mean	72.02
std	141.01
min	0.0
25% 1	8.0
50%	24.0
75%	93.0
max	7533.0

Table 2: Summary of Text

Figure 3: Number of Unique Authors

Figure 3 illustrates the count of unique authors for each year, revealing a trend parallel to that observed in the average number of posts per year. A noticeable dip is discernible in the years 2020 and 2021. Specifically, the average number of unique authors from 2018 to 2019 stood at 73,076.5. Subsequently, during the pandemic era (2020 - 2021), there was a significant decline, with the average number of unique authors plummeting to 51,613.5—a decrease of approximately 29.4%. However, in the post-pandemic period, there was a remarkable resurgence, averaging 79,327 - an increase of about 53.5% from the pandemic-era average.

The Table 1 represents the post counts of various travel-related subreddits, providing insight into the popularity of different travel topics within the Reddit community. The "travel" subreddit leads with 338,639 posts - 48.13%. Subreddits like "backpacking," "CampingandHiking," and "solotravel" also demonstrate significant engagement, with post counts of 52,525 (7.46%), 44,292 (6.29%), and 38,831 (5.52%), respectively. On the other hand, subreddits like "hitchhiking," "couchsurfing," and "longtermtravel" have lower post counts at 2,741 (0.39%), 2,630 (0.37%), and 800 (0.11%) respectively.

The average post length stands at 72.02 characters(Table 2); however, the standard deviation of 141.01 indicates considerable variability, with posts ranging widely from the mean. The median post length is 24 characters, while the maximum post length is 7,533 characters. signifies the presence of some significantly longer posts.

4 Methodology

4.1 Model

BERTopic can also produce a Hierarchical Clustering plot (Figure 10 in Appendix), which is useful to provide an overview of the extracted topics as well as their clusters in multiple levels. By inspecting the clusters, we came up with 4 groups to present the variation of travel types from the list of topics: backpacking, cruise, roadtrip and hiking&camping. Then, we selectively added topics to these pre-defined groups based on their similarities and our impressions as follows:

Group 1 - Backpacking: Topic 1,37,73,74,77

Group 3 - Roadtrip: Topic 3,70,76,91.

Group 4 - Cruise: Topic 4.

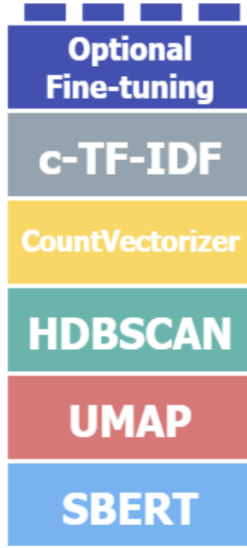Group 8 - Camping&Hiking: Topic 8,39,58,66,71,84.

Figure 4: BERTopic Structure

In the current landscape of Natural Language Processing (NLP), the prevalent trend involves the adoption of pre-trained models such as BERT[DCLT18], which consistently outperforms across various tasks. The advantages of BERT, particularly its proficiency in handling raw texts without extensive preprocessing [AJ21], make it an ideal choice for the project at hand. Notably, considering that a significant portion of the documents comprises succinct content, often totaling less than 24 characters, BERT's efficacy is further highlighted.

One prominent choice for our project is BERTopic[Gro22], a topic modeling approach that sets itself apart from LDA-based models by assigning each post to a single topic rather than a distribution of topics. The BERTopic methodology involves a series of four key steps for topic extraction from the corpus. Initially, it converts each document into embeddings using a pre-trained sentence transformer. Subsequently, it employs dimensionality reduction techniques to cluster posts into topics. Finally, the approach employs CounterVectorization and c-TF-IDF to refine topics and represent each topic effectively. The determination of the number of topics is based on the size of the clusters.

Through a systematic exploration of various cluster sizes, ranging from 30 (default) to 1000, a thorough selection process of resulting topics was conducted. The decision was reached to proceed with the list of topics derived from the minimum cluster size of 1000 for further research. This approach resulted in the identification of 93 distinct topics, forming the basis for subsequent investigations.

We generated a cross table (Figure 5) to assess the results of our topic modeling. The analysis indicates that the groups share names with the subreddits and most documents within each group are found in their respective subreddits. Using these groups will give us a better understanding of changes among the travel types, which could be explained by the following reasons. First, we may miss a lot of topic-related documents from more general subreddits such as /travel. As we can see from the table, topic 3 from Roadtrip Group has over 4700 documents in /travel subreddit, which is half of the number of documents in /roadtrip subreddit. Second, although the majority of the posts should be relevant to the subreddits, some documents could not provide us with a clear indication of travel types, such as general or random questions (Figure 6). Thus, using these topic groups could potentially provide us with better insights into changes among different types of travel.

4.2   Sentiment Analysis

Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral. We utilized the results from the analysis to enrich our findings from the BERTopic. Despite its current reputation for not producing practical results, it can still be used to provide certain insights into the topics extracted from BERTopic. Furthermore, samples from the results of this analysis, specifically documents with significant high/low sentiment scores or close to the boundaries between 3 classes (Positive, Negative, Neutral), will be inspected manually to ensure the quality of the result.

5   Results and Discussion

5.1   Changes in Topic Frequency Over Time

From the topic grouping result, we investigated the possible changes in trends in each type of travel. First, we compare the number of posts with the topic and the proportion of posts with the topic within the total number of posts each month. First of all, Figure 7 and Table 3 shows that the number of cruise topic posts and backpacking topic post decreased a lot(-33.9%, -17.1%) the during the pandemic

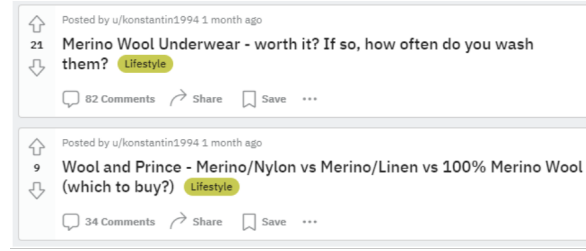| subreddit	Camping and Hiking	Cruise	backpacking	onebag	roadtrip	travel	All
Topic							
1	1462	78	**2667**	**14661**	85	3926	25802
3	2387	20	1456	13	**9251**	4718	21673
4	11	**17958**	27	7	24	1157	19557
8	**5705**	16	2906	13	400	1910	11739
37	767	29	**517**	**821**	15	365	3020
39	**849**	29	724	27	171	597	2991
58	**518**	1	210	1	351	634	1883
66	**789**	0	515	11	7	37	1481
70	23	2	12	0	**525**	415	1404
71	**454**	7	199	0	135	460	1368
73	234	4	**304**	**444**	10	172	1304
74	84	29	86	**394**	19	491	1285
76	37	5	64	4	**463**	409	1263
77	532	0	**576**	37	3	23	1263
84	**447**	0	258	0	98	280	1154
91	22	1	21	13	**583**	252	1017

Figure 5: Selected Topics across Subreddits

Figure 6: Questions with unclear Topics in onebag subreddit

comparing to pre-pandemic. It is reasonable because during the pandemic the cruise did not operate and additionally we could not travel to another country(sometimes in the country, too.), so people's interest in cruise might decreased. On the other hand, the number of posts of two year period with road trips does not decrease much (-8.4%) during the pre-pandemic and pandemic period.This makes sense because during the pandemic domestic travels like road trips are not much restricted. Therefore, people's interest in these types of travel might stay in a similar level. We think these results also show our assumption that the number of posts(or proportion of posts) shows people's (Reddit users) interest. As we explained, the number of posts about cruises decreased during the period which may explain the decrease in people's interest, while the number of posts about road trips stayed at a similar level during the pandemic which then may suggest that people's interest were similar.

Time Period	Cruise	Camping & Hiking	Backpacking	Road Trip	All Posts
Pre-Pandemic	5,883 (2.22%)	8,737 (3.30%)	9,315 (3.52%)	8,473 (3.20%)	264,663 (100%)
Pandemic	3,886 (2.08%)	7,054 (3.78%)	7,726 (4.14%)	7,760 (4.14%)	186,675 (100%)
Post-Pandemic	9,788 (3.88%)	4,884 (1.94%)	15,633 (6.20%)	9,124 (3.62%)	252,330 (100%)

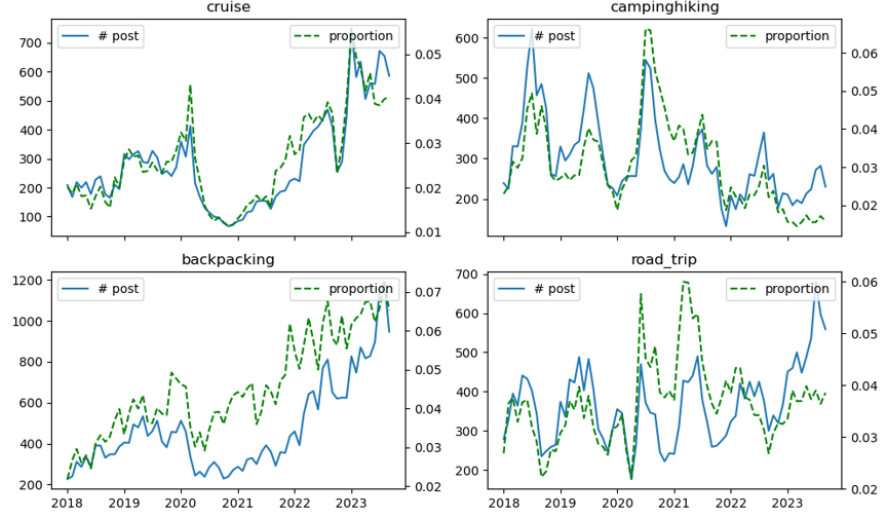Table 3: # of Posts of each Topic in Three Periods

Figure 7: The Number of post of each Topic over Months

One interesting finding is that the number of posts (also proportion) with cruise or backpacking topics increased dramatically after the pandemic even compared to before the pandemic(+66.4%, +67.8%). While the number of posts (also proportion) with camping&hiking topics decreased, and number of posts (also proportion) with road trip topics remained at a similar level(-44.6%, +7.7%). Also, the ANOVA test results (Table 4) for the average number of posts each month before and after the pandemic are all significant(at a 1% level) on these 4 topics.

Test Groups	P-value
Cruise (Pre-pandemic vs Pandemic vs Post-pandemic)	$4.17 * 10^{-14}$
Cruise (Pre-pandemic vs Post-pandemic)	$3.46 * 10^{-8}$
Campinghiking (Pre-pandemic vs Pandemic vs Post-pandemic)	$2.98 * 10^{-5}$
Campinghiking (Pre-pandemic vs Post-pandemic)	$6.16 * 10^{-6}$
Backpacking (Pre-pandemic vs Pandemic vs Post-pandemic)	$7.98 * 10^{-5}$
Backpacking (Pre-pandemic vs Post-pandemic)	$5.87 * 10^{-6}$
Road trip (Pre-pandemic vs Pandemic vs Post-pandemic)	$2.07 * 10^{-4}$
Road trip (Pre-pandemic vs Post-pandemic)	$3.54 * 10^{-3}$

Table 4: AVONA Result between Period of each Topic

From these results, we concluded that there is a possibility that after the pandemic people(at least Reddit users) become more interested in cruise travel and backpacking and less interested in camping & hiking.

5.2 Changes in Word Frequency

To better understand how people's interests in different topics changed and figure out why, we checked how often certain words were used for each topic group during different periods. First, we pre-processed the words by removing common words, special characters, and emojis, and applying lowercase to them. Then, we calculated the percentage of each word's usage in each period as follows:

$$WordFrequency = \frac{\text{number of documents contain } \mathbf{w}}{\text{total documents of each topic}}$$

Second, we calculated the difference between selected periods and sorted the outputs into 2 categories: a list of the top 100 words in each topic group with the highest differences in frequency, whose 50

	Cruise	Camping & Hiking	Backpacking	Road Trip
Words used more	kids, family, recommend	tents, stove	airline, flight, cabin, flying	flight, vegas, kids, airport
Words used less	wife, honeymoon, fiancee	lake, national, park, mountain	camping, hiking, yellowstone	shorts, camping

Table 5: Top Words Used Less/More between before and after COVID-19

words were used more after COVID-19, and the others were used less after COVID-19. We can see some interesting choices of words (Table 5) as follows:

(The complete output can be referred to in Table 7 in the Appendix.)

For Group 8 - Hiking&Camping: there were more discussions on pieces of equipment, and words suggesting nature were less used.

For Group 4 - Cruise: family-oriented terms emerged more frequently after COVID-19. (family, kids)

For Roadtrip and Backpacking groups: more discussions on flights or popular cities as destinations could be found. It is reported that flights to those popular destinations can be costly, which leads to a suggestion that people might become more open to more expensive trips post-COVID-19.

5.3   Change in Sentiment Analysis

We delved into the sentiment scores generated by VADER by examining 50 posts from each category within every topic group and then cross-referenced these results with our manually assigned labels. It became apparent that documents with notably positive sentiment often comprised positive reviews from past trips or reflected users' enthusiasm while seeking recommendations and advice. Documents categorized as having neutral sentiments predominantly consisted of pieces offering advice. On the other hand, posts with elevated negative sentiments tended to revolve around venting or recounting unfavorable experiences.

Despite the challenges posed by the pandemic, sentiments towards backpacking and road trips remained consistently positive, with no significant fluctuations in the proportions of posts of all 3 classes. However, the cruise topic experienced a notable decline in positive sentiment during the pandemic, accompanied by a slight uptick in negative and neutral sentiments. Post-pandemic, the number of positive discussions on the Cruise topic increased by 10% compared to the previous period and was 5% more than the pre-COVID period. Along with our previous finding regarding the Cruise topic, the result suggests that more and more people, especially those with family, gained more interest in cruise traveling and reached out to others on Reddit. According to a report by Reuters [OV23], there will be over 35 billion passengers by the end of 2023, and it is expected to have another increase of 6% in 2024. We assume that more families started to have an interest in cruise travel post-pandemic, which may explain the increase in discussions, excitedly, on Reddit, as a source of recommendation. Additionally, sentiment scores shed some light on the matter by showing that the majority of these discussions are favorably positive.

On the other hand, camping was distinctive in having more neutral sentiments than positive ones both before and during the pandemic, with comparable figures. Firstly, the number of posts on the camping and hiking topic significantly decreased after the pandemic stabilized (around 2022). This suggests that users who entered this topic largely left after the pandemic. Therefore, it is assumed that only people with a deeper interest in camping and hiking might have remained after 2022, leading to a prevalence of discussions on more specific topics (like equipment). Table 6 shows the number of unique authors and the average post per author. As you can see average post per author increased post-pandemic slightly. Additionally, 3362 out of 3577 are longtime members of the topic.

6   Conclusion

We have shown that based on the different analysis utilizing extracted topics by BERTopic, COVID-19 might have resulted in certain changes in people' references to 4 selected types of traveling. By cross

Period	# Post	# Unique Author	Avg Post/Author
Pre-Pandemic	8737	6818	1.281461
Pandemic	7054	5394	1.307749
Post-Pandemic	4884	3577	1.36539

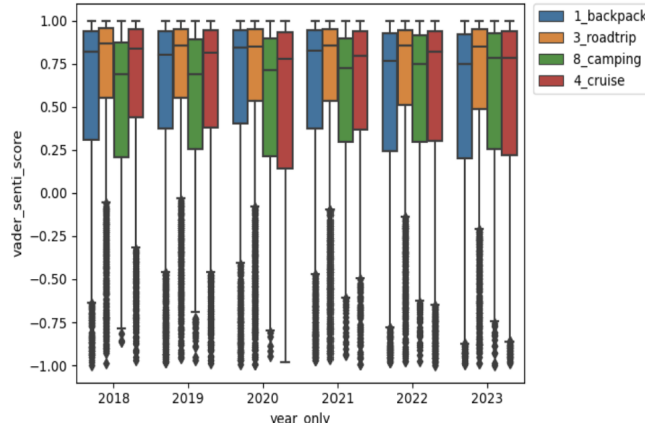Table 6: # Post, # Unique Users, and Average Post per Author

Figure 8: Sentiment Distribution Box-Plot of Each Topic

referencing some reports [OV23] as a better source of truth, we could point out that the trends from the expressed interests by Reddit users could mirror real-life trends to a certain extent such as the increasing demands in cruise traveling with their families.

Limitations: One limitation we encountered is that our topics groups might have overlooked some documents that demand a more in-depth understanding of the subjects. Since many of the documents are rather short, certain layers of information may not have been fully explored. This is partly because these posts are shared within subreddits with specific topics, where members are presumed to possess a certain level of knowledge on those topics. To enhance our coverage, we could explore the discussions of each post to further the analysis or consult experts to possibly provide better groups from clustering.

On the other hand, there is a possibility that the trends could be affected by other factors or external events during the time periods of our research. Some trends such as the significant decrease in the interaction rate of camping and hiking could not be explained well by our current analysis, whether it could be a consequence of the pandemic.

7   Future Scope

Figure 9: Topic Similarity before and after COVID

This project has shown that COVID-19 also made an impact on the concepts of travel of people to a certain extent. We planned to observe this phenomenon further to understand whether there exists a shift of interests on the individual level. In Figure 10, we focused on users who interacted multiple times during two periods: Pre-COVID and Post-COVID. We extracted sets of topics from these users in 2 periods, and compared them. As we can see, The majority of users tend to change the topics within the subreddit. We may be able to capture the potential change in interest better by utilizing the BERT pre-trained model [ML19] to group the documents into specific and more well-defined topics, which allows us to use all documents more efficiently.

8 Appendix

Travel-related Subreddit Explanation[2]

r/travel - The preeminent travel community with more than 7 million subscribers and a treasure trove of travel information.

r/shoestring: Combat the notion that you need a bulging bank account to tour the world by finding creative ways to keep travel costs down.

r/wanderlust: A place for travel discussion and information, and a community of people who share your love for travel.

r/solotravel: A useful resource for solo travelers, with stories and experiences from millions of solo travelers.

r/travelhacks: Tips and tricks from experienced travelers, and a good place for beginner travelers to ask questions.

r/backpacking: One of the largest travel subs, for people who understand the liberating feeling of having everything they own on their back.

r/campingandhiking: A resource for travelers who love camping and hiking, with knowledgeable people and exceptional stories and photos.

r/roadtrip : Your go-to resource for all things related to road trips. This is the place for information and inspiration on your next road trip adventure.

r/vagabond: A community dedicated to sharing low budget travel stories and experiences.

r/travelphotos: A place to remember why travel is so important, with eclectic travel photos from all over the world.

r/adventures: Find exciting travel experiences that will increase your heart rate.

r/longtermtravel - a subreddit for those interested in making long-term travel a lifestyle.

r/hitchhiking - a subreddit for those interested in hitchhiking, sharing their nomadic and adventurous experiences.

r/flights - a subreddit focused on flights, with information on airlines, frequent flyer miles, reward programs, and more.

r/onebag - a subreddit committed to travelling with minimal belongings, discussing what you actually need to travel.

r/couchsurfing - A subreddit for couchsurfers to discuss their experiences and the pros and cons of the Couchsurfing community and website.

r/airbnb - A subreddit for Airbnb guests and hosts to discuss their experiences and best practices for using the site.

r/travelpartners - A subreddit for travelers looking for travel companions and to share their travel experiences.

r/cruise - A subreddit for cruise enthusiasts to share stories, news, tips, and more about their experiences.

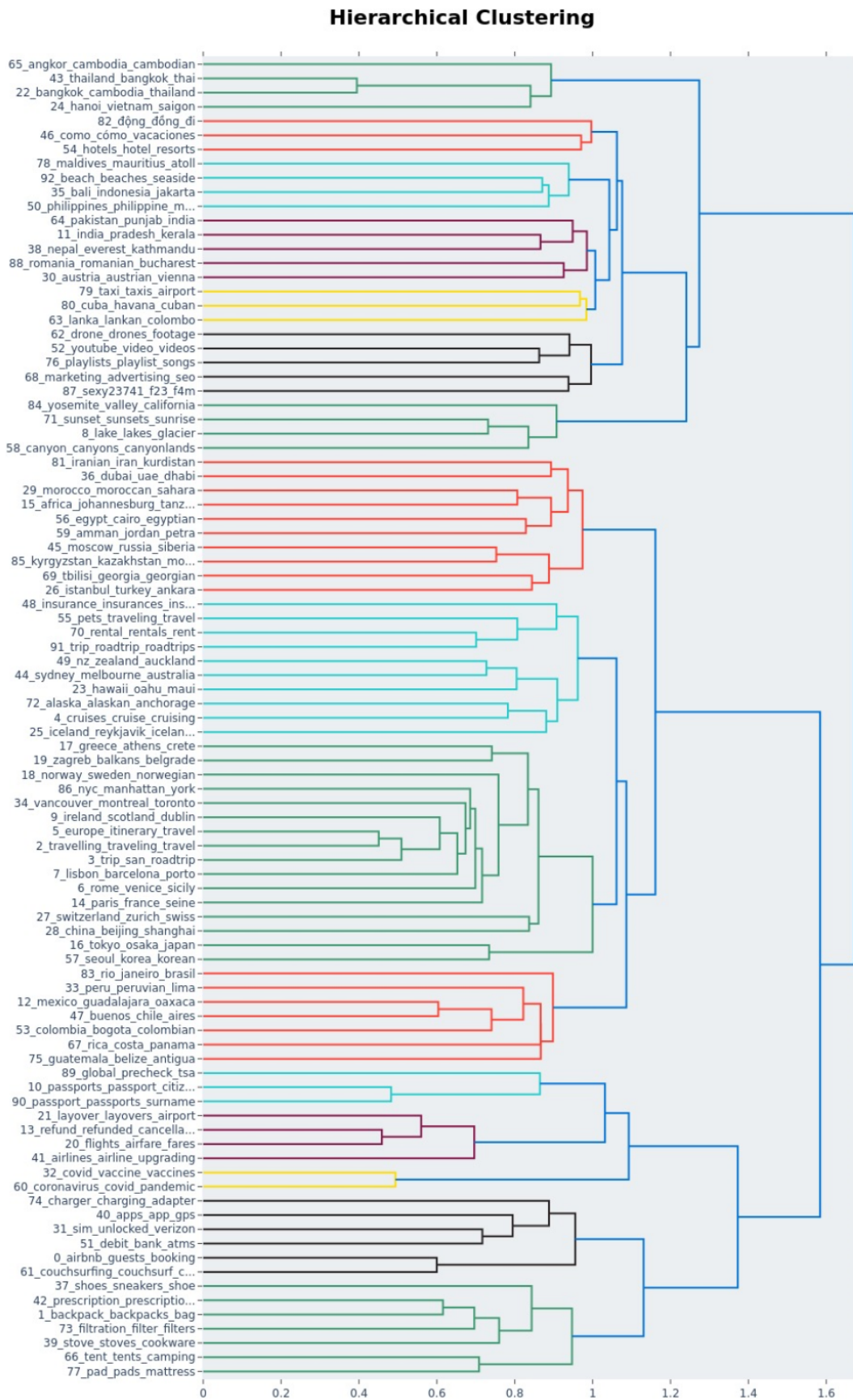[2]https://www.reddit.com/r/travel/comments/1100hca/the_definitive_list_of_travel_subreddits_to_help/

Hierarchical Clustering

Figure 10: Topic Clustering by BERTopic

8-Camping&Hiking

(a) More Used Words After COVID-19

Word	Freq Diff
im	0.0689
like	0.0537
tent	0.0499
backpacking	0.0483
would	0.0476
anyone	0.0427
camping	0.0421
food	0.0385
trip	0.0331
good	0.0328
get	0.0318
dont	0.0311
also	0.0293
looking	0.0293
know	0.0291
go	0.0271
want	0.0271
ive	0.027
meals	0.0264
one	0.0262
days	0.0247
something	0.0244
much	0.0243
recommendations	0.0237
use	0.0237
people	0.0229
find	0.0225
need	0.0223
really	0.0217
thanks	0.0213
two	0.0208
weight	0.0204
tents	0.0202
stove	0.0199
going	0.0197
time	0.0197
best	0.019
solo	0.0189
could	0.0188
lot	0.0174
make	0.0174
person	0.0171
small	0.017
coffee	0.0168
love	0.0167
well	0.0163
using	0.0161
bit	0.0159
used	0.0158
think	0.0157

(b) Less Used Words After COVID-19

Word	Freq Diff
lake	0.0525
national	0.0443
park	0.0356
view	0.0346
yosemite	0.0241
sunset	0.0213
mt	0.0197
canyon	0.0196
beautiful	0.0166
co	0.0159
sunrise	0.0159
colorado	0.015
trail	0.015
glacier	0.0143
near	0.0111
hike	0.0107
oc	0.0104
taken	0.0104
forest	0.0102
weekend	0.0094
np	0.0087
wa	0.008
ca	0.0077
mountain	0.0076
falls	0.0072
valley	0.0072
mount	0.0064
banff	0.006
took	0.006
mountains	0.0058
shot	0.0057
antelope	0.0055
alberta	0.0054
tahoe	0.0053
az	0.0049
pass	0.0049
rocky	0.0049
hiked	0.0048
moraine	0.0046
picture	0.0046
rainier	0.0045
river	0.0045
utah	0.0043
campsite	0.0042
louise	0.0042
arizona	0.0041
top	0.0041
yesterday	0.0041
gorgeous	0.0039
pic	0.0038

4-Cruise

(a) More Used Words After COVID-19

Word	Freq Diff
cruise	0.0549
ncl	0.0419
celebrity	0.041
alaska	0.0389
msc	0.0348
princess	0.0347
im	0.0343
would	0.0326
line	0.0311
food	0.0257
hi	0.0226
cruises	0.0224
dont	0.0207
anyone	0.0202
solo	0.0198
lines	0.0196
book	0.0193
virgin	0.019
cruising	0.0187
kids	0.0185
thank	0.018
one	0.0174
covid	0.0173
like	0.0172
family	0.017
good	0.0168
booked	0.0157
better	0.0155
booking	0.0151
know	0.015
us	0.015
love	0.0148
day	0.0144
alaskan	0.0143
looking	0.0143
trip	0.0142
recommend	0.0141
year	0.014
people	0.0139
time	0.0137
itinerary	0.0135
options	0.0132
encore	0.0128
end	0.0127
everyone	0.0127
holland	0.0123
never	0.0123
club	0.0119
havent	0.0117
want	0.0117

(b) Less Used Words After COVID-19

Word	Freq Diff
carnival	0.0739
seas	0.0514
cozumel	0.0304
ampxb	0.0212
caribbean	0.0206
grand	0.0205
cayman	0.0157
guys	0.0156
nassau	0.0144
beach	0.014
going	0.0139
allure	0.0135
oasis	0.0133
excursion	0.0132
bahamas	0.0127
harmony	0.0125
wife	0.0122
drink	0.0118
st	0.0118
rc	0.0117
amp	0.0115
escape	0.0114
symphony	0.0114
thanks	0.0113
edit	0.0111
tour	0.0106
honeymoon	0.0105
leaving	0.0099
island	0.0097
tips	0.009
royal	0.0089
liberty	0.0088
western	0.0088
juan	0.0085
maarten	0.0085
adventure	0.0081
january	0.0081
norwegian	0.0081
beverage	0.008
card	0.008
cheers	0.008
rccl	0.0078
fiance	0.0076
march	0.0073
turk	0.0072
questions	0.0071
ship	0.0071
jamaica	0.0069
vista	0.0069
freeport	0.0068

3-Roadtrip

(a) More Used Words After COVID-19

Word	Freq Diff
route	0.0309
car	0.0191
rental	0.019
stops	0.0188
xb	0.0174
itinerary	0.0134
drive	0.0133
hotel	0.0133
recommendations	0.0131
thank	0.0122
visit	0.0121
kids	0.0118
safe	0.0117
rent	0.0115
tia	0.0109
driving	0.0107
roadtrip	0.0105
october	0.0097
thoughts	0.0091
add	0.0085
nature	0.0085
cities	0.0084
recs	0.0084
state	0.0078
vegas	0.0078
airport	0.0074
avoid	0.0074
usa	0.0074
chill	0.0072
flight	0.007
dallas	0.0065
lol	0.0065
big	0.0064
las	0.0064
santa	0.0064
barbara	0.0063
museum	0.0063
partner	0.0063
better	0.0062
mid	0.0062
vs	0.0062
la	0.006
hello	0.0059
stop	0.0059
overnight	0.0058
walking	0.0058
worth	0.0058
cars	0.0057
fly	0.0057
rentals	0.0057

(b) Less Used Words After COVID-19

Word	Freq Diff
camping	0.0555
thanks	0.0415
looking	0.0411
guys	0.0341
know	0.0333
ampxb	0.0324
go	0.0323
im	0.03
hiking	0.0285
places	0.0274
camp	0.0272
anyone	0.0271
suggestions	0.027
hey	0.024
friends	0.0234
ive	0.0225
national	0.0223
going	0.022
hike	0.022
grand	0.021
want	0.0205
good	0.02
canyon	0.0197
day	0.0196
ill	0.019
utah	0.0166
near	0.016
budget	0.0158
really	0.0157
two	0.0156
zion	0.0155
let	0.0154
time	0.0151
girlfriend	0.0148
id	0.0144
new	0.0144
plan	0.0144
friend	0.0143
much	0.0143
set	0.0143
could	0.0142
may	0.0142
park	0.014
well	0.014
edit	0.0138
appreciated	0.0136
somewhere	0.0135
road	0.0134
yellowstone	0.0133
great	0.0132

1-Backpacking

(a) More Used Words After COVID-19

Word	Freq Diff
l	0.0591
bag	0.0478
personal	0.0466
item	0.0406
recommendations	0.0392
looking	0.0339
carry	0.0337
luggage	0.0315
hi	0.0289
airlines	0.0274
carryon	0.0267
xb	0.0246
flight	0.0222
size	0.0222
fit	0.0216
backpack	0.0209
aer	0.0208
flying	0.0201
use	0.0199
suitcase	0.0198
would	0.0198
thank	0.0193
seat	0.0192
cm	0.0187
osprey	0.0187
dont	0.0185
bags	0.0182
dimensions	0.0179
air	0.0162
laptop	0.016
also	0.0159
one	0.0159
hello	0.0151
sling	0.0151
mini	0.0143
small	0.0143
max	0.0141
pocket	0.0138
cabin	0.0137
flights	0.0136
ryanair	0.0136
open	0.0134
checked	0.0132
patagonia	0.0132
baggage	0.0131
find	0.013
cotopaxi	0.0125
hole	0.0125
bellroy	0.0123
europe	0.0123

(b) Less Used Words After COVID-19

Word	Freq Diff
ampxb	0.052
gear	0.0295
guys	0.0248
tortuga	0.0237
setout	0.0219
pair	0.021
boots	0.0196
wool	0.0196
url	0.0187
shoes	0.0186
merino	0.017
edit	0.0165
list	0.0155
socks	0.0151
uniqlo	0.0151
hiking	0.0148
camera	0.0135
review	0.0132
pants	0.0126
first	0.0124
amp	0.0123
two	0.0121
shirts	0.012
camping	0.0113
shirt	0.0113
packing	0.0112
jacket	0.011
divide	0.0109
right	0.0109
shorts	0.0108
jeans	0.0101
hey	0.0099
goruck	0.0098
stuff	0.0098
underwear	0.0098
clothing	0.0096
winter	0.0096
gr	0.0094
weeks	0.0093
little	0.0092
synapse	0.0091
pairs	0.0089
asia	0.0088
pretty	0.0088
charger	0.0087
battery	0.0085
heres	0.0085
months	0.0083
outlier	0.0082
far	0.0081

Table 7: Word Frequency Difference for Pre-COVID-19 and Post-COVID-19

References

[Adm22]   International Trade Administration. Ita data release: January 2022 international air passenger travel. *Travel and Tourism*, 2022.

[AJ21]   Esam Alzahrani and Leon Jololian. How different text-preprocessing techniques using the bert model affect the gender profiling of authors. 2021.

[DCLT18]   Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[DSP21]   Denise Capasso da Silva Deborah Salon, Matthew Wigginton Conway and Ram M. Pendyala. The potential stickiness of pandemic-induced behavior changes in the united states. Technical report, 2021.

[Gro22]   Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[HH18]   Md Mehedi Hasnat and Samiul Hasan. Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. *Transportation Research Part C: Emerging Technologies*, 96:38–54, 2018.

[ML19]   Naman Goyal Marjan Ghazvininejad Abdelrahman Mohamed Omer Levy Ves Stoyanov Luke Zettlemoyer Mike Lewis, Yinhan Liu. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. 2019.

[oH21]   National Institute of Health. Mental health during the covid-19 pandemic. 2021.

[OV23]   Doyinsola Oladipo and Granth Vanaik. Travelers ready to set sail on cruises at record levels next year, Dec 2023.

[RAM+17]   Taha H. Rashidi, Alireza Abbasi, Mojtaba Maghrebi, Samiul Hasan, and Travis S. Waller. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, 75:197–211, 2017.

[SBPC16]   Chihyung Michael Ok Seunghyun Brian Park and Bongsug Kevin Chae. Using twitter data for cruise tourism marketing and research. *Journal of Travel & Tourism Marketing*, 33(6):885–898, 2016.

[SLR21]   James Manyika Sven Smit Kweilin Ellingrud Susan Lund, Anu Madgavkar and Olivia Robinson. The future of work after covid-19. 2021.