Evaluating Changes in Mobility Behavior Throughout the COVID-19 Pandemic Using Twitter

Jenny Wang
Wellesley College
Wellesley, Massachusetts, USA
jw10@wellesley.edu

ABSTRACT

Mobility sustained dramatic impacts during the COVID-19 pandemic. Government lockdowns, mask mandates, and risk of exposure to the virus led to shifts in travel behavior which have yet to return to "normal". Recent studies primarily focus on surveys and polling techniques to understand public sentiment toward transportation and mobility. Existing literature suggests that social media platforms such as Twitter hold valuable information which can be extracted in order to help improve transport management systems and evaluate transit ridership opinion. This study builds on extant research by exploring Twitter's potential as a consistent source of user mobility data. In this work, we present a highly curated dataset of 9,123 tweets about pandemic and travel that we are making available to the research community via our COVID-19+Transportation GitHub repository. Leveraging a large Twitter dataset reflecting COVID-19 chatter, we filter tweets based on mobility and transportation keywords [6]. In addition to releasing this dataset for research use, we also propose a framework for analyzing shifts in balance from early-pandemic to late-pandemic attitudes regarding several modes of transportation—motorized (cars, rideshare), non-motorized (biking, walking), and public transit—using shift diagrams and sentiment analysis tools. We use text mining techniques to extract coherent themes related to transportation during the pre-vaccine and post-vaccine period, then conduct sentiment analysis to assess public opinion over time. Overall we find that, among opinionated tweets, sentiment regarding transportation and COVID-19 is consistently negative from January 2020 to April 2022. This study seeks to inform future policymaking by utilizing discourse on Twitter to better understand early and late pandemic attitudes in regard to transport, travel behavior, and the longer-term impacts of COVID-19 on the transportation industry.

ACM Reference Format:
Jenny Wang. 2022. Evaluating Changes in Mobility Behavior Throughout the COVID-19 Pandemic Using Twitter. In *Proceedings of Data and Text Mining for the Web (CS 315 '22).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

CS 315 '22, May 16, 2022, Wellesley, MA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Transportation networks provide essential support to the growth of trade, communication, and socioeconomic development. Although transportation underpins the complex systems which allow for economic and social interactions, it is little acknowledged when compared to other influences such as telecommunications, production technologies, and the Internet [14]. However, its key role in the development of society, in generating mobility, and in providing access to essential resources outside the home suggest that further understanding of the interactions between people and agents of transportation is critical.

Mobility is often evaluated through three categories: motorized, non-motorized, and public transportation [4]. During the COVID-19 pandemic, the state of mobility was dramatically impacted. Abdullah et al. found a significant change in the primary purpose for travel during the pandemic, as well as a shift from public transport to private transport (cars, motorbikes) and non-motorized (biking, walking) modes of transportation. Researchers speculate that restrictions imposed by authorities and fear of infection are the main reasons for short-term adjustments to travel behavior and mode preferences [1]. However, the long-term effects of the pandemic on mobility habits are uncertain.

A poll conducted in February 2022 suggested that most adults believe the worst of the pandemic is over but disagree on when a "return to normal" should happen and what it means [24]. COVID-19 continues to permeate daily life in spite of the growing trend toward reopening national economies. On March 10, 2022, the CDC announced that they continue to recommend masks on public transportation and transportation hubs, but will revisit the policy in mid-April 2022. Throughout this period of uncertainty, policymakers will seek out guidance from both health professionals and the general public to inform future action. We hope to employ Twitter as a data source because it is a platform where users share their general day-to-day opinions on all aspects of life. The findings from this study will provide supplementary data and analysis which decision-makers can use to understand mobility behavior, feelings toward travel, and the longer-term impacts of COVID-19 on the transportation industry. We choose to collect and examine worldwide data in order to also provide this resource to international communities.

Research Questions

We were motivated to contribute a dataset which intersects mobility and the COVID-19 pandemic, and to answer the following research questions:

RQ1 : Which themes surrounding transportation and mobility behavior are trending the most throughout the pandemic?

RQ2 : Are there distinct shifts in sentiment about modes of transportation (non-motorized, motorized, and public transit) between the early-pandemic and late-pandemic periods?

[RQ1] explores the distribution of themes and topics within the transport-related tweets. Our goal with this research question is to determine how conversations about transportation and mobility changed throughout the COVID-19 pandemic. For example, if top conversations show that interest is shifting towards non-motorized modes of transportation such as walking and biking, then government authorities can prioritize building walkable cities. Alternatively, if findings indicate an increased preference toward ridesharing services, this could create an opportunity to consider collaborative solutions to issues like congestion and pollution. Recent studies suggest that ridesharing services may increase emissions and contribute to congestion, disputing previous beliefs that Uber and Lyft reduce personal vehicle usage and could simultaneously reduce congestion and increase public transit usage [27]. If consumer preferences are shifting away from public transit, government authorities can work with this knowledge to better understand whether it is more socially optimal to reinvigorate the appeal of public transit or to redirect efforts toward other means of mobility.

[RQ2] seeks to analyze cross-sectional and time series variation in the data. It examines general topics and sentiment toward non-motorized, motorized, and public transit throughout the entire pandemic and uses key events to explain peaks and troughs. While [RQ1] focuses on observing mode-specific trends during the pandemic, [RQ2] utilizes natural language processing techniques to understand the polarity levels in Twitter conversations about transportation and mobility.

2 LITERATURE REVIEW

2.1 Traditional Data Collection Methods

Considerable prior literature focuses on the collection of survey data to motivate transportation planning. Transport planners traditionally base their decisions with the needs of the "average" user in mind using questionnaires which cannot always capture attributes which are specific to gender, race, or age [26].

For instance, Abdullah et al. used the following options to answer a question which measured changes in commuting behaviors due to COVID-19:

(1) I never go to office/college and I work/study at home.
(2) I go to office/college/work place less often (less than 3 times per week).
(3) I go to office/college few days per week and work/study from home for the rest of the time.
(4) Lost my job/not studying these days
(5) I go to work on-call.
(6) Nothing changed.

These options limit a respondent's ability to explain nuances in their personal circumstances [1]. This causes a loss in context which could be crucial toward understanding how to approach user needs. Twitter creates an opportunity to observe a user's specific thoughts, opinions, and preferences in real time. Our dataset aggregates opinions about transportation in relation to the COVID-19 pandemic from tweets. Given the evolving nature of the coronavirus, it is not unlikely that a resurgence could occur. Thus, it is important to acknowledge how pandemic measures in the past have affected people's lives before future policies are enacted.

2.2 Using Social Media to Crowdsource User Feedback

The dramatic increase in social media discourse creates an opportunity to supplement traditional methods with alternative data collection techniques. Analysis of social networking platforms have been used to measure emotion and sentiment in various industries. De Choudhury and Counts (2013) used Linguistic Inquiry and Word Count (LIWC) to perform sentiment analysis on a Twitter-like microblogging platform called OfficeTalk in order to understand affect among employees in the workplace [8]. Chen et al. matched geotagged tweets with Foursquare venues to create sentiment profiles for individual stores in chain businesses (e.g. Starbucks locations in San Francisco) using a logistic regression based sentiment analyzer [7].

Within the sphere of transportation planning, the use of social media to analyze user feedback is still relatively novel. A study in traffic management by Noaeen and Far (2020) sought to utilize the dynamism of social media to explore how its data could support urban traffic management systems [19]. They used an unsupervised topic modeling approach–Biterm Topic Modeling (BTM)–to find key traffic-related topics and used the Google Vision API to classify Twitter images by traffic-related content. Their overall results found that while social media analysis is effective in providing crowdsourced information from the end-users of traffic management systems, there are limitations to developing this data into real-time systems that can provide planners with time and location-specific traffic information. In contrast to Noaeen and Far's study into Twitter's efficacy in providing immediate insights, Vasquez-Henriquez et al. (2019) collected tweets over a six-month time period to analyze crowdsourced transportation sentiment by users who self-reported their location as Santiago, Chile. Their work sought to characterize mode and gender differences in transport perception in order to provide insights into a wider range of metrics which survey data cannot capture [26]. They associated users and tweets to latent features interpreted as modes of transportation using a semi-supervised method called Topic-Supervised Non-Negative Matrix Factorization and classified multiple heuristics which predicted gender labels using Stochastic Gradient Descent [18]. Vaquez-Henriquez et al.'s work suggests that, while social media analysis may fall short in regards to measuring real-time data, it can help transportation planners better understand the daily travel experience over time.

2.3 Analysis of Travel Behavior During the COVID-19 Pandemic

New concerns for transportation and society emerged with the onset of the COVID-19 pandemic. Fear of infection, stay-at-home orders, and social distancing led to plummeting demand for travel. In March 2021, the TRIP (Transportation Research: Interdisciplinary Perspectives) journal presented a series of international insights into the pandemic's impact on issues such as changes in travel behavior, changes in transit operations, and demand for other mobility services such as rideshare and bikeshare [16]. The overall findings

showed that COVID-19's short-term impacts led to significant reductions in the airline, cruise ship, and public transit industries and that reductions in mobility through quarantine, isolation, and social distancing decreased the spread of the disease [16, 23, 13, 9].

Researchers around the world used a variety of theoretical and analytical methods to approach changes in travel perception and behavior due to the COVID-19 pandemic. An empirical study by Nurhadi and Suryadari (2021) surveyed a sample of rail passengers in Jakarta. Their results found that perceptions of health, psychological, and time risks were concerns of potential passengers or passengers when using public train stations and public transportation facilities [20]. Another research paper by Schaefer et al. (2021) analyzed a sample of over 4,000 participants in the Hanover region of Germany to determine the substitution effect of bikes and cars for public transport during the pandemic. The results found that (1) local light rail and bus are substituted by bike, car and working from home, (2) women have a higher level of fear of infection than men have during public transport use and therefore reduce public transport use more, and (3) income displays a positive effect on increased car use while cycling is independent of socio-economic indicators but instead driven by the eco-consciousness of users, among other outcomes [22]. These results show similar findings to the study conducted by Abdullah et al. (2020), which sampled 1,203 responses from various countries around the world. Furthermore, Brough et al. (2021) highlighted the socioeconomic disparities in travel behavior during the COVID-19 pandemic using data from SafeGraph, which tracked locations of 100,000 mobile devices in King County Washington in order to measure changes in travel intensity between census block groups [3].

Using surveys and questionnaires, [1], [20], and [22], were able to hone in on binary (i.e. "yes/no") or multiple choice responses to mobility behavior. Using SafeGraph, [3] was able to measure correlations between travel intensity and socioeconomic characteristics of neighborhood locations. The preceding papers study the relationship between transportation and COVID-19 through non-social media perspectives. Meanwhile, to take advantage of the open-ended opinions and real-time data from social media platforms, Habib and Anik (2021) proposed a framework to analyze public discourse regarding transport systems in Twitter [12]. Their methodology involved categorizing tweets into themes and subthemes, then applying text mining (visualized using word clouds) and topic modeling using the Latent Dirichlet Allocation (LDA) technique. This study underscored the substantial amount of information that can be gleaned from frequently occurring topics. However, data was collected using limited search queries—public transit and COVID-19, car and COVID-19, bicycle and COVID-19, and reopening and COVID-19—and spanned a one-month period between May 15, 2020 and June 15, 2020. Our research builds on [12] by expanding the transportation keyword search set and sampling tweets throughout the entire pandemic. The synthesis of these results, obtained from both theoretical and analytical methods, help guide our research and uphold the validity of our social media feedback.

3 DATA AND METHODS

This section presents the methodology used to conduct our analysis. We first extracted tweets from a large COVID-19 dataset, filtered the tweets by 130 manually selected transportation keywords, then preprocessed the remaining tweets for analysis. Due to time and resource constraints, the transportation keyword collection process was fairly unrefined. We discuss this further in the *Data Concerns and Limitations* section. Using randomly selected date ranges throughout the COVID-19 pandemic, we collected 16,789 tweets from the COVID-19 dataset which contained our transportation keywords. After filtering out irrelevant tweets, we obtained 9,123 tweets related to COVID-19 and transportation.

3.1 Transportation Keywords

The transportation keywords consists of a list of 130 unique words categorized as "public transport", "motorized", "non-motorized", or "other". The list of public transport keywords contains common transit words like bus, subway, and railroad, as well as location-specific words like Monorail, BART, or silver line. Similarly, the motorized words contain common vehicle words, ridesharing buzzwords (Uber, Lyft), and car brands (Ford, Subaru). In the "non-motorized" category, we include words like walk, pedestrian, and bicycle, while in the "other" category, we use words that are loosely related to mobility, such as shipping, logistics, and city. We chose to include an "other" category in order to expand the range of the COVID-19+Transportation dataset. For instance, we capture keywords like "freight", "cargo", and "truck" so that future researchers can study how COVID-19 affected supply chain. For the purpose of answering this paper's research questions, we do not focus on the "other" category in our analysis.

Figure 1: Transportation Keywords By Category

public transport	motorized	non-motorized	all
transportation	vehicle	walking	infrastructure
transit	road	running	shipping
bus	automobile	bike	truck
subway	auto	pedestrian	ferry
line	traffic	lane	passenger
rail	car	bicycle	freight
...	...	...	...

We note that some keywords that are included have multiple synonyms, some of which appear frequently in everyday language. For instance, the words line and running are part of the "public transport" and "non-motorized" categories, respectively. This causes tweets such as the following to be included in our raw dataset:

> Elton John will host a benefit special that will pay tribute to front **line** health care workers and first responders amid the coronavirus pandemic, and seek donations

> So, my ex is COVID+ and my son started **running** a slight fever this evening and so did his younger half brother. As far as I know my daughter is just fine right now. The boys both always have issues with

> bronchitis every winter so I'm trying not to freak out right now.

The first tweet contains the word "line" within the phrase "front line" to describe essential workers in the healthcare industry who must physically show up to their jobs, while the second tweet uses the word "running" to describe suffering from a body temperature higher than normal. Neither of these words are related to transportation. However, it is possible that we could capture a tweet that says this:

> If you're not from Boston you don't quite get how exciting it is to see new trans on the orange or red line to put it in perspective, the entire existing orange line fleet is from the early 70s and the red line fleet is from the 60s through the 80s

The tweet's misspelling of the word "trains" means that only the "line" keyword led to its inclusion in our dataset. Thus, we make an effort to remove any unrelated tweets during the preprocessing stage.

3.2 Data Collection

We utilized the Tweepy API, an open-source Python package to access tweets from the COVID-19 TweetIDs GitHub repository. This repository contains an ongoing collection of tweet IDs associated with COVID-19 starting from January 21, 2020. We obtained our raw dataset of transportation-related tweets in three steps. First, we accessed the tweet ID text files from the COVID-19 dataset. Second, we "rehydrated" each tweet ID to get all fields of the Tweet object in JSON form. Using the "text" field, we filtered the tweets by transportation keywords. Finally, we stored our COVID-19+Transportation tweets into directories organized by YYYY-MM. Since the COVID-19 TweetIDs dataset currently contains over 2 billion tweets, we randomly selected short date ranges between January 21, 2020 and April 1, 2022 to perform our data collection on.

3.2.1 Accessing COVID-19 Dataset. The COVID-19 Tweet-ID files were stored in folders that indicated the year and month of the collection (YEAR-MONTH). Individual Tweet-ID files were uploaded from 0:00-23:00, representing each hour in the day. Each file name followed the same structure, with a prefix "coronavirus-tweet-id-" followed by the YEAR-MONTH-DATE-HOUR. Because technical difficulties occurred during brief time frames, some corresponding files were missing. We covered all available date and time combinations by iterating over all files using the YEAR-MONTH-DATE-HOUR pattern and ignoring any URLs in which there was a 404 Not Found Error.

3.2.2 Rehydrating Tweet IDs. Twitter's Terms of Service requires that public datasets release only the tweet IDs of collected Tweets. To access the text fields of our COVID-19 tweets, we used a tool called Twarc. **Twarc** is a Python library that collects and archives Twitter data via the Twitter API. The benefit of using Twarc was that it gave us two methods called hydrate and dehydrate which could generate tweet JSON from a .txt file of tweet IDs and generate tweet IDs from a file of tweets. In our analysis, we first generated json files of COVID-19+Transportation tweets by iterating through the COVID-19 dataset and collecting text containing our transportation keywords. After generating json files, we used Twarc to dehydrate the tweets into a compact text file of tweet IDs. While attempting to retrieve the tweets, we noticed that a significant proportion were inaccessible. Of the tweet IDs requested, we accessed an average of only 0.19%. Furthermore, only an average of 2.34% of tweets contained a transportation keyword. This means we requested roughly 375 million tweets to obtain our 16,789 tweets.

3.2.3 Storing COVID-19+Transportation Tweets. Initially, we saved the COVID-19+Transportation tweets in directories named according to collection time ranges. Each tweet file followed the naming convention *covid-mobility-tweet-starting-yyyy-MM-dd HH:mm:ss.json*. We sorted each tweet into its respective file by day, then structured these tweet files into directories organized by year and month. In order to share this dataset for public use, we dehydrated these tweets into their unique IDs and made them available on GitHub through our *COVID-19-Transportation-TweetIDs* repository.

3.3 Data Cleaning and Preprocessing

3.3.1 Data Cleaning. After obtaining our 16,789 raw tweets, we used two classifiers to determine if tweets were relevant or irrelevant to our COVID-19+Transportation dataset. We first created a classifier to remove non-English tweets. Using the "lang" field in each tweet's json body, we filtered out non-English tweets. For any tweet which did not contain a language field, we used the **langid** Python library to classify its language. We also cleaned tweets by verifying words with multiple synonyms, as noted in section 3.1. We built a classifier to determine if tweets containing the keywords "line", "run", "running", "Ford", or "Lincoln" occurred alongside another transportation keyword. If not, we labeled the tweet irrelevant and manually checked it for relevance. After the cleaning stage, 9,123 tweets remained.

3.3.2 Data Preprocessing. Tweets often contain informal texts using slang, emojis, abbreviations, and URLs. Hence, traditional preprocessing techniques trained on conventional texts are less effective on Twitter data. Our tweets were preprocessed using similar techniques to those described in [15]. We first lowercased all our tweets, then tokenized them using the **NLTK TweetTokenizer** library. Next, we removed English stopwords using the NLTK English corpus. Afterwards, we focused on removing unnecessary punctuation and stand-alone numbers which did not add meaning to our tokenized tweets. For example, the number "19" occurred quite frequently in our corpus because users split the term "COVID-19" into the words "COVID 19". We also chose to remove user mentions because we found that observing commonly retweeted usernames did not help address our research questions. Because we sampled from a COVID-19 dataset and saved tweets based on transportation keywords, we knew that our tweets would relate to COVID-19 and transportation in some regard. Hence, to observe heterogeneity in our COVID-19 and transportation topics, we also removed the COVID-19 keywords used to generate the large COVID-19 Twitter dataset we sampled from. We chose to keep retweeted tweets in our corpus because greater retweets reflects interest in the topic.

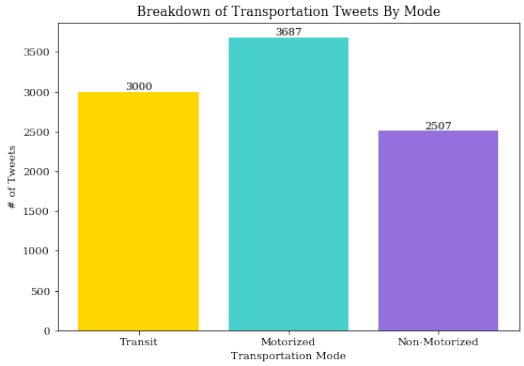


Figure 2: Breakdown of Transportation Tweets By Mode

4 ANALYSIS AND RESULTS

4.1 RQ1: Text Analysis

Our corpus of 9,123 tweets was used to analyze [RQ1]: which themes surrounding transportation and mobility behavior are trending the most throughout the pandemic? We analyze these themes by category, since each mode of transportation presents its own unique advantages and challenges. Because tweets could contain keywords from multiple categories, we chose to count these tweets in each related category. Figure 2 shows a breakdown of tweets collected per category. Each tweet category contains between 2,500 and 3,700 tweets.

4.1.1 Word Clouds. Figure 3 displays word clouds corresponding to each mode of transit. These word clouds were generated after removing the COVID-19 keywords. We found that the highest occurring words in the transit category were "people", "day", "train", "new", "service", "home", "work", and "school". These top terms suggest that many people tweeted about their experiences related to their routes between home, work, and school. It is less clear whether tweets in the motorized category were related to specific topics; again, "people" appears at the top of the list, followed by "new", "one", "time", and "day". In the non-motorized category, "people", "think", "everyone", "one", and "around" occur most frequently. However, the prevalence of sentimental words such as "lighthearted", "happy", and "painful" suggest that discussion about COVID-19 and the non-motorized category were more associated with emotional experiences during the pandemic.

4.1.2 Shift Graphs. Figure 4 shows pairwise comparisons between the corpus of tweets prior to the vaccine and the corpus of tweets after the vaccine's roll-out. We generated word shift graphs for each mode of transportation. The word shift framework is used to visualize differences between two texts according to a measurement like word frequency, sentiment, or information content [11]. Gallagher et al. presented "Generalized Word Shift Graphs" in 2020 to visualize quantitative measurements of variation between two texts. We used the associated **Shifterator** Python library to construct a proportional shift diagram.

We defined the cut off between *pre-vaccine* and *post-vaccine* to be March 8, 2021 because it coincided with the day the CDC announced that fully vaccinated people can gather indoors without masks. This announcement suggested that public health officials were beginning to loosen strict pandemic safety measures due to reported vaccine efficacy. Furthermore, the announcement indicated that COVID-19 cases were declining and hence, that perceived risks of socializing were decreasing. We expected that dividing the pre- and post-vaccine periods would cause us to observe more conversations reflecting "fear of infection" sentiment prior to March 2021, and less so after this time.

In the shift graph for transit, our early-pandemic results suggest that conversations about COVID-19 and transportation focused on shutdowns to transportation access, with particular focus on cities. In tweets later in the pandemic, our results showed greater considerations for railway employees and their risk as essential workers. Furthermore, the post-vaccine period accumulated more conversations about the issue of mask mandates on public transit. An interesting finding was that words like "weapons" and "assault" appeared heavily in the pre-vaccine period. Upon further inspection, we found that a frequently retweeted tweet about China's harsh lockdown measures contributed to the prevalence of these terms in the pre-vaccine period.

In the shift graph for motorized transportation, we found that a highly retweeted tweet dominated most of the pre-vaccine motorized tweets:

> My father died of Covid alone in a hospital. I had to say goodbye to him over a phone. Trump got a joyride to sooth his desperate need for attention, while endangering the lives of the Secret Service people in the car with him. To hell with him and all who enable him.

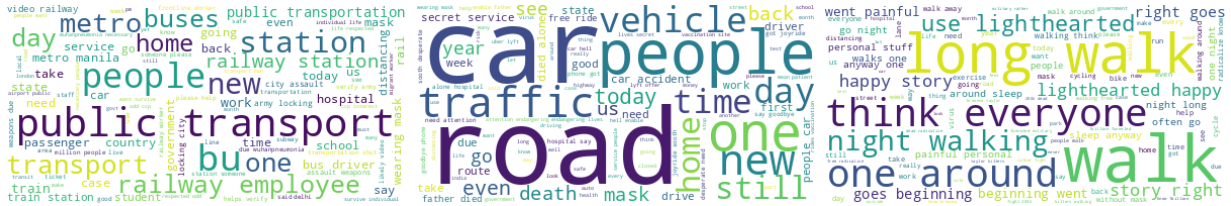
This tweet explained a significant proportion of the word differences between the pre-vaccine and post-vaccine periods. In general, we found that the motorized category was the most difficult to obtain substantial patterns from. Keywords like "car", "road", and "drive" are used ubiquitously in everyday life, meaning it was far more likely to capture data about day-to-day life rather than meaningful information about motorized transportation and COVID-19. For example:

> "I'm a Brit living in France, I can jump in my car and be in Spain within 4 hours, no checks whatsoever and covid is everywhere in France and at higher levels than the UK. Nothing to do with health all to do with politics.

4.2 RQ1: Findings

Given the time constraint on this project, we were able to capture only limited data. Depending on the size of the COVID-19 file and the number of retrievable tweets, it could take 12 hours to obtain a single day's worth of tweets. For instance, it took nearly two hours to obtain 11 tweets from 1:00-1:59 AM on January 18th, 2021. We expect that collecting more data would help with both discovering patterns and training classifiers to remove irrelevant tweets during preprocessing.

We experienced similar ambiguity with the non-motorized tweets. The non-motorized category was dominated by several highly retweeted tweets which focused on particular mobility topics. Although we captured a diverse assortment of relevant tweets, they



(a) Transit (b) Motorized (c) Non-Motorized

Figure 3: Word Cloud by Transportation Mode

were overshadowed by the highly retweeted tweets. Thus, we recommend that future studies consider how to capture relevant "popular tweets" which may skew the dataset, while reinforcing the status of equally relevant but less visible tweets.

In summary, our analysis of RQ1 found that people tended to tweet about their experiences related to their routes between home, work, and school and that there were noticeable shifts in conversations between the pre- and post-vaccine periods.

4.3 RQ2: Sentiment Analysis

To characterize sentiment, we followed the framework of the TweetEval sentiment analysis benchmark which uses a RoBERTA-base model trained on about 60 million tweets [2]. According to Barbieri et al., the language model *RoBERTA* was chosen because it is one of the top performing models in the General Language Understanding Evaluation (GLUE) benchmark [17]. The TweetEval benchmark contains several classification tasks including emoji recognition, hate speech detection, stance detection, and sentiment analysis. On sentiment analysis, the RoBERTA model's validation results performed with an average accuracy of 73%. We implemented our analysis using the Hugging Face **transformers** library and the prebuilt TweetEval model, *Twitter-roBERTa-base for Sentiment Analysis* , from the Cardiff NLP group at Cardiff University [2].

Figure 5 shows the monthly frequency of opinionated transportation tweets over the COVID-19 pandemic by mode of transportation. Our findings indicated that emotionally-charged tweets related to COVID-19 and transportation were consistently more negative than positive over time for all categories of transportation.

However, we noticed that more variation in sentiment occurred in the non-motorized category, which could be attributed to non-mobility related tweets captured by the keyword search. We noticed that generalized sentiment toward non-motorized transportation methods seemed to improve during the pandemic. According to a recent study about city infrastructure, it is possible that residents in major cities who frequently traveled by car or bus opted to use non-motorized forms of transit to abide by social distancing measures [10]. This phenomenon likely led to increased interest in non-motorized forms of transport—which also serve as sources of exercise and immersion in the outdoors—and thus higher positive sentiment. Since non-motorized transport was actually made *more* accessible during the pandemic, we chose to focus the majority of our analysis on the public transit and motorized categories.

In RQ2, we asked: are there distinct shifts in sentiment about modes of transportation (non-motorized, motorized, and public transit) between the early-pandemic and late-pandemic periods? We found that the proportion of sentiment did not rise or fall in a linear fashion, but rather, followed a periodic trend. The amplitude peaked in September 2020 for both transit and motorized transportation tweets. Furthermore, both graphs showed a brief period in May 2021 where the proportion of positive tweets increased and negative tweets decreased. We explored possible explanations for this phenomenon using historical COVID-19 data. This increase in net positivity occurred just weeks after the United States surpassed 200 million vaccinations, COVID-19 cases were trending downward, and COVID-19 vaccines were becoming increasingly available worldwide [21]. The sudden spike in net negative sentiment occurred in June 2021, when the Delta variant became the dominant variant and initiated a third wave of COVID-19 infections [5].

4.4 RQ2: Findings

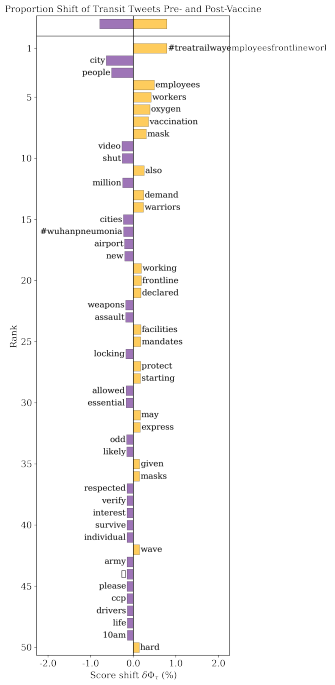
After observing the periodic trends in transportation and COVID-19 sentiment, we were curious about how much this negativity could be explained by the pandemic's outlook. First, we hypothesized that news about higher COVID-19 death tolls and higher infection rates would lead to greater concern about contracting the virus [1]. Second, we hypothesized that higher COVID-19 death tolls and higher infection rates would cause governments to tighten "Do Not Travel" advisories and mask requirements for mass transit. These hypotheses implied that sentiment toward travel, through both commuting and pleasure, would be considered more unsafe, and thus, negative. We also expected the reverse—that decreased death tolls and decreased infection rates would lead to increases in positive tweets related to COVID-19 and transportation.

The resulting comparison for COVID-19 death tolls and transit tweets is shown in Figure 6. We overlaid a graph of the *United States Biweekly confirmed COVID-19 deaths per million people* from Our World in Data over the % Polar Transit and Motorized Tweets across the same time range [21]. Our results aligned quite precisely with our expectation and suggested that transportation and COVID-19 Twitter discourse is highly correlated with pandemic outcomes. While we did observe some differences in sentiment between the modes of transportation, the most significant indicator of polarity related to overall trends in sentiment about COVID-19.

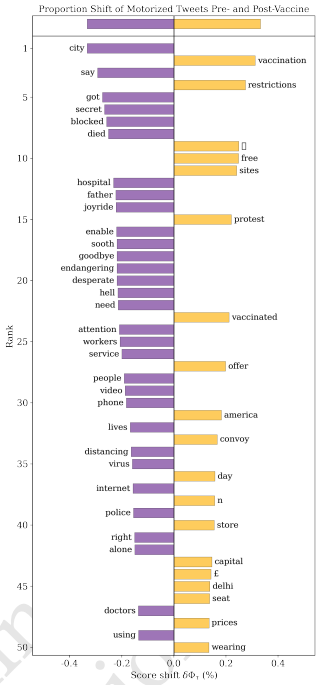
5 DISCUSSIONS

5.1 Data Concerns and Limitations

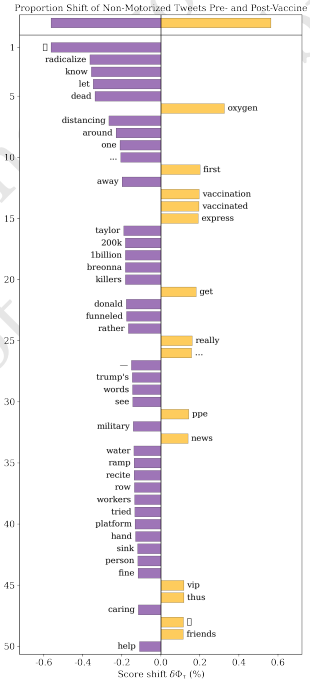
The COVID-19 dataset is an ongoing repository of 2.4 billion tweets. While this is a large dataset, it collects only 1% of the total Twitter



(a) Transit

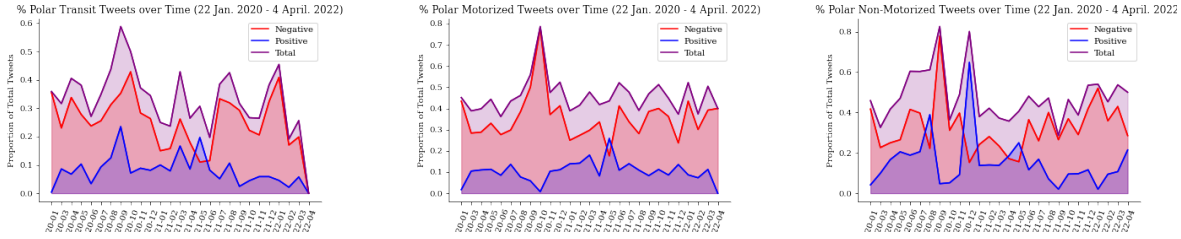


(b) Motorized



(c) Non-Motorized

Figure 4: Shift Graphs by Transportation Mode



(a) Transit **(b) Motorized** **(c) Non-Motorized**

Figure 5: Polarity Over Time by Transportation Mode

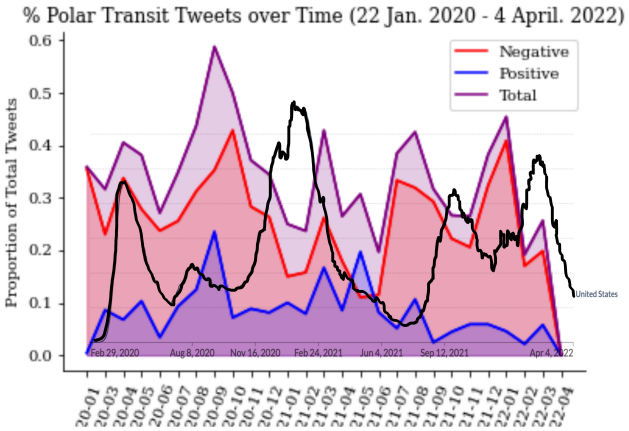


Figure 6: COVID-19 Death Total Count and Polar Transit Tweets Over Time

volume from the Twitter API (Chen, Lerman, and Ferrara 2020). This means that our ability to collect COVID-19+Transportation data is limited by the volume of tweets collected by the authors. On top of that, we noticed that our access to tweets was highly limited by the Twitter API. Throughout the data collection process, we kept log files which counted the fraction of transportation tweets obtained, the total tweets accessed, and the date and time of the log. Upon observing the logs of roughly 1800 files containing tweet IDs collected from the COVID-19 dataset, we found that an average of only 0.19% were accessible. More concretely, out of 100,000 tweet IDs, an average of 190 tweets were accessible. Furthermore, an average of 2.34% of tweets contained a transportation keyword. The exact cause of the tweet inaccessibility is unknown, however we speculate that these tweets may have been since deleted, tweeted by private accounts, or related to Twitter's spam filter [25].

Since the COVID-19 dataset only captures tweets containing COVID-19 words, we are potentially missing relevant data about transportation and mobility during this time period that does not use coronavirus-specific language. Since the COVID-19 dataset is the only avenue with which we can access historical records on COVID-19 Twitter discourse, we must rely on high data accumulation to prevent this possible source of bias. Moreover, the unique transportation keywords used to determine whether or not a tweet belonged in the COVID-19+Transportation dataset were manually selected by a single individual. In the future, we recommend having multiple individuals generate keyword lists and then aggregating the results. This would also help prevent possible bias in the search results obtained.

5.2 Ethical Concerns

All studies which utilize social media posts suffer from the same issue of data usage without informed consent. Our analysis does not specify any identifying information related to individual tweets. We also do not record any usernames or user attributes in our analysis. Following Twitter's Terms of Service, our COVID-19+Transportation dataset only makes the unique tweet identifier publicly available. Only those adhering to the Twitter Developer policy are able to access information related to the tweet and user itself. Regardless, the tweets we retrieved and analyzed contain not only relevant data about the tweet but also identifying information information about the user, such as name, user ID, follower count, and location (on occasion). Future consumers of the COVID-19+Transportation dataset must be concerned with the private information they choose to include in their research.

6 CONCLUSIONS

In this paper, we built a dataset of COVID-19 and Transportation tweets to serve as a contribution to the public and private transportation industries as they grapple with the uncertainty of the COVID-19 pandemic. The publicly available dataset includes Tweet IDs, which are unique identifiers tied to specific tweets, organized by time. Our preliminary analysis suggests that Twitter does hold a great deal of information related to transportation which, if used properly, can inform future policy. Our textual analysis on our small dataset found that themes relating to public transit were most discernible. It was difficult to interpret the data related to motorized and non-motorized modes of transportation, but we expect that the accumulation of more data would help mitigate the ambiguity.

Our sentiment analysis showed that sentiment toward transportation and COVID-19 evolved according to the state of the pandemic. Opinionated messages were tweeted at greater proportions at times when COVID-19 death tolls were high, and furthermore, these opinionated tweets were primarily net negative. We hope that these preliminary results encourage future work which utilizes Twitter to inform public opinion on transportation. As we do not see an end to the pandemic for the foreseeable future, we hope that other researchers with greater computing resources can continue to gather rich historical COVID-19 and transportation data.

REFERENCES

[1] Muhammad Abdullah et al. "Exploring the impacts of COVID-19 on travel behavior and mode preferences". en. In: *Transportation Research Interdisciplinary Perspectives* 8 (Nov. 2020), p. 100255. ISSN: 2590-1982. DOI: 10.1016/j.trip.2020.100255. URL: https://www.sciencedirect.com/science/article/pii/S2590198220301664 (visited on 03/17/2022).

[2] Francesco Barbieri et al. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650. DOI: 10.18653/v1/2020.findings-emnlp.148. URL: https://aclanthology.org/2020.findings-emnlp.148.

[3] Rebecca Brough, Matthew Freedman, and David C. Phillips. "Understanding socioeconomic disparities in travel behavior during the COVID-19 pandemic". en. In: *Journal of Regional Science* 61.4 (2021), pp. 753–774. ISSN: 1467-9787. DOI: 10.1111/jors.12527. URL: http://onlinelibrary.wiley.com/doi/abs/10.1111/jors.12527 (visited on 03/20/2022).

[4] Helmut Brunner et al. "Evaluation of various means of transport for urban areas". In: *Energy, Sustainability and Society* 8.1 (Mar. 2018), p. 9. ISSN: 2192-0567. DOI: 10.1186/s13705-018-0149-0. URL: https://doi.org/10.1186/s13705-018-0149-0 (visited on 03/17/2022).

[5] *CDC Museum Covid-19 Timeline*. Jan. 2022. URL: https://www.cdc.gov/museum/timeline/covid19.html#.

[6] Emily Chen, Kristina Lerman, and Emilio Ferrara. "Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set". In: *JMIR Public Health Surveill* 6.2 (May 2020), e19273. ISSN: 2369-2960. DOI: 10.2196/19273. URL: http://www.ncbi.nlm.nih.gov/pubmed/32427106.

[7] Francine Chen et al. "Social Media-based Profiling of Business Locations". In: *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*. GeoMM '14. New York, NY, USA: Association for Computing Machinery, Nov. 2014, pp. 1–6. ISBN: 978-1-4503-3127-2. DOI: 10.1145/2661118.2661119. URL: https://doi.org/10.1145/2661118.2661119 (visited on 03/16/2022).

[8] Munmun De Choudhury and Scott Counts. "Understanding affect in the workplace via social media". en. In: *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*. San Antonio, Texas, USA: ACM Press, 2013, p. 303. ISBN: 978-1-4503-1331-5. DOI: 10.1145/2441776.2441812. URL: http://dl.acm.org/citation.cfm?doid=2441776.2441812 (visited on 03/20/2022).

[9] Jonas De Vos. "The effect of COVID-19 and subsequent social distancing on travel behavior". en. In: *Transportation Research Interdisciplinary Perspectives* 5 (May 2020), p. 100121. ISSN: 2590-1982. DOI: 10.1016/j.trip.2020.100121. URL: https://www.sciencedirect.com/science/article/pii/S2590198220300324 (visited on 03/20/2022).

[10] Annie Doubleday et al. "How did outdoor biking and walking change during COVID-19?: A case study of three U.S. cities". en. In: *PLoS One* 16.1 (Jan. 2021), e0245514.

[11] Ryan J. Gallagher et al. "Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts". In: *EPJ Data Science* 10.1 (Jan. 2021), p. 4. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-021-00260-3. URL: https://doi.org/10.1140/epjds/s13688-021-00260-3.

[12] Muhammad Ahsanul Habib and Md Asif Hasan Anik. "Impacts of COVID-19 on Transport Modes and Mobility Behavior: Analysis of Public Discourse in Twitter". en. In: *Transportation Research Record* (Aug. 2021). Publisher: SAGE Publications Inc, p. 03611981211029926. ISSN: 0361-1981. DOI: 10.1177/03611981211029926. URL: https://doi.org/10.1177/03611981211029926 (visited on 03/17/2022).

[13] Hirohito Ito, Shinya Hanaoka, and Tomoya Kawasaki. "The cruise industry and the COVID-19 outbreak". en. In: *Transportation Research Interdisciplinary Perspectives* 5 (May 2020), p. 100136. ISSN: 2590-1982. DOI: 10.1016/j.trip.2020.100136. URL: https://www.sciencedirect.com/science/article/pii/S2590198220300476 (visited on 03/20/2022).

[14] Donald G. Janelle and Michel Beuthe. "Globalization and research issues in transportation". en. In: *Journal of Transport Geography* 5.3 (Sept. 1997), pp. 199–206. ISSN: 09666923. DOI: 10.1016/S0966-6923(97)00017-3. URL: https://linkinghub.elsevier.com/retrieve/pii/S0966692397000173 (visited on 03/17/2022).

[15] Elías Jónsson. "An Evaluation of Topic Modelling Techniques for Twitter". In: 2016.

[16] Karl Kim. "Impacts of COVID-19 on transportation: Summary and synthesis of interdisciplinary research". en. In: *Transportation Research Interdisciplinary Perspectives* 9 (Mar. 2021), p. 100305. ISSN: 2590-1982. DOI: 10.1016/j.trip.2021.100305. URL: https://www.sciencedirect.com/science/article/pii/S2590198221000129 (visited on 03/20/2022).

[17] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

[18] Kelsey MacMillan and James D. Wilson. "Topic supervised non-negative matrix factorization". In: *arXiv:1706.05084 [cs, stat]* (July 2017). arXiv: 1706.05084. URL: http://arxiv.org/abs/1706.05084 (visited on 03/20/2022).

[19] Mohammad Noaeen and Behrouz Far. "The Efficacy of Using Social Media Data for Designing Traffic Management Systems". In: Aug. 2020, pp. 11–17. DOI: 10.1109/CrowdRE51214.2020.00009.

[20] Nurhadi and R. T. Suryadari. "Understanding changes in perceptions and behaviour of train passengers during the Covid 19 pandemic". en. In: *IOP Conference Series: Earth and Environmental Science* 824.1 (July 2021). Publisher: IOP Publishing, p. 012107. ISSN: 1755-1315. DOI: 10.1088/1755-1315/824/1/012107. URL: https://doi.org/10.1088/1755-1315/824/1/012107 (visited on 03/20/2022).

[21] Hannah Ritchie et al. "Coronavirus Pandemic (COVID-19)". In: *Our World in Data* (2020). URL: https://ourworldindata.org/coronavirus.

[22] Kerstin J. Schaefer, Leonie Tuitjer, and Meike Levin-Keitel. "Transport disrupted – Substituting public transport by bike or car under Covid 19". en. In: *Transportation Research Part A: Policy and Practice* 153 (Nov. 2021), pp. 202–217. ISSN: 0965-8564. DOI: 10.1016/j.tra.2021.09.002. URL: https://www.sciencedirect.com/science/article/pii/S0965856421002263 (visited on 03/20/2022).

[23] Joseph B. Sobieralski. "COVID-19 and airline employment: Insights from historical uncertainty shocks to the industry". en. In: *Transportation Research Interdisciplinary Perspectives* 5 (May 2020), p. 100123. ISSN: 2590-1982. DOI: 10.1016/j.trip.2020.100123. URL: https://www.sciencedirect.com/science/article/pii/S2590198220300348 (visited on 03/20/2022).

[24] Mellisha Stokes et al. *KFF COVID-19 Vaccine Monitor: February 2022*. en-US. Mar. 2022. URL: https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-february-2022/ (visited on 03/21/2022).

[25] Christopher Torres-Lugo et al. "Manipulating Twitter Through Deletions". In: (2022). DOI: 10.48550/ARXIV.2203.13893. URL: https://arxiv.org/abs/2203.13893.

[26] Paula Vasquez-Henriquez, Eduardo Graells-Garrido, and Diego Caro. "Characterizing Transport Perception using Social Media: Differences in Mode and Gender". In: *Proceedings of the 10th ACM Conference on Web Science*. WebSci '19. New York, NY, USA: Association for Computing Machinery, June 2019, pp. 295–299. ISBN: 978-1-4503-6202-3. DOI: 10.1145/3292522.3326036. URL: https://doi.org/10.1145/3292522.3326036 (visited on 03/16/2022).

[27] Jacob W. Ward, Jeremy J. Michalek, and Constantine Samaras. "Air Pollution, Greenhouse Gas, and Traffic Externality Benefits and Costs of Shifting Private Vehicle Travel to Ridesourcing Services". In: *Environmental Science & Technology* 55.19 (Oct. 2021). Publisher: American Chemical Society, pp. 13174–13185. ISSN: 0013-936X. DOI: 10.1021/acs.est.1c01641. URL: https://doi.org/10.1021/acs.est.1c01641 (visited on 03/21/2022).