6-28-2022

Community-Based Behavioral Understanding of Mobility Trends and Public Attitude through Transportation User and Agency Interactions on Social Media in the Emergence of Covid-19

Md Rakibul Alam
Florida International University, malam071@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

COMMUNITY-BASED BEHAVIORAL UNDERSTANDING OF MOBILITY TRENDS AND PUBLIC ATTITUDE THROUGH TRANSPORTATION USER AND AGENCY INTERACTIONS ON SOCIAL MEDIA IN THE EMERGENCE OF COVID-19

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CIVIL ENGINEERING

by

Md Rakibul Alam

2022

To: Dean John L. Volakis
College of Engineering and Computing

This dissertation, written by Md Rakibul Alam, and entitled Community-Based Behavioral Understanding of Mobility Trends and Public Attitude through Transportation User and Agency Interactions on Social Media in the Emergence of Covid-19, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Mohammed Hadi

_____
Wensong Wu

_____
Nazife Ganapati

_____
Xia Jin, Co-Major Professor

_____
Arif Mohaimin Sadri, Co-Major Professor

Date of Defense: June 28, 2022

The dissertation of Md Rakibul Alam is approved.

_____
Dean John L. Volakis
College of Engineering and Computing

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2022

DEDICATION

To my beloved parents, siblings, daughter, and the better half

ACKNOWLEDGMENTS

First of all, I would like to thank the Almighty for the continuous blessings and the opportunity to complete the research work. I am grateful to my parents for their constant sacrifice, support, and encouragement in pursuing my doctoral degree.

I would like to convey my heartiest gratitude to my honorable supervisor Dr. Arif Mohaimin Sadri, for his excellent supervision and constant support in this dissertation. In every phase of this dissertation, his guidance, resourceful insights, and wisdom directed me on the way to completing my work in time. I would also like to express my sincere gratitude to my committee members Dr. Xia Jin, Dr. Mohammed Hadi, Dr. Wensong Wu, and Dr. N. Emel Ganapati, for their valuable guidance and suggestions for my research.

ABSTRACT OF THE DISSERTATION

COMMUNITY-BASED BEHAVIORAL UNDERSTANDING OF MOBILITY TRENDS AND PUBLIC ATTITUDE THROUGH TRANSPORTATION USER AND AGENCY INTERACTIONS ON SOCIAL MEDIA IN THE EMERGENCE OF COVID-19

Md Rakibul Alam

Florida International University, 2022

Miami, Florida

Professor Arif Mohaimin Sadri, Co-Major Professor

Professor Xia Jin, Co-Major Professor

The increased availability of technology-enabled transportation options and modern communication devices (smartphones, in particular) is transforming travel-related decision-making in the population differently at different places, points in time, modes of transportation, and socio-economic groups. The emergence of COVID-19 made the dynamics of passenger travel behavior more complex, forcing a worldwide, unparalleled change in human travel behavior and introducing a new normal into their existence. This dissertation explores the potential of social media platforms (SMPs) as a viable alternative to traditional approaches (e.g., travel surveys) to understand the complex dynamics of people's mobility patterns in the emergence of COVID-19. In this dissertation, we focus on three objectives. First, a novel approach to developing comparative infographics of emerging transportation trends is introduced by natural language processing and data-driven techniques using large-scale social media data. Second, a methodology has been developed to model community-based travel behavior under different socioeconomic and demographic factors at the community level in the emergence of COVID-19 on Twitter,

inferring users' demographics to overcome sampling bias. Third, the communication patterns of different transportation agencies on Twitter regarding message kinds, communication sufficiency, consistency, and coordination were examined by applying text mining techniques and dynamic network analysis.

The methodologies and findings of the dissertation will allow real-time monitoring of transportation trends by agencies, researchers, and professionals. Potential applications of the work may include: (1) identifying spatial diversity of public mobility needs and concerns through social media platforms; (2) developing new policies that would satisfy the diverse needs at different locations; (3) introducing new plans to support and celebrate equity, diversity, and inclusion in the transportation sector that would improve the efficient flow of goods and services; (4) designing new methods to model community-based travel behavior at different scales (e.g., census block, zip code, etc.) using social media data inferring users' socio-economic and demographic properties; and (5) implementing efficient policies to improve existing communication plans, critical information dissemination efficacy, and coordination of different transportation actors to raise awareness among passengers in general and during unprecedented health crises in the fragmented communication world.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1
INTRODUCTION

1.1 Background

In 2020, different cities and nations imposed various policies and restrictions to prevent the spread of COVID-19, which caused a globally unprecedented change in human travel behavior. With the increased availability of technology-enabled transportation options, and modern communication devices (smartphones, in particular), this variation in mobility behavior caused by this pandemic has diversified across different places, points of time, modes of transportation, and socio-demographic & economic groups. Figure 1-1 shows the variation in resultant changes in mobility over time for two different places during the pandemic. The study of these human mobility behavior dynamics is considered transportation trend analysis [1].

Several studies reflect this complex and dynamic environment of transportation trends by exploring the impact of sociodemographic & economic factors on travel demand [2–7], the influence of the built environment on the mobility pattern [8–11], and attitude toward emerging mobility options (shared mobility, autonomous and connected vehicle, electric vehicle) across different social groups [12–17]. Moreover, this pandemic made the dynamics of passenger travel behavior more complex. Few studies have been published to understand this pandemic's impact on passenger mobility trends [18–21]. It was found that the uneven spread of COVID-19 among various demographic groups causes heterogeneity in mobility patterns [22–26]. Several studies used cell phone data, which is expansive, to understand and visualize the effectiveness of control measures [27–29]. The primary data

sources used in these studies are surveys (e.g., travel surveys) that feature representative populations and detailed information about travel mode and trip purposes. Surveys data has some limitations, such as variabilities across countries in data collection method and availability, lack of real-time engagement of the respondents, expansive and time-consuming as trend analysis requires periodic data collection. As a result, it is still unclear how this pandemic, demographic trends, behavior shifts, and technology advancements may work together and influence passenger travel patterns.

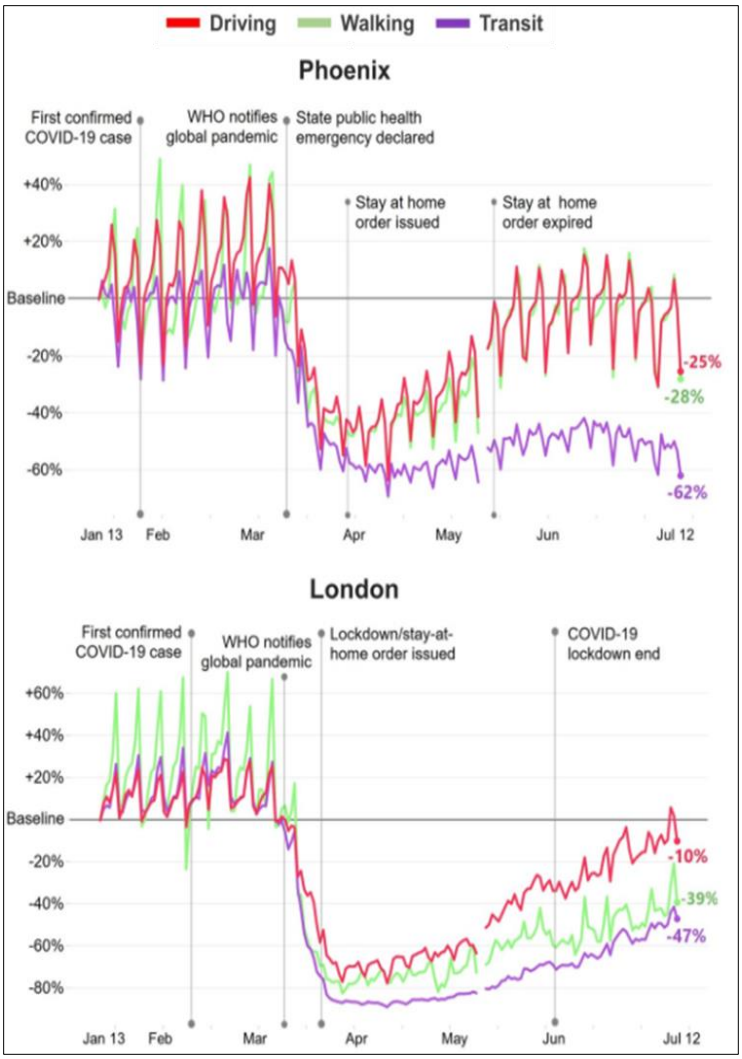

Figure 1-1: COVID-19 Policy Events and the Resultant Changes in Mobility over Time (Jan 13-Jul 12, 2020) [30].

Due to the different forms of travel restrictions during COVID-19, with nothing to do, people have become more active on Social Media Platforms (SMPs) (e.g., Twitter) [31]. Various transportation agencies also increased their activities on SMPs to increase public awareness. These interactions among these agents form a transportation system that varies over space and time at different geographic scales (Fig. 2). This communication network has massive potential in behavioral research. It can be used in understanding the dynamics of transportation trends. Moreover, SMPs are increasingly leveraged to overcome the drawbacks of surveys and other sources of travel data as it serves the need for a more unified, less privacy-invading, and simply accessible method to understand the dynamics of travel patterns fully. Also, social media data incorporates the spatiotemporal feature and real-time engagement of the respondents.

An in-depth understanding of the changing transportation and mobility trends in the emergence of COVID-19 is needed to design the nation's transportation infrastructure better and make policies to meet people's mobility needs in the future during such crisis events. This necessity lays the foundation of the research motivation of this study. The novelty of this study is in demonstrating the capability of large-scale social media data by exploring the communication pattern of the transportation social network system and capturing the dynamics of transportation trends in the emergence of COVID-19. This study will not only contribute methodologically but also is expected to produce a higher quality of results as it will deal with the people's spontaneous real-time engagement. This study presents a comprehensive approach to exploring how SMPs (Twitter) can be used to understand public and agency (transportation actors) perceptions and attitudes towards transportation and mobility trends. This study also investigates how these different

transportation actors interact with each other during this crisis moment of COVID-19 using text mining and network science principles. Finally, this study attempts to understand community concerns towards transportation trends in the emergence of COVID-19 on Twitter through tweet classification, inferring Twitter users' demographics to overcome sampling bias and econometric analysis.

1.2 Research Objective

This proposed study's central research objective is to understand the dynamics of the transportation trends and indicators in the spatio-temporal platform at the emergence of COVID-19. The dissertation focuses on the following specific objectives:

I. Develop a model that can capture emerging transportation trends based on social media interactions with enriched space and time information using sentiment and topic analysis.

II. Develop a methodology to model community-based travel behavior and access public attitudes towards different mobility trends under different socio-economic and demographic factors in the emergence of COVID-19 on Twitter.

III. Develop a model that can detect the long-term communication pattern among transportation actors, as well as their interaction on social media platforms in the emergence of COVID-19 in terms of communication consistency and coordination on Twitter at various stages of the pandemic.

1.3 Conceptual Framework

A conceptual framework is developed to depict the proposed methodology to understand the dynamics of the transportation trends and indicators in the emergence of COVID-19.

Figure 1-2 presents the essence of the communication pattern within and among different communities and how agencies can influence the communities. At the top, different agency reacts and disseminate information differently to address an event. They interact and flow information regarding an adopted policy or event within themselves and different communities. On the other hand, different communities of people react and interact differently regarding this policy or event within themselves and with the agencies. The different community has different community characteristics (V_i), such as travel behaviors (e.g., mode choices) which vary (ΔV_i) among communities based on their socio-economic and demographic properties. Community characteristics also depend on geographic location and scale. So, it is evident that three kinds of interactions are present in the SMPs.

1) Interaction within different communities of people
2) Interaction within different agencies
3) Interaction among different agencies with different communities of people

In this dissertation, the first two social media interactions have been extensively studied to understand the dynamics of the transportation trends and indicators in the emergence of COVID-19. Research questions 1 and 2 of this study focus on the first type of interaction, and research question 3 is based on the second type of interaction.

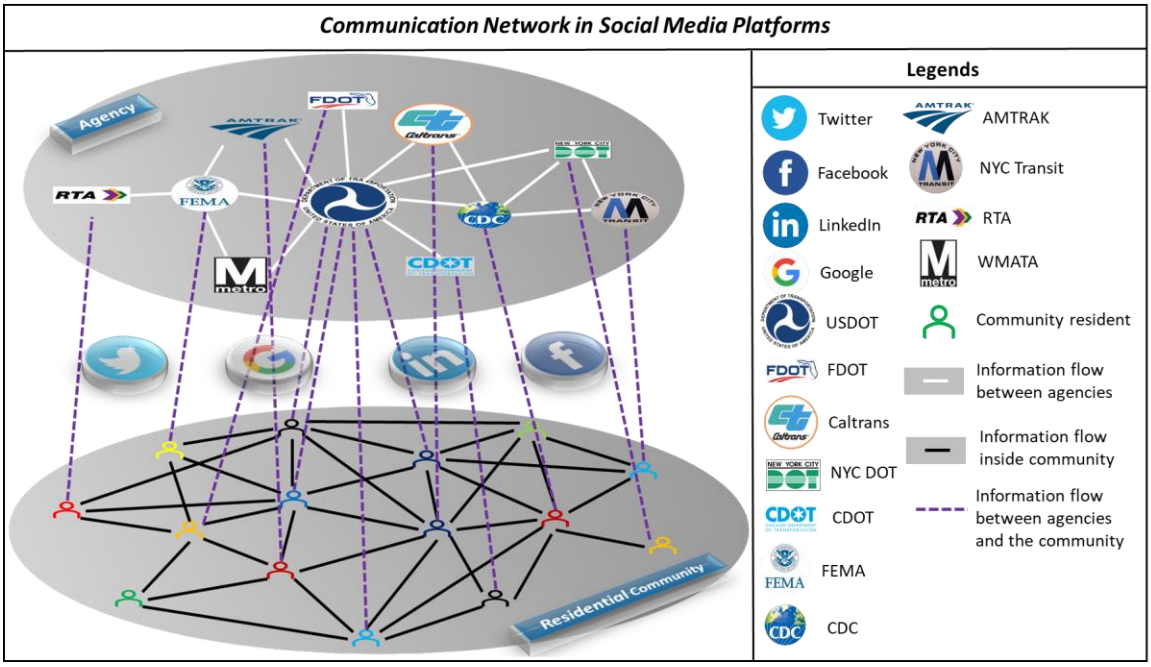

Figure 1-2: Framework of the Communication Pattern within and among the Communities with Agencies.

1.4 Contributions of Dissertation

The significant contributions of this dissertation to existing literature and civil engineering are listed below:

- Developed a structured method to mine and analyze large-scale public interactions from SPMs enriched with time and location information.
- Developed comparative infographics showing the spatial-temporal variation of transportation trends and mobility indicators using natural language processing and data-driven techniques.
- Introduced an approach to increase topical relevancy of social media signals by unsupervised (sentiment analysis, topic modeling) and semi-supervised test classification techniques.

- Presented a novel method to reduce the sampling biases of social media signals by inferring users' socioeconomic and demographic attributes through machine learning [32] and national database mapping (e.g., American Community Survey [33])
- We have developed a novel approach to performing econometric analysis using social media data to understand the public attitudes on different mobility indicators (e-commerce, ridesharing, telework, transit) under different socio-economic and demographic factors.
- Investigated, for the first time, the long-term communication pattern among transportation actors, as well as their interaction on SMPs in the emergence of COVID-19 in terms of communication consistency and coordination on Twitter at various stages of the pandemic.

This dissertation makes significant methodological contributions by introducing different novel approaches using large-scale social media signals reducing sampling biases to capture and model public attitude towards different mobility indicators. Moreover, this dissertation makes major theoretical contributions by investing the communication pattern among transportation actors and their interaction on SMPs for the first time. This dissertation also recommends that agencies increase social media activity and interaction on SMPs to provide consistent public information.

1.5 Structure of Dissertation

The dissertation consists of a total of five chapters. Chapter 1 includes the introduction, research objectives, the conceptual framework, and significant contributions of the research work. Chapter 2 adopts a novel approach to understanding public opinion and

identifying emerging transportation trends based on social media interactions with enriched space and time information using sentiment and topic analysis. Chapter 3 proposes a methodology to understand community concerns towards different mobility indicators in the emergence of COVID-19 on Twitter through tweet classification, inferring Twitter users' demographics to overcome sampling bias, and econometric analysis. Chapter 4 investigates, for the first time, the long-term communication pattern among transportation actors, as well as their interaction on social media platforms in the emergence of COVID-19 in terms of communication consistency and coordination on Twitter at various stages of the pandemic. Chapter 5 summarizes the findings and potential applications of research findings, provides recommendations for future studies, and lists the limitations of the research.

CHAPTER 2
IDENTIFYING PUBLIC PERCEPTIONS TOWARD EMERGING TRANSPORTATION TRENDS ON TWITTER

2.1 Introduction and Motivation

With the rapid expansion of modern technologies, the wide availability of spatial data and smartphone apps, and emerging transportation options, the landscape of transportation demand and supply are changing. The increased availability of technology-enabled transportation options (e.g., ridesharing) and modern communication devices (smartphones, in particular) are transforming travel-related decision-making in the population differently at a different level. These complex dynamics of emerging mobility behaviors are also expected to be influenced by individual lifestyles and different social (e.g., education), economic (e.g., employment), and demographic (e.g., gender) factors. National Cooperative Highway Research Program (NCHRP) explores the effect of these factors on travel demand [2–4]. Several studies reflect the dynamics of mobility patterns influenced by the built environment [8–11]. Moreover, numerous studies have focused on exploring the impact of different socioeconomic factors, such as age, gender, and income levels, on travel behavior [5–7, 34–36].

The primary data sources used in the abovementioned studies are surveys (e.g., travel surveys) that feature representative populations and detailed information about travel mode and trip purposes. Survey data has some limitations, such as variabilities across countries in data collection method and data availability, lack of real-time engagement of the respondents, expansive and time-consuming as trend analysis requires periodic data collection making. Nowadays, people have become more active on Social Media Platforms

(SMPs) (e.g., Twitter) [31]. SMPs are increasingly leveraged to overcome the drawbacks of surveys and other sources of travel data as it serves the need for a more unified, less privacy-invading, and simply accessible method to understand the dynamics of travel patterns fully. Moreover, Social media data incorporates the spatiotemporal feature and real-time engagement of the respondents.

An in-depth understanding of the changing emerging transportation and mobility trends is needed to design the nation's transportation infrastructure better to meet people's mobility needs over the following decades. This necessity lays the foundation of the research motivation of this pilot study. The novelty of this pilot study is in demonstrating the capability of large-scale social media data using natural language processing techniques to capture emerging transportation trends. This study will not only contribute methodologically but also is expected to produce a higher quality of results as it will deal with the people's spontaneous real-time engagement.

In this study, for the first time, Twitter data has been used to track emerging transportation trends over a large geological scale with a large volume of data (~13M tweets) which would not be achievable with survey techniques. We explored emerging travel trends in North America using data obtained from Twitter for around 20 days from Dec 16th, 2019, to Jan 4th, 2020. The main purpose of this study is to understand public opinion and identify emerging transportation trends based on social media interactions with enriched space and time information. This study aims to achieve the following objectives:

- Identify spatio-temporal characteristics of relevant social media interactions on shared mobility, vehicle technology, built environment, user fees, e-commerce, and

telecommuting, which can give an understanding of the spatial and temporal distribution of the relevant tweets describing the emerging transportation trends;

- Measure public sentiments and perceptions on emerging transportation trends through natural language processing such as sentiment analysis, which can allow the classification of tweets based on sentiment scores (highly positive, positive, neutral and negative, highly negative);
- Explore spatio-temporal differences in user sentiments by classifying sentiment scores on transportation and mobility indicators which can make sense of the spatial and temporal distribution of tweets concerning their sentiment direction;
- Extract emerging transportation topics and user concerns from social media interactions through Latent Dirichlet Allocation (LDA) which is a machine learning approach to identify the patterns of the filtered relevant tweets to recognize the emerging transportation trends

2.2 Background and Related Work

Though SMPs are relatively new fields for research, researchers have used them in various cases such as travel demand forecasting, mobility pattern identification, disaster management, mass transit evaluation, and traffic incident management.

There are several studies where SMPs have been used to forecast travel demand. Golder & Macy [37] and Yin et al. [38] investigated the capacity, scope, and application of various SMPs to derive information on household daily travel. Tasse & Hong discussed the opportunities of using geotagged social media data instead of traditional survey data to

understand people's mobility patterns, the average distance traveled, and the overall spatial distribution of urban areas[39].

SMPs have been applied to understand mass human mobility patterns too. These studies have established Location-based Social Networking (LBSN) data as a strong proxy not only for tracking and predicting human movement, identifying mobility patterns, and recognizing various geographic and economic factors that affect human mobility patterns at aggregate levels across different geographical scales [40–42] but also to model user activity patterns too. Hasan & Ukkusuri presented a novel approach to understanding urban human activity and mobility patterns using large-scale location-based data characterizing temporal and spatial aspects of the mobility and activity patterns[43, 44].

Opinion mining has been performed in a few studies to show people's attitudes towards public transit, which can affect how stakeholders think about future transit investments [45]. Pender et al. [46] applied crowdsourcing techniques to derive transit service information that can satisfy the increased demand and expectation for real-time information dissemination. Luong & Houston [47] also used social media data to study public attitudes about light rail transit services in Los Angles.

Recent studies have also extracted traffic data from social media for transportation network operation and management purposes. Tian et al. [48] validated traffic incidents posted on social media by checking camera footage data and found that tweets about severe incidents tended to be more accurate. Steur [49] showed the correlation between accidents and the frequency of tweets near the incident locations. Several studies also show the potentiality of SMPs in disaster management. Wang & Taylor [50, 51] studied the perturbation and

resilience of human mobility patterns during and after tropical storms and confirmed the correlation of daily human trajectories between steady-state and perturbation states and the high inherent resiliency of human mobility using Twitter data. Researchers also have focused on detecting influential social media users and explored their network features to understand the spread of targeted information in major disasters [52, 53].

In summary, SMPs have been utilized to retrieve relevant information for demand prediction, pattern recognition, transit evaluation, incident management, and disaster management. No study has used SMPs to infer public opinions and perceptions toward emerging transportation trends. As such, this pilot study presents a comprehensive approach to exploring how SMPs (Twitter) can be used to understand public perceptions and attitudes towards emerging transportation and mobility trends using natural language processing and data-driven methods.

2.3 Methodology

2.3.1 Data Collection and Preparation

The research team created a Twitter Developer Account using Twitter Apps (apps.twitter.com/) to retrieve data through Twitter Streaming API (Application Programming Interfaces). Python programming language was used to collect the data, and associated Python libraries have been used. The main focus of this study is English geotagged tweets as tweet geographic information is a potential parameter for spatio-temporal analysis. The location-based data collection method produced a more suitable and reliable dataset that serves the study's goal. As a result, tweets from North America and its surrounding area (as most of the people in this region speak English) are collected using a

location-bounding box for around 20 days (Dec 16th, 2019- Jan 4th, 2020) which covered the USA, Canada, Mexico, Cuba, Puerto Rico, and part of Guatemala and Greenland (Figure 2-1).

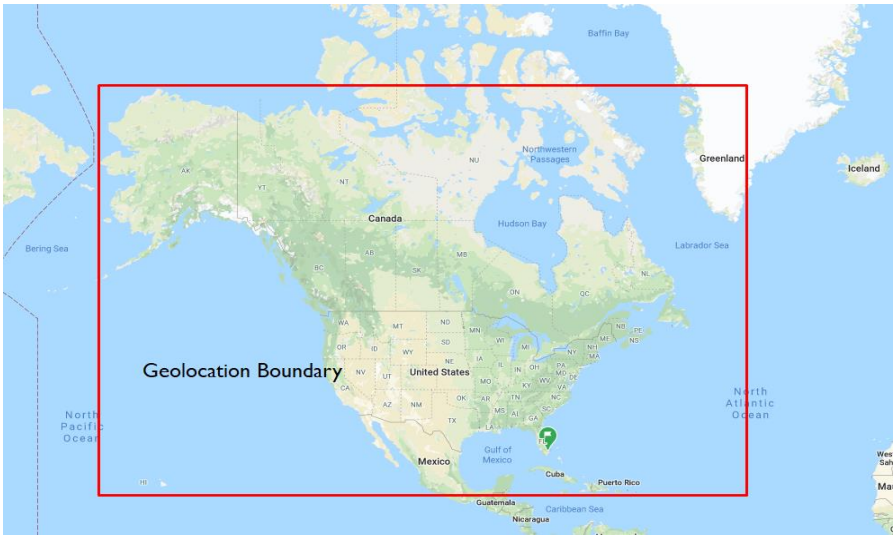

Figure 2-1: Bounded Box Used for the Data Collection for North America.

The raw data contains approximately 12.9M tweets. Approximately 100% of tweets are geotagged and mainly in English (~ 77%), with around 0.97 M unique users. Tweets retrieved from the streaming API contain additional information such as user id, profile information, and creation time along with the tweet text. Only tweet texts were considered for analysis in this study. Given the inherent ambiguity of tweets (e.g., non-standard spelling, inconsistent punctuation, and/or capitalization), additional preprocessing steps are performed to extract clean tweet text, which is suitable for analysis. Noises (such as html tags, character codes, emojis, stop words, etc.) were removed from the text data, and tweets were tokenized, which is the process of breaking down an expression, sentence, paragraph, or even an entire text document into smaller units like individual words or phrases. Tokens are the names given to each of these smaller units.

For this pilot study, we focus on relevant tweets related to six major categories of emerging mobility trends:

1. **Shared Mobility:** shared, mobility, carpool, car, uber, lyft
2. **Vehicle Technology:** autonomous, automated, self-driving, connect, connected
3. **Built Environment:** walk, gym, cycle, activity, sidewalks, bypass, access, bus, station
4. **User Fees:** toll, express, lane, mileage, price, gas, gallon, fee, fare, tax, booth
5. **Telecommuting:** telecommute, job, flexible, hours, dollar, commute, telework, mobile, remote
6. **Ecommerce:** ecommerce, amazon, deliver, delivery, walmart, publix, ebay, fedex, ups

Relevancy was established if the tweet contained at least one of the keywords identified for this study. In total, 205 keywords in six major categories were identified as relevant to emerging trends. Although this approach may filter out some relevant tweets, it ensures that all tweets involving these keywords were included in the filtered dataset for further analysis. After filtering the dataset, a total number of 1.25 M (9.68% of the total tweets) relevant English tweets were obtained for this study. Table 2-1 presents the keywords used to filter relevant tweets for each category. The percentage value represents the percentage of tweets containing specific keywords concerning the whole dataset.

Table 2-1. Complete List of Keywords Used for Keyword-Based Data Collection.

Category	Relevant keywords	Tweet Count
Shared Mobility (44 words)	shared, mobility, carpool, car, uber, lyft, tnc, share, zipcar, waze, juno, driver, passenger, ride, maas, e-hail, ehail, carclubs, bicycle, via, uberpool, hail, scooter, flexdrive, vehicle, zebra, flexwheels, e-scooter, escooter, lime, wheels, spin, bird, mobi, bike, evo, gogo, jax, rental, curb, wingz, birdj, traffic, fdot	170,289 (1.31%)
Vehicle Technology (26 words)	autonomous, automated, self-driving, connect, connected, v2v, v2i, v2x, tesla, electric, hybrid, google, drive, platoon, airbags, energy, phonefob, vpa, telematics, ai, b2v, eascy, automation, artificial, intelligence, map	74,144 (.60%)
Built Environment (49 words)	built environment, walk, gym, cycle, activity, sidewalks, bypass, access, bus, station, stop, transit, mile, metro, rail, mover, land, work, office, shop, school, bank, airport, flight, plane, restaurant, park, malls, theater, bar, pick-up, pickup, drop-off, dropoff, atm, fitbit, train, subway, universal, disney, hyperloop, everglades, tour, tourist, arrive, depart, destination, eta, home	631,697 (4.87%)
User Fees (20 words)	toll, express, lane, mileage, price, gas, gallon, fee, fare, tax, booth, market, charge, payment, tariff, dues, levy, duty, liter, litre	66,668 (.51%)
Telecommuting (32 words)	telecommute, job, flexible, hours, dollar, video-conference, videoconference, commute, telework, mobile, remote, workplace, technology, home-sourced, home sourced, e-work, ework, outwork, operation, mode, labor, regime, freelance, screen, voice, chat, video, phone, yammer, zoom, virtual, employee	344,868 (2.66%)
Ecommerce (34 words)	ecommerce, amazon, deliver, delivery, walmart, publix, ebay, fedex, ups, browse, purchase, e-business, ebusiness, online, trade, internet, sale, retail, transaction, paperless, macy's, macys, wish, lowe's, lowes, best buy, bestbuy, target, home depot, homedepot, etsy, rakuten, groupon, ebates	142,101 (1.10%)

2.3.2 Spatial and Temporal Analysis

Twitter allows users to share their location from where the user posted the tweet, which is a confined area generated automatically with the tweet if the location of the user's device

remains enabled. Geolocational information and timestamp of tweets were extracted from the ‘place’ and ‘created_at’ fields, respectively. Temporal or time series analysis is one of the best techniques to understand the internal patterns (trends, temporal variation) within data over time. Heatmaps were produced to represent the correlation between the most frequently used words in relevant tweets and the dates when they were tweeted. This illustrates the daily variation of popular words that have been tweeted, which provides insight into the temporal variation of the most popular and unpopular trends over time. Another type of heatmaps, plotting the inter-relationship between the most frequently used words and tweet location, was also created. It is a very efficient way to understand the spatial variation of the popularity of transportation trends. For this reason, geotagged tweets were considered a source to improve situational awareness and understanding real-world transportation trends.

2.3.3 Sentiment Ratings

Sentiment analysis or opinion mining is the computational study of opinions, sentiments, and emotions. It tries to infer people’s sentiments based on their language expressions expressed in a text. It usually uses a sentiment lexicon to provide sentiment scores on the generated corpus (a textual body clustered by required class or cluster). The analysis focuses on individual sentence targets to determine whether a sentence expresses an opinion or not (often called subjectivity classification), and if so, whether the opinion is positive or negative (called sentence-level sentiment classification). Assume an opinionated document (tweet) be *w*, which expresses on a subject or a group of subjects. Generally, $w = (w_1, w_2, \ldots w_i, \ldots, w_n)$ where w_i is a word or sentence. An opinion passage

on a feature *f* of an object *o* evaluated in w is a group of consecutive sentences in *w* that expresses a positive or negative opinion on *f*.

Additionally, sentiments also contain subjectivity. A subjective sentence expresses some personal feelings or beliefs. Sentence-level sentiment classification involves two definite tasks with a single assumption [54]. These are stated below:

- Task: Given a sentence s, two subtasks are performed:
    1. Subjectivity Classification: Determine whether *w* is a subjective sentence or an objective sentence,
    2. Sentence-level sentiment classification: If *w* is subjective, determine whether it expresses a positive or negative opinion.
- Assumption: The tweet *w* expresses a single opinion from a single opinion holder

In this study, we used a Python package called VADER[55], which detects the sentiment value of a short text for analyzing the sentiments of relevant tweets about emerging transportation trends. Using a pre-defined list of words, VADER assigns a final compound score to each of the input words, which is the sum of all the lexicon ratings which have been standardized to range between -1 and 1 [56].

To decide on a range to categorize highly negative, negative, neutral, positive tweets, and highly positive, a heatmap of the sentiment scores was produced and used to gauge roughly where scores were landing -1 to -0.6 (highly negative), -0.6 to -0.2 (negative), -0.2 to 0.2 (neutral), 0.2 to 0.6 (positive), and 0.6 to 1.0 (highly positive) were ultimately set as the bounds for the three categories. VADER considers currently frequently used slang and informal writings - multiple punctuation marks, acronyms, and an emoticon to express how

a person is feeling, which makes VADER great for social media text. Some real tweets were presented here as examples to demonstrate the categories:

(1) “thank you for creating vision for sustainability and leading the way not only electric cars but also solar autonomous software energy storage among other accomplishments im looking forward seeing what you and your team create”-Highly Positive (Score 0.7992);

(2) “loves tesla though it’s the worst drive during holiday who knew”-Positive (Score 0.3182);

(3) “bosch finally making lidar sensors for autonomous cars” – Neutral (Score 0);

(4) “They’d stop fighting long enough maybe we’d all have autonomous self-driving cars the road now” – Negative (Score -0.296);

(5) “autonomous cars are highly susceptible risk being commandeered visual spoofing attacks” – Highly Negative (Score -0.6461).

2.3.4 Topic Mining

Topic modeling is a machine learning technique that analyzes text data automatically to classify cluster terms for a series of documents. The topic mining technique is applied in this study to identify the patterns of the filtered tweets to recognize the emerging transportation trends. Latent Dirichlet Allocation (LDA) or topic modeling approach[54] was applied is applied in this study. LDA used a probabilistic latent semantic analysis model to recognize the patterns of the posted tweets. Though the topic model has been used popularly in machine learning, recently, it has been applied in transportation studies [44, 57, 58].

The probabilistic procedure for the document (tweet) generating is adopted in LDA which starts with choosing a distribution ψ_k over words in the vocabulary for each topic *k* ($k \epsilon$ *1, K*) . Here, ψ_k is selected from a Dirichlet distribution $Dirichlet_v(\beta)$. After that, another distribution θ_d over *K* topics is sampled from a different Dirichlet distribution $Dirichlet_k(\alpha)$ to generate a document *d* (a set of word w_d). Thus, a topic is assigned for each word in $\boldsymbol{w_d}$ followed by selecting each word w_{di} based on θ_d .

For LDA, initial sampling is done on a particular topic z_{di} ϵ *1, K* from a multinomial distribution $Multinomial_k(\theta_d)$ in generating each word w_{di} . Finally, the word w_{di} is selected from the multinomial distribution $Multinomial_v(\psi_{z_{di}})$**.** Figure 2-2 shows the graphical representation of LDA where Sun & Yin [59] . The inference of LDA models can be done by applying the variational expectation-maximization (VEM) algorithm [60] or through Gibbs sampling [61]. The posterior of document-topic distribution θ_d and topic-word distribution ψ can be efficiently inferred by both methods which allow us to discover the latent thematic structure from a large collection of documents[59].

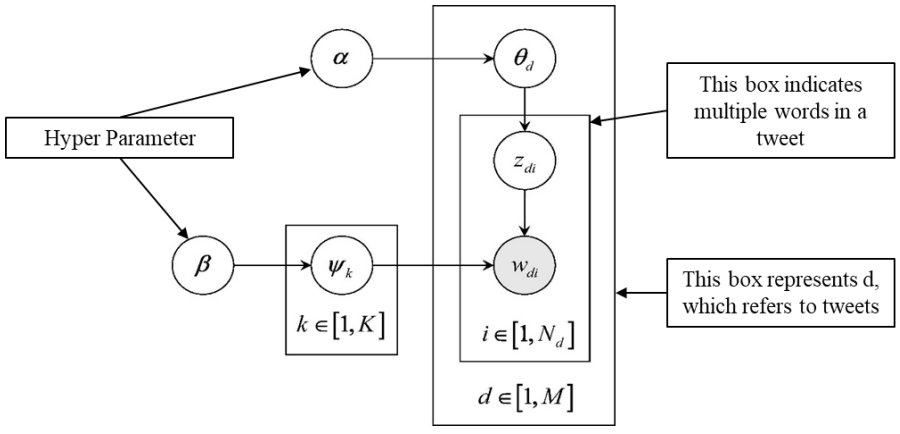

Figure 2-2: Graphical Model Representation of LDA [59].

The key steps involved in the Tweet data analysis are summarized in Figure 2-3.

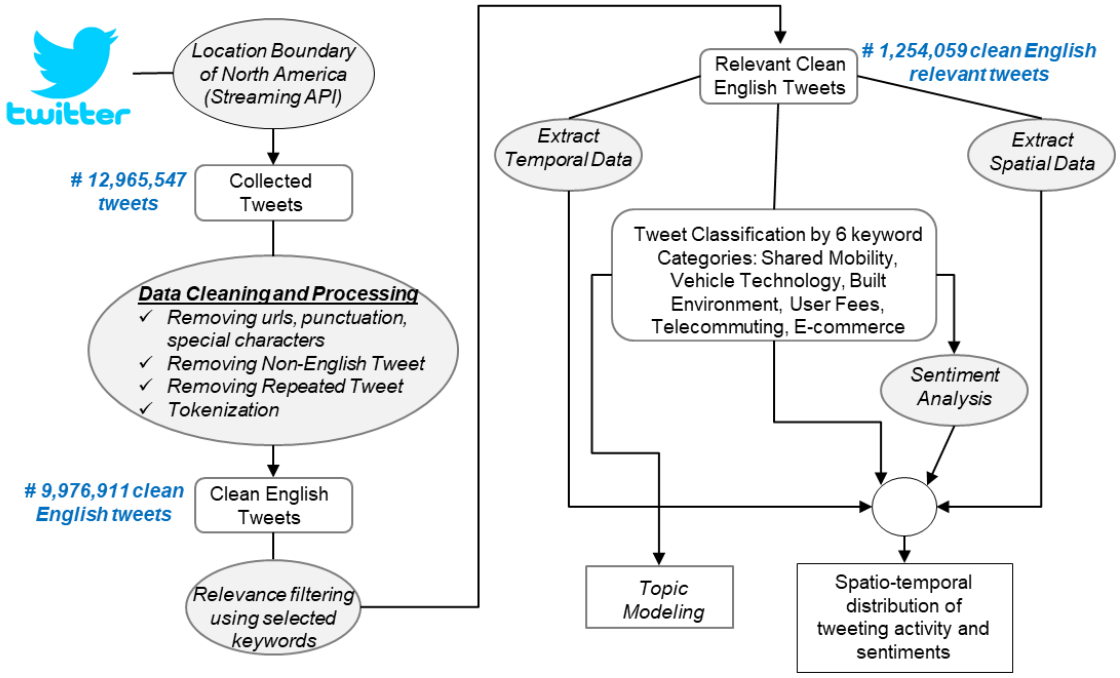

Figure 2-3: Framework for Data Collection, Preparation, and Analysis.

2.4 Results

After data processing and cleaning, a total number of 1.25 M relevant English tweets were obtained for further analysis. Figure 2-4 presents the main components and characteristics of the dataset. There are mainly three kinds of location information that can be extracted from a tweet:

- Profile Location is the location of residence of the person (set by the account holder) who posted the tweet.
- Tweet originating city (Tweet_location) is the boundary of the location from where the user posted the tweet. This feature is generated automatically with the tweet if the location of the user's device remains enabled. The tweet containing this information is called geotagged tweet, which is the focus of this study.

- The exact tweet location is the exact point location from where the user posted the tweet. This feature is generated automatically with the tweet when the user does check-in or tag that place while posting the tweet.

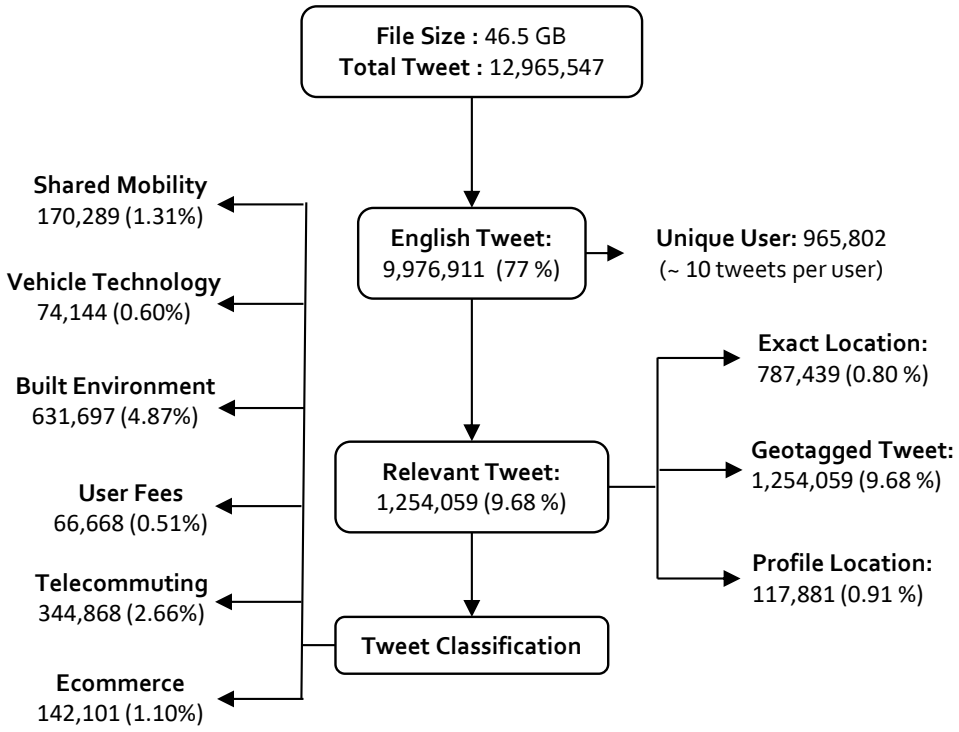

Figure 2-4: Description of the Dataset.

2.4.1 Spatio-Temporal Heatmaps of Tweets

Spatio-temporal distribution of tweeting activities can broaden the understanding of the credibility and representativeness of the datasets being used for the analyses over space and time. Due to the limitation of measuring the statistically significant difference mathematically over different categories across different places, visual inspection was adopted, and almost identical spatio-temporal distribution patterns were observed across all categories, i.e., shared mobility, vehicle technology, built environment, user fees,

telecommuting, and e-commerce (Figure 2-5). Figure 2-5 is a two-dimensional representation of tweeting activities based on tweet originating dates and the most frequent 50 locations (state-level).

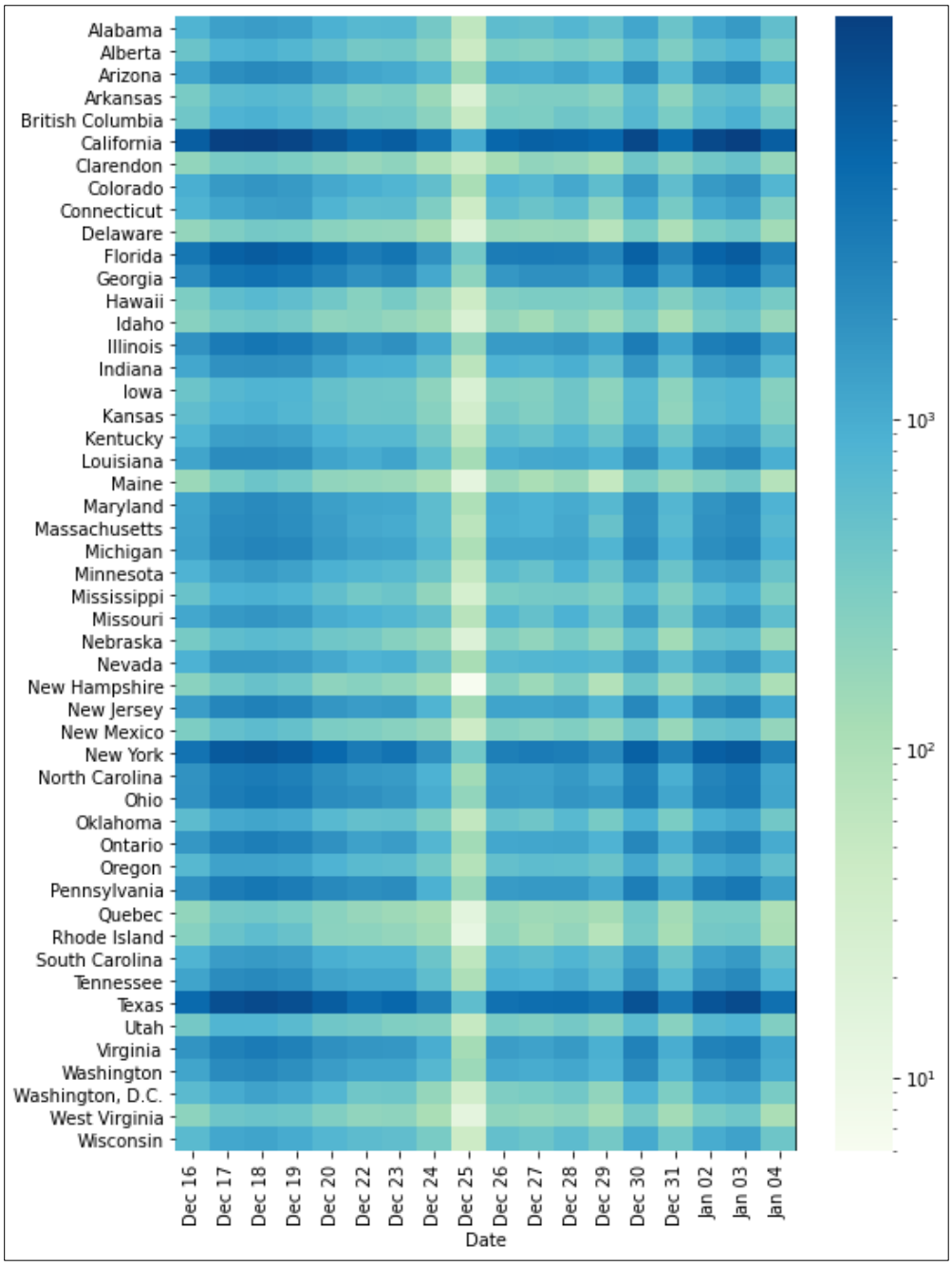

Figure 2-5: Spatio-temporal Distribution of Relevant Tweets (Top 50 location).

Places such as California, Florida, Georgia, Illinois, New York, North Carolina, Ohio, Pennsylvania, Texas, Virginia, and Washington were among the most frequent locations and generated ~10K tweets daily on emerging transportation trends. People from these locations were likely to be more expressive of emerging mobility trends through social media interactions as evident from Twitter. In contrast, places like Alberta, Clarendon, Delaware, Hawaii, Maine, Nebraska, New Hampshire, New Mexico, Quebec, Rhode Island, Utah, and West Virginia generated as low as only ~1.5K tweets per day on emerging trends. Other locations in Figure 2-5 represent moderate levels of concern among social media users (~3K-10K tweets on average). Locations that do not appear in Figure 2-5 were inactive, with less than 100 tweets a day. These findings indicate spatial diversity of the transportation-related needs and concerns people express through social media channels and the need to utilize such information to develop new policies meeting the diverse needs people may have in different locations. Moreover, the temporal patterns for almost all locations indicate people were less expressive of such concerns during and immediate before/after a government holiday such as Christmas and New Year.

2.4.2 Temporal Heatmaps of Tweet Keywords

To delve deeper into the understanding of social media interactions in different categories, i.e., shared mobility, vehicle technology, built environment, user fees, telecommuting, and e-commerce, temporal heatmaps of tweet keywords were generated (Figure 2-6 a-f).

The word frequencies in the heatmaps indicate that people tweeted more about user fees and e-commerce, followed by vehicle technology, telecommuting, built environment, and shared mobility. This indicates the potential to utilize such information to rank people's

social media interactions and leverage social sharing platforms to promote user interests in emerging trends based on similar word clustering. A closer look at the word heatmaps by categories shows the following findings:

Shared Mobility:

- ‘via’ is highly prominent. It is a commonly used word, also an emerging ridesharing platform
- ‘car’, ‘share’, ‘ride’, ‘driver’ also showed strong presence, followed by ‘traffic’, ‘uber’, ‘vehicle’, ‘bird’, ‘shared’, and ‘bike’
- ‘Uber’ was more popular than ‘Lyft’
- Emerging platforms such as ‘Waze’, ‘Zipcar’, ‘escooter’, ‘uberpool’ were found less frequent on Twitter
- ‘bike’ and ‘bicycle’ showed less prominence compared to ‘car’. This is indicative of the need to leverage social media for bike-sharing

Vehicle Technology:

- ‘energy’ was highly prominent. This is a commonly used word, also a fuel-efficient transportation platform
- ‘drive’, ‘google’, ‘intelligence’, ‘connect’ also showed strong presence, followed by ‘tesla’, ‘electric’, ‘map’, ‘connected’, and ‘hybrid.’
- ‘electric’ was more popular than ‘hybrid’
- emerging platforms such as ‘automation’, ‘artificial’, ‘automated’, ‘autonomous’ were found less frequent on Twitter.

- ‘hybrid’ and ‘autonomous’ showed less prominence relative to ‘energy’. This is indicative of the need to leverage social media for hybrid and autonomous transport

Built Environment:

- ‘work’ was highly prominent.
- ‘home’, ‘stop’, ‘school’, ‘office’ also showed strong presence, followed by park’, ‘walk’, ‘bar’, ‘gym’, ‘station’, and ‘shop’
- ‘park’ was more popular than ‘gym’
- emerging platforms such as ‘subway’, ‘transit’, ‘bus’, ‘sidewalks’ were found less frequent on Twitter
- ‘pickup’ and ‘dropoff’ showed less prominence. This is indicative of the need to leverage social media for online shopping

User Fees:

- ‘tax’ was highly prominent.
- ‘market’, ‘gas’, ‘price’, ‘charge’ also showed strong presence, followed by ‘lane’, ‘express’, ‘duty’, and ‘booth’
- Financial activities such as ‘dues’, ‘levy’, ‘liter’, ‘tariff’ were found less frequent on Twitter
- ‘toll’ and ‘tariff’ showed less prominence relative to ‘tax’. This is indicative of the need to leverage social media for the charge on using bridge or road and the duty on imports and exports

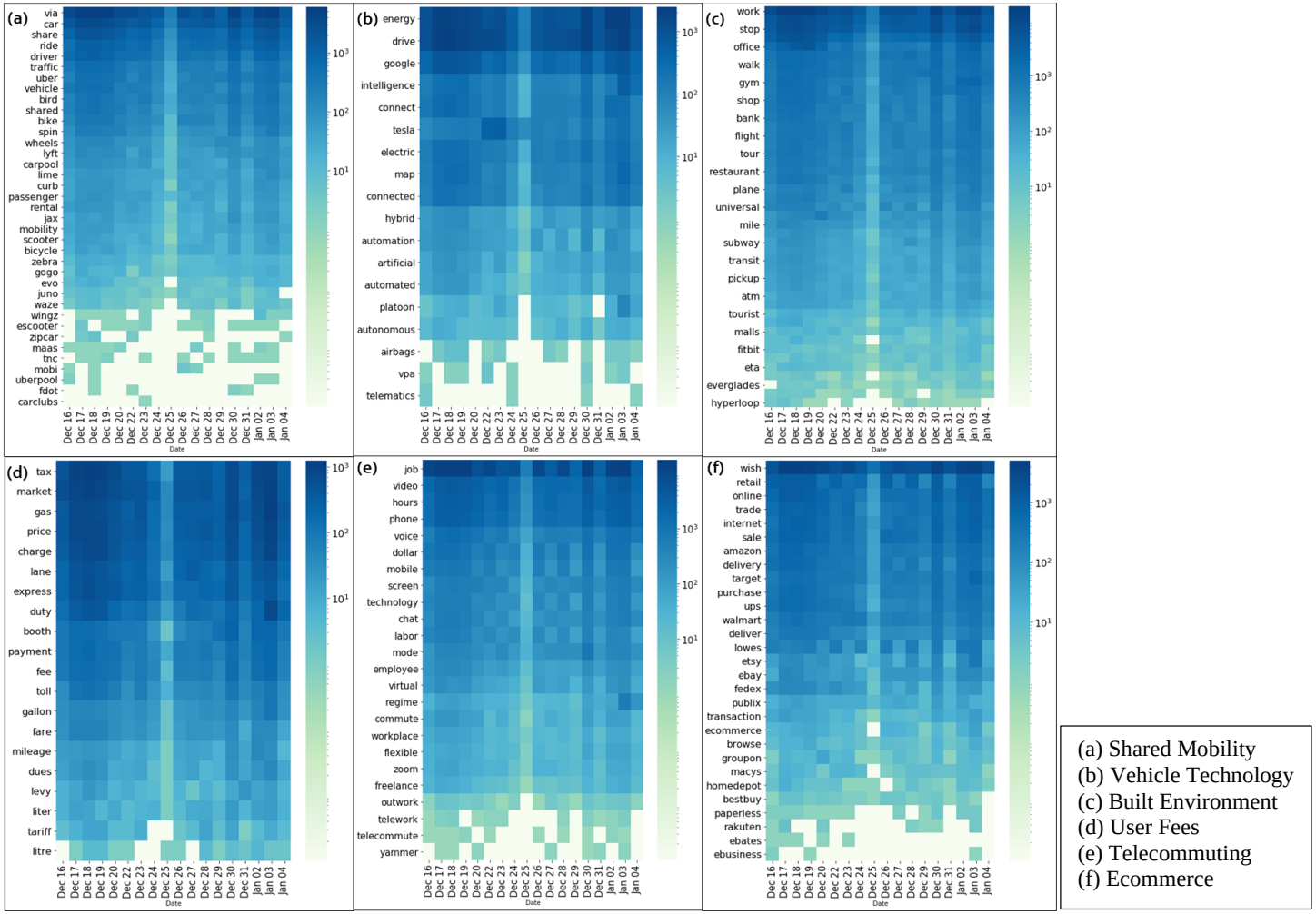

Figure 2-6: Word Frequency over Time for Six Keyword Categories.

Telecommuting:

- ‘job’ is highly prominent. This is also an important telecommuting platform
- ‘video’, ‘hours’, ‘phone’, ‘voice’ also showed strong presence, followed by ‘dollar’, ‘mobile’, ‘screen’, and ‘technology’
- emerging platforms such as ‘freelance’, ‘outwork’, ‘telework’, ‘yammer’ was found less frequent on Twitter
- ‘zoom’ showed less prominence relative to ‘phone’. This is indicative of the need to leverage social media for zoom meeting

Ecommerce:

- ‘wish’ was highly prominent. This is a popular e-commerce platform
- ‘retail’, ‘online’, ‘trade’, ‘internet’ also showed strong presence, followed by ‘sale’, ‘amazon’, ‘delivery’, ‘target’, ‘shared’, and ‘bike’
- ‘Walmart’ was more popular than ‘Publix’
- platforms such as ‘Macy's’, ‘home depot’, ‘BestBuy’, ‘paperless’ were found less frequent on Twitter
- ‘Walmart’ and ‘target’ were less frequently relative to ‘amazon’. This is indicative of the popularity of ‘amazon’ over ‘Walmart’ and ‘target’ as an e-commerce platform

2.4.3 Sentiment Analysis

The VADER python package performed sentiment analyses of tweets, and corresponding user sentiments were reported as highly negative, negative, neutral, positive, and highly

positive. While the heatmaps of tweeting keywords provided the significance of individual keywords representing social media user concerns on transportation and mobility trends, the combined effects of multiple words in each tweet were analyzed to quantify user emotions or sentiments based on such interactions. Sentiment or opinion mining results are presented in Figure 2-7 and Figure 2-8 for each category, i.e., shared mobility, vehicle technology, built environment, user fees, telecommuting, and e-commerce. While Figure 2-7(a) shows the distribution of relative sentiments i.e., the percentage distribution of five different sentiment types for all the relevant tweets, Figure 2-7(b) presents the distribution of relative sentiments, i.e., the percentage distribution of five different sentiment types for all the six categories. Figure 2-8 presents the percentage distribution of five different sentiment types at the top 50 tweeting locations (state level) for all the six categories, i.e., sentiments over space.

A few key observations from Figure 2-7 (a, b) are summarized here:

- Overall, around one-third of the relevant tweets are positive, and about one-fifth expressed highly positive views. Around 24% of tweets showed a negative view (negative and highly negative).
- In all the categories, most of the tweets are positive, and the least of the tweets are highly negative.
- While shared mobility, vehicle technology, telecommuting, and ecommerce have a higher proportion of positive tweets than negative tweets, built environment and user fees showed opposite scenarios.

- Vehicle technology and ecommerce have the highest proportion of positive tweets among all the categories.
- User fees has the least proportion of positive tweets and the highest proportion of negative and highly negative tweets among all the categories.

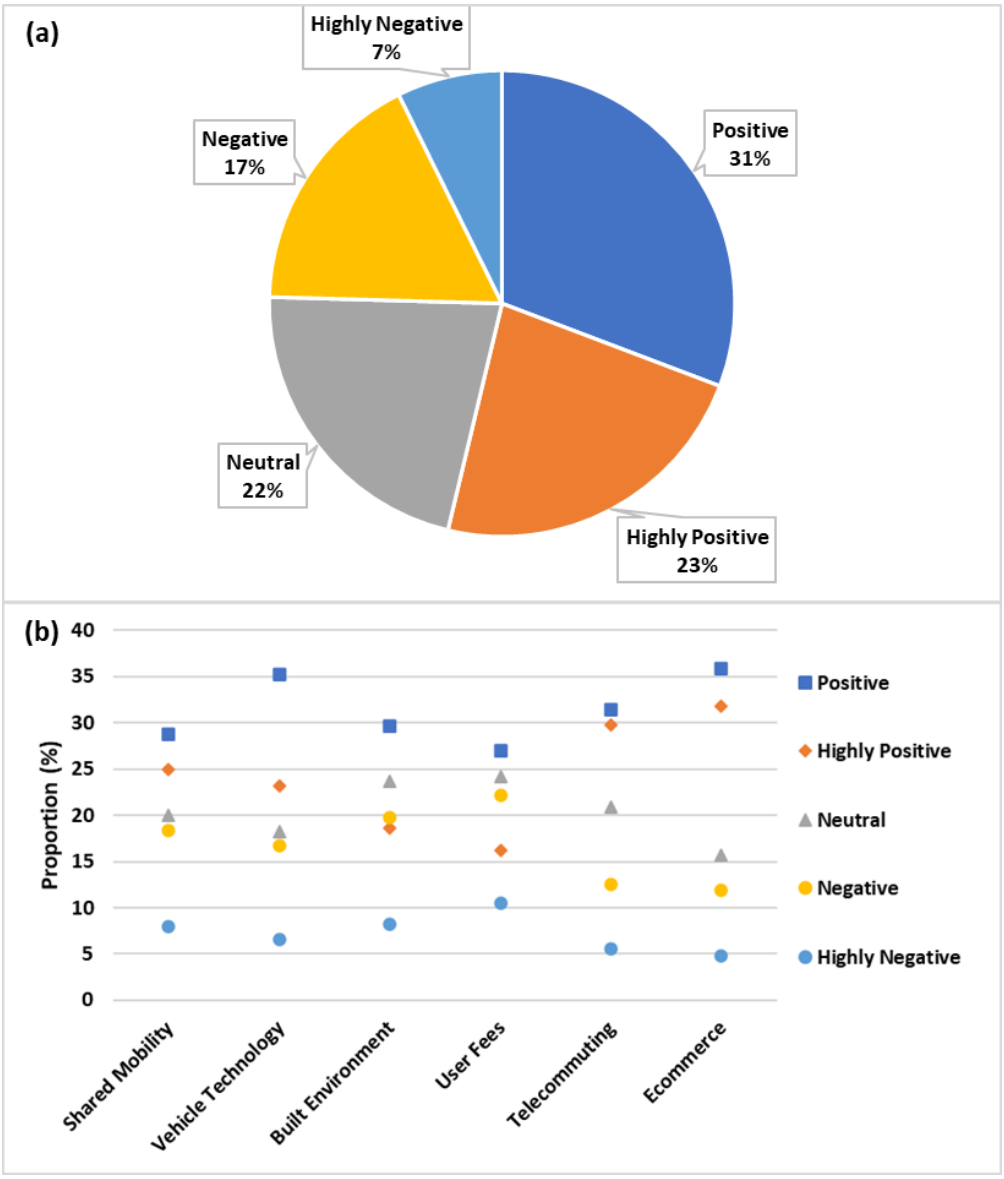

Figure 2-7: Sentiment Analysis over Time for Six Categories. (a) Shared Mobility, (b)Vehicle Technology, (c) Built Environment, (d) User Fees, (e) Telecommuting, (f) E-commerce.

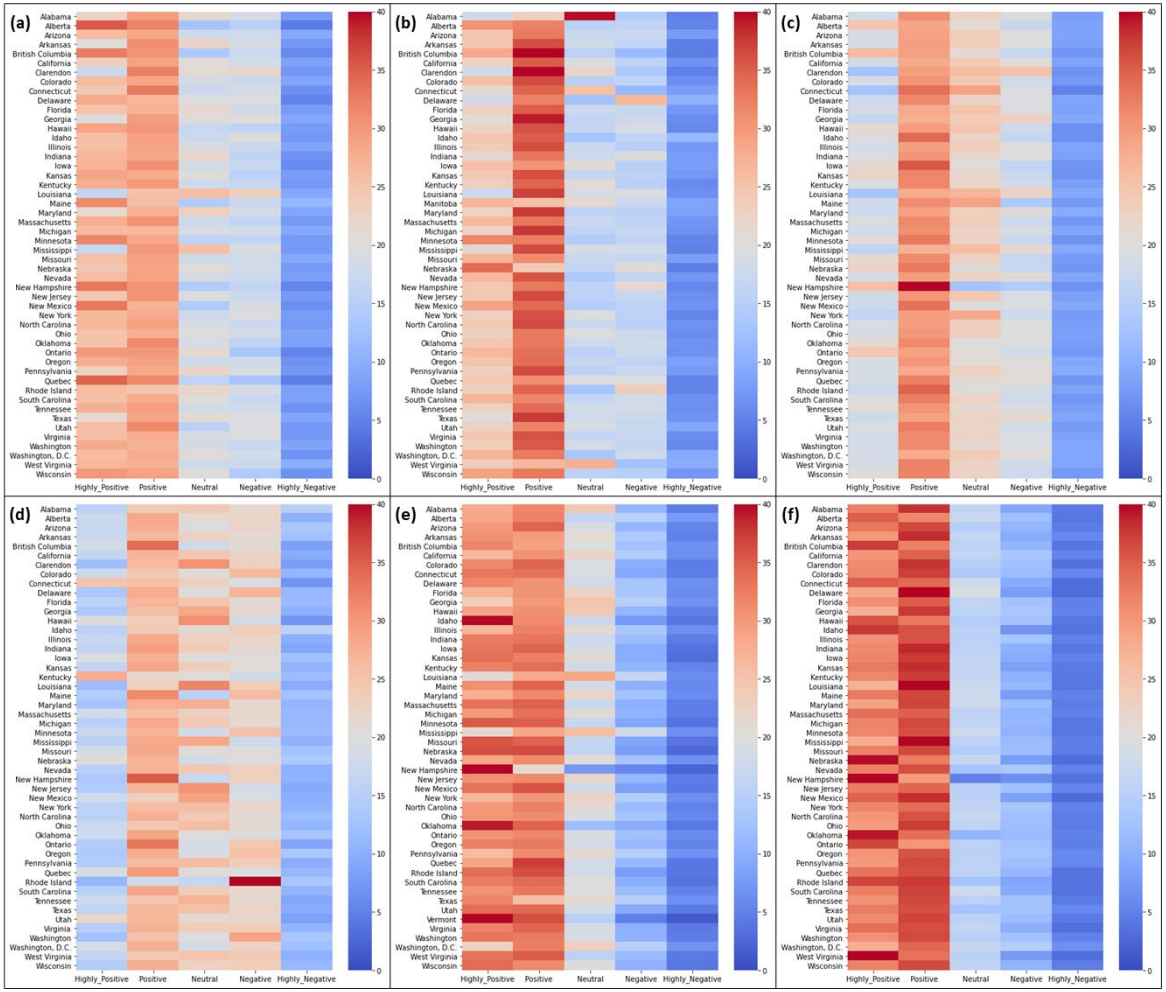

Figure 2-8: Sentiment Analysis over Space for Six Categories. (a) Shared Mobility, (b)Vehicle Technology, (c) Built Environment, (d) User Fees, (e) Telecommuting, (f) E-commerce.

A few key observations from Figure 2-8 (a-f) are summarized here:

- Most of the tweets on shared mobility showed positive and highly positive views in almost all the places. However, places such as Arkansas, Clarendon, Georgia, Louisiana, and Mississippi showed some exceptions, generating a relatively higher proportion of neutral and negative tweets.
- Though tweets on vehicle technology also showed an almost similar trend like shared mobility in different places, places such as Alabama, Connecticut, Delaware, Rhode Island, and West Virginia generated a relatively higher proportion of neutral and negative tweets.
- In the case of the built environment, though most of the tweets are positive over different places, there is also a higher proportion of neutral and negative tweets in many places concerning other categories (except user fees).
- In most places, tweets are more likely positive, neutral, and negative on user fees. Even places like Rhode Island, Washington, and Colorado generated a higher proportion of negative tweets than other sentiment types.
- Telecommuting and E-commerce showed similar patterns in different places. In all places, tweets mainly showed positive and highly positive views, and there are a tiny proportion of neutral, negative, and highly negative tweets.
- Overall, most locations showed a more positive attitude towards shared mobility, vehicle technology, telecommuting, and e-commerce, whereas they were relatively more negative about the built environment and user fees.

These findings indicate the need to design and implement more dedicated and targeted efforts to improve public satisfaction with certain transportation aspects based on quantitative evidence observed through social media interactions.

2.4.4 Topic Modeling

Topic modeling analysis was applied to investigate how different combinations of words in the data may constitute social interaction topics of transportation trends. While sentiment analyses helped quantify positive, neutral, or negative attitudes of social media users, topic models typically provide more insights on the actual topics in text data. Topic coherence means the average /median of the pairwise word-similarity scores of the words in the topic and has been used to specify the number of unique topics [62]. A good topic modeling depends on higher coherence, which depends on two predefined parameters: (a) Number of topics; (b) Number of iterations. The optimal number of topics and iterations was estimated after several trials. The tentative generated topics for six categories are presented in Figure 2-9. For other categories (user fees, vehicle technology, built environment, telecommuting, and e-commerce), the tentative number of generated topics was found at 6, 5, 5, 8, and 5, respectively.

A total of 17 topics related to emerging transportation tend to have been reported (Table 2-2). Table 2-2 reports the topic modeling coherence score for each category as well as the probable interaction topics with their probability in that category and the five most frequent associated words contributing to the formation of a topic with their probability at that topic (in brackets). Only the top 5 words were reported here for illustration purposes.

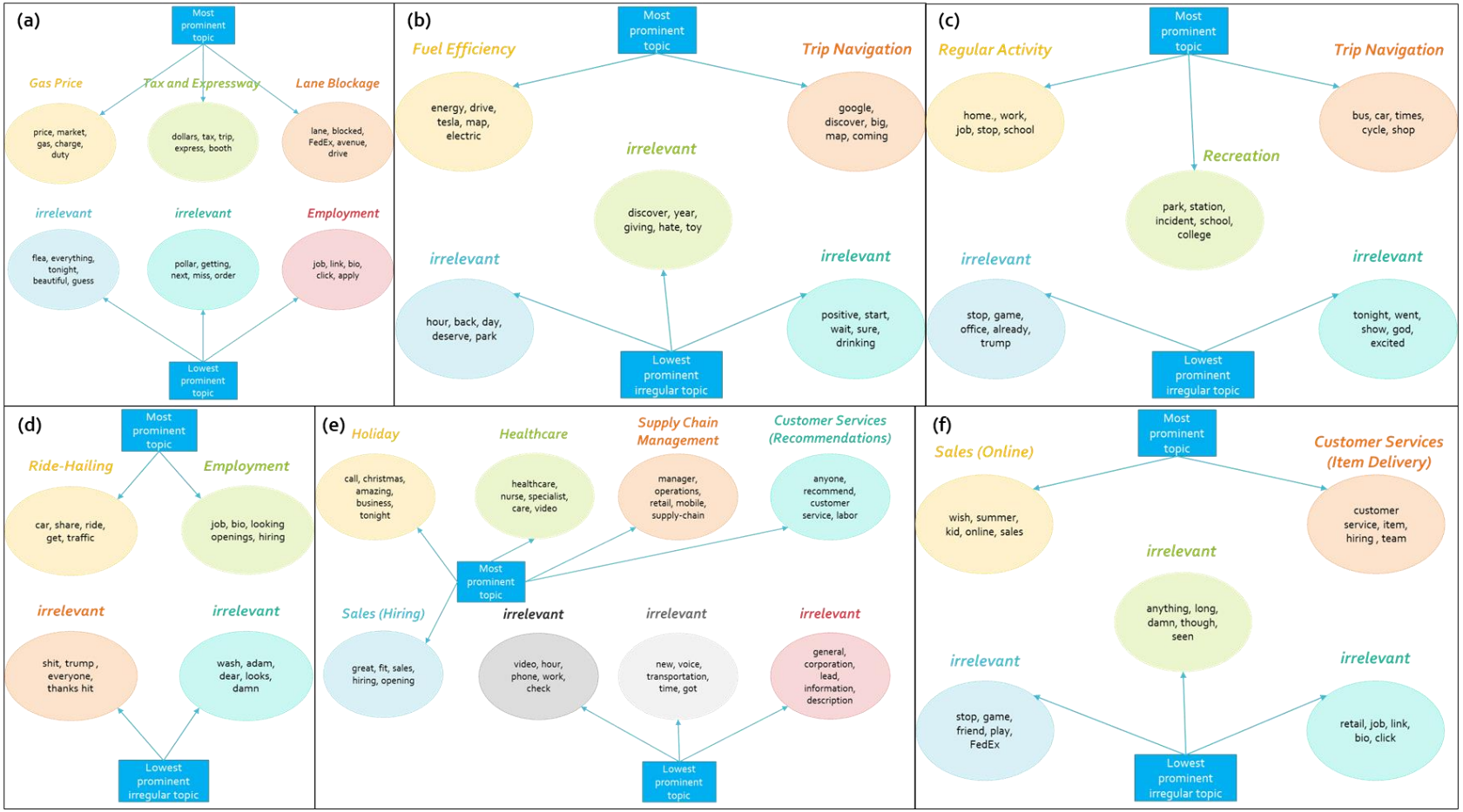

Figure 2-9: Tentative generated topics for Six Categories. (a) User Fees (b) Vehicle Technology (c) Built Environment (d) Shared Mobility (e) Telecommuting (f) E-commerce.

People primarily mentioned ride-hailing and employment opportunities as part of shared mobility. On vehicle technology, interactions mainly included topics on fuel efficiency and trip navigations. Regular activities on a day-to-day basis are among the built environment topics, in addition to shopping and recreational activities. Under the user fees category, people were more concerned about gas price, tax, and expressways, along with their probable frustration with lane blocks while driving. On telecommuting, people talked more about the holiday season and healthcare activities, and customer services related to item delivery were among the predominant topics on e-commerce. Such topics and associated words provide better insights on how to identify and connect to social media users based on their topics of interest and the use of specific keywords that can maximize influence.

Table 2-2. Emerging Transportation Trends Related Most Coherent Topics.

Trend Category (Coherence score)	**Interaction Topics**	**Topic Probability**	**Most probable words incoherent topic**
Shared Mobility (0.363)	Ride-Hailing	0.472	Car (0.016), share (0.011), ride(0.011), get(0.006), traffic (0.005)
	Employment Opportunity	0.192	Job(0.027), bio(0.026), looking(0.017), openings (0.017), hiring(0.12)
Vehicle Technology (0.321)	Fuel Efficiency	0.561	Energy(0.026), drive(0.016), tesla(0.008), map(.005), electric(.005)
	Trip Navigation	0.168	Google(0.039), discover(0.03), big(0.022), map(.009), coming(0.008)
Built Environment (0.325)	Daily Activities	0.569	home(0.012), Work(0.011), job(0.023), stop(0.007), school(0.007)
	Shopping	0.152	Bus (0.017), car(0.016), times(0.007), cycle(0.007), shop (0.005)

	Recreation	0.063	Park(0.023), station (0.023), incident (0.01), school (0.007), college (0.007)
User Fees (0.338)	Gas Price	0.582	Price(0.01), market(0.01), gas(0.01), charge(0.007), duty(0.005)
	Tax and Expressway	0.064	dollars(0.023), tax(0.019), trip(0.009), express(0.006), booth(0.006)
	Lane Blockage	0.065	lane(0.018), blocked(0.013), FedEx(0.013), avenue(0.012), drive(0.012)
Telecommuting (0.353)	Holiday	0.5	call(0.032), Christmas(0.026), amazing(0.023), business(0.022), tonight(0.021)
	Healthcare	0.303	Healthcare(0.06), nurse(0.01), specialist(0.044), care(0.045), video (0.022)
	Supply Chain Management	0.055	manager(0.036), operations (0.021), retail(0.036), mobile(0.026), supply chain(0.023)
	Customer Services (Recommendations)	0.077	anyone(0.137), recommend(0.134), customer service(0.048), labor(0.043)
	Sales (Hiring)	0.266	Great(0.06), fit(0.042), sales(0.041), hiring(0.040), opening(0.033)
Ecommerce (0.390)	Sales (Online)	0.523	Wish(0.023), summer(0.01), kid(0.01), online(0.006), sale(0.006)
	Customer Services (Item Delivery)	0.192	customer service(0.025), item(0.01), hiring (.078), team(0.05)

2.4.5 Study Limitations

This study's results showed that there seems to be significant potential for using social media data to develop models for identifying emerging transportation indicators and long-term planning purposes. However, it is acknowledged that small events that are retweeted

several times may affect the collected dataset. Moreover, due to user privacy issues that limit personal information availability, there is usually insufficient information on social media users to detect biases in any given subject's sample population.

Twitter users include people, news organizations, and companies, and, perhaps most troubling, are not always human. Previous research has shown that Twitter includes many bots that automatically send tweets, mainly to promote a product or a political campaign [63]. The elimination of these tweets is not achieved in this study, but several methods for finding them have been proposed [64–66]. So, special caution is required to the biases associated with social media data.

Another limitation is that Twitter data was not able to collect all the tweets during that period as the streaming API was used for collecting tweets because that specific API does not allow collecting all data. To make this type of online social media research more authentic and comprehensive, a different type of paid Twitter API (Power track, Enterprise) and other social media platforms (Facebook, LinkedIn, etc.) can also be used for future research, which will collect most of the tweets.

2.5 Conclusions and Discussions

Transportation researchers, in recent times, used SMPs extensively for problems related to travel demand forecasting, activity pattern modeling, transit service assessment, traffic incident, and disaster management, among others. Nevertheless, there is still much more to explore how such information can contribute to understanding public perception and attitude towards emerging transportation trends and mobility indicators. As such, this study aims to mine and analyze large-scale public interactions from SPMs enriched with time

and location information and develop comparative infographics of emerging transportation trends and mobility indicators using natural language processing and data-driven techniques.

About 13M tweets for about 20 days (Dec 16th, 2019- Jan 4th, 2020) were collected using Twitter API. Tweets closely aligned with emerging transportation and mobility trends such as shared mobility, vehicle technology, built environment, user fees, telecommuting, and e-commerce were identified. Data analytics captured spatio-temporal differences in social media user interactions and concerns about such trends, as well as topics of discussions formed through such interactions. California, Florida, Georgia, Illinois, New York, North Carolina, Ohio, Pennsylvania, Texas, Virginia, and Washington are among the highly visible cities discussing such trends. Key observations from sentiment analysis indicate that around one-third of the relevant tweets are positive, and about one-fifth expressed highly positive views.

Moreover, around 24% of tweets showed negative views (negative and highly negative). People carried more positive views on shared mobility, vehicle technology, telecommuting, and e-commerce while being more negative about user fees, and built environment. Analysis of sentiment over space showed that most locations showed a more positive attitude towards shared mobility, vehicle technology, telecommuting, and e-commerce, whereas relatively more negative on the built environment and user fees.

Topic modeling analysis identified 17 topics related to transportation trends. Ride-hailing, fuel efficiency, trip navigation, daily as well as shopping and recreational activities, gas price, tax, and product delivery were among the topics. Specifically, people primarily

mentioned ride-hailing and employment opportunities as part of shared mobility. On vehicle technology, interactions mainly included topics on fuel efficiency and trip navigations. Regular activities on a day-to-day basis are among the built environment topics, in addition to shopping and recreational activities. Under the user fees category, people were more concerned about gas price, tax, and expressways, along with their probable frustration with lane blocks while driving. On telecommuting, people talked more about the holiday season and healthcare activities, and customer services related to item delivery were among the predominant topics on e-commerce. Such topics and associated words provide better insights on how to identify and connect to social media users based on their topics of interest and the use of specific keywords that can maximize influence. The above-listed topics and information can help transportation planners and policymakers systematically make better and timely decisions while facing future transportation demand for emerging technology. This will lead to a step forward in understanding the need for a modern transportation system to reduce dependency on fossil fuel, controlling climate changes, reducing traffic jams and accidents while increasing the reliability of the transportation system.

The social media data-driven framework presented in this study would allow real-time monitoring of transportation trends by agencies, researchers, and professionals. Potential applications of the work may include: (i) identifying spatial diversity of public mobility needs and concerns through social media channels; (ii) developing new policies that would satisfy the diverse needs at different locations; (iii) leveraging SMPs to promote user interests on emerging trends based on similar word clustering; (iv) design and implement more efficient strategies to improve and influence public interest and satisfaction. While

data biases may exist in such an approach, large-scale observations from SMPs would help predict convincing patterns with heightened statistical power.

CHAPTER 3
COMMUNITY-BASED MOBILITY BEHAVIOR ANALYSIS IN THE EMERGENCE OF COVID-19 OVERCOMING SAMPLING BIAS OF SOCIAL MEDIA DATA

3.1 Introduction

The emergence of major disruptive events seriously impacts human behavior and decision-making. The current COVID-19 outbreak has substantially altered people's travel and purchasing habits worldwide [67–70]. New York City (NYC) is at the center of this pandemic in the U.S. According to data, subway ridership plummeted by 91%, and vehicular traffic over Metropolitan Transportation Authority (MTA) Bridges and tunnels declined by 68% in April 2020 compared to April 2019 [71]. In 2020, the imposition of various policies and restrictions to prevent the spread of COVID-19 caused an unprecedented decline in the vehicle and public transit systems in NYC. With the broader accessibility of technology-enabled transportation options and modern communication devices (particularly smartphones), the variance in travel behavior caused by this pandemic has become more diverse across different places, points of time, modes of transportation, and socio-demographic and economic groups.

Twitter is among the most microblogging sites in the United States, with 199 million daily active users [72]. It is a microblogging website that allows users to communicate their opinions, activities, and ideas in 280-character messages known as "tweets." Geo-tagged tweets include tweet text, hashtags, and geo-location. Those are regarded as check-in data since they indicate the tweet's posting location [73]. In conjunction with tweet content, hashtags can give important information about traffic incidents. Data collected from social

media may be a valuable tool for demonstrating popular opinion on socioeconomic problems. The development of SMPs and people's increased involvement with online media have provided an excellent opportunity for transportation service providers to collect real-time information from SMPs users while paying as little as possible[74].

The primary sources of data used in most studies related to transportation trends (e.g., shared-mobility) [5, 9–11] are surveys (e.g., travel surveys) that feature representative populations and detailed information about travel mode and trip purposes. Survey data has some limitations, such as variabilities across countries in data collection method and data availability, lack of real-time engagement of the respondents, expansive and time-consuming as trend analysis requires periodic data collection making. SMPs can overcome the drawbacks of surveys and other travel data sources as they serve the need for more unified, less privacy-invading, simply accessible, and unified data to fully understand the dynamics of travel patterns. However, it is generally recognized that research, including social media data, faces a looming issue in the form of sample bias. Compared to the general population, the Twitter user population shows enormous disparities. As a result, when utilized directly for travel behavior research, social media data contains biases and mistakes to some extent. Moreover, social media data contains many noises, making it challenging to keep the topical relevance of the research. Nevertheless, text classification using supervised[75, 76], unsupervised [77, 78], and semi-supervised[79, 80] can handle this drawback of social media data significantly.

This study's goal is to demonstrate the value of large-scale social media data, particularly Twitter data, for community label travel behavior in the emergence of COVID-19. For this

purpose, around two months (April. 12, 2020 – May 31, 2020) worth of Twitter data (~1.35 M tweets) from NYC have been retrieved and used in the analysis. The first objective is to capture reliable signals from the Twitter stream concerning the pandemic. Once the pandemic-relevant signals had been extracted, tweets concerning different travel indicators (e-commerce, ridesharing, telework, transit) were classified through a semi-supervised machine learning approach. Semi-supervised approaches are being considered because they allow us to use a minimal quantity of labeled data, minimizing the initial labeling work necessary to develop a classifier. Then this study reduced the sampling biases of Twitter data by stratified random sampling after extracting user-level (e.g., gender, race) and tweet-level (e.g., per capita income) attributes from national databases (Social Security Network[81], census[82], and American Community Survey[83]). Finally, this study developed a multinomial logistic regression model to understand the public concern on different mobility indicators (e-commerce, ridesharing, telework, transit) under the influence of different socioeconomic and demographic factors at the community level.

3.2 Background and Related Work

Though SMPs are relatively new fields for research, during the previous decade, an increasing number of research on the subject of sentiment analysis on Twitter have been conducted[84–86]. Pender et al. [46] applied crowdsourcing techniques to derive transit service information that can satisfy the increased demand and expectation for real-time information dissemination. Collins et al. proposed a novel approach to assess rider satisfaction in Chicago train lines using social media data, considering people's opinions and metrics[87]. Luong & Houston [47] also used social media data to study public

attitudes about light rail transit services in Los Angles. Nik Bakht et al. [88] used not only social media data but also news sources to assess public attitudes regarding transportation planning.

But in the field of behavioral research, the use of social media data in opinion mining raised questions regarding sample representativeness to understand population sentiments. There is a big difference between the population of Twitter users and the general population. Therefore, when social media data is used directly for travel behavior analysis, there is a certain degree of bias and error. Several researchers have attempted to reduce the sampling biases by different sampling techniques, such as simple random sampling[89–91] and constructed week sampling[89, 92, 93]. In these studies, the sample was created by collecting data at different predefined selected days and times and at different frequencies. Researchers also utilized convenience sampling to conduct a social media content analysis based on their objectives[94–96], such as how Twitter has been used by local television for branding. But none of the studies used users' demographic attributes to remove the sampling biases to make population-level travel behavior predictions.

On the other hand, there is growing interest in predicting the attributes of social media users. Researchers have studied recently to predict age[97–99], sex[100, 101], and race/ethnicity[102, 103]. Few other works identify demographics from web browsing histories[104]. Population-level statistics were found to be predicted in a few studies. Eisenstein et al. predict zip-code statistics of race/ethnicity, income, and other factors using Census data[105]; Schwartz et al. [106] and Culotta [107] predict county health statistics

similarly using Twitter. However, none of the previous research attempted to anticipate or assess at the user level.

Topical irrelevancy is one of the significant drawbacks of working with social media data. Though SMPs provide a cheap way to extract a large amount of real-time data easily, most of the data lack topical relevancy. The intrinsic ambiguity (e.g., non-standard spelling, inconsistency) of SMPs data creates biases and misleading results when utilized directly for travel behavior research. Researchers have been using different machine learning approaches across domains to classify SMPs data to reduce its inherent lack of topical relevancy[75, 80, 108–110].

Compared to the existing studies, the core contribution of this study is to categorize tweets and incorporate the user demographics (e.g., gender) to make a suitable sample by reducing sampling biases. This study further develops an econometric model incorporating socio-economic and demographic attributes to explore people's travel behavior. In brief, this study introduces a way to interpret social media statistics with the socio-economic and demographic statistics.

3.3 Data and Methods

3.3.1 Data Collection and Description

We collect four types of data for this study. The description of the datasets is given below:

3.3.1.1 Twitter Data

The study period of this research is chosen in such a way so that the impact of the “first wave” of the pandemic in NYC can be captured. According to NYC mobility statistics, the city's transportation system had the most significant decrease during the first wave of

COVID-19 since tight anti-pandemic measures were implemented[111]. Data showed that the daily trip started decreasing from mid-March and remained significantly low during April and May[112]. So, tweets from NYC and its adjacent regions were collected through Twitter Streaming API (Application Programming Interfaces) using a location-bounding box that included all five boroughs (i.e., county-level administrative divisions) of NYC and its surrounding areas, constrained by about (40.49, -74.25) and (40.92, -73.70) coordinates from April 12, 2020, to May 31, 2020. Python programming language was used to collect the data using associated Python libraries. This study mainly depends on geotagged tweets as tweet geographic information is a potential parameter for spatial filtering, sampling, and analysis. No additional features or keywords were used to collect the tweets. The raw data contains approximately 1.35 M tweets. Approximately 100% of tweets are geotagged and mainly in English (~ 80.44%), with around 62,200 unique users.

3.3.1.2 Social Security Administration Data

Social Security Administration (SSA)[81] collects all names from social security card applications for births in the United States that occurred after 1879. We have collected these data from the SSA website to identify the gender demographics of the Twitter data by comparing the first name. It is worth noting that many persons born before 1937 never applied for a Social Security card; thus, their names are not in the database. The SSA data contains 63,152 male and 37,212 female names.

3.3.1.3 Census Data

The United States Census Bureau compiles annual estimates of race and Hispanic origin shares for each county in the country according to the respondent's surname. These

estimates are based on the most recent decennial census as well as estimates of population changes (deaths, births, and migration) since that time. Respondents can choose from one of six racial groups on the census questionnaire: White, Black, or African American, American Indian, Alaska Native, Asian, Native Hawaiian, and Other Pacific Islander, or Other, which creates a wide range of variations. While race/ethnicity is a complicated topic, we simplify it for the sake of this study by focusing on only four categories: Asian, Black, Hispanic, and White. We use the 2010 estimates for this study[82]. The last names according to different races were collected from the census bureau to identify a user's race/ethnicity in the Twitter data. The census dataset contains 162,254 surnames from different races/ethnicity.

3.3.1.4 Socio-economic Factors

NYC has a total of 5 counties and around 6807 census block groups. The socio-economic characteristics vary among different counties and different block groups. In this study, we collected five relevant socio-economics factors on the census block groups level from American Community Survey (ACS)[83] to understand their influence on the public concern towards different travel-related options. The collected census block group-level factors are:

- PCI: Per capita income in the past 12 months (in 2020 inflation-adjusted US dollars)
- MTT: Mean travel time to work (minutes), workers aged 16 years+, in the past 12 months
- BPL: Proportion of people living under the poverty limit in the past 12 months
- UE: Proportion of unemployed people in the past 12 months

- HS: Proportion of people who completed at least a high school degree

3.3.2 Data Cleaning

Tweeter data has been preprocessed before performing the analysis. The demographic and socio-economic factors are also infused with tweeter data to prepare a fine-tuned dataset. Multiple steps have been adopted in the study for analysis. Figure 3-1 describes the general framework of the study.

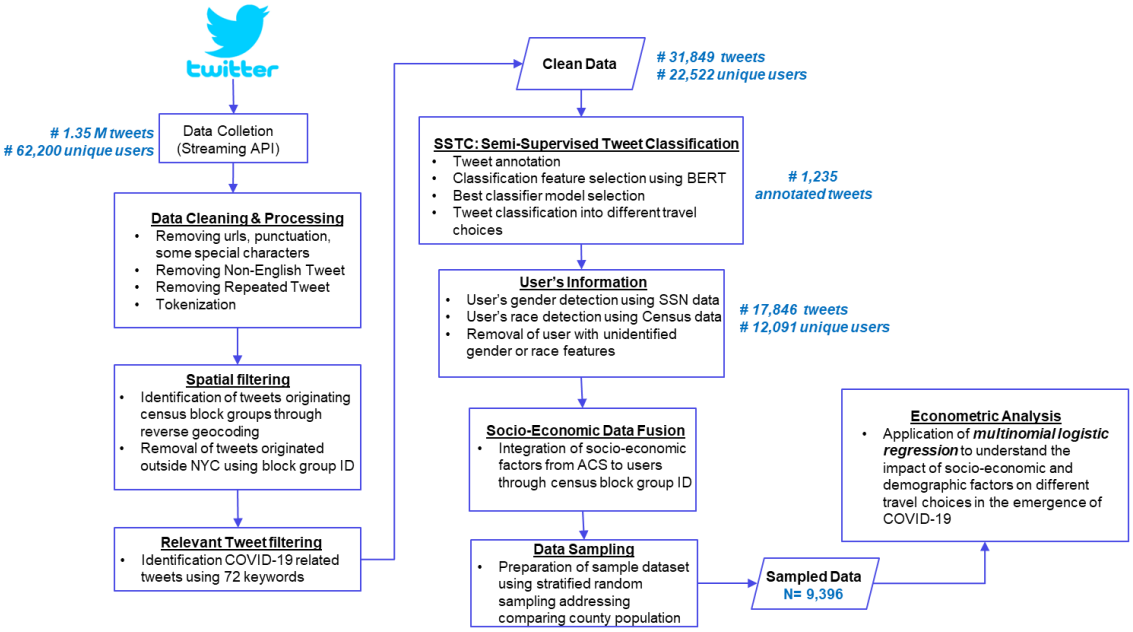

Figure 3-1: Conceptual Framework of this Study.

Tweets retrieved from the streaming API contain additional information such as user id, username, profile information, and tweet location (co-ordinates of the neighborhood) along with the tweet text. Tweet texts, user information, and location information were considered for analysis in this study. Given the inherent ambiguity of tweets (e.g., non-standard spelling, inconsistent punctuation, and/or capitalization), additional preprocessing steps are performed to extract clean tweet text and username, which is suitable for analysis.

Noises (such as Html tags, character codes, emojis, stop words, etc.) were removed from the text data and username, and tweets were tokenized, which is the process of breaking down an expression, sentence, paragraph, or even an entire text document into smaller units like individual words or phrases. Tokens are the names given to each of these smaller units.

Though the data was collected using a rectangle bounding box, it contained many tweets generated outside NYC. The census block group id of the tweet originating places was generated by reverse geocoding using census geocode 0.5.2 python package[113]. Using the census block group information, tweets that originated outside NYC were removed from the dataset. Finally, the COVID-19-related tweets were filtered using a keyword list of 72 terms which is given below:

- **COVID-19 related keywords:** *breathe, face, breathing_problem, 2019_ncovnyc, breathing, 2019_ncov, social, spread, COVID-19nyc, corona, wuhan_outbreak, home, distancing, vaccine, wearing_mask, sars-cov-2, social_distancing, COVID19, stay_at_home, distance, pandemic, social_distance, virus, sars, spreading, shelter, lockdown, sarscov, support, sanitizer, wuhan_virus, health, quarantine, deaths, order, safe, tested, masks, cases, facecover, wuhanvirus, face_mask, positive, coronavirus, wearing_masks, immunity, isolation, COVID, socialdistance, facecovering, herd, infect, orders, wuhanoutbreak, ncov, hygiene, herdimmunity, COVID-19, crisis, wuhan, stayhome, wearing, death, face_masks, stayathome, ppe, n95, outbreak, mask, wear, stay_home*

COVID-19 relevancy of the tweets was established if the tweet contained at least one of the keywords identified for this study. Although this approach may filter out relevant tweets, it ensures that all tweets involving these keywords were included in the filtered

dataset for further analysis. At this step's end, a dataset of the relevant clean tweet was obtained with 31,849 tweets.

3.3.3 SSTC: Semi-supervised Tweet Classification

This step aims to classify cleaned relevant tweets obtained from previous steps according to various mobility indicators (e-commerce, ridesharing, telework, transit, others) from a given text input, where only one mobility indicator may be present. Being relatively challenging and expensive to annotate many tweets (n=31,849) manually, semi-supervised approaches are being considered because they allow us to use a minimal quantity of labeled data, minimizing the initial labeling work necessary to develop a classifier [114].

3.3.3.1 Data Labeling

1235 tweets from dataset of clean tweets (n=31,849) were randomly selected and manually annotated as 'E-commerce', 'Ride sharing', 'Telework', 'Transit', 'Others'. Two human annotators labeled the tweets. To ensure that the right labels of the travel trends were retrieved, those labels were considered when both annotators agreed on it. Each tweet can have only one label out of the five possible labels. Tables 3-1 and 3-2 show the distribution of annotated tweets and a few example tweets related to different travel choices.

Table 3-1. Distribution of Annotated Tweets.

Mobility Indicator	Number of tweets
E-commerce	97
Telecommuting	143
Ridesharing	72
Transit	161
Others	762
Total	1235

Table 3-2. Annotated Example Tweets.

Example tweet	Mobility Indicator
• *@edible i placed an order on friday, for delivery yesterday. the order still has not arrived. i left a vm and sent an email. silence. this is very time sensitive. please help.* • *im scared to order stuff online during this corona shit.* • *so i wont be ordering online groceries during a pandemic again. substitute my whole cart why dont ya ?*	Ecommerce
• *looking for those who needs extra income! opportunity to work online! #unemployment #nyc #brooklyn #queens #florida #miami #nj #ct #covid19* • *the coming mental health crisis as remote working surges* • *im home and already got like 10 absences from online school.*	Telework
• *@itzsuds lol i know exactly where you are. uber works, and so do masks and gloves.* • *@nytimes uber, lyft, and others have been suffering from their own "pandemic" since they started operating. $30b thrown away to develop a not for profit! unless your on the top floor. ride sharing propaganda has warped the minds of policy makers (some contributions haven't hurt either)* • *@uber .@uber_support been attempting to correspond w uber support through the chat app regarding free rides for medics. they continue to tell me i dont qualify for a promotion that ive sent screenshots of proof for.*	Ridesharing
• *yes, i wear my hospital badge now on the subway so that people wont come near me #workseverytime #socialdistancing* • *covid-19 update this mta bus did not pick me up @ 179th st subway station* *(f line)* • *my friend jessica took this picture on saturday, april 11 on the r train. so i ask again, how is this being addressed in the midst of nyc's #covid19 crisis @nycmayor, @nygovcuomo, @nyctsubway and @mta? this was exactly how the 2 train i took saturday morning looked.*	Transit

3.3.3.2 Classification Feature Extraction

Training a text classifier model in machine learning refers to supplying it with training data that includes both inputs and correct responses so that the algorithm may identify the pattern to map the input features to the target/output features. The tweets need to be

embedded, or features should be extracted for each tweet. For this purpose, we employed the uncased base model[115] of original English-language Bidirectional Encoder Representations from Transformers (BERT)[116], which is an open-source machine learning framework for natural language processing (NLP). BERT is intended to assist computers in understanding the meaning of ambiguous words in the text by establishing context via the use of surrounding material. BERT base has a total of 12 encoders with 12 bidirectional self-attention heads. The BERT framework was pre-trained with BooksCorpus[117] with 800M words and English Wikipedia with 2,500M words. It generated 768 hidden units/features for each tweet. The model's output contains a single travel choice; thus, the annotated labels were converted into multiclass formats where different labels are mutually exclusive.

3.3.3.3 Best Classifier Model Selection

A classification algorithm for text can be used to classify tweets, in this case, into the categories of e-commerce/telework/ridesharing/transit/others automatically. Firstly, a variety of popular and powerful supervised classification algorithms were applied to the data, namely: Logistic regression (LR), k-nearest neighborhood (KNN), random forest (RF), support vector machine (SVM), and multinomial naïve bayes (MNB). Python implementations found in the Natural Language ToolKit (NLTK) and Sci-Kit Learn [118] were used in this study.

To achieve our goal, we have created a semi-supervised technique suitable for small to medium-sized datasets. Semi-supervised learning seeks to leverage the combined knowledge from labeled and unlabeled data to outperform the classification performance

achieved by eliminating the unlabeled data and using supervised learning or deleting the labels and using unsupervised learning. The manually annotated dataset (n=1,235) was converted to a feature matrix of size 1235 X 768 using BERT, which was used to train the classifiers and find the best model. Lastly, the best model was applied to the remaining unlabeled dataset to get a labeled dataset (n=31,849) which is the basis of SSTC (Figure 3-2).

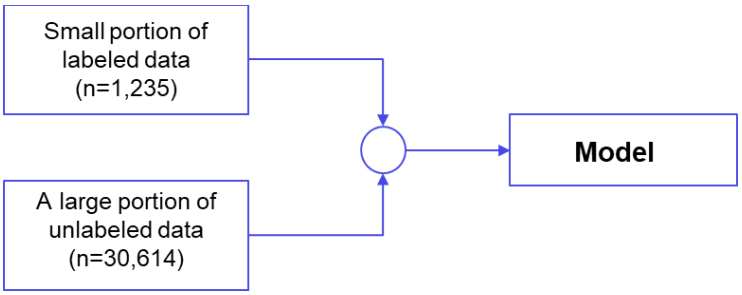

Figure 3-2: Conceptual Framework of Semi-supervised Text Classification.

According to this approach, a single input/tweet can be assigned to a single type of travel activity. Let A = $\{\lambda_i\}$ be the set of labels for different types of travel choices, where i = 1, …, |A|. In our case, |A|= 5. The purpose of the tweet classification model, m, is that given the input tweet, X, the model must predict a travel activity type, $Y \in$ A.

$$m: X \rightarrow Y \quad \text{.....(3-1)}$$

Three generally used performance metrics were chosen to evaluate the tweet classification model: accuracy, precision, and recall. Model Accuracy was calculated by using Eq. (3-2) expressed as:

$$Accuracy = Correctly\ classified\ tweets / Total\ tweet\ number \quad \text{.....(3-2)}$$

Also, precision is calculated as:

$$Precision = True\ Positive / (True\ Positive + False\ Positive) \quad \text{.....(3-3)}$$

The recall is calculated as:

$$Recall = True\ Positive/(True\ Positive + False\ Negative) \quad \text{(3-4)}$$

Here, a true positive result is one in which the model accurately predicts the positive class, whereas a true negative result is one in which the model adequately predicts the negative class. On the other hand, a false positive is an outcome in which the model predicts the positive class inaccurately, and a false negative is an outcome in which the model predicts the negative class inaccurately.

3.3.4 Gender and Race Identification

Gender-Race (GR) model was developed using a supervised machine learning approach [32, 119] to identify the race and gender of Twitter data. The model was trained by SSN (first name) and Census data (last name) to identify race and gender, respectively. GR model was applied to the final dataset (n=31,849), which was received by SSTC, to identify the race and gender. As the tweet username is self-reported, many improper names were discovered. These improper usernames contain lots of noises that make it impossible to identify gender and race by the GR model (Figure 3-3). The output of this step was used for econometric analysis to understand the influence of people's demographic on their travel-related choices.

3.3.5 Stratified Random Sampling

Depending upon the purpose of the studies and the characteristics of the datasets, we considered stratified random sampling as a perfect candidate for this study to represent the population. We proposed a noble approach to avoid the sampling bias of tweeter data by incorporating the demographic parameter (e.g., population) in the sampling process. In this

study, data were sampled according to the county demographic distribution (e.g., population) of the ACS data [83] by using stratified sampling. First, users will be divided into different subgroups according to their county. Then, simple random sampling (SRS) will be used to sample users from each subgroup. The number of samples drawn from different subgroups is determined by the proportion of this corresponding subgroup in the ACS survey data.

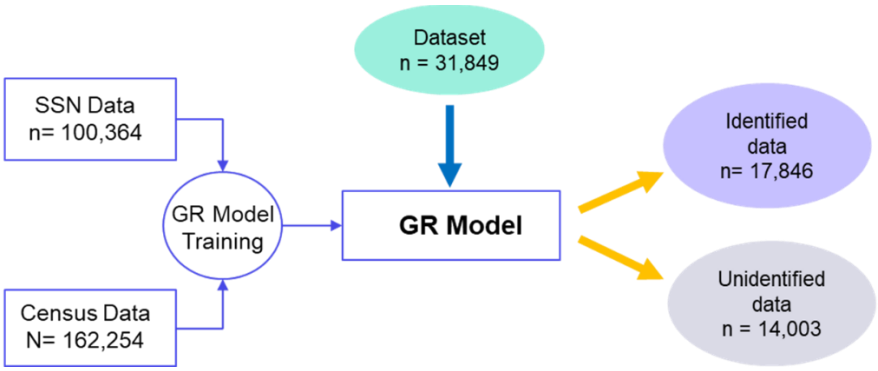

Figure 3-3: Conceptual Framework of Gender-Race Identification Model

3.3.6 Choice Model for Analyzing Mobility Indicators

To understand the public attitude on different mobility indicators (e-commerce, ridesharing, telework, transit, others) under the influence of different socio-economic and demographic actors, an econometric model was developed by applying the multinomial logit model to the classified tweets in this study. As the tweeting class is a categorical response variable in this study, the logit model was considered a perfect candidate for this model development.

The logit model is the most popular type of random utility model derived from consumer economics theory, and it was initially developed by McFadden[120, 121]. A person *i*

chooses one choice among discrete alternatives by evaluating their associated features *X* in utility maximization behavior. The person *i* choose the alternative *m* that provides the largest utility:

$$U_{im} > U_{ik}\ ; m \neq k \quad \text{(3-5)}$$

The utility can be classified into observed utility V_{im} and an unobserved utility ε_{im} to observe the complete utility of a person. There are two sets of attributes for $V_{im:}$

1) covariates associated with both the individual and the alternative X_{im}
2) decision-maker characteristics, S_i[121]

The observed utility (*V*) is a value determined by a linear function of the attributes employed, and it describes the popularity of an option constrained by the provided model specification as follows:

$$V_{im} = V(X_{im}, S_i) \quad \text{(3-6)}$$

On the other hand, unobserved utility ε_{im} remains unobserved by the researchers, which results from the specification of the observed utility V_{im}. As a result, researchers consider the unobserved terms to be stochastic. The logit model is explicitly constructed by assuming that each unobserved term, ε_{im}, is an IID extreme value, i.e., Gumbel and type 1 extreme values. By combining two utilities, we can calculate the likelihood of person *i* selecting alternative *j* by solving the mathematical formula:

$$P_{ij} = \frac{e^{\beta' X_{im}}}{\sum_{k=1}^{M} e^{\beta' X_{ik}}} \quad \text{(3-7)}$$

where X_{ik} is the vector of observed explanatory variables to choose a given alternative, and β' is the parameters for the observed utility. For more technical details about logit models in the discrete choice method[121].

3.4 Results

3.4.1 Tweet Classification

After data cleaning, a total number of 31,849 English tweets with 22,522 unique users were obtained. Using BERT, each of the tweets (labeled/ unlabeled) was embedded, and 768 features were extracted for each tweet. The labeled tweet dataset (n=1,235) was split (4:1) in a stratified way to address the data imbalance (Table 3-1) and create training/ testing data (Table 3-3). The BERT vector data of the training dataset (n=989) was used as input in the classifier model.

Table 3-3. Distribution of Training and Testing Tweets

Mobility indicator	**Tweet count**	**Training (80%)**	**Testing (20%)**
E-commerce	97	78	19
Telework	143	114	29
Ridesharing	72	58	14
Transit	161	129	32
Others	762	610	152
Total	1235	989	246

After training each model with 989 annotated tweets, testing was performed on 246 tweets for each model. The performance measures of the models are given in Table 3-4. It is seen that SVM performs better with our dataset. SVM was later used to classify the remaining unlabeled clean dataset (n=30,614) in SSTC. The results of the classification using SVM are given below in table 3-5.

Table 3-4. Model performance values (accuracy, precision, recall) (A higher score of accuracy, precision, or recall measure indicates better performance).

Model	**Accuracy**	**Precision**	**Recall**
Logistic Regression (LR)	0.52	0.55	0.51
K-nearest neighborhood (KNN)	0.51	0.53	0.54
Random Forest (RF)	0.56	0.58	0.63
Support vector machine (SVM)	**0.69**	**0.71**	**0.75**
Multinomial Naïve Bayes (MNB)	0.45	0.46	0.49

Table 3-5. Semi-supervised text classification results

Mobility indicator	Tweet count	Percentage (%)
Others	21,223	66.64
Telework	3,603	11.3
Transit	2,624	8.23
E-commerce	2,331	7.32
Ridesharing	2,068	6.51

3.4.2 Demographic Distribution of Users and Tweets

After data cleaning, a total number of 31,849 English tweets with 22,522 unique users were obtained. GR model, which was trained by SSN and Census data GR model, was used in idenication of the gender and race of the users from the Twitter username. The unidentified users were later removed from the dataset. Finally, a dataset of 17,846 tweets (unique user= 12,091) was obtained.

Around 59% of tweets are posted by males, which indicates the higher popularity of Twitter among males than females (Figure 3-4a). Moreover, the dominance of the white individuals among different races in Twitter usage in the dataset can be easily visible in Figure 3-4b.

Whites are the highest user group, followed by Hispanic, Asian, and Black in the clean dataset. Moreover, Figure 3-5a presents the tweet distribution over the counties in terms of frequency. Though most of the users and tweets belong to New York County, Bronx County leads in case of average tweets by individuals, followed by Queens and Kings (Figure 3-5b). People in Richmond are least likely to be active on Twitter to share their concerns.

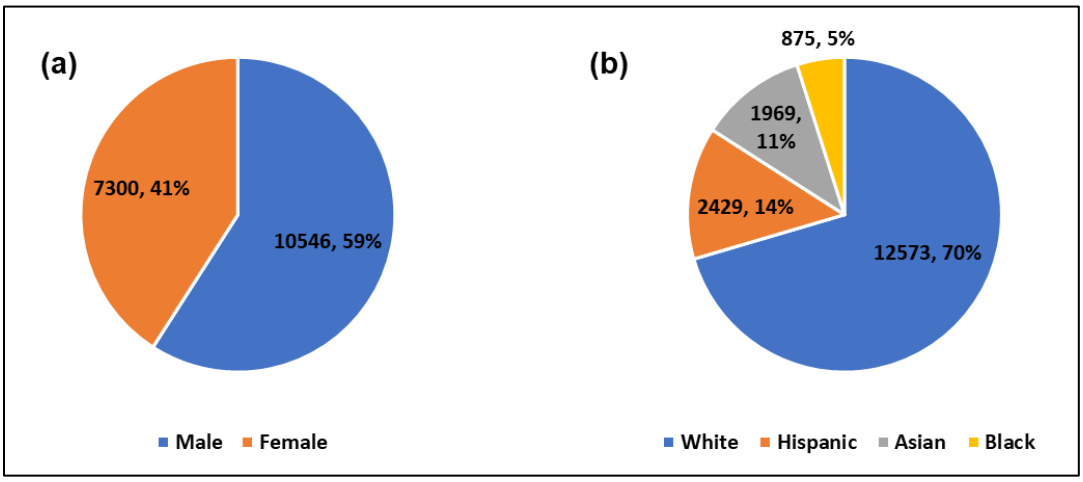

Figure 3-4: User demographic distribution. (a) over the gender, (b) over the race

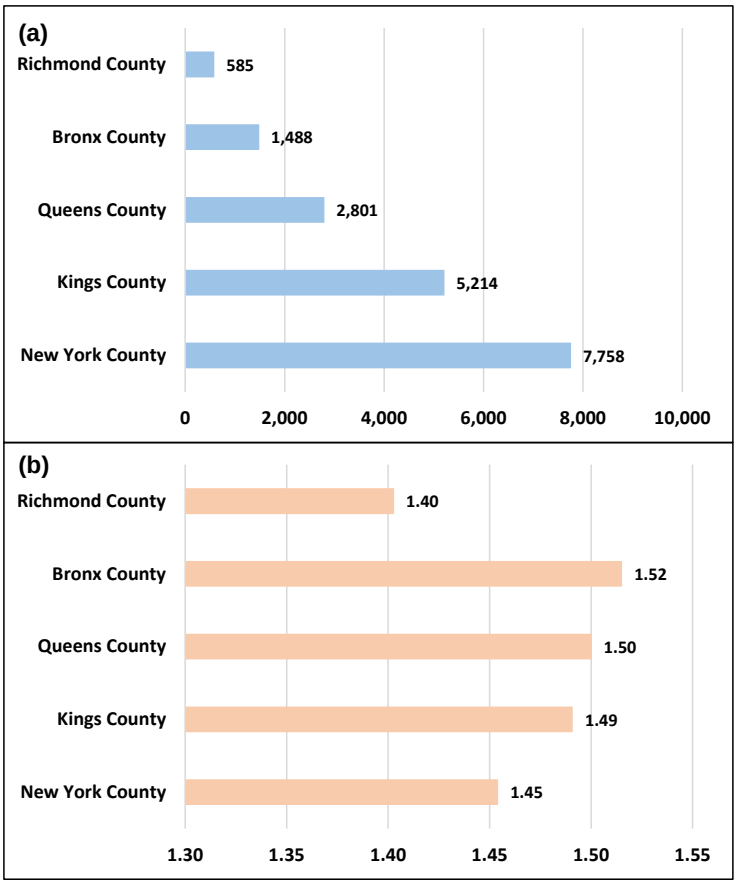

Figure 3-5: County (a) total tweet count and (b) average tweet per user distribution.

3.4.3 Data Sampling

One of the main contributions of this study is to introduce a stratified simple random sampling process, different from the existing studies on Twitter data[89–91, 93, 122], to gather a representative sample dataset that can be considered a representative subset of the NYC. We used the county population demographic of NYC in creating the sample dataset (Figure 3-6). So, we pulled data by stratified simple random sampling from the list of classified tweets to assure that the population proportion in the sampled dataset reflects the county population of NYC, which provides a strong correlation (r =0.99). Eventually, we ended up with a sample dataset of 9,396 tweets. This stratified sampling process can be perfectionated more by incorporating racial and gender joint distribution of NYC.

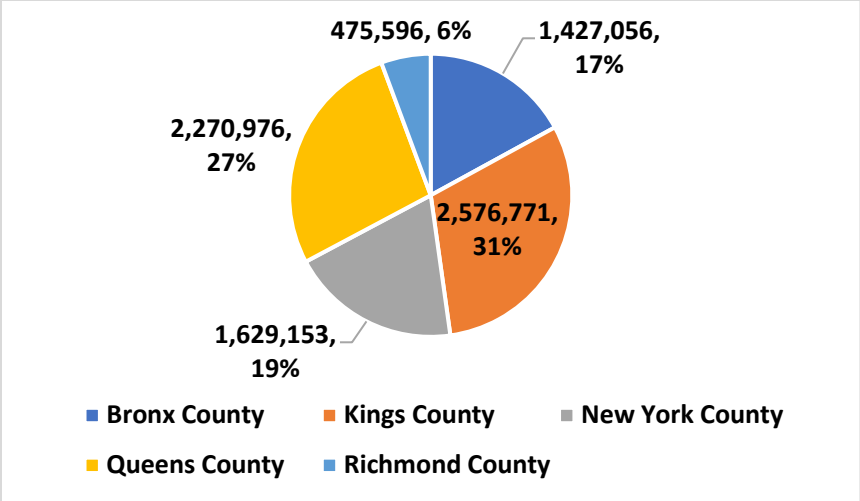

Figure 3-6: County Population distribution in NYC (2020)[123].

3.4.4 Choice Model Results

3.4.4.1 Mobility Indicator Choice (MIC) Model

We applied multinomial logit to understand public concerns about transportation choices (e-commerce, ridesharing, telework, transit, others) under different socioeconomic and demographic factors. In this step, we integrated the county-level socioeconomic attributes (i.e., PCI, MTT) into the sampled dataset of 9,396 tweets to prepare a suitable dataset for econometric analysis. In Table 3-6, we have detailed descriptions of the variables used in this model development. A total of 11 variables are listed in this table. The last variable (Mobility Indicator) in this table is the response variable. The other variables are predictors. All the variables in this model are categorical except PCI, MTT, HS, BPL, and UE.

The results of the final model are presented in Table 3-7. For every unit change in the explanatory variable, the logit of each mobility indicator (in log-odds units) compared to the reference mobility indicator (others) varies by its corresponding parameter estimate, assuming all other factors remain constant. The values in parentheses denote the associated

factor's p-value or Pr(>|z|). Variables that were not significant (at the 90% significance level) for a given travel activity were not reported. The model's McFadden Pseudo R^2 value is 0.15012. The p-value of Hosmer-Lemeshow goodness of fit test (HL test) is 0.123 which suggests the model being good fit as p-value is greater than 0.05 [124].

Table 3-6. Descriptive Statistics of Key Variables.

Variable Description		**Mean or %**	**Minimum**	**Maximum**
User's Gender				
	1: Female 0: Male	40 %	0	1
User's Race				
	White	69.1 %	0	1
	Asian	10.9 %	0	1
	Black	4.9%	0	1
	Hispanic	15.1 %	0	1
County				
	New York	19.7%	0	1
	Kings	32.2%	0	1
	Queens	27.0%	0	1
	Bronx	14.9%	0	1
	Richmond	6.2%	0	1
Tweet Time				
	Morning (5 am- 12 pm)	9.9%	0	1
	Afternoon (12 pm-5 pm)	25.4%	0	1
	Evening (5 pm-9 pm)	23.7%	0	1
	Night (9 pm- 5 am)	41.0%	0	1
Weekday				
	1: Weekend 0: Weekday	30.5%	0	1
User's Socioeconomic attribute				
PCI: Per Capita Income ($)		$ 45,502	$ 2,758	$ 354,695
MTT: Mean Household Travel Time (mins)		43.53 mins	7 mins	100 mins
HS: High School Education Rate (%)		0%	48.56%	100%
BPL: Below Poverty Limit Rate (%)		0%	8.5%	100%
UE: Unemployment Rate (%)		0	42.43%	100%
Mobility Indicator				
	Ecommerce	12.5%	0	1
	Telework	19.2%	0	1
	Ridesharing	10.5%	0	1
	Transit	12.1%	0	1
	Others	45.7%	0	1

Table 3-7. MIC Model Results

Parameter	Ecommerce	Ride sharing	Telework	Transit
Intercept	-1.369 (1.30E-08)	-1.982 (1.37E-12)	-0.829 (6.78E-05)	-0.906 (0.0001328)
Female	0.185 (0.010)	-	0.181 (0.003)	0.129 (0.076)
PCI	0.061 (0.081)	-0.274 (0.0008)	0.150 (0.017)	-0.212 (0.018)
MTT	-0.020 (0.075)	-	-0.027 (0.062)	0.043 (0.055)
HS	0.121 (0.084)	-	0.174 (0.016)	-
BPL	-	-0.128 (0.051)	-	0.244 (0.0006)
Unemployed	-0.419 (0.013)	-	0.337 (0.027)	0.355 (0.075)
Asian	-	0.435 (0.042)	0.269 (0.097)	-
Evening (5pm-9pm)	0.056 (0.066)	-	-0.003 (0.098)	-
New York County	-	0.715 (0.011)	-	-0.085 (0.077)
Queens County	-	0.430 (0.045)	-	-
Bronx County	0.477 (0.008)	0.574 (0.006)	0.311 (0.043)	0.510 (0.007)
Weekend	-0.202 (0.009)	-	-0.234 (0.0004)	-0.247 (0.001)
Other Information:				
Number of Cases		9,396		
McFadden Pseudo R^2		0.15		
Hosmer-Lemeshow Test: p-value		0.123		

3.4.4.2 Impacts of Model Characteristics

According to the model, females are likelier to tweet about e-commerce and work from home. Females are also found to tweet more about transit in the emergence of the

pandemic. The lockdown of NYC introduces new normal where females are more involved in telework and online shopping. Moreover, being a transit-oriented city, covid hit NYC very severely and quickly. All these findings are consistent with the literature[125–129].

Higher-income people are more likely to tweet about online shopping and telework compared with ridesharing and transit, which is consistent with the finding in the literature, which showed that Higher-income people were less likely to shop in-store and work at the office, but lower-income people (who were more likely to take the subway or bus) were more likely[125, 130].

Travel time to work negatively impacted people's attitudes on Twitter to share their concerns regarding online shopping and telework. The people whose travel time to work is higher are more likely to share the covid-related concern using transit which might be the daily transportation system they depend on to go to work[129]. The people with higher travel time to work have to go to the workplace using a transit system, which is correlated with in-store shopping during the pandemic in the literature[125].

The more the people are educated, the more likely they will discuss e-commerce and telework, as previous studies have shown that education positively correlates with income[131]. This also indicates the increased involvement in online shopping and telework (working from home, online school) for higher educated people, which is consistent with the literature[132, 133].

According to the literature, transit services are the primary mode of transportation for poor people [133]. People are living below the poverty level likely to tweet more about transit

and less likely to tweet about ridesharing services. For this reason, people share concerns about covid contamination while using transit services.

Unemployed people are more likely to tweet about transit services (regarding the risk of covid contamination as transit as their primary mode of transportation) and telework (probably being furloughed/sagged people looking for online work or gaining skills through online studies). Moreover, they are less likely to share their concern regarding online shopping. These findings are consistent with the previous studies[134, 135].

Among different races, Asian are found to be more likely to tweet about the rise of sharing and telework, indicating their involvement in these travel choices, which goes with the finding in the literature[136, 137]. Moreover, people living in Bronx County are likely to tweet all the travel choices, whereas people living in New York County discuss more ridesharing and less about transit. These indicate the involvement of the people of Bronx to adopt all sorts of travel choices. On the other hand, New York County is the heart of the business of NYC; people living there are expected to be educated, employed, and affluent who tends to avoid transit, not take the risk of covid contamination, and prefer shared mobility services.

In the evening, people are likelier to tweet about online shopping and less likely to tweet about telework. Similarly, on weekends people are less likely to discuss e-commerce, telework, or transit on Twitter. These outcomes are consistent with the studies showing that people are more involved in e-commerce in the evening and less on weekends [138].

3.4.4.3 Sentiment Choice (SC) Model

In the emergence of the pandemic, people's activity-travel patterns changed significantly due to the increase in online shopping and working from home. Table 3-7 also corroborated this new trend as people stayed home to save themselves from covid exposure. Though Table 3-7 provides the propensity of people twitting activity regarding different travel choices, it does not provide whether people are expressing positive or negative opinions while twitting towards these different travel choices. To understand people's sentiments toward the increasing e-commerce and telework, sentiment analysis [78] was deployed to classify the tweets related to e-commerce and telework into positive and negative signals. A multinomial logit model was developed using the same explanatory variables in Table 3-6 and the updated response variable presented in Table 3-8 to investigate the people's sentiment toward telework and e-commerce under different socioeconomic and demographic factors.

The results of the sentiment choice model are presented in Table 3-9. The values in parentheses denote the associated factor's p-value or Pr(>|z|). The model's McFadden Pseudo R^2 value is 0.0913. Variables that were not significant (at the 90% significance level) for a given travel activity were not reported. The p-value of HL test is 0.0834 which suggests the model being good fit as p-value is greater than 0.05. Though this model results in low R^2, the results provide a better understanding of people's attitudes towards e-commerce and telework.

Table 3-8. Descriptive Statistics of Response Variables in Sentiment Choice Model.

Mobility Indicator			
Ecommerce (Positive)	6.7%	0	1
Ecommerce (Negative)	6.0%	0	1
Telework (Positive)	10.0%	0	1
Telework (Negative)	9.0%	0	1
Others	68.3%	0	1

Table 3-9. SC Model Results

Parameter	**Ecommerce (Negative)**	**Ecommerce (Positive)**	**Telework (Negative)**	**Telework (Positive)**
Intercept	-2.567 (3.042E-14)	-2.558 (1.11E-15)	-1.641 (8.828E-10)	-2.256 (8.882E-16)
Female	-0.179 (0.061)	-	-	0.192 (0.011)
PCI	-	-	-0.208 (0.094)	-
MTT	-	-	-0.019 (0.081)	-
HS	-	-	-0.290 (0.024)	-
UE	-0.511 (0.047)	-0.742 (0.001)	-	-
Night	-0.577 (0.0001)	-	-	-
Asian	-	-	-	0.395 (0.065)
Hispanic	--	-	-0.403 (0.036)	
Bronx	0.610 (0.022)	-	-	-
Queens	-	-0.451 (0.054)	-	-
Weekend	-	0.178 (0.074)	-0.258 (0.003)	-
Other Information:				
Number of Cases	9396			
McFadden Pseudo R^2	0.091			
Hosmer-Lemeshow Test: p-value	0.083			

3.4.4.4 Impacts of SC Model Characteristics

According to the model, females are likelier to tweet positively regarding telework and less likely to tweet negatively regarding e-commerce. Moreover, people are less likely to tweet pessimism about e-commerce at night, which indicates their active participation in online shopping at that time. This is consistent with the literature showing the increase in e-commerce activity [139] and the reluctance of people to join the in-person workforce after the lockdown was relaxed [140].

Unemployed people are more likely to tweet positively about telework and less likely to tweet about e-commerce, which supports the results of the Travel choice model. Asian and Hispanic people are more likely to tweet positively about telework, indicating their active participation in telecommuting. People living in the Bronx are more likely to tweet negatively about e-commerce, whereas people in Queens show the opposite behavior. This is an indication to improve e-commerce services in the Bronx. Lastly, during the weekend, people are showing a positive vibe toward e-commerce on Twitter and a less negative vibe regarding telework.

3.5 Study Limitations

This study paves a path for creating a representative sample of the population to capture public concern about any transportation-related issues. Despite such a contribution to the literature, this study is not beyond limitations. It is acknowledged that small events retweeted several times may affect the collected dataset. Another problem is that it is impossible to remove the sampling bias to create a perfect population sample due to issues with user privacy that limit the availability of structured personal information. Previous

research has shown that Twitter includes many bots that automatically send tweets to promote a product or a political campaign [63]. This study does not eliminate these tweets, but several methods for finding them have been proposed in some literature [64–66]. So, caution is required regarding social media data's biases. Another limitation is that the streaming API cannot collect all the tweets during the data collection period, but it collects sampled data distributed by Twitter through SRS. Recently, the tweeter has announced introducing a new API to collect 100% of data for academic and research purposes, which may make this online social media research more authentic and comprehensive[141]. Data was collected using academic track API on randomly selected five days (from April 12th, 2020, and April 16th, 2020) within the study period and from the same study region of the NYC bounding box of this research to compare the data collected by Academic track (AT) API and nonacademic streaming (NAS) API (the API used in this research). AT provided 186,729 geotagged tweets, whereas NAS provided 173,724 geotagged tweets during the same five-day period. In other words, AT provided only 7.5% more geotagged data than NAS. As the geotagged data collected by AT was not significantly high, the relevant tweet frequency of the geotagged data collected from both APIs would be similar. Future studies should collect the data using both APIs for comparative analysis to investigate the variances in the results.

Semi-supervised machine learning approach was used to classify tweets related to different indicators, which has some limitations of producing inconsistent results at different iterations and low accuracy. Cross-validation can increase the model accuracy and precision while working with small amount of labeled data dataset. As there is imbalance in the dataset, future studies should consider stratified cross-validation to increase the

classification model performance[142, 143]. Future studies should also consider supervised learning, which has received much attention [144–148] and given good results.

The socioeconomic and household information used in this study are average values collected from ACS at the census block group level, which may contain variations within a county census block group resulting in a lower R^2 value. Future studies should consider conducting a survey among Twitter users, incorporating other national databases (e.g., National Household Travel Survey [149]) and other statistical models (e.g., mixed logit model [150], structural equation model [151]) to improve the model. Moreover, future research should also consider different modeling frameworks incorporating other user attributes (e.g., age, race, income) to predict public sentiments towards different mobility indicators.

3.6 Conclusions and Discussions

Transportation researchers, in recent times, used SMPs extensively for problems related to travel demand forecasting, activity pattern modeling, transit service assessment, traffic incident, and disaster management, among others. Nevertheless, there is still much more to explore how such information can contribute to understanding public perception and attitude towards emerging transportation trends and mobility indicators (e.g., shared mobility). Representing public perception of any transportation issues requires creating a suitable population sample. This sampling is also required to develop a predictive model to grab the future transportation trend. As such, this study aims to introduce a new methodology that can classify tweets according to different travel choices and make a population sample of the tweeter dataset by removing biases through inferring

demographic user attributes compared with SSN and Census data. This study also performed an econometric analysis to show the impacts of different socioeconomic and demographic factors on transportation trends by combining Twitter data with a custom selection of socioeconomic variables from ACS[83].

The multinomial logit model was developed to perform the econometric analysis investigating the popularity of different travel choices in SMPs in the emergence of COVID-19. Personal characteristics such as race and gender showed significant impacts. Females are more likely to discuss online shopping activities, work from home, and transit service SMPs. Asian people are more likely to tweet about the rise of sharing and telework, which indicates the relatively higher involvement of Asian people in ridesharing services and telework compared with other races. Moreover, various generic characteristics such as per capita income, education, poverty, and unemployment had different impacts on travel choices. Overall, people with higher education, high income, and employment are likely to tweet about telework and online shopping. Some tweet characteristics such as time and location significantly impacted peoples' tweeting behavior regarding travel-related concerns. It was also found that people living in Bronx County are likely to tweet about all the travel choices, whereas people living in New York County discuss more ridesharing and less about transit. Another interesting finding showed that people are more likely to discuss e-commerce in the evening and less likely on weekends than other travel choices.

Moreover, the sentiment choice model provides in-depth insights into public satisfaction with online shopping and e-commerce. The model supports the public satisfaction with e-commerce and telework among the user base with higher income, higher education, and

higher travel time to work. Females are found to be significantly satisfied with these two trends as they can now spend more time at home with their family and manage the household chores simultaneously. Though this model results in low goodness of fit (R^2=0.0913), the results provide a better understanding of people's attitudes towards e-commerce and telework, which the different e-commerce-based companies can use to improve and spread their services. Transportation planners also can be helpful by using these outputs to understand the activity-travel pattern in the emergence of covid.

Our analysis shows that Twitter can be used as an effective source to capture travel-related concerns, draw population samples inferring socio-demographic attributes, and predict public perceptions of transportation choices at any geographical scale. Potential applications of the work may include: (i) capturing different travel-related signals from SMPs with high topical relevancy; (ii) incorporating exiting national databases to investigate and model community travel behavior at a different level of resolution in the emergency period (e.g., hurricane, pandemic) within a short period before conducting the survey; (iii) identifying and predicting spatial diversity of different travel-related needs and concerns through social media channels; (iv) developing new policies that would satisfy the diverse needs of emerging mobility at different locations; (v) design and implement more efficient strategies to improve and influence public interest and satisfaction towards different travel options.

CHAPTER 4
EXAMINING THE COMMUNICATION PATTERN OF TRANSPORTATION AGENCIES ON TWITTER AT DIFFERENT PHASES OF COVID-19

4.1 Introduction

SMPs enable the widespread transmission of information quickly and easily, resulting in a massive volume of digital material. Active users of SMPs such as Facebook, Twitter, Reddit, Instagram, and others outweigh frequent consumers of traditional news sources such as newspapers, television, and internet portals. SMPs bring news to people who would not have had access to it otherwise [152]. Social signals, derived from messages posted on social networking sites, track our everyday actions and generate massive volumes of data for traffic and transportation studies [153]. Transportation actors (Transportation and transit agencies) may utilize SMPs to provide traffic-related information to commuters [154, 155]. Agencies like the State Department of Transportation (DOT) are increasingly embracing social media channels to provide legit information to passengers [156, 157].

Twitter is among the most microblogging sites in the United States, with 199 million daily active users [72]. Most state DOTs have Twitter accounts that provide essential information like traffic congestion, wrecks, incidents, and planned road construction [158]. No relevant study examines transportation actors' communication patterns and interaction dynamics in SMPs. However, several prior studies sought to assess an account's success by looking at its follower count, number of retweets, number of mentions, the geographical spread of followers, and so on [159, 160]. Kocatepe et al. [159] evaluated the impact of Twitter accounts using a case study of FDOT District-3 Twitter accounts. Bregman and Watkins [161] advised transportation groups to get started with social media or improve current

initiatives. In other domains, such as public health, researchers have studied the Twitter account of various government agencies (public health) and stakeholders to access risk and crisis communications during the early stages of COVID-19 [162].

The potential to examine transportation actors' communication in online contexts during a somewhat long-term outbreak (e.g., COVID-19) has a unique and historically unprecedented prospect. So, our following research objective is to investigate the long-term communication pattern among transportation actors, as well as their interaction on SMPs in the emergence of COVID-19 in terms of communication consistency and coordination on Twitter, at various stages of the pandemic in the United States. The following research questions will be addressed in this paper:

- Did different transportation actors communicate information on SMPs consistently and adequately before and after the emergence of COVID-19?
- What kinds of information do the transportation actors discuss on SMPs before and after the emergence of COVID-19?
- Did the transportation actors interact with each other appropriately before and after the emergence of COVID-19?

This study analyzed data collected from the Twitter account of 395 different transportation actors from January 1st, 2018, to April 3rd, 2021. The sufficiency of the tweeting activity of different transportation actors was examined at different phases of the pandemic. A set of network measures were used to assess communication cooperation inside and among transportation actors. Finally, text mining and network analysis have been used during different phases of the pandemic to extract communication patterns and network

connectivity among transportation actors. The findings on communication adequacy, coherence, and coordination will guide transportation actors in communicating effectively within fragmented social networking settings.

4.2 Background and Related Work

Transportation organizations have been using social media channels to disseminate and gather information during regular crises [155]. Timely updates, citizen participation, marketing, research opportunities, and customer happiness, among other things, are enticing more service providers/agencies to adopt social media platforms as a networking tool [161]. Misra et al. [163] investigated the best practices for leveraging social media data and discovered that virtually every state department of transportation, many public transit agencies, and airports have a social media presence, showing a significant shift in how agencies engage with their consumers.

Recent research has looked at how to understand the interactions between user behaviors, network characteristics, and the attention received in social media, as well as how to identify variables for successful crisis communication in emergency circumstances [148, 149]. Researchers studied posted tweets during various disasters to extract valuable information about the disaster [139–141], user behavior on an individual level [48, 139, 142], examine the dependability of uploaded messages [143], and connect to well-known statistics [144], raise awareness [145], evaluate the damage [146], and even for earthquake detection [147]. Researchers also have focused on detecting influential social media users and explored their network features to understand the spread of targeted information in major disasters [27, 28.

Though SMPs are relatively new fields for research, researchers have used them in various cases in the transportation domain. There are several studies where social media have been used to forecast travel demand. Golder & Macy [37] and Yin et al. [38] investigated the capacity, scope, and application of various social media platforms to derive information on household daily travel. SMPs have been applied to understand mass human mobility patterns [41, 42] and to model user activity patterns [43, 44, 164]. Opinion mining has been performed in a few studies to show people's attitudes towards public transit, which can affect how stakeholders think about future transit investments [45].

In summary, SMPs have been utilized to retrieve relevant information in various sectors of the transportation domain. However, very few attempts have been made to explore the enormous potential to understand the dynamics of communication patterns of various transportation actors in the emergence of such crisis moments like COVID-19. Most of the case studies on social media risk communication are emerging across hazard types (e.g., hurricanes and infectious diseases) as attention on the use of SMPs in extreme disasters grows [165, 166]. None of the available studies have investigated the use of SMPs deeply to understand the dynamics of communication interaction among transportation actors and their long-term social media messages. So, this study presents a comprehensive approach to exploring how SMPs (Twitter) can be used to understand various transportation actors' perceptions and attitudes towards information dissemination and how they interact with each other in general and during this crisis moment of COVID-19 using text mining and network science principles.

4.3 Data and Methods

4.3.1 Data Description and Preprocessing

Twitter User Timeline API [167] has been used to collect tweets posted by the official account of 18 federal transportation agencies, 247 state-level transportation agencies and their different regional branches (i.e., Department of Transportation or DOT), 14 city-level transportation agencies (i.e., DOT) and, 116 transit agencies (64 local bus agencies, 25 light rail agencies, 7 heavy rail agencies and, 20 commuter rail agencies). Transit agencies have been selected from the corresponding ridership table in the American Public Transportation Association's (APTA) Quarterly "Public Transportation Ridership Reports" [168]. Table 4-1 listed our studied agencies, and their Twitter usernames.

Table 4-1. The Studied Agencies and Their Twitter Accounts.

Agency Type	Twitter Usernames
Federal (18)	DOTInspectorGen, DOTMARAD, FAANews, FAASafetyBrief, FMCSA, FTA_DOT, ITS_USDOT, NHTSAgov, PHMSA_DOT, Research_USDOT, SeawayUSDOT, SecretaryPete, TransportStats, USDOT, USDOT_, USDOTFHWA, USDOTFRA, VolpeUSDOT
State DOT (247)	All the state DOTs main and regional tweeter handle. An example: ArizonaDOT
City DOT (14)	BmoreCityDOT, CharlotteDOT, ChicagoDOT, DDOTDC, dobetterddot, LADOTofficial, ladottransit, NYC_DOTr, NYSThruway, PANYNJ, RideDDOT, seattledot, ThruwayTraffic, CDot
Local Bus (64)	Few examples: VIA_Transit, VTA, wmata, wmataRAC
Light Rail (25)	Few examples: MetrolinkVC, MLineTrolley, NewOrleansRTA
Heavy Rail (7)	metrolaalerts, NYCTSubway, PATHTrain, RidePATCO, SFBART, statenislandr, trenurbanopr
Commuter Rail (20)	ACE_train, Amtrak, Caltrain, CapitolCorridor, HLalerts, LIRR, MBTA_CR, Metra, MetroNorth, MusicCityStar, NMRXExpress, northstarlink, RideRail, RideSunRail, SLEalerts, smarttrain, southshoreline, Tri_Rail, TrinityMetro, VaRailXpress

Data were collected from January 1st, 2018, to April 3rd, 2021 (1189 days in total) to capture the communication pattern of the abovementioned agencies. Data has been analyzed using the python programming language. To investigate the communication behavior of different transportation actors at different pandemic stages (i.e., pre-pandemic, during the pandemic, states reopening) the study timeframe has been divided into three phases identifying two boundary dates between phase #1 and phase #2, and phase #2 and phase #3 respectively.

The national emergency was declared on March 13th, 2020, following the significant risk caused by COVID-19 to the public health and safety of the nation. However, different states and territories started issuing mandatory stay-at-home orders from March 1st, 2020, and by May 31st, 2020. Across states and territories, the government took actions differently in response to the outbreak. Considering the national announcement, March 12th, 2020, has been identified as the end of phase #1 in this study. On the other hand, states started reopening partially from April 26th, 2020. By August 28th, 2020, all the states have been reopened (except New Mexico) [169]. So, August 28th, 2020, has been considered as the beginning of phase #3. Total 868, 284 tweets have been collected from the studied agencies during these three phases. Figure 4-1 shows the collected data description at different phases and Figure 4-2 describes the framework for data collection, preparation and analysis.

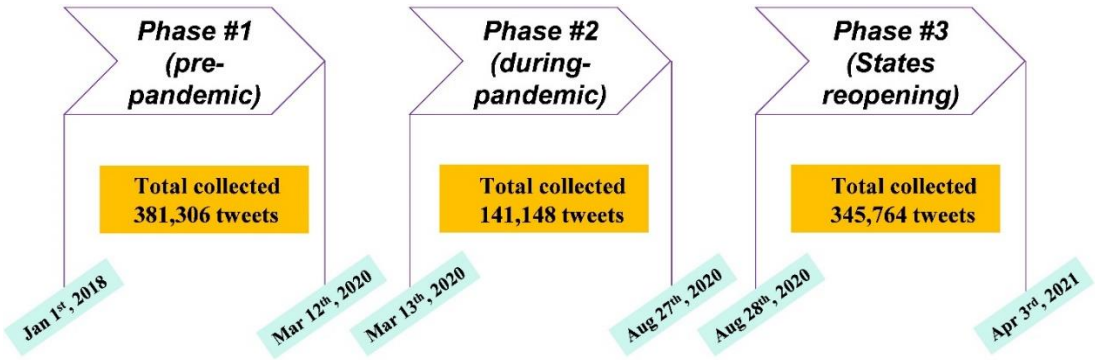

Figure 4-1: Phases of Data Collection.

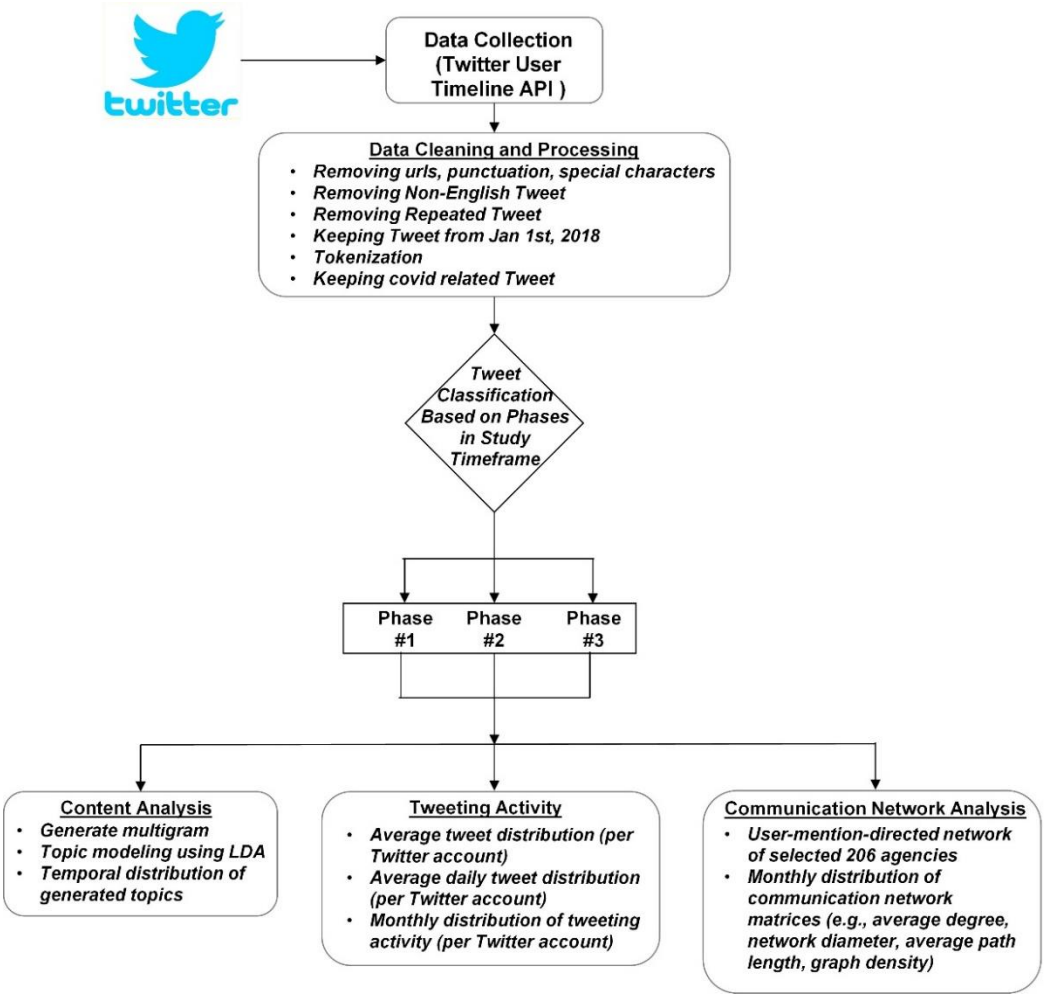

Figure 4-2: Framework for Data Collection, Preparation, and Analysis.

4.3.2 Generalized Topic Model

To identify the patterns of tweets posted by different types of agencies, giving information to the traveler, Latent Dirichlet Allocation (LDA) or topic modeling approach [60] was applied in this study. Topic modeling is a machine learning technique that analyzes text data automatically to classify cluster terms for a series of documents. LDA used a probabilistic latent semantic analysis model to recognize the patterns of the posted tweets. Though the topic model has been used popularly in machine learning, it was recently applied in transportation studies [44, 57, 58].

The probabilistic procedure for the document (tweet) generating is adopted in LDA, which starts with choosing a distribution ψ_k Over words in the vocabulary for each topic k ($k \epsilon$ *1, K*) . Here, ψ_k is selected from a Dirichlet distribution $Dirichlet_v(\beta)$. After that, another distribution θ_d over K topics is sampled from a different Dirichlet distribution $Dirichlet_k(\alpha)$ to generate a document d (a set of word w_d). Thus, a topic is assigned for each word in $\boldsymbol{w_d}$ followed by selecting each word w_{di} based on θ_d .

For LDA, initial sampling is done on a particular topic z_{di} ϵ *1, K* from a multinomial distribution $Multinomial_k(\theta_d)$ in generating each word w_{di} . Finally, the word w_{di} is selected from the multinomial distribution $Multinomial_v(\psi_{z_{di}})$**.** Figure 4-3 shows the graphical representation of LDA where Sun & Yin [59] summarized the processes. The inference of LDA models can be done by applying the variational expectation-maximization (VEM) algorithm [60] or through Gibbs sampling [61]. The posterior of document-topic distribution θ_d and topic-word distribution ψ can be efficiently inferred

by both methods which allow us to discover the latent thematic structure from a large collection of documents [59].

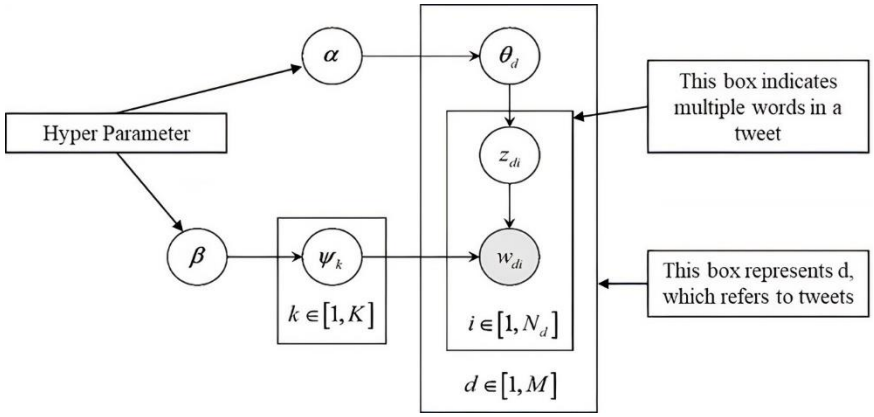

Figure 4-3: Graphical Model Representation of LDA [59].

4.3.3 N-gram Topic Model

LDA traditionally depends on the bag of words assumption, which often results in inscrutable lists of topical unigrams, single words inferred as representative of their topics. However, word order and phrases are often crucial to capturing the meaning of the text in many text mining tasks. In this study, we incorporated the n-gram model and LDA to generate more interpretable topics consisting of different n-grams (i.e., unigrams, bigrams, trigrams). The details on estimating model parameters can be found in Wang et al. [170] and Lindsey et al. [171]. The n-gram LDA (NGLDA) used in this study derived topics and topical phrases. This probabilistic model sampled each topic and generated words in their textual sequence to determine if a topic is a unigram or a bigram. Finally, the word was chosen randomly from a topic-specific unigram or bigram distribution. Thus, this model can model "social distancing" as a special meaning phrase in the "COVID-19 safety measures" topic, but not in the 'social science' topic.

4.3.4 Topic Variation over User-group and Time

It is possible to analyze how each inferred topic differs across agency categories and varies with time by using the posterior document-topic distribution θ_d and agency information (i.e., agency name, agency category) of each tweet or document d. This study aimed to capture the temporal topic variation and topic variation at each agency group level. Given this multi-level (i.e., time, agency group) nature of the analysis, the direct approach of Griffiths and Steyvers [172] was adopted. To investigate temporal trends over a different group of agencies, a time-insensitive topic model was created using NGLDA, then ranked and aggregated tweets based on their timestamps for each agency group level.

4.3.5 Aggregate Network Analysis of Communication Coordination

Dynamic network analysis has been performed to examine the interaction pattern among different agencies. During events (e.g., the West Virginia water crisis [173]), it has also proved its usefulness in risk and crisis communications. The directed communication networks among the agencies have been extracted to understand the information flow by analyzing the retweeting (RT) and mentioning (@) relationships using Gephi [174]. For example, the information goes from B to A if Agency A retweets (RT) a post from Agency B, and the information flows from A to B if Agency A mentions (@) Agency B in a tweet.

Different metrics (i.e., graph density, average degree, average path length, network diameter) of communication networks were calculated. The general frequency of retweeting and mentioning among studied agencies is represented by the average degree and graph density. A higher degree or graph density alludes to more coordination amongst agencies, implying that the public receives more consistent information. The diameter is

the most significant number of links connecting two agencies, whereas the average path length is the average linkages between all agency pairs. A better-linked communication network is indicated by shorter path lengths or diameters [175].

4.4 Results

4.4.1 Temporal Analysis of Tweeting Activity

The importance of an agency on the social web relies on the extent and speed of its capability of information dissemination. If an agency can reach a larger audience group in a short period while spreading information, it is considered a very crucial actor in communication media. Consistent tweeting activities (i.e., posting, sharing, retweeting) over time and social networks (i.e., follower count, list count) are the keys for transportation actors to make an agency a significant actor in the social web. In this study, overall tweeting activities (posting tweet+ sharing tweet+ retweeting) of different agency groups have been studied over time.

Figure 4-4(a) represents the average tweet distribution (per agency Twitter account) of different studied agency groups during the study period. Local Bus, Heavy Rail, and Light Rail are the top three active communication transportation actors in social media as each agency in these groups has a tweeting frequency of 2464, 2399, and 2342 accordingly over the study timeframe. Commuter Rail, City DOT, and State DOT were found to be moderately active on Twitter media, with an average tweeting count of 2252, 2166, and 2145. Federal agencies seemed to be the least active among all the agency groups, as each agency in this group tweeted just 1673 times on average over the study period.

Figure 4-4(b) shows the average tweet distribution of each group for a single agency Twitter account for a single day over three study phases. It had been found that all the groups (except Federal) showed gradually increased tweeting activity over the phases. Agencies of the Federal group seemed to have unvarying average daily tweeting activity of 1.4, 1.7, and 1.9 in Phase #1, #2, and #3 accordingly. All the remaining groups showed consistent daily twitting activity (per account), varying from 1.4 to 1.7 in phase #1. In Phase #2 and #3, increased daily tweeting activity (per account) was experienced for these groups varying from 2.2 to 6.9 and from 3.4 to 8.4 accordingly. Though Local Bus (1.7 counts) was the top active agency group on Twitter in Phase #1, Heavy Rail kept this top position for the following two phases tweet counts of 6.9 and 8.4. The inconsistent tweeting behavior of the heavy rail agency group was noticeable by the increased tweeting activity of more than 300% in Phase #2 than Phase #1.

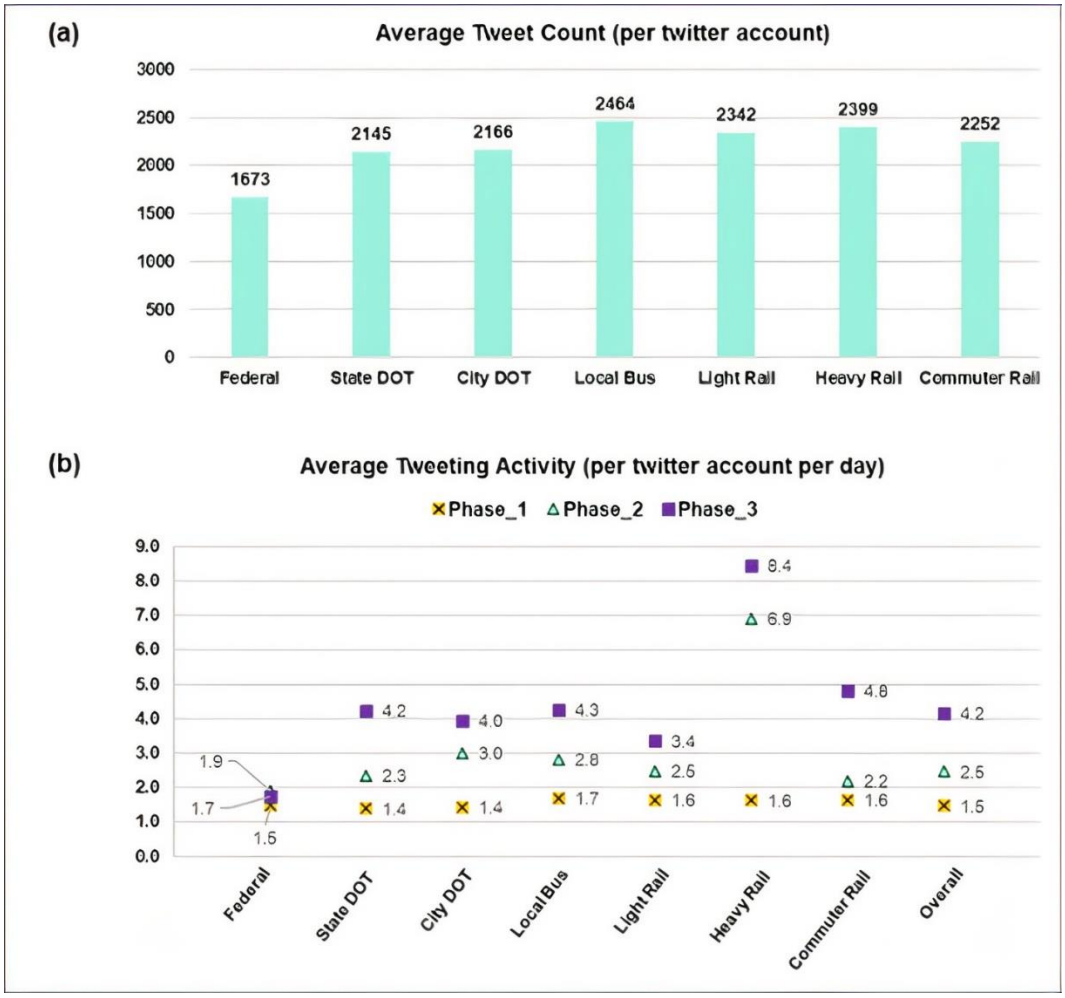

Figure 4-4: Activity by Different Groups of Agencies Twitter Accounts (a) Average Tweet Distribution (Per Twitter Account) for Different Groups of Agencies, (b) Average Daily Tweet Distribution (Per Twitter Account) for Different Groups of Agencies.

To understand this erratic behavior, the resolution of the analysis has been increased to a month over the study timeframe. Figure 4-5(a) shows the monthly distribution of tweeting activity (per Twitter account) for all agency groups. From May 2019 to Dec 2019 (8 months), agencies Heavy rail groups showed no tweeting activity. Before that period, this group was found to be consistent with other groups in tweeting behavior. From the last three months of Phase #1 to the end of the study time frame, this group showed gradually increased tweeting activity than other groups. But this group showed the highest tweeting

activity (tweet frequency of around 700) during Mar 2021 (Phase #3). Figure 4-5(b) is the same as Figure 4-5(a), excluding the Heavy Rail group, to get a better visualization of the remaining groups. According to Figure 4-5(b), the Federal group maintains consistent monthly tweeting activity with around 50 tweets per month over the three phases. State DOT and Commuter Rail group showed similar tweeting behavior (tweeting frequency of around 75 per month) in Phase #1 and #2. However, they increased their tweeting activity in the last few months of Phase #3 as their highest tweeting frequency was around recorded as around 200 and 275, accordingly. Though Local bus and Light rail have consistent tweeting activity of 75-100 per month in most of the duration of period #1, they increased their tweeting activity in the last few months of Phase #1. However, their tweeting activity has decreased in phase #2. Light rail did not increase their tweeting activity in phase #3 significantly. Local bus showed high tweeting activity with around 200 tweets per month in the last few months of Phase #3. On the other hand, City DOT's monthly tweet distribution increased uniformly over the three phases.

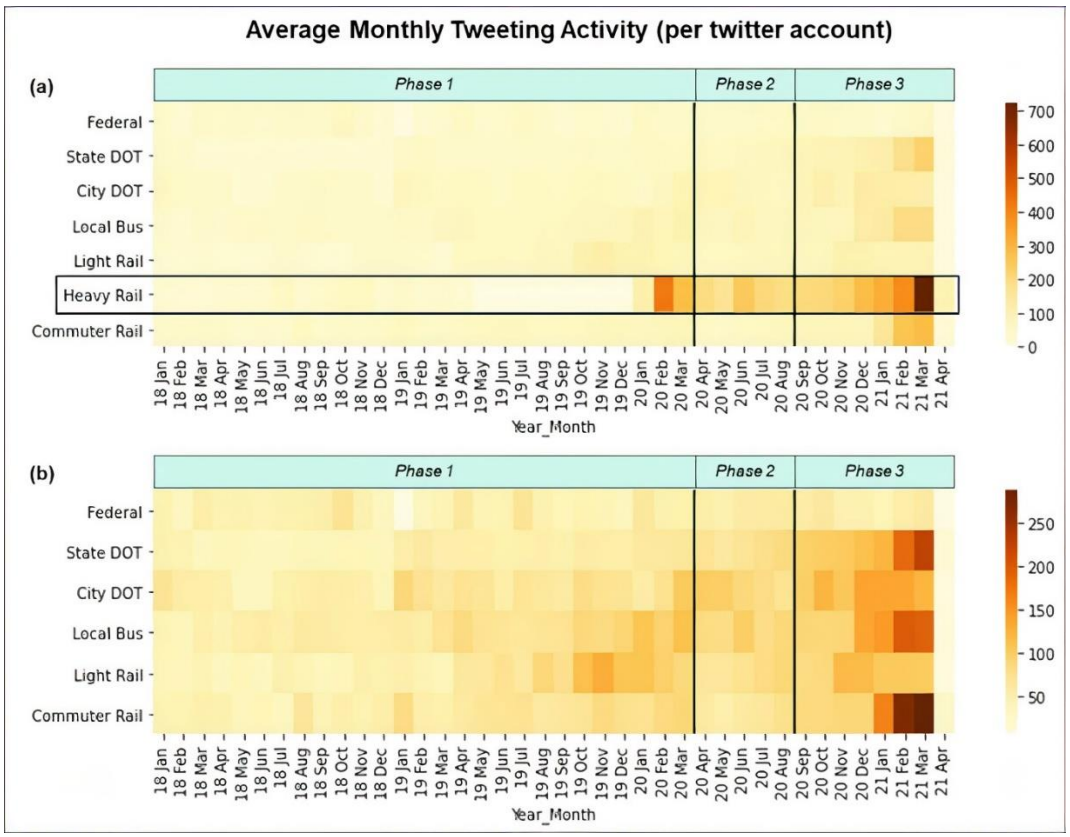

Figure 4-5: The Monthly Distribution of Tweeting Activity (Per Twitter Account) for Different Agencies: (a) Including Heavy Rail Group, (b) Excluding Heavy Rail Group.

4.4.2 Content Analysis over Temporal Platform

NGLDA model was applied to investigate how different combinations of n-grams in the data may constitute social interaction topics in each agency group. NGLDA model not only discovered the topics in each agency group but also classified each tweet by its dominant topic. This tweet classification technique allowed us to analyze the topics in a temporal platform over different period. These kinds of study approaches were expected to help to delve deeper into the understanding of the dynamics of social media communication patterns of different agency groups. The findings of this analysis are listed below accordingly for each agency group.

Federal Agency Group:

Figure 4-6(a) represents the distribution of tentative 6 topics from the posted tweets by the Federal agency group. The monthly probabilistic topic distribution is presented in Figure 4-6(b). The following information can be drawn from those two figures.

- Safety programs, project management, research and development, vehicle technology, work zone/ pedestrian safety, and traffic information are found to be the most frequent topics.
- Though the safety program was the top discussed topic in all the phases.
- Project management, research and development, and vehicle technology were similarly popular during phase#1. However, these topics were found to be rarely discussed during phase #2 and replaced by work zone/ pedestrian safety and traffic information.
- In phase #3, safety program, research and development, vehicle technology, and traffic information were the top discussed topics.

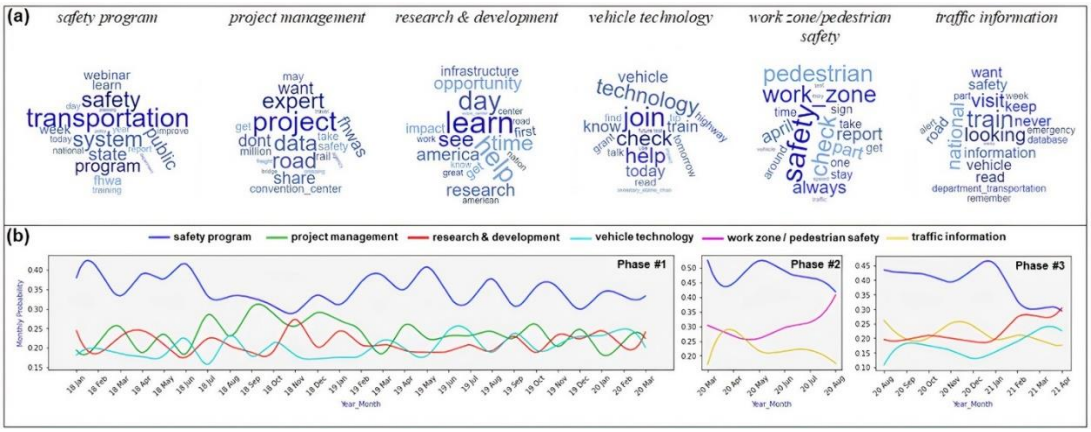

Figure 4-6: Topic Distribution for Federal Agency Group. (a) Tentative Generated Topics, (b) Probabilistic Topic Distribution over Months.

State DOT Agency Group:

Figure 4-7(a) represents the distribution of tentative 6 topics from the posted tweets by the State DOT agency group. The monthly probabilistic topic distribution is presented in Figure 4-7(b). The following information can be drawn from those two figures.

- More active in spreading information about winter travel, safety recommendation, lane closure, accident/ crash, project construction, and travel information.
- In phase #1, winter travel followed a periodic pattern as it was discussed more during Dec to Feb, remained less discussed during the middle of the year.
- On the other hand, lane closure and project construct also followed the period pattern but in the opposite way to winter travel in phase #1. This interpreted the correlation between these two topics, and it also can be inferred that during the winter period few amounts of road maintenance and construction have been performed than another month of the year in phase #1.
- Safety recommendations and lane closure were the top two discussed topics in phase #2.
- In phase #3, winter travel and lane closure were the top two discussed topics, where accident/ crash and travel information were discussed rarely.

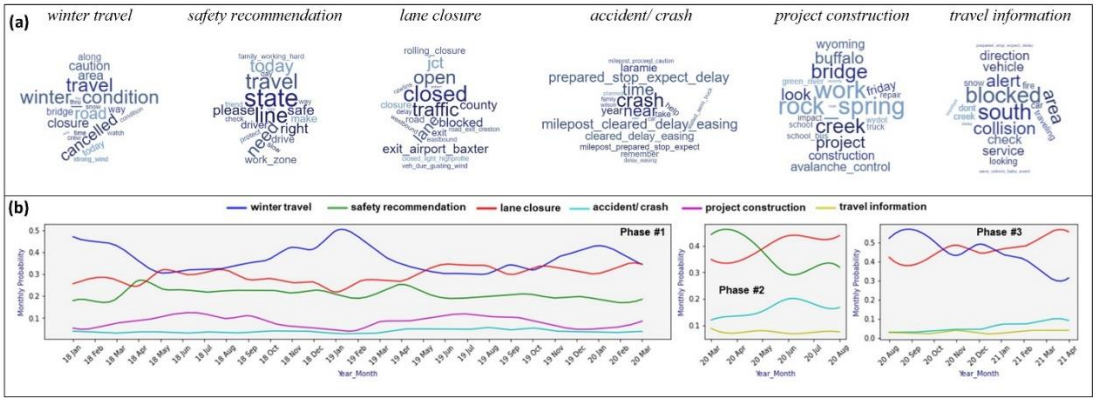

Figure 4-7: Topic Distribution for State DOT Agency Group. (a) Tentative Generated Topics, (b) Probabilistic Topic Distribution over Months.

City DOT Agency Group:

Figure 4-8(a) represents the distribution of tentative 5 topics from the posted tweets by the City DOT agency group. The monthly probabilistic topic distribution is presented in Figure 4-8(b). The following information can be drawn from those two figures.

- Active in spreading information about crash and lane closure, street parking, safety measure, recreation, and travel information.
- In phase #1 and #3, crash and lane closure and travel information were found to be discussed.
- Safety measures seemed to be the hot topic at the beginning of Phase #2 for more likely to spread information regarding COVID.

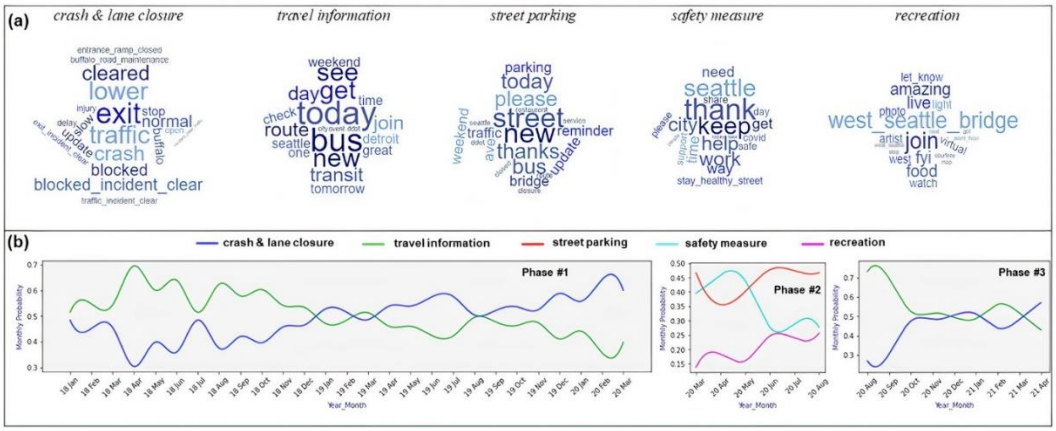

Figure 4-8: Topic Distribution for City DOT Agency Group. (a) Tentative Generated Topics, (b) Probabilistic Topic Distribution over Months.

Local Bus Agency Group:

Figure 4-9(a) represents the distribution of tentative 7 topics from the posted tweets by the Local bus agency group. The monthly probabilistic topic distribution is presented in Figure 4-9(b). The following information can be drawn from those two figures.

- More active in spreading information about station service, schedule announcement, customer, detour, greetings & information, COVID safety measures, and transit service
- In phase #1, station service, schedule announcement, customer, detour, and greetings & information were discussed where station service was the hot topic.
- On the other hand, station service, greetings & information, and COVID safety measures have been discussed in phase #2.
- Station service, greetings & information, COVID safety measures, and transit service were the focused topic in phase #3 where the transit station was the hot topic.

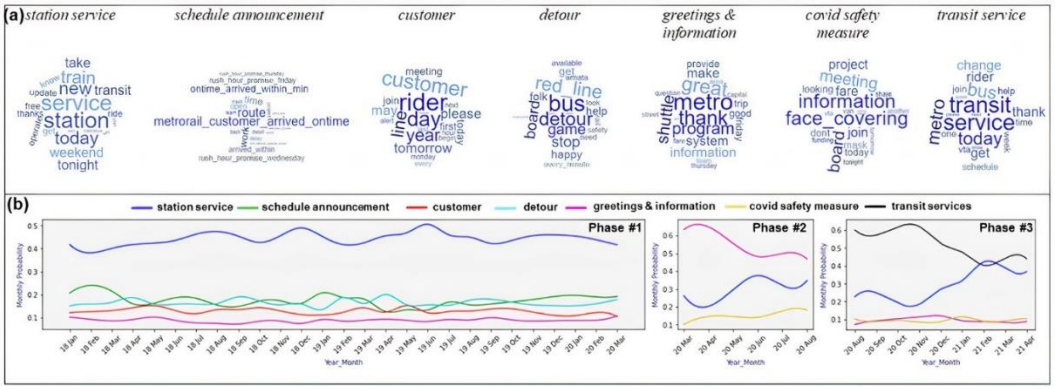

Figure 4-9: Topic Distribution for Local Bus Agency Group. (a) Tentative Generated Topics, (b) Probabilistic Topic Distribution over Months.

Light Rail Agency Group:

Figure 4-10(a) represents the distribution of tentative 5 topics from the posted tweets by the Light rail agency group. The monthly probabilistic topic distribution is presented in Figure 4-10(b). The following information can be drawn from those two figures.

- More active in spreading information about schedule & service, station service, delay & reroute, line blocking, and safety measures.
- In phase #1, schedule & service, station service, and delay & reroute were discussed where schedule & service was the hot topic at the beginning of this phase.
- On the other hand, schedule & service, delay & reroute, and line blocking have been discussed in phase #2. In this phase delay & reroute was the top topic.
- Schedule & service, delay & reroute, and safety measures (COVID related safety issues) were the focused topic in phase #3 where delay & reroute was also the hot topic.

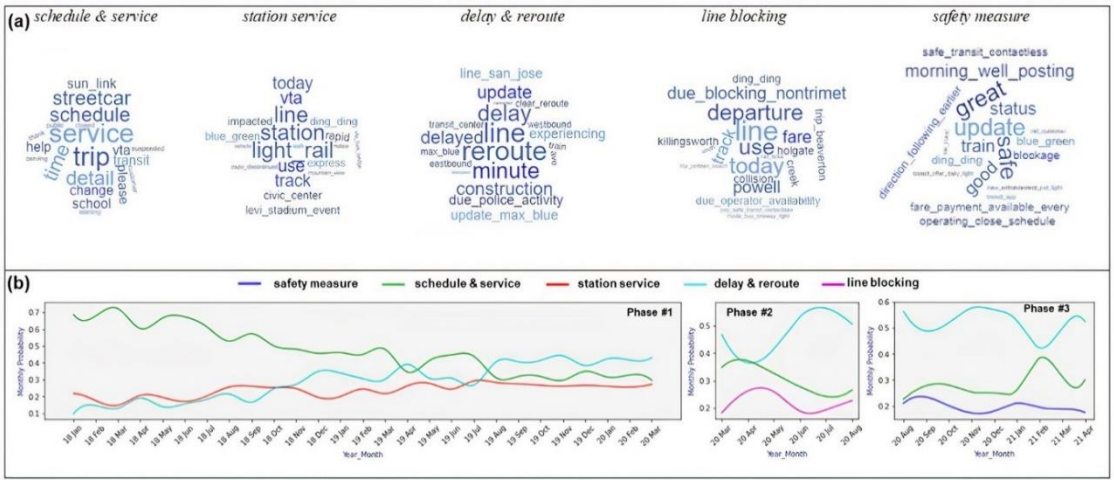

Figure 4-10: Topic Distribution for Light Rail Agency Group. (a) Tentative Generated Topics, (b) Probabilistic Topic Distribution over Months.

Heavy Rail Agency Group:

Figure 4-11(a) represents the distribution of tentative 5 topics from the posted tweets by the Heavy rail agency group. The monthly probabilistic topic distribution is presented in Figure 4-11(b). The following information can be drawn from those two figures.

- More active in spreading information about track work & maintenance, schedule update, passenger service, special update, and safety measures.
- In phase #1, schedule update, passenger service, and track & work maintenance were discussed topics until May 2019. During that scheduled update was a hot topic where the remaining two topics were rarely discussed.
- Heavy Rail did not show any tweeting activity for the next 6 months. From Dec 2019, this group resumes its tweeting activity discussing almost equally of the three topics.
- On the other hand, schedule updates, COVID safety measures, and track & work maintenance have been discussed in phase #2.

- Lastly, track work & maintenance, schedule update, passenger service, and safety measures were focused topics in phase #3.

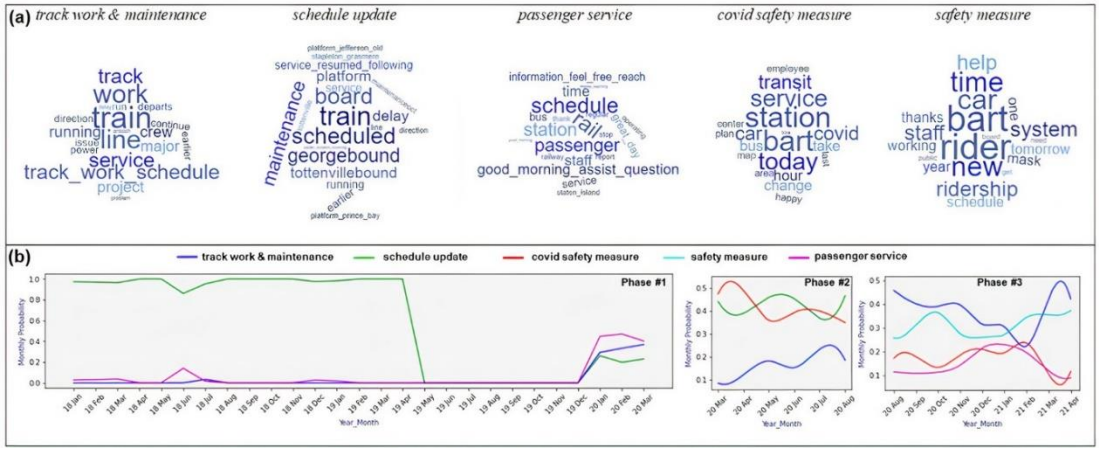

Figure 4-11: Topic Distribution for Heavy Rail Agency Group. (a) Tentative Generated Topics, (b) Probabilistic Topic Distribution over Months.

Commuter Rail Agency Group

Figure 4-12(a) represents the distribution of tentative 5 topics from the posted tweets by the Commuter rail agency group. The monthly probabilistic topic distribution is presented in Figure 4-12(b). The following information can be drawn from those two figures.

- Active in spreading station service, schedule update, delay update, online forum, and safety measures.
- In phase #1, station service, delay update, and online forum were discussed topics where station service was the hot topic.
- On the other hand, delay update, schedule update, and safety measure (e.g., COVID related safety measures) have been discussed in phase #2 and #3.

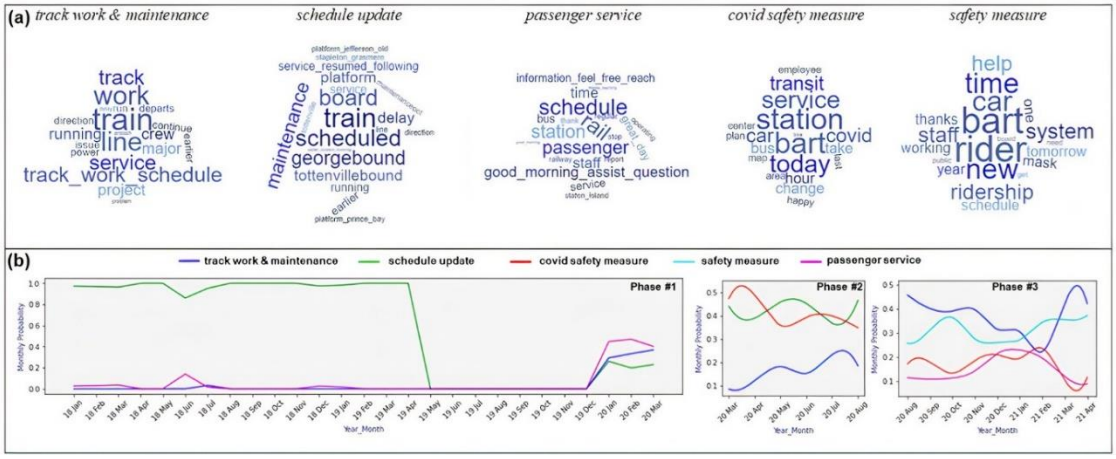

Figure 4-12: Topic Distribution for Commuter Rail Agency Group. (a) Tentative Generated Topics, (b) Probabilistic Topic Distribution over Months.

4.4.3 Dynamic Communication Networks Analysis

Dynamic user-mention network analysis was employed to investigate how information flows among different transportation actors during the study period. We also incorporated 14 health agencies (e.g., CDC, FEMA) in our analysis to see the interaction pattern between them and different transportation actors. In the network analysis, we excluded the regional state DOTs and lowly online active local bus agencies. The network analysis was performed on a total of 206 nodes (agencies). The aggregated communication network was constructed during the study period (Figure 4-13). Different colors represent distinct groups of communication actors. The size of nodes is determined by the degree of each node (i.e., the level of the agency connects with other agencies). The curve linking edges represents the information flows among the communication actors. The color of the edge is determined by the parent communication actor from which the information is flowing. The average degree of the aggregated network is 8.704, suggesting an overall connected communication network among actors in terms of mentioning and retweeting. The low graph density (0.042) with respect to the average degree suggests that the aggregate

network is poorly coordinated and connected. The network diameter (8) demonstrates the shortest distance between the most distant nodes in the network. On average, a communication actor's message needs to travel two links to reach another actor, indicated by the average path length of 2.694. The FTA's Twitter account has the highest degree, followed by the USDOT, FHWA, NHTSA, and CDC. For state agencies, VA DOT, ME DOT, and NC DOT have higher degrees than others. All the city DOTs were found to have a degree of similar level. In the case of local bus agencies, the MTA, MBTA, COTABus, and RidePSTA showed higher connectivity than the remaining agencies. Metrolink, SFBART, Amtrak were the highest interactive agencies in light rail, heavy rail, and commuter rail groups, respectively. Among the different federal health agencies, CDCgov, fema, and who showed higher interaction with different transportation actors.

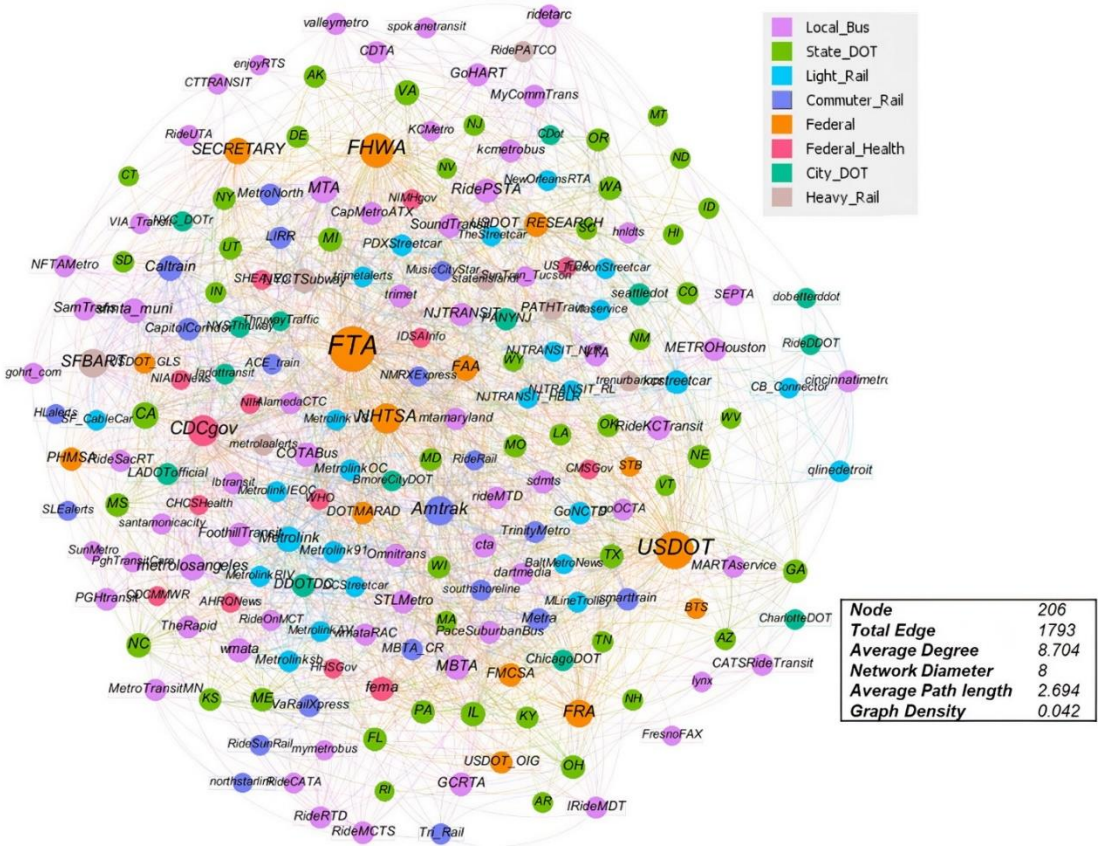

Figure 4-13: User-Mention-Directed Network of Selected Agencies.

The dynamics of the monthly communication networks among actors over the 40 consecutive months during the study period were further examined (Figure 4-14). Average degree, network diameter, and average path length showed a similar trend over the study period. At the beginning of phase #1, all the three matrices show higher values followed by gradual ups and downs until the beginning of phase #3. This suggests that the agencies' connectivity did not significantly over this time. At the start of phase #2, the average degree increased, whereas the average path length decreased. This is an indication that the agencies were coordinating with each other closely during this time. The higher values of the three matrices at the end of phase #3 present a poor coordinated network though the connectivity increased (higher average degree). The graph density almost remained stable

throughout the study period also proves the existence of a poorly connected network. The increase in graph density at the end of phase #3 suggests higher connectivity among the agencies.

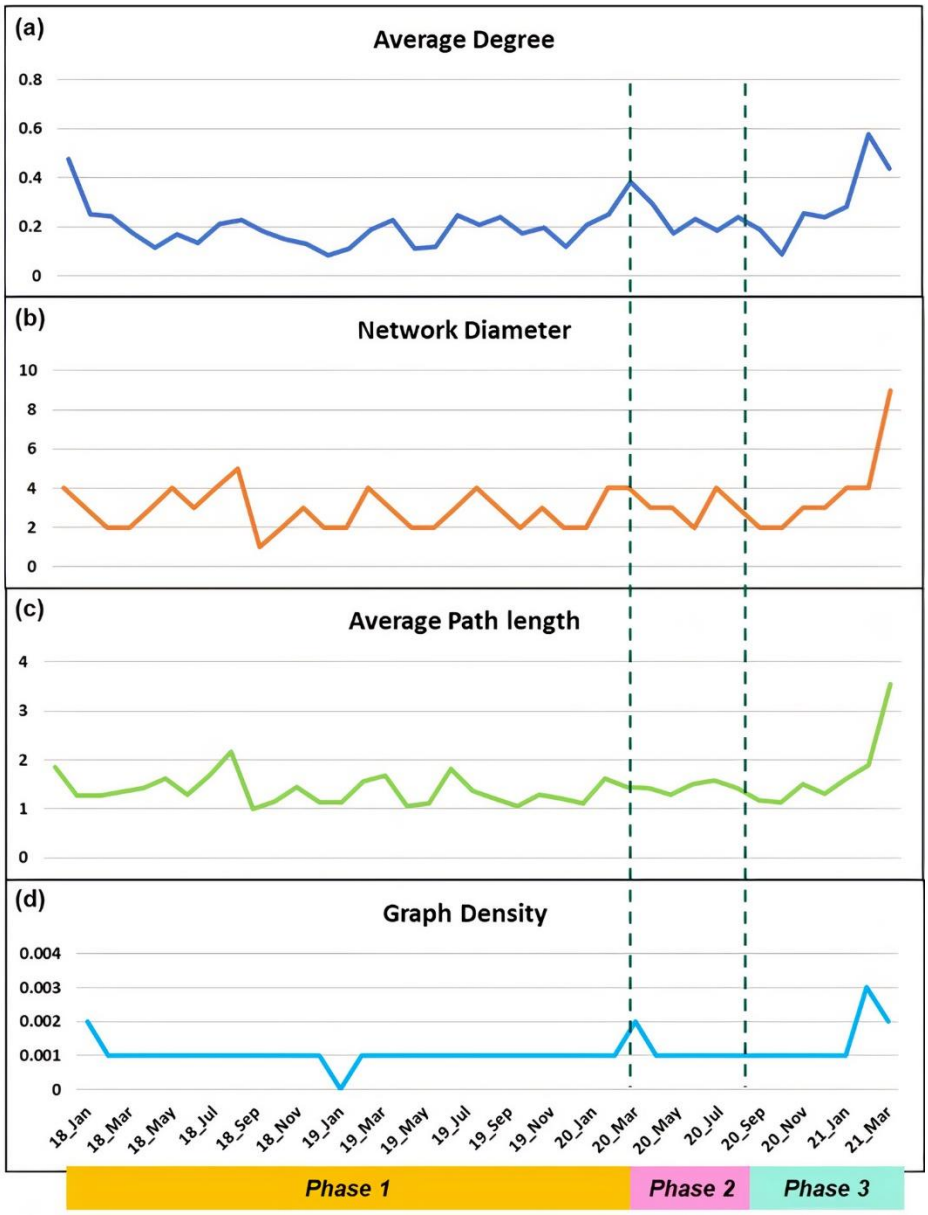

Figure 4-14: Monthly Changes in Communication Network Matrices.

4.5 Conclusions and Discussions

The rapid advancement of communication devices and programs (i.e., cellphones, SMPs) has brought the new uprising called social revolution. Nowadays, people are more likely to be active in virtual life than in real life. Transportation agencies should explore this platform more and invest proper resources to reach people of different sectors and extend public engagement. Most transportation agencies have official Twitter accounts nowadays. Also, due to COVID, the public lifestyle has changed a lot, and people started adopting the new normal. Many agencies tweeted regarding the COVID safety measures. However, there is still much more to explore how transportation agencies can contribute to disseminating information and influence public perception and attitude. As such, this study aims to investigate the long-term communication patterns among transportation and actors, as well as their interaction on social media platforms, in the emergence of COVID-19.

Temporal analysis of tweeting activity of the different groups of agencies showed that Local Bus, Heavy Rail, and Light Rail are the top three active transportation actors in social media over the study timeframe. Federal agencies seemed to be the least active among all the agency groups during the study period. Moreover, Agencies of the Federal group seemed to have unvarying average daily tweeting activity (per account) at all three phases of the study. On the other hand, all the remaining groups showed a consistent increase in daily twitting activity at different phases. Regarding average daily tweeting activity (per account), the local bus agency group was top in phase #1, and the heavy rail agency group was at the top in phases #2 and #3. However, the inconsistent tweeting behavior of the heavy rail agency group was noticeable by the increased tweeting activity of more than

300% in phase #2 than in phase #1. This unusual tweeting activity of the heavy rail agency group can be explained as no agency from this group tweeted at all for over eight months (May 2019-Dec 2019) of phase #1. These findings suggest that the agencies of the federal group need to take more committed and targeted measures to be more active in online platforms through tweeting and retweeting, conveying information to the public. Similarly, the agencies of the Heavy rail group need to be more consistent in their online platforms to keep people's trust in them.

Content analysis identified different discussed topics among different groups of agencies as well as the probabilistic temporal distribution of topics over the three phases of this pandemic. It was found that the Federal agency group mainly discussed safety programs, work zone safety, project management, etc. Safety program was the top discussed topic during all three phases of the study. This group of agencies was less active on Twitter and did not discuss that much regarding COVID-related safety information. The state DOT agency group discussed safety issues and project-related information similarly to the federal agencies. However, they and the city DOT agency group also expressed concern about lane closure, accidents, and travel information. COVID-related topics were not among the top discussed topics by state and city DOTs. The abovementioned three groups of agencies should share more information regarding taking safety precautions for the public while traveling during the pandemic.

On the other hand, the transit agencies (local bus, light rail, heavy rail, and commuter rail) mainly discussed station services, passenger services, schedule information, route or line blockage, delay, and greetings over the three phases of the study. They also showed

concern about sharing COVID safety-related information during phase #2 and #3. As the transit agencies directly serve the people, the COVID-related concerns they shared in phases #2 and #3 are not enough. They should be more constructive and vocal in disseminating information regarding safety measures during COVID (mask, social distance, vaccination, etc.)

A better-coordinated network amongst agencies is essential to disseminate more consistent information among general people. A higher degree or graph density and shorter path lengths or diameters allude better coordination and a better-linked communication network amongst agencies. The aggregate user-mention network analysis results suggest poor coordination and interaction amongst different transportation agencies. The low graph density (0.042) with respect to the average degree (8.704) indicates that the studied transportation and federal health agencies is not well coordinated. Dynamic user-mention network analysis suggests that during the study period, even during the pandemic (phase #2), the transportation actors interacted with the health agencies at a very low intensity. This is evident as just three health agencies (CDCgov, FEMA, and WHO) showed higher interaction with different transportation actors among the 14 studied health agencies. Among different agencies, FTA, USDOT, NHTSA, FHWA, and CDCgov showed higher connectivity which recommends increasing the Twitter activity for the remaining agencies. Overall, the network analysis results suggest that the transportation actors should take proper coordinated steps to interact with each other more on Twitter to speed up the information flow, which will eventually serve the public. They should also keep close ties with the different health agencies to share accurate information and recommendation for the public to travel safely in such a time of the pandemic. Transportation agencies should

be more active and interact among themselves to improve the communication network coordination and connection, which will assure consistent information sharing among the people.

This study's results showed that there seems to be significant potential for using social media data to understand the communication pattern of different transportation actors in the emergence of COVID-19. However, there are a few drawbacks to the study that can be addressed in future research. First, because this study focuses on the communication of transportation players, only tweets made by their official accounts were examined. Future empirical research might also look at the public's behavioral responses to inadequate, irregular, and incoherent communication throughout the pandemic's life cycle. As this is exploratory research on Twitter-based communication, it is not suitable to make conclusions regarding the behavioral impacts of specific social media (e.g., Twitter) users' platforms or message type dissemination. Second, the research analyzes all the tweets posted by transportation actors. Future research can be focused on just the communication for COVID-19 to identify specific strategies for good, congruent, and effective risk communication for different transportation actors. Lastly, Twitter data was only used since Twitter is one of the top micro-blogging sites in the United States and has accessible APIs. In the future, as data from additional social media platforms become accessible, a cross-platform examination may yield more detailed results.

For the first time, this study introduced a social media data-driven framework to examine the communication pattern of different transportation actors. This study provides transportation actors with an updated understanding of their role in disseminating

information on social media and in interacting with the health agencies to contribute to raising awareness among passengers during such a pandemic period. The research findings of this study will also lead to fundamental knowledge of social media communication in large-scale hazards (e.g., pandemics) by bridging public health and pandemic emergency management. The outcomes of this study will potentially improve existing communication plans, critical information dissemination efficacy, and coordination of different transportation actors in general and during unprecedented health crises in the fragmented communication world.

CHAPTER 5
CONCLUSIONS

Because of the epidemic's severity, practically all nations affected by COVID-19 have taken many safeguards. It is thought that travel patterns and mode preferences alter significantly during pandemic scenarios compared to typical (pre-pandemic) situations, owing to these limitations imposed by the government and individual fear of infection. COVID-19 had an impact not just on health but also on the economy, the environment, and social life, which are the different factors that influence the dynamics of mobility behavior. This epidemic, demographic changes, behavioral transformations, and technological developments all interacted to impact passenger travel behaviors, which are still very ambiguous. The unequal distribution of COVID-19 among diverse demographic groups was discovered to promote variation in mobility patterns. In such cases, the traditional travel surveys would not be effective in understanding the complex dynamics of people's mobility patterns in the emergence of COVID-19 due to having some limitations, such as variabilities across countries in data collection methods and data availability, lack of real-time engagement of the respondents, expansive and time-consuming as trend analysis requires periodic data collection making. SMPs can be used as a viable alternative in understanding the complex dynamics of people's mobility patterns in the emergence of COVID-19, overcoming the drawbacks of surveys and other sources of travel data as it serves the need for a more unified, less privacy-invading, and simply accessible method to understand the dynamics of travel patterns fully. In this dissertation, three studies are presented developing methods to understand the dynamics of the transportation trends and

indicators in the spatio-temporal platform at the emergence of COVID -19. The objectives of this dissertation are as follows:

The first objective of this dissertation is to introduce a novel approach to understanding public opinion and identifying emerging transportation trends based on social media interactions with enriched space and time information using sentiment and topic analysis. The second objective of this dissertation is to develop a methodology to model community-based travel behavior and access public attitudes towards different mobility trends under different socio-economic and demographic factors in the emergence of COVID-19 on Twitter. The final objective of this dissertation is to investigate the long-term communication pattern among transportation actors, as well as their interaction on social media platforms in the emergence of COVID-19 in terms of communication consistency and coordination on Twitter at various stages of the pandemic.

5.1 Summary of Major Results

This dissertation demonstrated the significant potential of using social media data in transportation planning, identifying user needs, requirements, and concerns, which are critical aspects for satisfying the general public's transportation needs, particularly during future pandemic situations. The summary of significant results of this dissertation is listed below:

- In the second chapter of this dissertation, we presented a method to capture emerging transportation trends and indicators (e.g., shared mobility, e-commerce, telecommuting, etc.) using social media data. We developed a model that can capture spatio-temporal differences in social media user interactions and concerns about such

trends, as well as topics of discussions formed through such interactions. Our findings show that the proposed data-driven approach using sentiment analysis and topic modeling can identify the spatial and temporal variations of different emerging mobility trends.

- In chapter three, we use Twitter data, ACS data, SSN data, and lastly, census data to develop a method to model community-based travel behavior using social media data in the emergence of COVID-19. Using tweet classification and users' demographics, we have attempted to overcome the sampling biases of the Twitter data. Lastly, we introduce the multinomial logit model to perform an econometric analysis under different socio-economic and demographic factors. Results showed that it is possible to model community-based travel behavior and analyze the impact of different factors (e.g., race, gender, etc.) on individual travel-related attitudes.
- In chapter 4, we introduced an approach that can capture the communication pattern in the long term in the emergence of COVID-19 among different transportation agencies (e.g., federal, heavy rail, bus, etc.). We have also identified the dynamics of the interaction among different transportation agencies to access their communication consistency and coordination on Twitter at various stages of the pandemic. Our study indicates that the transit agencies (e.g., bus, rail, etc.) became more consistent and coherent than federal or state-level agencies (e.g., FDOT, USDOT, etc.) to raise public awareness in the emergence of COVID-19.

5.2 Potential Applications of Research Findings

This dissertation makes significant methodological contributions by introducing different novel approaches to using large-scale social media signals reducing sampling biases to capture and model public attitude towards different mobility indicators. Moreover, this dissertation makes significant theoretical contributions by investing the communication pattern among transportation actors and their interaction on SMPs for the first time. This dissertation also recommends that agencies increase social media activity and interaction on SMPs to provide consistent public information. The social media data-driven framework presented in this study would allow real-time monitoring of mobility indicators by agencies, researchers, and professionals. Potential applications of the work may include:

- Capturing different travel-related signals from SMPs with high topical relevancy.
- Identifying and predicting spatial diversity of different travel-related needs and concerns through social media channels.
- Incorporating exiting national databases to investigate and model community travel behavior at a different level of resolution in the emergency period (e.g., hurricane, pandemic) within a short period before conducting the survey.
- Leveraging SMPs to promote user interests on emerging mobility behaviors.
- Developing new policies that would satisfy the diverse needs of emerging mobility at different locations.
- Designing and implementing more efficient strategies to improve and influence public interest and satisfaction towards different mobility behaviors.

- Improving existing communication plans, critical information dissemination efficacy, and coordination of different transportation agencies in SMPs in general and during emergency period (e.g., hurricane, pandemic).

5.3 Limitations and Future Research Directions

This dissertation shows the significant potential for using social media data to develop models for identifying transportation trends and long-term planning purposes. However, this dissertation is not beyond limitations. Some drawbacks related to social media data, in general, include differences in penetration rate across various places, unequal distribution across different age groups, and so on.

Previous research has shown that Twitter includes many bots that automatically send tweets to promote a product or a political campaign [63]. This study does not eliminate these tweets, but several methods for finding them have been proposed in some literature [64–66]. So, caution is required regarding social media data's biases. Another limitation of this dissertation is that data collection uses streaming API, which cannot collect all the tweets during the data collection period, but it collects sampled data distributed by Twitter through SRS. Recently, the tweeter has announced introducing a new API to collect 100% of data for academic and research purposes. However, because the relevant tweet frequency of the data obtained by the academic track was not considerably higher, the relevant tweet frequency of the data collected by both APIs would be identical. However, as in this dissertation, only geotagged tweets were considered; future research should also incorporate non-geotagged data along with geotagged data using both APIs for comparative analysis to evaluate differences in results.

Semi-supervised machine learning approach was used to classify tweets related to different indicators, which has some limitations of producing inconsistent results at different iterations and low accuracy. Cross-validation can increase the model accuracy and precision while working with a small amount of labeled data dataset. As there is an imbalance in the dataset, future studies should consider stratified cross-validation to increase the classification model performance[142, 143]. Future studies should also consider supervised learning, which has received much attention [144–148] and given good results.

Individual sociodemographic characteristics collected from the census, SSN, and ACS, were incorporated into our data-driven models, although such data is not at the individual level. Future studies should consider conducting a survey among Twitter users, incorporating other national databases (e.g., National Household Travel Survey [149]) and other statistical models (e.g., mixed logit model [150], structural equation model [151]) to improve the model. Moreover, future research should also consider different modeling frameworks incorporating other user attributes (e.g., age, race, income) to predict public sentiments towards different mobility indicators.

In case of examining the communication patterns of different transportation agencies in the emergence of the pandemic, the analysis solely looks at tweets from their official accounts. Future empirical studies might investigate the public's behavioral responses to insufficient, irregular, and inconsistent communication throughout the pandemic's life cycle. Furthermore, this study examines all the tweets sent by transportation players during various pandemic stages. Future studies can focus solely on pandemic-related tweets to

better understand COVID-19 communication and establish techniques for adequate, consistent, and effective risk communication for various transportation players. Finally, Twitter data was only used because Twitter is one of the most popular microblogging sites in the United States and offers APIs that may be accessed. As data from more social media platforms become available, a cross-platform investigation may produce more comprehensive conclusions in the future.

REFERENCES

[1] Praharaj, S.; King, D.; Pettit, C.; Wentz, E. Using Aggregated Mobility Data to Measure the Effect of COVID-19 Policies on Mobility Changes in Sydney, London, Phoenix, and Pune. *Findings*, **2020**, *17590*.

[2] Meyer, M.; Flood, M.; Keller, J.; Lennon, J.; McVoy, G.; Dorney, C.; Leonard, K.; Hyman, R.; Smith, J. *Strategic Issues Facing Transportation, Volume 2: Climate Change, Extreme Weather Events, and the Highway System: Practitioner's Guide and Research Report*; 2014.

[3] Popper, S. W.; Kalra, N.; Silberglitt, R.; Molina-Perez, E.; Ryu, Y.; Scarpati, M. *Strategic Issues Facing Transportation, Volume 3: Expediting Future Technologies for Enhancing Transportation System Performance*; 2013.

[4] Zmud, J.; Barabba, V. P.; Bradley, M.; Kuzmyak, J. R.; Zmud, M.; Orrell, D. *Strategic Issues Facing Transportation, Volume 6: The Effects of Socio-Demographics on Future Travel Demand*; 2014.

[5] Cheng, L.; Chen, X.; Lam, W. H. K.; Yang, S.; Wang, P. Improving Travel Quality of Low-Income Commuters in China: Demand-Side Perspective. *Transp. Res. Rec.*, **2017**, *2605* (1), 99–108.

[6] Figueroa, M. J.; Nielsen, T. A. S.; Siren, A. Comparing Urban Form Correlations of the Travel Patterns of Older and Younger Adults. *Transp. Policy*, **2014**, *35*, 10–20.

[7] Scheiner, J.; Holz-Rau, C. Gendered Travel Mode Choice: A Focus on Car Deficient Households. *J. Transp. Geogr.*, **2012**, *24*, 250–261.

[8] Cheng, L.; De Vos, J.; Shi, K.; Yang, M.; Chen, X.; Witlox, F. Do Residential Location Effects on Travel Behavior Differ between the Elderly and Younger Adults? *Transp. Res. part D Transp. Environ.*, **2019**, *73*, 367–380.

[9] Wang, D.; Zhou, M. The Built Environment and Travel Behavior in Urban China: A Literature Review. *Transp. Res. Part D Transp. Environ.*, **2017**, *52*, 574–585.

[10] Lin, T.; Wang, D.; Guan, X. The Built Environment, Travel Attitude, and Travel Behavior: Residential Self-Selection or Residential Determination? *J. Transp. Geogr.*, **2017**, *65*, 111–122. https://doi.org/10.1016/j.jtrangeo.2017.10.004.

[11] Wang, D.; Lin, T. Built Environment, Travel Behavior, and Residential Self-Selection: A Study Based on Panel Data from Beijing, China. *Transportation (Amst).*, **2019**, *46* (1), 51–74. https://doi.org/10.1007/s11116-017-9783-1.

[12] Shaheen, S.; Cohen, A.; Bayen, A. The Benefits of Carpooling. **2018**.

[13] Shaheen, S. A.; Cohen, A. P.; Zohdy, I. H.; Kock, B. *Smartphone Applications to Influence Travel Choices: Practices and Policies*; United States. Federal Highway Administration, 2016.

[14] NACTO. Shared Micromobility in the US: 2018. *Natl. Assoc. City Transp. Off.*, **2019**.

[15] Langbroek, J. H. M.; Franklin, J. P.; Susilo, Y. O. Electric Vehicle Users and Their Travel Patterns in Greater Stockholm. *Transp. Res. Part D Transp. Environ.*, **2017**, *52*, 98–111.

[16] Auld, J.; Sokolov, V.; Stephens, T. S. Analysis of the Effects of Connected–Automated Vehicle Technologies on Travel Demand. *Transp. Res. Rec.*, **2017**, *2625* (1), 1–8.

[17] Pudane, B. Time Use and Travel Behaviour with Automated Vehicles. **2021**.

[18] Shakibaei, S.; De Jong, G. C.; Alpkökin, P.; Rashidi, T. H. Impact of the COVID-19 Pandemic on Travel Behavior in Istanbul: A Panel Data Analysis. *Sustain. cities Soc.*, **2021**, *65*, 102619.

[19] Parady, G.; Taniguchi, A.; Takami, K. Travel Behavior Changes during the COVID-19 Pandemic in Japan: Analyzing the Effects of Risk Perception and Social Influence on Going-out Self-Restriction. *Transp. Res. Interdiscip. Perspect.*, **2020**, *7*, 100181.

[20] Shamshiripour, A.; Rahimi, E.; Shabanpour, R.; Mohammadian, A. K. How Is COVID-19 Reshaping Activity-Travel Behavior? Evidence from a Comprehensive Survey in Chicago. *Transp. Res. Interdiscip. Perspect.*, **2020**, *7*, 100216.

[21] Abdullah, M.; Dias, C.; Muley, D.; Shahin, M. Exploring the Impacts of COVID-19 on Travel Behavior and Mode Preferences. *Transp. Res. Interdiscip. Perspect.*, **2020**, *8*, 100255.

[22] Schmitt-Grohé, S.; Teoh, K.; Uribe, M. *COVID-19: Testing Inequality in New York City*; National Bureau of Economic Research, 2020.

[23] Borjas, G. J. *Demographic Determinants of Testing Incidence and COVID-19 Infections in New York City Neighborhoods*; National Bureau of Economic Research, 2020.

[24] Abedi, V.; Olulana, O.; Avula, V.; Chaudhary, D.; Khan, A.; Shahjouei, S.; Li, J.; Zand, R. Racial, Economic, and Health Inequality and COVID-19 Infection in the United States. *J. racial Ethn. Heal. disparities*, **2021**, *8* (3), 732–742.

[25] Boneva, T.; Golin, M.; Adams-Prassl, A.; Rauh, C. Inequality in the Impact of the Coronavirus Shock: Evidence from Real Time Surveys. *Inst. Labor Econ. Discuss. Pap. Ser.*, **2020**, No. 13183.

[26] McLaren, J. Racial Disparity in COVID-19 Deaths: Seeking Economic Roots with Census Data. *BE J. Econ. Anal. Policy*, **2021**.

[27] Bengtsson, L.; Gaudart, J.; Lu, X.; Moore, S.; Wetter, E.; Sallah, K.; Rebaudet, S.; Piarroux, R. Using Mobile Phone Data to Predict the Spatial Spread of Cholera. *Sci. Rep.*, **2015**, *5*, 1–5. https://doi.org/10.1038/srep08923.

[28] Wesolowski, A.; Qureshi, T.; Boni, M. F.; Sundsøy, P. R.; Johansson, M. A.; Rasheed, S. B.; Engø-Monsen, K.; Buckee, C. O.; Singer, B. H. Impact of Human Mobility on the Emergence of Dengue Epidemics in Pakistan. *Proc. Natl. Acad. Sci. U. S. A.*, **2015**, *112* (38), 11887–11892. https://doi.org/10.1073/pnas.1504964112.

[29] Wesolowski, A.; Eagle, N.; Tatem, A. J.; Smith, D. L.; Noor, A. M.; Snow, R. W.; Buckee, C. O. Quantifying the Impact of Human Mobility on Malaria. *Science (80-.).*, **2012**, *338* (6104), 267–270. https://doi.org/10.1126/science.1223467.

[30] COVID-19 - Mobility Trends Reports - Apple https://covid19.apple.com/mobility (accessed Jul 12, 2022).

[31] Pacheco, E. COVID-19's Impact on Social Media Usage https://www.thebrandonagency.com/blog/covid-19s-impact-on-social-media-usage/ (accessed Jul 5, 2021).

[32] Predict, T.; Name, G.; Census, U.; Predicts, D. Package ' Predictrace .' **2021**.

[33] American Community Survey (ACS) https://www.census.gov/programs-surveys/acs (accessed Jul 5, 2022).

[34] Qi, Y.; Sarker, M. A. A.; Imran, M.; Pokhrel, R. Motor Vehicle Crashes among the Older Population. **2020**.

[35] Sarker, M. A. A. Multiple Logistic Regression Analysis to Evaluate Older People Traffic Safety in Illinois. Southern Illinois University at Edwardsville 2020.

[36] Sarker, M. A. A.; Rahimi, A.; Azimi, G.; Jin, X. Investigating Older Adults' Propensity toward Ridesourcing Services. *J. Transp. Eng. Part A Syst.*, **2022**, *148* (9), 4022054.

[37] Golder, S. A.; Macy, M. W. Digital Footprints: Opportunities and Challenges for Online Social Research. *Annu. Rev. Sociol.*, **2014**, *40*, 129–152.

[38] Yin, Z.; Fabbri, D.; Rosenbloom, S. T.; Malin, B. A Scalable Framework to Detect Personal Health Mentions on Twitter. *J. Med. Internet Res.*, **2015**, *17* (6), e138.

[39] Tasse, D.; Hong, J. I. Using Social Media Data to Understand Cities. **2014**.

[40] Cheng, Z.; Caverlee, J.; Lee, K.; Sui, D. Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the International AAAI Conference on Web and Social Media*; 2011; Vol. 5.

[41] Jurdak, R.; Zhao, K.; Liu, J.; AbouJaoude, M.; Cameron, M.; Newth, D. Understanding Human Mobility from Twitter. *PLoS One*, **2015**, *10* (7), 1–16. https://doi.org/10.1371/journal.pone.0131469.

[42] Noulas, A.; Scellato, S.; Lambiotte, R.; Pontil, M.; Mascolo, C. A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS One*, **2012**, *7* (5). https://doi.org/10.1371/journal.pone.0037027.

[43] Hasan, S.; Ukkusuri, S. V. Location Contexts of User Check-Ins to Model Urban Geo Life-Style Patterns. *PLoS ONE*. 2015. https://doi.org/10.1371/journal.pone.0124819.

[44] Hasan, S.; Ukkusuri, S. V. Urban Activity Pattern Classification Using Topic Models from Online Geo-Location Data. *Transp. Res. Part C Emerg. Technol.*, **2014**, *44*, 363–381. https://doi.org/10.1016/j.trc.2014.04.003.

[45] Schweitzer, L. Planning and Social Media: A Case Study of Public Transit and Stigma on Twitter. *J. Am. Plan. Assoc.*, **2014**, *80* (3), 218–238.

[46] Pender, B.; Currie, G.; Delbosc, A.; Shiwakoti, N. Social Media Use during Unplanned Transit Network Disruptions: A Review of Literature. *Transp. Rev.*, **2014**, *34* (4), 501–521. https://doi.org/10.1080/01441647.2014.915442.

[47] Luong, T. T. B.; Houston, D. Public Opinions of Light Rail Service in Los Angeles, an Analysis Using Twitter Data. *IConference 2015 Proc.*, **2015**.

[48] Tian, Y., Zmud, M., Chiu, Y.-C. Carey, D., Dale, J., Smarda, D., Lehr, R., James, R. Quality Assessment of Social Media Traffic Reports – a Field Study in Austin, Texas. In *Transportation Research Board 95th Annual Meeting*; Washington D.C., 16-6852, 2016.

[49] Steur, R. Twitter as a Spatio-Temporal Information Source for Traffic Incident Management. *Geogr. Inf. Manag. Appl.*, **2014**.

[50] Wang, Q.; Taylor, J. E. Quantifying Human Mobility Perturbation and Resilience in Hurricane Sandy. *PLoS One*, **2014**, *9* (11), 1–5. https://doi.org/10.1371/journal.pone.0112608.

[51] Wang, Q.; Taylor, J. E. Resilience of Human Mobility under the Influence of Typhoons. *Procedia Eng.*, **2015**, *118*, 942–949. https://doi.org/10.1016/j.proeng.2015.08.535.

[52] Sadri, A. M.; Hasan, S.; Ukkusuri, S. V.; Cebrian, M. Crisis Communication Patterns in Social Media during Hurricane Sandy. *Transp. Res. Rec.*, **2018**, *2672* (1), 125–137. https://doi.org/10.1177/0361198118773896.

[53] Roy, K. C.; Hasan, S.; Sadri, A. M.; Cebrian, M. Understanding the Efficiency of Social Media Based Crisis Communication during Hurricane Sandy. *Int. J. Inf. Manage.*, **2020**, *52* (August 2018), 102060. https://doi.org/10.1016/j.ijinfomgt.2019.102060.

[54] McDonald, D. D. Natural Language Generation. *Handb. Nat. Lang. Process.*, **2010**, *2*, 121–144.

[55] GitHub - cjhutto/vaderSentiment: VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains. https://github.com/cjhutto/vaderSentiment (accessed Aug 1, 2021).

[56] Hutto, C.J. and Gilbert, E. VADER: A Parsimonious Rule-Based Model For. *Eighth Int. AAAI Conf. Weblogs Soc. Media*, **2014**, 18.

[57] Farrahi, K.; Gatica-Perez, D. Discovering Routines from Large-Scale Human Locations Using Probabilistic Topic Models. *ACM Trans. Intell. Syst. Technol.*, **2011**, *2* (1), 1–27. https://doi.org/10.1145/1889681.1889684.

[58] Huynh, T.; Fritz, M.; Schiele, B. Discovery of Activity Patterns Using Topic Models. *UbiComp 2008 - Proc. 10th Int. Conf. Ubiquitous Comput.*, **2008**, 10–19. https://doi.org/10.1145/1409635.1409638.

[59] Sun, L.; Yin, Y. Discovering Themes and Trends in Transportation Research Using Topic Modeling. *Transp. Res. Part C Emerg. Technol.*, **2017**, *77*, 49–66. https://doi.org/10.1016/j.trc.2017.01.013.

[60] Blei, D. M.; Ng, A. Y.; Jordan, M. I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, **2003**, *3*, 993–1022.

[61] Griffiths, T. L.; Steyvers, M. Finding Scientific Topics. *Proc. Natl. Acad. Sci. U. S. A.*, **2004**, *101* (SUPPL. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101.

[62] Ahmed, M. A.; Sadri, A. M.; Pradhananga, P.; Elzomor, M.; Pradhananga, N. Social Media Communication Patterns of Construction Industry in Major Disasters. In *Construction Research Congress 2020: Computer Applications - Selected Papers from the Construction Research Congress 2020*; American Society of Civil Engineers: Reston, VA, 2020; pp 678–687. https://doi.org/10.1061/9780784482865.072.

[63] Howard, P. N.; Kollanyi, B. Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum. *SSRN Electron. J.*, **2017**. https://doi.org/10.2139/ssrn.2798311.

[64] Chu, Z.; Gianvecchio, S.; Wang, H.; Jajodia, S. Who Is Tweeting on Twitter: Human, Bot, or Cyborg? *Proc. - Annu. Comput. Secur. Appl. Conf. ACSAC*, **2010**, 21–30. https://doi.org/10.1145/1920261.1920265.

[65] Clark, E. M.; Williams, J. R.; Jones, C. A.; Galbraith, R. A.; Danforth, C. M.; Dodds, P. S. Sifting Robotic from Organic Text: A Natural Language Approach for Detecting Automation on Twitter. *J. Comput. Sci.*, **2016**, *16*, 1–7. https://doi.org/10.1016/j.jocs.2015.11.002.

[66] Dickerson, J. P.; Kagan, V.; Subrahmanian, V. S. Using Sentiment to Detect Bots on Twitter: Are Humans More Opinionated than Bots? *ASONAM 2014 - Proc. 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min.*, **2014**, 620–627. https://doi.org/10.1109/ASONAM.2014.6921650.

[67] Klein, B.; LaRock, T.; McCabe, S.; Torres, L.; Privitera, F.; Lake, B.; Kraemer, M. U. G.; Brownstein, J. S.; Lazer, D.; Eliassi-Rad, T. Assessing Changes in Commuting and Individual Mobility in Major Metropolitan Areas in the United States during the COVID-19 Outbreak. *Northeast. Univ. Netw. Sci. Inst.*, **2020**, *29*.

[68] Gao, S.; Rao, J.; Kang, Y.; Liang, Y.; Kruse, J. Mapping County-Level Mobility Pattern Changes in the United States in Response to COVID-19. *SIGSpatial Spec.*, **2020**, *12* (1), 16–26.

[69] Leatherby, L.; Gelles, D. How the Virus Transformed the Way Americans Spend Their Money. *New York Times*, **2020**, *11*.

[70] Knoll, C. Panicked Shoppers Empty Shelves as Coronavirus Anxiety Rises. *New York Times*, **2020**.

[71] Gao, J.; Bernardes, S. D.; Bian, Z.; Ozbay, K.; Iyer, S. Initial Impacts of COVID-19 on Transportation Systems: A Case Study of the US Epicenter, the New York Metropolitan Area. *arXiv Prepr. arXiv2010.01168*, **2020**.

[72] Twitter by the Numbers (2021): Stats, Demographics & Fun Facts https://www.omnicoreagency.com/twitter-statistics/ (accessed Jul 26, 2021).

[73] Rashidi, T. H.; Abbasi, A.; Maghrebi, M.; Hasan, S.; Waller, T. S. Exploring the Capacity of Social Media Data for Modelling Travel Behaviour: Opportunities and Challenges. *Transp. Res. Part C Emerg. Technol.*, **2017**, *75*, 197–211. https://doi.org/10.1016/j.trc.2016.12.008.

[74] Qi, B.; Costin, A.; Jia, M. A Framework with Efficient Extraction and Analysis of Twitter Data for Evaluating Public Opinions on Transportation Services. *Travel Behav. Soc.*, **2020**, *21*, 10–23.

[75] Joshi, R.; Tekchandani, R. Comparative Analysis of Twitter Data Using Supervised Classifiers. In *2016 International conference on inventive computation technologies (ICICT)*; IEEE, 2016; Vol. 3, pp 1–6.

[76] Çelenli, H. İ.; Öztürk, S. T.; Şahin, G.; Gerek, A.; Ganiz, M. C. Document Embedding Based Supervised Methods for Turkish Text Classification. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)*; IEEE, 2018; pp 477–482.

[77] Alam, M. R.; Sadri, A. M. Examining the Communication Pattern of Transportation and Transit Agencies on Twitter: A Longitudinal Study in the Emergence of COVID-19 on Twitter. *Transp. Res. Rec.*, **2022**, 03611981221082564.

[78] Alam, M. R.; Sadri, A. M.; Jin, X. Identifying Public Perceptions toward Emerging Transportation Trends through Social Media-Based Interactions. *Futur. Transp.*, **2021**, *1* (3), 794–813.

[79] Edo-Osagie, O.; Smith, G.; Lake, I.; Edeghere, O.; De La Iglesia, B. Twitter Mining Using Semi-Supervised Classification for Relevance Filtering in Syndromic Surveillance. *PLoS One*, **2019**, *14* (7), e0210689.

[80] Karisani, P.; Karisani, N. Semi-Supervised Text Classification via Self-Pretraining. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*; 2021; pp 40–48.

[81] Popular Baby Names https://www.ssa.gov/oact/babynames/limits.html (accessed Jul 30, 2021).

[82] Decennial Census by Decades https://www.census.gov/programs-surveys/decennial-census/decade.2010.html (accessed Jul 30, 2021).

[83] American Community Survey (ACS) https://www.census.gov/programs-surveys/acs (accessed Jul 29, 2021).

[84] Pak, A.; Paroubek, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREc*; 2010; Vol. 10, pp 1320–1326.

[85] Go, A.; Bhayani, R.; Huang, L. Twitter Sentiment Classification Using Distant Supervision. *CS224N Proj. report, Stanford*, **2009**, *1* (12), 2009.

[86] Davidov, D.; Tsur, O.; Rappoport, A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *Coling 2010: Posters*; 2010; pp 241–249.

[87] Collins, C.; Hasan, S.; Ukkusuri, S. V. A Novel Transit Rider Satisfaction Metric: Rider Sentiments Measured from Online Social Media Data. *J. Public Transp.*, **2013**, *16* (2), 2.

[88] Nik Bakht, Mazdak, Sherif N. Kinawy, and T. E. E.-D. News and Social Media as Performance Indicators for Public Involvement in Transportation Planning: Eglinton Crosstown Project in Toronto, Canada. In *Transportation Research Board 94th Annual Meeting*; Location: Washington DC, United States, 2015; Vol. No. 15-011.

[89] Giglietto, F.; Selva, D. Second Screen and Participation: A Content Analysis on a Full Season Dataset of Tweets. *J. Commun.*, **2014**, *64* (2), 260–277.

[90] Takahashi, B.; Tandoc, E. C.; Carmichael, C. Communicating on Twitter during a Disaster: An Analysis of Tweets during Typhoon Haiyan in the Philippines. *Comput. Human Behav.*, **2015**, *50*, 392–398. https://doi.org/10.1016/J.CHB.2015.04.020.

[91] Cavazos-Rehg, P. A.; Krauss, M. J.; Sowles, S.; Connolly, S.; Rosas, C.; Bharadwaj, M.; Bierut, L. J. A Content Analysis of Depression-Related Tweets. *Comput. Human Behav.*, **2016**, *54*, 351–357. https://doi.org/10.1016/J.CHB.2015.08.023.

[92] Harlow, S.; Johnson, T. J. The Arab Spring| Overthrowing the Protest Paradigm? How the New York Times, Global Voices and Twitter Covered the Egyptian Revolution. *Int. J. Commun.*, **2011**, *5*, 16.

[93] Artwick, C. G. News Sourcing and Gender on Twitter. *Journalism*, **2014**, *15* (8), 1111–1127.

[94] Greer, C. F.; Ferguson, D. A. Using Twitter for Promotion and Branding: A Content Analysis of Local Television Twitter Sites. *J. Broadcast. Electron. Media*, **2011**, *55* (2), 198–214.

[95] Waters, R. D.; Jamal, J. Y. Tweet, Tweet, Tweet: A Content Analysis of Nonprofit Organizations' Twitter Updates. *Public Relat. Rev.*, **2011**, *37* (3), 321–324. https://doi.org/10.1016/J.PUBREV.2011.03.002.

[96] Adams, A.; McCorkindale, T. Dialogue and Transparency: A Content Analysis of How the 2012 Presidential Candidates Used Twitter. *Public Relat. Rev.*, **2013**, *39* (4), 357–359. https://doi.org/10.1016/J.PUBREV.2013.07.016.

[97] Schler, J.; Koppel, M.; Argamon, S.; Pennebaker, J. Effects of Age and Gender on Blogging. AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. Stanford, CA 2006.

[98] Rosenthal, S.; McKeown, K. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre-and Post-Social Media Generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*; 2011; pp 763–772.

[99] Nguyen, D.; Smith, N. A.; Rose, C. Author Age Prediction from Text Using Linear Regression. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*; 2011; pp 115–123.

[100] Burger, J. D.; Henderson, J.; Kim, G.; Zarrella, G. *Discriminating Gender on Twitter*; MITRE CORP BEDFORD MA BEDFORD United States, 2011.

[101] Rao, D.; Yarowsky, D.; Shreevats, A.; Gupta, M. Classifying Latent User Attributes in Twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*; 2010; pp 37–44.

[102] Rao, D.; Paul, M.; Fink, C.; Yarowsky, D.; Oates, T.; Coppersmith, G. Hierarchical Bayesian Models for Latent Attribute Detection in Social Media. In *Fifth International AAAI Conference on Weblogs and Social Media*; 2011.

[103] Pennacchiotti, M.; Popescu, A.-M. A Machine Learning Approach to Twitter User Classification. In *Fifth international AAAI conference on weblogs and social media*; 2011.

[104] Goel, S.; Hofman, J. M.; Sirer, M. I. Who Does What on the Web: A Large-Scale Study of Browsing Behavior. In *Sixth International AAAI Conference on Weblogs and Social Media*; 2012.

[105] Eisenstein, J.; Smith, N. A.; Xing, E. Discovering Sociolinguistic Associations with Structured Sparsity. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*; 2011; pp 1365–1374.

[106] Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Lucas, R. E.; Agrawal, M.; Park, G. J.; Lakshmikanth, S. K.; Jha, S.; Seligman, M. E. P. Characterizing Geographic Variation in Well-Being Using Tweets. In *Seventh International AAAI Conference on Weblogs and Social Media*; 2013.

[107] Culotta, A. Estimating County Health Statistics with Twitter. In *Proceedings of the SIGCHI conference on human factors in computing systems*; 2014; pp 1335–1344.

[108] Bandhakavi, A.; Wiratunga, N.; Padmanabhan, D.; Massie, S. Lexicon Based Feature Extraction for Emotion Text Classification. *Pattern Recognit. Lett.*, **2017**, *93*, 133–142.

[109] Li, Z.; Gurgel, H.; Dessay, N.; Hu, L.; Xu, L.; Gong, P. Semi-Supervised Text Classification Framework: An Overview of Dengue Landscape Factors and Satellite Earth Observation. *Int. J. Environ. Res. Public Health*, **2020**, *17* (12), 4509.

[110] Anjaria, M.; Guddeti, R. M. R. Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning. In *2014 sixth international conference on communication systems and networks (COMSNETS)*; IEEE, 2014; pp 1–8.

[111] Trips by Distance | Tyler Data & Insights https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv (accessed Apr 17, 2022).

[112] Zuo, F.; Wang, J.; Gao, J.; Ozbay, K.; Ban, X. J.; Shen, Y.; Yang, H.; Iyer, S. An Interactive Data Visualization and Analytics Tool to Evaluate Mobility and Sociability Trends During COVID-19. 2020.

[113] censusgeocode · PyPI https://pypi.org/project/censusgeocode/ (accessed Apr 18, 2022).

[114] Edo-Osagie, O.; Smith, G.; Lake, I.; Edeghere, O.; De La Iglesia, B. Twitter Mining Using Semi-Supervised Classification for Relevance Filtering in Syndromic Surveillance. *PLoS ONE*. 2019. https://doi.org/10.1371/journal.pone.0210689.

[115] TensorFlow Hub https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/4 (accessed Jun 26, 2022).

[116] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv Prepr. arXiv1810.04805*, **2018**.

[117] Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE international conference on computer vision*; 2015; pp 19–27.

[118] Hardeniya, N. *NLTK Essentials*; Packt Publishing, 2015.

[119] Brandt, J.; Buckingham, K.; Buntain, C.; Anderson, W.; Ray, S.; Pool, J.-R.; Ferrari, N. Identifying Social Media User Demographics and Topic Diversity with Computational Social Science: A Case Study of a Major International Policy Forum. *J. Comput. Soc. Sci.*, **2020**, *3* (1), 167–188.

[120] McFadden, D. The Measurement of Urban Travel Demand. *J. Public Econ.*, **1974**, *3* (4), 303–328.

[121] Train, K. E. *Discrete Choice Methods with Simulation*; Cambridge university press, 2009.

[122] Kim, H.; Jang, S. M.; Kim, S.-H.; Wan, A. Evaluating Sampling Methods for Content Analysis of Twitter Data. *Soc. Media+ Soc.*, **2018**, *4* (2), 2056305118772836.

[123] 2020 Data Release New and Notable https://www.census.gov/programs-surveys/acs/news/data-releases/2020/release.html#XYZ (accessed Apr 21, 2022).

[124] Fagerland, M. W.; Hosmer, D. W. A Generalized Hosmer–Lemeshow Goodness-of-Fit Test for Multinomial Logistic Regression Models. *Stata J.*, **2012**, *12* (3), 447–453.

[125] Drummond, J.; Hasnine, M. S. *Online and In-Store Shopping Behavior During COVID-19 Pandemic: Lesson Learned from a Panel Survey in New York City*; 2022.

[126] Koch, J.; Frommeyer, B.; Schewe, G. Online Shopping Motives during the COVID-19 Pandemic—Lessons from the Crisis. *Sustainability*, **2020**, *12* (24), 10247.

[127] Chen, Y.; Jiao, J.; Bai, S.; Lindquist, J. Modeling the Spatial Factors of COVID-19 in New York City. *Available SSRN 3606719*, **2020**.

[128] Wang, D.; He, B. Y.; Gao, J.; Chow, J. Y. J.; Ozbay, K.; Iyer, S. Impact of COVID-19 Behavioral Inertia on Reopening Strategies for New York City Transit. *Int. J. Transp. Sci. Technol.*, **2021**, *10* (2), 197–211.

[129] Sahraei, M. A.; Kuşkapan, E.; Çodur, M. Y. Public Transit Usage and Air Quality Index during the COVID-19 Lockdown. *J. Environ. Manage.*, **2021**, *286*, 112166.

[130] Barbour, N.; Menon, N.; Mannering, F. A Statistical Assessment of Work-from-Home Participation during Different Stages of the COVID-19 Pandemic. *Transp. Res. Interdiscip. Perspect.*, **2021**, *11*, 100441.

[131] Herrera-Escobar, J. P.; Seshadri, A. J.; Rivero, R.; Toppo, A.; Al Rafai, S. S.; Scott, J. W.; Havens, J. M.; Velmahos, G.; Kasotakis, G.; Salim, A. Lower Education and Income Predict Worse Long-Term Outcomes after Injury. *J. Trauma Acute Care Surg.*, **2019**, *87* (1), 104–110.

[132] Mittal, A.; Mantri, A.; Tandon, U.; Dwivedi, Y. K. A Unified Perspective on the Adoption of Online Teaching in Higher Education during the COVID-19 Pandemic. *Inf. Discov. Deliv.*, **2021**.

[133] Wang, S.; Cheah, J.-H.; Lim, X.-J.; Leong, Y. C.; Choo, W. C. Thanks COVID-19, I'll Reconsider My Purchase: Can Fear Appeal Reduce Online Shopping Cart Abandonment? *J. Retail. Consum. Serv.*, **2022**, *64*, 102843.

[134] Escudero-Castillo, I.; Mato-Díaz, F. J.; Rodriguez-Alvarez, A. Furloughs, Teleworking and Other Work Situations during the COVID-19 Lockdown: Impact on Mental Well-Being. *Int. J. Environ. Res. Public Health*, **2021**, *18* (6), 2898.

[135] Aryani, D. N.; Nair, R. K.; Hoo, D. X. Y.; Hung, D. K. M.; Lim, D. H. R.; Chew, W. P.; Desai, A. A Study on Consumer Behaviour: Transition from Traditional Shopping to Online Shopping during the COVID-19 Pandemic. *Int. J. Appl. Bus. Int. Manag.*, **2021**, *6* (2), 81–95.

[136] Dubay, L.; Aarons, J.; Brown, K. S.; Kenney, G. M. How Risk of Exposure to the Coronavirus at Work Varies by Race and Ethnicity and How to Protect the Health and Well-Being of Workers and Their Families. *Washingt. DC Urban Inst.*, **2020**.

[137] Zhang, W.; Liu, L. Exploring Non-Users' Intention to Adopt Ride-Sharing Services: Taking into Account Increased Risks Due to the COVID-19 Pandemic among Other Factors. *Transp. Res. Part A Policy Pract.*, **2022**, *158*, 180–195.

[138] When Are People Most Likely to Buy Online? - SaleCycle https://www.salecycle.com/blog/stats/when-are-people-most-likely-to-buy-online/ (accessed Jun 27, 2022).

[139] Jensen, K. L.; Yenerall, J.; Chen, X.; Yu, T. E. US Consumers' Online Shopping Behaviors and Intentions during and after the COVID-19 Pandemic. *J. Agric. Appl. Econ.*, **2021**, *53* (3), 416–434.

[140] Liu, Z.; Van Egdom, D.; Flin, R.; Spitzmueller, C.; Adepoju, O.; Krishnamoorti, R. I Don't Want to Go Back: Examining the Return to Physical Workspaces during COVID-19. *J. Occup. Environ. Med.*, **2020**, *62* (11), 953–958.

[141] Twitter API for Academic Research | Products | Twitter Developer Platform https://developer.twitter.com/en/products/twitter-api/academic-research (accessed Jun 27, 2022).

[142] Purushotham, S.; Tripathy, B. K. Evaluation of Classifier Models Using Stratified Tenfold Cross Validation Techniques. In *International Conference on Computing and Communication Systems*; Springer, 2011; pp 680–690.

[143] Zeng, X.; Martinez, T. R. Distribution-Balanced Stratified Cross-Validation for Accuracy Estimation. *J. Exp. Theor. Artif. Intell.*, **2000**, *12* (1), 1–12.

[144] Dilrukshi, I.; De Zoysa, K.; Caldera, A. Twitter News Classification Using SVM. In *2013 8th International Conference on Computer Science & Education*; IEEE, 2013; pp 287–291.

[145] Read, J. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Proceedings of the ACL student research workshop*; 2005; pp 43–48.

[146] Yerva, S. R.; Miklós, Z.; Aberer, K. What Have Fruits to Do with Technology? The Case of Orange, Blackberry and Apple. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*; 2011; pp 1–10.

[147] Nishida, K.; Banno, R.; Fujimura, K.; Hoshide, T. Tweet Classification by Data Compression. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*; 2011; pp 29–34.

[148] Sriram, B.; Fuhry, D.; Demir, E.; Ferhatosmanoglu, H.; Demirbas, M. Short Text Classification in Twitter to Improve Information Filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*; 2010; pp 841–842.

[149] Federal Highway Administration. 2009 National Household Travel Survey. **2010**, 82.

[150] Hole, A. R. Mixed Logit Modelling in Stata An Overview. **2013**, No. September, 1–43.

[151] Westland, J. C. An Introduction to Structural Equation Models. *Stud. Syst. Decis. Control*, **2015**, *22*, 1–8. https://doi.org/10.1007/978-3-319-16507-3_1.

[152] Mitchell, A.; Kiley, J.; Gottfried, J.; Guskin, E. The Role of News on Facebook: Common yet Incidental. *Pew Res. Cent.*, **2013**.

[153] He, K.; Xu, Z.; Wang, P.; Deng, L.; Tu, L. Congestion Avoidance Routing Based on Large-Scale Social Signals. *IEEE Trans. Intell. Transp. Syst.*, **2015**, *17* (9), 2613–2626.

[154] Eirkis, D.; Eirkis, M. Friending Transit: How Public Transit Agencies Are Using Social Media to Expand Their Reach and Improve Their Image. *Mass Transit*, **2010**.

[155] Bregman, S. Uses of Social Media in Public Transportation. Transit Cooperative Research Program (TCRP) Synthesis 99. *Transp. Res. Board, Washingt.*, **2012**.

[156] Wojtowicz, J.; Wallace, W. A. Use of Social Media by Transportation Agencies for Traffic Management. *Transp. Res. Rec.*, **2016**, *2551* (1), 82–89.

[157] Yeh, D.; Schmitt, P. Social Media Use by State Departments of Transportation and Other Government Agencies. *Transp. Synth. Rep.*, **2011**.

[158] Rahman, R.; Roy, K. C.; Abdel-Aty, M.; Hasan, S. Sharing Real-Time Traffic Information with Travelers Using Twitter: An Analysis of Effectiveness and Information Content. *Front. Built Environ.*, **2019**, *5*, 83.

[159] Kocatepe, A.; Lores, J.; Ozguven, E. E.; Yazici, A. The Reach and Influence of DOT Twitter Accounts: A Case Study in Florida. In *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*; 2015; Vol. 2015-Octob, pp 330–335. https://doi.org/10.1109/ITSC.2015.63.

[160] Anger, I.; Kittl, C. Measuring Influence on Twitter. In *ACM International Conference Proceeding Series*; 2011. https://doi.org/10.1145/2024288.2024326.

[161] Bregman, S.; Watkins, K. E. *Best Practices for Transportation Agency Use of Social Media*; CRC Press, 2019.

[162] Wang, Y.; Hao, H.; Platt, L. S. Examining Risk and Crisis Communications of Government Agencies and Stakeholders during Early-Stages of COVID-19 on Twitter. *Comput. Human Behav.*, **2021**, *114*, 106568.

[163] Mishra, D.; Qian, Y.; Kazmee, H.; Tutumluer, E. Investigation of Geogrid-Reinforced Railroad Ballast Behavior Using Large-Scale Triaxial Testing and Discrete Element Modeling. *Transp. Res. Rec.*, **2014**, *2462* (1), 98–108.

[164] Hossain, E.; Shariff, M. A. U.; Hossain, M. S.; Andersson, K. A Novel Deep Learning Approach to Predict Air Quality Index. In *Proceedings of International Conference on Trends in Computational and Cognitive Engineering*; Springer, 2021; pp 367–381.

[165] Yao, F.; Wang, Y. Tracking Urban Geo-Topics Based on Dynamic Topic Model. *Comput. Environ. Urban Syst.*, **2020**, *79*, 101419. https://doi.org/10.1016/J.COMPENVURBSYS.2019.101419.

[166] Hao, H.; Wang, Y. Leveraging Multimodal Social Media Data for Rapid Disaster Damage Assessment. *Int. J. Disaster Risk Reduct.*, **2020**, *51*, 101760. https://doi.org/10.1016/J.IJDRR.2020.101760.

[167] Twitter Developers https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/introduction (accessed Jul 19, 2021).

[168] Public Transportation Association, A.; Dickens, M. PUBLIC TRANSPORTATION RIDERSHIP REPORT. **2020**.

[169] Status of lockdown and stay-at-home orders in response to the coronavirus (COVID-19) pandemic, 2020 - Ballotpedia https://ballotpedia.org/Status_of_lockdown_and_stay-at-home_orders_in_response_to_the_coronavirus_(COVID-19)_pandemic,_2020#Defining_critical_industries.2C_essential.2C_and_nonessential_businesses (accessed Jul 19, 2021).

[170] Wang, X.; McCallum, A.; Wei, X. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. In *Seventh IEEE international conference on data mining (ICDM 2007)*; IEEE, 2007; pp 697–702.

[171] Lindsey, R.; Headden, W.; Stipicevic, M. A Phrase-Discovering Topic Model Using Hierarchical Pitman-Yor Processes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*; 2012; pp 214–222.

[172] Griffiths, T. L.; Steyvers, M.; Blei, B.; Blei, J. Finding Scientific Topics. Pnas 101 (SUPPL. 1): 5228–5235. 2004.

[173] Getchell, M. C.; Sellnow, T. L. A Network Analysis of Official Twitter Accounts during the West Virginia Water Crisis. *Comput. Human Behav.*, **2016**, *54*, 597–606.

[174] Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Third international AAAI conference on weblogs and social media*; 2009.

[175] Tabassum, S.; Pereira, F. S. F.; Fernandes, S.; Gama, J. Social Network Analysis: An Overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, **2018**, *8* (5), e1256.

VITA

MD RAKIBUL ALAM

20012- 2017	B.Sc. in Civil Engineering Bangladesh University of Engineering and Technology Dhaka, Bangladesh
2017- 2018	Design Engineer Equal Engineers and Consultants Dhaka, Bangladesh
2018- 2019	Ph.D. Student/Graduate Assistant Civil, Environmental, And Construction Engineering University of Central Florida, Orlando, Florida
2019- Present	Ph.D. Student/Graduate Assistant Civil and Environmental Engineering Florida International University, Miami, Florida
2021	M.S. in Civil Engineering Civil and Environmental Engineering Florida International University, Miami, Florida
2019- 2020	Social Coordinator, Institute of Transportation Engineers (ITE) and Women's Transportation Seminar (WTS) Student Chapter, FIU

PUBLICATIONS AND PRESENTATIONS

Alam, M. R., Sadri, A. M., & Jin, X. (2021). Identifying Public Perceptions toward Emerging Transportation Trends through Social Media-Based Interactions. Future Transportation, 1(3), 794-813. https://doi.org/10.3390/futuretransp1030044.

Alam, M. R., Sadri A. M., Examining the Communication Pattern of Transportation and Transit Agencies on Twitter: A Longitudinal Study in the Emergence of COVID-19 on Twitter. https://doi.org/10.1177/03611981221082564.

Jin, X., Alam, R., Sadri, A., & Zhang, L. (2020). Identifying and Tracking Emerging Transportation Trends and Indicators. Publisher: Florida Department of Transportation. https://rosap.ntl.bts.gov/view/dot/56907.

Alam, M. R., Sadri A. M. (2021, Jan). Examining the Communication Pattern of Transportation and Transit Agencies on Twitter: A Longitudinal Study in the Emergence of COVID-19 on Twitter. *(TRB 2022)*

Alam, M. R., Sadri A. M., Jin, X. (2021, Jan). Identifying Public Perceptions Toward Emerging Transportation Trends through Social Media-Based Interactions: A pilot study. *(TRB 2022)*