AN ANALYSIS OF RIDESHARING TRIP TIME PRE- AND AMID-COVID-19 PANDEMIC USING ADVANCED TEXT MINING TECHNIQUE – THE USA VS INDIAN CASE STUDY ACROSS DIFFERENT AGE AND GENDER GROUPS

Wenxiang Xu, Graduate Research Assistant, Hangzhou Innovation Institute, Beihang University, 18, Chuanghui Street, Binjiang District, Hangzhou, Zhejiang, 310052, China, (+86)198-2183-4125, xwxtom.163.com

Anae Sobhani, Ph.D., Assistant Professor, Barney School of Business, Harford University, Beatrice Fox Auerbach Hall, 200 Bloomfield Ave, West Hartford, CT 06117, United States, (+1) 617-580-1734, sobhani@hartford.edu

Ting Fu, Corresponding Author, Ph.D., Associate Professor, School of Transportation Engineering, Tongji University, 4800 Cao'an Highway, Shanghai, 201804, China, (+86)138-1681-0642, Fax: (+86)021-69585717, tingfu@tongji.edu.cn

Amir Mahdi Khabooshani, Research Assistant, School of Computer and Information Technology Engineering, Sadjad University of Technology, 64 Jalal Al Ahmad St, Mashhad, Iran, (+98)939-858-6881, a.mkh000@gmail.com

Aminreza Vazirinasab, Graduate Research Assistant, School of Urban Planning, Faculty of Fine Arts, University of Tehran, 16th Azar St., Enghelab Sq., Tehran, Iran, (+98)913-340-3449, amin.vazirinasab@ut.ac.ir

Sina Shokoohyar, Assistant Professor, Erivan K. Haub School of Business, Saint Joseph's University, 5600 City Ave, 352 Mandeville Hall, Philadelphia, PA 19131, 610-660-2217, Sina.Shokoohyar@sju.edu

Ahmad Sobhani, Senior Data Scientist, Amazon, Washington DC-Baltimore Area, ahmadsobhani@oakland.edu

Behnaz Raouf, Transportation Planner, Fairfax County, Washington DC-Baltimore Area, +15713989355, behnaz.rauofhassanzadeh@fairfaxcounty.gov

ABSTRACT

In this study, our aim was to investigate the opinions of users on the Twitter application pre- and post-pandemic about travel time in ridesharing. Our results show that trip time happening, such as mornings and other times, became more important after COVID-19 and users` concern about waiting time has increased. We also found that women and US users were more positive than others about ridesharing trip time in the pre-pandemic era. Our results can be used to analyze the sentiments of ridesharing industry users in order to compete with other companies by providing better services and designing a standard to improve travel time.

Keywords: ridesharing, trip time, topic modeling, sentiment analysis, Twitter data, COVID-19

INTRODUCTION

Overview

The increase in traffic congestion and the subsequent increase in travel time and pollution caused by it, as well as the growth of technology and the Internet, prompted researchers to create an integrated and flexible system to solve these problems by using a combination of public transportation and the Internet. The advent of ride-sharing systems is a response to the growing demand for technology-driven travel that has spread around the world over the past decade, and in America, Uber grew as a ride-hailing service system *(1)*. Although Lyft is also present in the US market, Uber dominates the current market share *(2)*. However, if we're being realistic, this system increases traffic and congestion, the Vehicle Kilometers Traveled (VKT) index, and the subsequent accidents when it takes the place of public transportation, notably traditional taxis *(3)*.

After a while, the ride-sharing system did not have much effect on traffic reduction *(4, 5)* and was not suitable for people with lower incomes *(6)*, and experts thought of creating a system that reduces traffic and VKT and costs less for different people. Users who are temporary owners of goods and services supplied by others are connected by the platform's service providers in sharing services *(7, 8)*. Ridesharing is defined as the sharing of an automobile trip by two or more people traveling to a particular destination *(9)*. Passengers have the option to use the service for all or a portion of their trip, according to Transport for London *(10)*. Different mobile applications are used to manage and schedule ride-sharing. A study found that ride-sharing services like Uber have the power to decrease car use, switch traffic from single occupancy to ride-sharing, and postpone travel plans during rush hours, hence reducing overall traffic congestion in urban areas *(11)*.

According to a paper, at first, during the COVID-19 epidemic, ride-sharing traffic was drastically cut. The drop in ride-hailing trips was significantly greater than the overall traffic volume drop. Second, during the epidemic, non-shared trips had significantly higher travel distances, yet travel times were not necessarily longer due to decreased overall traffic network congestion. Third, while the number of inter-census-tract short trips decreased, their distances grew *(12)*. However, due to the COVID-19 pandemic, many ride-sharing service providers experienced disruption because most large cities have been in lockdown *(13)*. According to Smith *(14)*, reports from the official Uber website, the number of Uber users and trips has been on the rise until 2019, but after the start of the COVID-19 pandemic, the growth rate of users and trips of this platform has sharply decreased. This rate has also decreased for the Lyft platform during COVID-19 *(15)*.

Research context

In this study, the decrease of users and travels in ride-sharing systems of different age and gender groups was examined in terms of travel time to determine what changes in travel time, waiting time, and the time of travel during the day were from the perspective of users. The pandemic forced many people to switch to private transportation, which led to a reduction in the number of public transport passengers and a significant change in travel behavior in urban areas *(16)*. The purpose of this study is to examine the views of a large community of users of ride-sharing in the Twitter application for the United States and India. The first one is a developed country, and the second is a developing country with many cultural differences that can help in crises such as pandemics to better manage ride-sharing systems and to make this system stable over time. 63800 data from Twitter were collected using Text Mining methods, specifying age, gender, and country in the pre-and amid- Pandemics. According to the data collected from USA and India our goal is to compare the two countries. The methods we chose for data analysis are the Bidirectional

Encoder Representations from Transformers model (BERT), the Valence Aware Dictionary and sEntiment Reasoner (VADER) for sentiment analysis, and Latent Dirichlet Allocation (LDA) for topic modeling.

LITERATURE REVIEW

Travel time has long been an important issue for everyone in transportation, and the occurrence of the COVID-19 pandemic may have an impact on it, especially for ride-sharing users, which is the subject of our discussion in the literature. According to a study in the United States, in addition to existing barriers, the disease has created new barriers to transportation *(17)*, leading to reduced mobility at the peak of the epidemic compared to before in many countries, such as China *(18)*, America *(19)*, Canada up to 50% *(20)*, Germany *(21)*, Greece up to 50% *(22)*, Japan (10 to 50% intra-provincial travel and 70% inter-provincial travel in Tokyo) *(23)*, and a 70% reduction in public transport travel in Tampere, Finland *(24)*. Although the Vehicle Miles Traveled (VMT) decreased due to telecommuting, the total number of trips during this period increased *(25)*, and most of the trips were made by car. One reason for the decline in travel can be attributed to strict government policies and regulations to control the disease *(26)*. Work away from home and participation in activities such as meeting friends and family, leisure activities, and bulk shopping also declined. However, men, lower-income groups, and healthcare professionals are more likely to experience an increase in out-of-home travel activities. According to a study conducted by Ye et al. *(27)* using Twitter data and social media mobility trends before and after COVID in New York, they concluded that in addition to government policies, social media also influences user mobility behaviors. It is a transition. They believe that trending tweets lead to changes in driving, transportation, and mobility. Another study of 15,776 English-language tweets found that during the pandemic, people did not use public transportation for fear of the disease and instead turned to private cars, bicycles, and walking *(28)*. In general, this disease has changed people's mobility choices, which has had a great impact on public transportation due to people's perception of its dangers. Two studies from China and the United States have shown how the disease has affected taxis and reduced their travel *(29, 30)*. Also, pedestrian travel increased, private cars were mainly used for transportation *(31)*, and demand for public transport such as buses and subways fell sharply *(32, 33)*. A survey of travel behavior for users of the GIRA shared bicycle system in Lisbon showed that the share of trips made by the system increased, while the percentage of combined trips by public transport decreased. Before the epidemic, the perceived safety of using public transport was similar to other modes. Now it is considered by far the most unsafe mode in Lisbon *(34)*. This is very important because if public transport is considered unsafe and unhealthy by people during this period, it will not be able to play the social roles it is supposed to play, including accessibility, sustainability, and equality *(35)*. According to data from 964 potential users of passenger sharing services, a study showed that tolerating ambiguity and environmental concerns of COVID-19 directly and positively affects the intention of potential users to accept passenger sharing services, and consumers who understand high health threats are less likely to accept ride-sharing services *(36)*.

But the views of different social groups of men and women of all ages may differ. For example, Young & Farber *(37)* believe that ride-hailing services are a wealthy phenomenon of the younger generation because younger riders use Uber more than older riders *(38)*. A study in Malaysia shows that young people aged 20 to 29 are more likely to prefer shared autonomous vehicles *(39)*. According to a study in Greece, older travelers generally maintained their mobility patterns despite their higher vulnerability to COVID-19 disease *(22)*. According to an online questionnaire for 26 Texas rural transportation leaders, they believe that the number of elderly riders is declining and that they may suffer from social isolation *(40)*. In India, it has been found that women generally prefer more security and less waiting time than comfort in ride sharing *(41)*.

The main issues discussed in this article are travel time in ride-sharing systems such as Uber and the effects of COVID-19 disease on travel time, and we also look at what users of the wider space, such as

Twitter, have to say about it, although there is little literature on it. To determine the importance of this issue, we examined the pre-and amid-COVID-19 literature. In fact, a shared future can reduce congestion, reduce Vehicle Miles Traveled (VMT), lead to better air quality, support economic growth *(42)*, and improve travel time for all road users *(9)*, as well as travel time by ride-sharing is more valuable than by train or bus in France *(43)*, For this reason, high-income people value their travel time more, and ride-sharing has been an attractive option for them more than other travel modes *(44)*. A study showed that the value of users' time and also travel time uncertainty has significant effects on the performance of sharing systems *(45)*. Accordingly, immediate and optimal compliance is one of the ways to succeed in ridesharing services *(46)*. For this reason, the ride subscription provider must provide users with information on arrival time, travel time, and financial benefits *(47)*. Otherwise, according to a survey in Delhi, India, the unpredictability of estimated travel time and waiting time will lead to the cancellation of 28.4% of requested trips *(48)*. Another study of 309 ride-sharing drivers found that users were upset by the length of their travel time *(49)*. A survey in China found that travel time was more valuable to tourists, and they were willing to pay more to shorten their travel time. Older residents are also likely to pay more to reduce waiting times *(50)*. In another study from China, by calculating the cost of travel time and waiting time, it was found that from 2016 to 2018, more than 1.7 billion hours of travel time were saved by ride-sharing users *(51)*. A study in India shows that for the people of Mumbai, travel comfort is more important than security and waiting time *(41)*. According to a survey of 4,365 people in the United States, reducing relative travel time by one minute per mile increased sharing by 33 percent *(42)*. A large portion of travel was canceled or rescheduled in a study on long travel duration in Canada during the COVID period (20). After the pandemic in China, we saw an improvement in the number of trips, although travelers tended to travel shorter distances *(18)*. But in Greece, the travel time became longer during COVID-19 *(22)*. In general, in the short term, COVID-19 has changed the departure time, mode, destination, and route choices *(25)*.

Method of Ridesharing Service

LDA *(52)* is a robust and popular topic model that presupposes communication between words and documents in a corpus represented by bag-of-words. LDA has been employed in a number of applications, including health (*53-54)*, e-petitions *(55)*, politics *(56)*, and the evaluation of social media strategy, in both long-length (e.g., abstracts) and short-length (e.g., tweets) datasets *(57)*. Pournarakis *(58)*, for example, utilized LDA for topic modeling of transportation services. This work created and implemented a Genetic Algorithm based on LDA to the task of sorting tweets into the various subjects, which outperformed the K-means clustering approach. Another pertinent study that uses Twitter data to examine ride-sharing services has been done, and the findings indicate that LDA topic modeling may well be able to extract the most prevalent topics from large datasets pretty quickly *(59)*.

In terms of sentiment analysis, BERT has lately gained popularity. It is intended to assist computers in understanding the sentiment of confusing language in the text by establishing context through the use of surrounding text. For instance, Sun et al. *(60)* developed an auxiliary sentence to transform (T)ABSA from a single sentence classification problem to a sentence pair classification task based on BERT. The outcome demonstrates that, using the SentiHood dataset, BERT-pair significantly beats other models in aspect detection and sentiment analysis. Language models historically could only read text input sequentially; they couldn't do both at once *(61)*. BERT is unique since it can concurrently read in both directions. The arrival of Transformers made this ability known as directionality possible *(62)*. The BERT framework, meanwhile, was pre-trained on text from Wikipedia and may be fine-tuned using a dataset of questions and answers. This method overcomes the supervised method's limitation on data volume and dataset transfer. In this work, the BERT was used as a reference model, and once the sentiment was retrieved, logistic regression was used for the sensitive and correlation analysis.

METHODOLOGY FRAMEWORK

The flowchart of the methodology is presented in Figure 1. The process has three parts: i) Data gathering and filtering. This part collects ridesharing time-related Tweets using the keyword search method of Twitter API. Then all tweets are cleaned and filtered based on error deleting, and text noise reduction methods. ii) Ride-sharing trip time-related topic modeling. This part extracted trip time-related keywords based on the word frequency and correlation analysis. Then the time-related texts are extracted based on those keywords. The trip time clusters and topics are modeled using LDA based on text distribution. To present the influence of passengers' characteristics and pandemics on the topics' changing, the topics' differences between each gender, age, and country group pre-and amid-pandemic are compared. iii) Sentiment analysis of ridesharing trip time. This part uses VADER and BERT to model the sentiment of each Twitter text, then, a more suitable model is chosen for further analysis based on model performance. In this part, the sentiment differences between pre-and amid-pandemic, gender groups, age groups, and countries are compared. Meanwhile, the difference in the gender groups pre- and amid-pandemic, age groups pre- and amid-pandemic, and USA/India pre- and amid-pandemic, are also compared. Then, the correlation and regression between each variable with sentiment are analyzed using the multi-logit regression model.

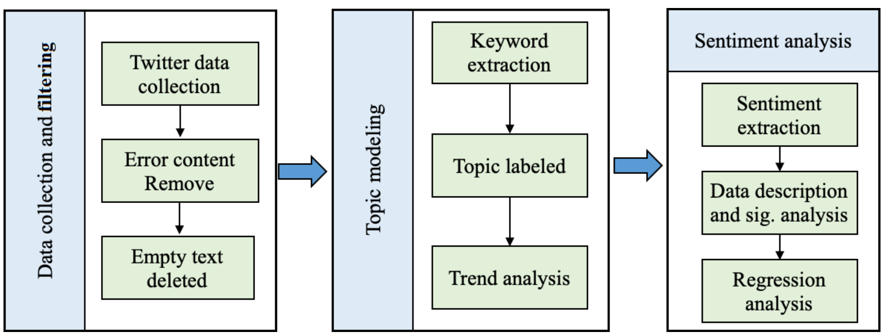


FIGURE 1 FLOWCHART OF THE METHODOLOGY FRAMEWORK

Data Collection and Filtering

The collection of tweets was done through an advanced Twitter scraping tool called Twitter Intelligence Tool *(63)*. The keywords "Uber Pool", "Uber Black", "Uber Comfort", "Uber X", "Lyft XL", "Lyft Shared", "Lyft Lux", "Lyft Black", "Ola KaaliPeeli", "Lyft Line", "BlaBlaCar Carpool", "Sride", "Ibibo Ryde", "Meru Carpool", "Ola Share", "Ola Carpool" and their derivative terms are also used to acquire the data in the form of text from 1 January 2019 to 04 May 2022. The data were selected through the following criteria: the location is the USA and India, and the text is English written; Duplicated texts were eliminated; Spam text was deleted using the detecting spammers method as *(64)*. Data issues, including missing information, no-sense characters, and noise in data, exist in the Twitter database. The no-sense characters, like emojis, emoticons, URL paths, numbers, punctuation marks, symbols, English stop words, non-alphabetical words, and tokens with less than one character, are removed from the sentences. Then, the dimensionality of the text is reduced based on Part of Speech Tagging method, and each sentence is changed into nouns, verbs, adverbs, and adjectives. The words are stemmed based on the Snowball method. Descriptions of the data are provided in **Table 1.**

TABLE 1 DESCRIPTION OF THE TWITTER DATA

Data type	Description
Users' characteristics	Gender, age, user name, user ID, followers.
Timestamp	The timestamp of each tweet publishes.
Location	The county and location of the user.
Tweet	The content of the tweet, the situation of the tweet (rewrite or not).
Sample of the tweet before and after the filter	Before filter: @Uber### I like 😊 and miss # uberpull much, prices are odeeeeeeeeer cheaper #uber. https://t.co/OOLOYLexyC After filter: I like and miss uberpool, these prices are cheaper.

Topic Modeling of Ridesharing Service

Time-related tweet extraction: The main focus of this paper was the ridesharing trip time. Therefore, the tweets related to our research are kept for further analysis. To extract the related tweets, the high-frequency words of the original dataset are extracted based on the flash-text model in Python. 2000 high-frequency words are collected from the dataset. Then, the words related to ride-sharing trip time, such as time cost, wait, tonight, etc., are selected from the 2000 words, and used for selecting the ridesharing trip time-related tweets. After the data selection, 6130 texts related to ridesharing trip time from 5593 Twitter users are kept for analysis. In this paper, the pandemic event threshold is chosen as March 2020. The data pre-pandemic includes 3621 texts, and amid-pandemic includes 2509 texts. When it comes to the comparison of gender, 3268 text belongs to males and 2862 to females. The mean users' age is 29.12, with an S.D. of 6.07. This paper chooses the age of 35 (only using this threshold, the data show the significance) to divide the users into younger and older categories (number of tweets belonging to younger = 4030, older = 2100). Note that the countries in this paper include the USA and India; 5330 text belongs to the USA; and 800 to India.

Keywords Extraction: This paper uses the LDA to extract ridesharing trip time keywords and clusters. The LDA is a dynamic statistical model that enables unobserved groups to explain why some areas of the data are similar by sets of observations. For instance, if observations are words gathered into documents, it is hypothesized that each document is a combination of a few different subjects and that each word's appearance is related to one of those topics.

To acquire the underlying structure of latent topics in the dataset based on LDA, Python's Gensim library *(65)* is utilized, which allows the execution of the algorithm with multi-threads, resulting in effectiveness and fast calculation. This paper concentrate on exploring the topic difference between the country, gender, and age groups pre- and amid-pandemic. Therefore, the data is divided into 7 groups (all tweets, USA user tweets, India user tweets, male user tweets, female user tweets, younger user tweets, and older user tweets). Then, 14 documents, which include 7 groups, pre-, and amid-pandemic from 2019.01 to 2022.05, are used for topic modeling based on the LDA model (5 clusters based) *(66)*. Finally, 35 clusters, each cluster has 20 keywords, are collected.

Topic labeling based on keywords combination: The result of LDA does not provide the documents' topics but only a distribution of probabilities for the different topics. Some studies explored the usage of several clustering techniques to arrange and group keywords into topics that are predefined, such as the Moreno *(67)* method, which involves the use of a K-mean clustering algorithm and a Genetic Algorithm combined with a local convergence algorithm to integrate the topics based on LDA results. However, those methods have the same disadvantage in grouping the topics; researchers still need to label the topics from the clustering result manually. As the main problem of the supervised method, researchers need to pre-define the number of topics, which reduces the informational value of text and the hidden variables of topics.

This paper proposes a new topic label method that includes 3 steps: cluster ordering, topic generation, and dataset labeling. The first step re-orders the clusters based on the coherence of each cluster. The coherence measures the score of a single cluster by measuring the degree of semantic similarity between high-scoring words in the cluster. The high score means the cluster has good performance in the model. Therefore, five clusters are sorted based on the enhancement of the coherence score. The topic generation step is the main step. The three most discussed topics in the latest research *(2)* including trip time cost, waiting time, and trip happen time, are chosen as reference topics at first in this step. Each keyword in the clusters is scored based on the correlation with three topics based on a 10-Likert scale (1–10). A 10 score means the keyword has a high probability of belonging to a certain topic. Once the keyword has got 1 score in all topics, a new topic will be created based on the meaning of this keyword. For example, the word "pandemic" does not belong to any topic. Therefore, the pandemic is created as the fourth topic. In the dataset labeling step, after all of the keywords are scored, topics are generated and labeled as topic-1..., topic-4. Then, each cluster is labeled as a topic based on the frequency of keywords belonging. Take the amid-pandemic, male, cluster 1 as an example, as can be seen in **Figure 2**, each keyword in the cluster is transferred to the topic label, and the cluster is labeled as topic 2 based on the keywords' belonging frequency. Note that only meaningful keywords are kept for topic modeling, and words such as can, much, etc., which have no sense are deleted. After each cluster is labeled, the difference between each group pre-and amid-pandemic is compared based on the hot topic difference and topic change trend.

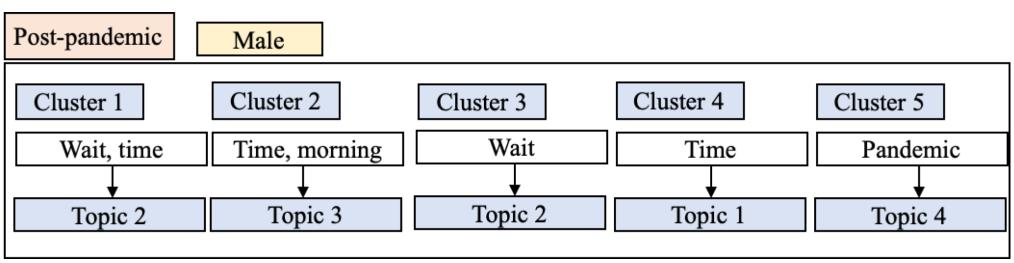


FIGURE 2 EXAMPLE OF TOPIC LABELING

Sentiment Analysis

Sentiment analysis can be used to classify the polarity of a given document; it can assign a score to a document to indicate whether the expressed opinion is positive, negative, or neutral. In this paper, the VADER and BERT model is used for extracting the sentiment of each text, then, the data description and significance pre- and amid-pandemic are analyzed. The logistic regression model is used for interpreting the correlation between passengers' characteristics, countries, pre-and amid-pandemic, and sentiment, and construing the regression model based on these variables. Therefore, the VADER and BERT models and logistic regression, which are related to this paper, are introduced as follows.

VADER model: VADER is a lexicon and rule-based sentiment analysis tool that is adjusted to sentiments conveyed on social media. VADER method employs a variety of A sentiment lexicon is a collection of lexical features (e.g., words) that are typically classified either positive or negative depending on their semantic orientation. VADER not only informs about the Positivity and Negativity scores but also of how positive or negative a sentiment is. The VADER lexicon operates remarkably well in the are of social media. The correlation coefficient illustrates that VADER (r = 0.881) operates as well as actual human raters (r = 0.888) at matching ground truth (collected group mean from 20 human raters for sentiment intensity of each tweet). In this article, the VADER model is constructed based on the Vader module installed in Python. Every lexical feature that had a non-zero mean rating, and whose standard deviation was less than 2.5 as determined by the aggregate of ten independent raters is kept. The sentiment intensity on a scale from –4 to +4. For instance, the term "okay" holds a positive valence of 0.9, "good" is at 1.9, and "great" is at 3.1, while "horrible" is at –2.5, the frowning emoticon ":(" is at –2.2, and both "sucks" and "sux" are at –1.5. More details and is available for download and can be found in *(68)*.

BERT Model: The sentiment model achieves a higher accuracy of 92% for sentiment when using the cased version of BERT analysis *(61)*. The model is combined of one or more input sequences, added with an initial token "CLS" and a token "SEP" to divide segments. Word embeddings are utilized to represent every token, concatenated with position embeddings and segment embeddings. Every model consists of two sublayers, one is a multi-head attention mechanism with A heads and hidden size H; the second is a fully connected layer with a position-wise feed-forward. The sublayer input is added to the normalized sum of each sublayer output. Two vectors S and E (which will be learned during fine-tuning) with shapes (1x768) have been defined. Afterwards a dot product of these vectors with the second sentence's output vectors from BERT was obtained, giving us some scores. SoftMax was applied over these scores to get probabilities. The purpose of training is the sum total of the log-likelihoods of the correct start and end positions.
In this paper, the BERT base model employs L = 12, A = 12, and H = 768. normally, the BERT takes an input of a sequence of no more than 512 tokens (which are lowered here to 128 dues to the small length of tweets). In this paper, the model parameter is set as learning rate: 0.0001, batch-size: 8, epochs 10, max-seq-length: 128.

Logistic Regression Model
Logistic regression is a commonly used model in transportation studies. Therefore, for brevity, logistic regression modeling is not provided in this paper. For more information, please see Agersti *(69)*.

RESULTS

Topic Modeling Performances and Results

Keywords distribution and results: Tweets related to ridesharing trip time are collected from the original dataset. After collection and ridesharing trip time dataset construction, the frequency of tweets in each month is shown in **Figure 3**. The results show that the number of tweets decreased significantly during 2020.01-04. This decrease occurred with a high probability due to the number of ridesharing users decreasing during the pandemic. Then, the number of tweets increased after 2021.02, and the trend shows a stable wave. This means that people share their problems and opinions on other transport platforms or other topics amid-pandemic at first, then come back to the ridesharing trip time topic.

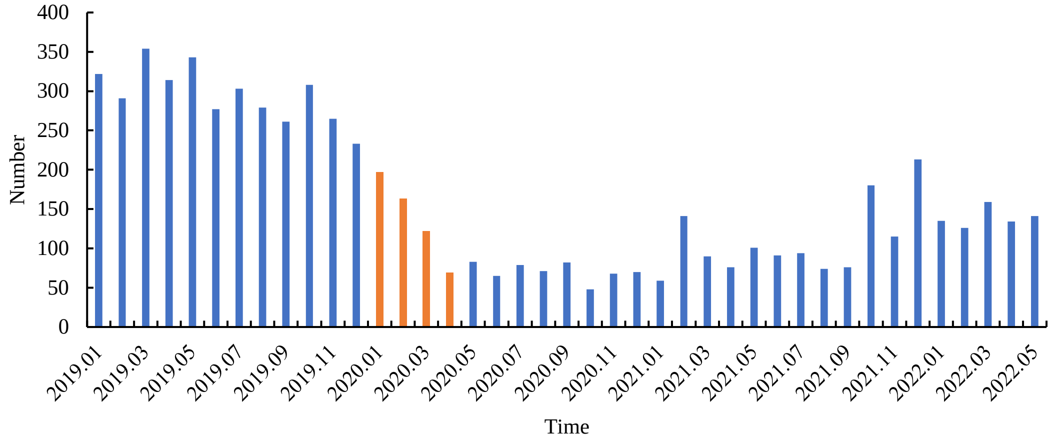


FIGURE 3 THE TWEET FREQUENCY EACH MONTH

Topic modeling and development trend analysis: As mentioned in the methodology section, to perform the task of extracting the most discussed topics in each group pre-and amid-pandemic, the clusters of seven groups (all-USA-India-male-female-younger-older) are extracted based on the LDA. Then, the clusters are sorted based on the coherence score, and the topics are labeled based on the proposed method. As a result, four topics are labeled. The details of the topic and its content can be seen in **Table 2**.

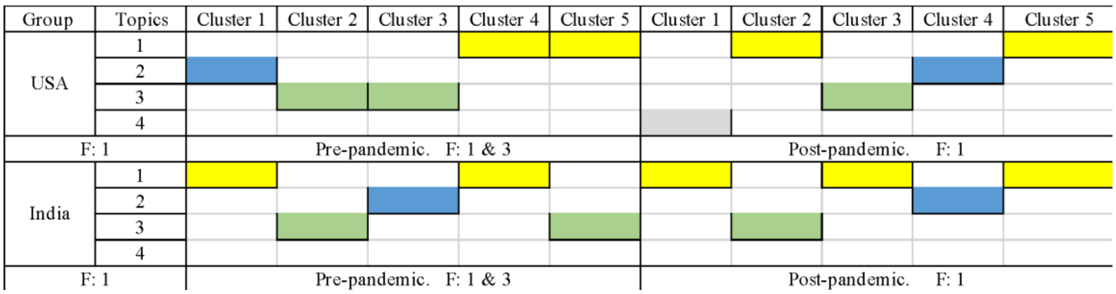
TABLE 2 DESCRIPTION OF LABELED TOPICS AND THEIR CONTENTS

Item	Label	Content	Description
Ridesharing trip time	1	Wait time	Wait time for the car
	2	Time cost	The time cost from entering the car to ending the trip
	3	Trip happen time	Trip time of day
	4	Pandemic	Topic related to pandemic

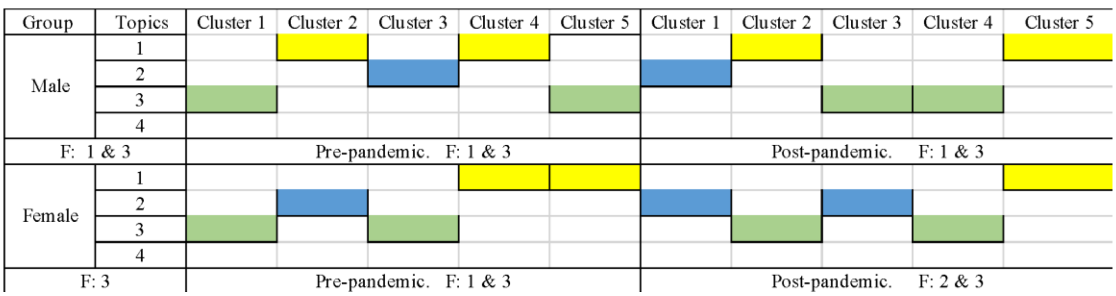
The topic distribution of five clusters is marked in the graph as shown in **Figure 4.** The left side of the graph shows the group and ridesharing trip time labels. The center of the graph shows the distribution of the topics. Each topic is assigned a color. In the graph, the F means the frequency of the topic in all periods, pre-pandemic and amid-pandemic. As a result, the topic of trip happens time (3) is the hottest topic in all periods, and the hot topic distribution has a difference pre-and amid-pandemic. The topic of "wait time" (1) gains more attention amid-pandemic. It indicated that passengers more care about the trip time happening, such as in the morning and other times. Meanwhile, the wait time may increase with the pandemic, therefore gaining more concern from passengers amid-pandemic.

In the country groups, both users from the USA and India paid attention to the topic "1". Users in the United States were more concerned about the pandemic than in India. In both groups, they more cared about topics "1" and "3" pre-pandemic, then decreased their concern about topic "3" amid-pandemic. It indicated that customers more cared about the service of the driver (pickup time) amid-pandemic. In the gender groups, both males and females paid attention to the topic "3", and males more cared about the topic "1" than females. Males' hottest topics are "1 & 3", pre-and amid-pandemic. Unlike males, the hottest topic among females is "1 & 3" pre-pandemic, which has changed to "2 & 3" amid-pandemic. It indicates that females are more concerned about the time cost amid-pandemic. In age groups, both of them have the hot topic "2", but the younger ones have more concern about topic "3". When it comes to the difference between pre-and amid-pandemic, both groups have significant differences. In the younger group, topic "2" gained more attention amid-pandemic than topic "3". It indicated that younger people are more concerned about the time cost amid-pandemic. When it comes to the older, the topic "2" is not popular during the pandemic, which shows that the older pay less attention to time cost amid-pandemic.

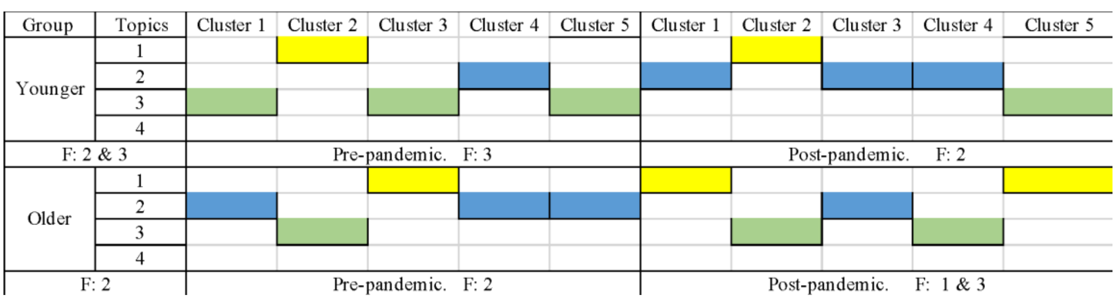
a) The trend graph of topics based on all tweet data



b) The trend graph of topics in the country group



c) The trend graph of topics in gender groups



d) The trend graph of topics in the age group

FIGURE 4 THE GRAPH OF RIDESHARING TRIP TIME TOPIC MODELING

Sentiment analysis of ridesharing service

Sensitive and significant analysis: The BERT model is approximated to the ground truth with high accuracy (87%) than the VADER model (61%). To further verify the applicability of the model, 400 tweets' sentiments are checked manually. The result shows that the BERT model still has high accuracy (72.3%) and less Mean Absolute Error (0.15) than VADER (accuracy = 52%, Mean Absolute Error =

0.32). The result means the BERT model has good performance in dealing with the sentiment analysis problem of Twitter data. Therefore, the sentiment result of the BERT model is used for further analysis. The sentiment result of BERT associated with the time series is shown in **Figure 5**. Based on the sentiment values, it is observed that there are more negative tweets than positive ones. This result means that users address the customer service platform (@ridesharing time) to tweet about complaints and problems with a negative expression with more frequency than using positive expressions. Meanwhile, the percentage of people with a positive attitude increases at the beginning of the pandemic and then decreases amid-pandemic. This result may be that the ridesharing company has put some discount policies for customers to respond to the pandemic at the beginning, which enhances the customer's positive sentiment.

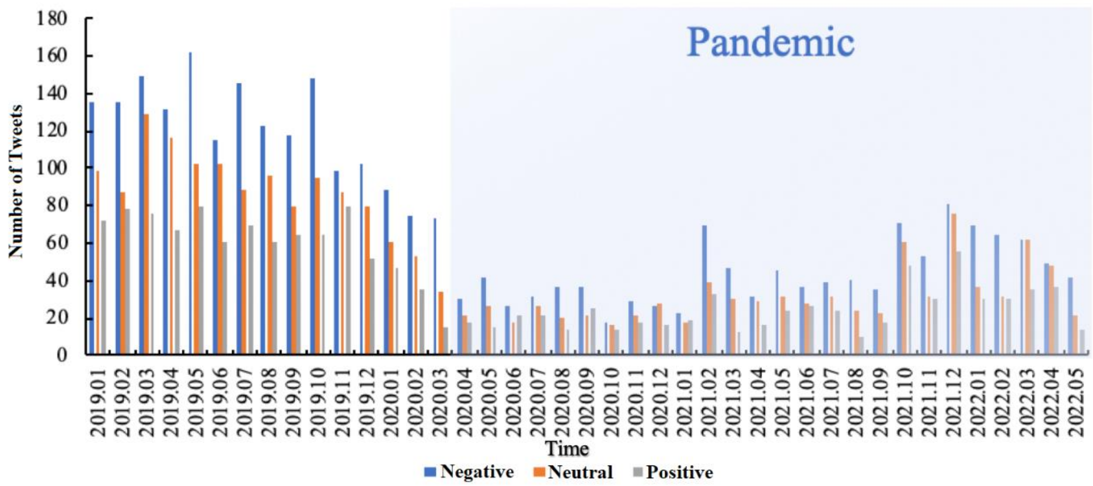
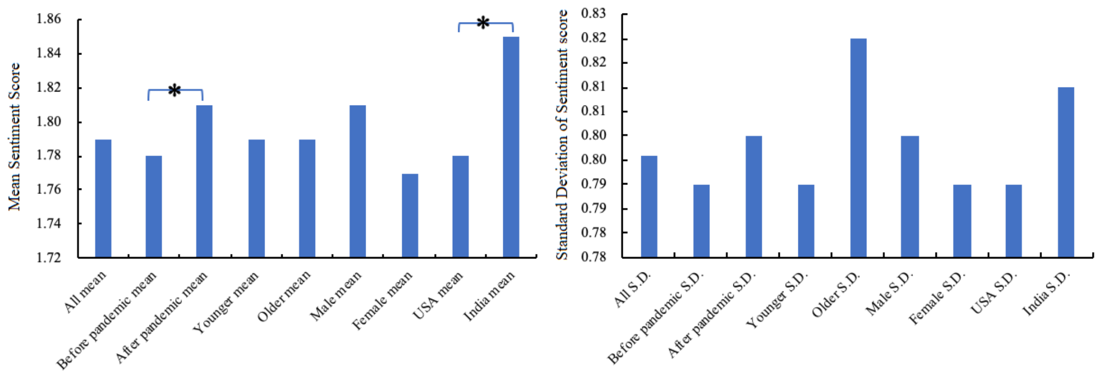
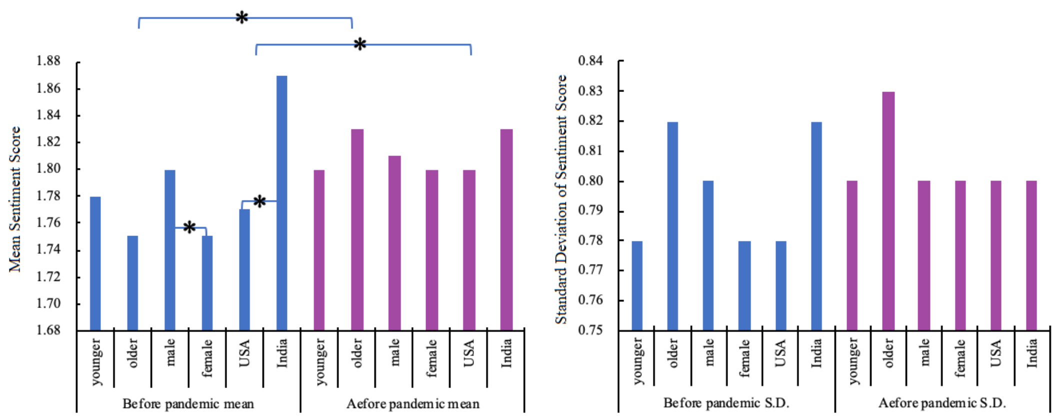


FIGURE 5 THE VOLUME OF RIDESHARING SENTIMENTS ASSOCIATED WITH THE TIME SERIES

To further analyze the sentiment performance, the sentiment is assigned values based on the positive enhancement, as the negative is 1, the neutral is 2, and the positive is 3. To explore the sentiment differences between groups (pandemic, gender, and age), the mean and standard deviation (S.D.) of each group are compared. As can be seen in **Figure 6 (a)**, the sentiment in the different pandemic periods and countries shows a significant difference. When comparing the mean of each group, customers show a more positive attitude amid-pandemic, and male and Indian customers have a more positive attitude toward ridesharing trip time. Meanwhile, as a result of S.D., the customer always keeps a more stable sentiment pre-pandemic. Similarly, females, younger people, and American customers have a more stable attitude toward ridesharing. Then, the difference between each group's pre-and amid-pandemic sentiment also is analyzed. As can be seen in **Figure 6 (b)**, the result shows that the sentiment of gender and country groups is significantly different pre-pandemic, but there are no significant differences in each group amid-pandemic. Meanwhile, female and American passengers are more positive about the ridesharing service in all periods. The older have a more positive attitude per-pandemic than the younger, but the opposite is true amid-pandemic. As a result of S.D., younger, males and American passengers keep a more stable attitude than others pre-pandemic, but only younger people keep a stable trend amid-pandemic. *Note that: * means that the difference was statistically significant at the significance level of 5% (0.01< p-value <0.05).*



a) The difference pre- and amid-pandemic, and customs' characteristics



b) The difference between customs' characteristics pre- and amid-pandemic

FIGURE 6 THE DESCRIPTION AND SIGNIFICANT ANALYSIS OF SENTIMENT IN EACH GROUP

Ridesharing trip time sentiment multi-logit regression model: This paper analyzed the correlation between each customer's characteristics, country, pre-and amid-pandemic variables, and sentiment to model the relationship between these variables. There is a significant correlation between sentiment and country (0.028**), but no correlation exists between sentiment and other variables. It indicated that the sentiment was performed differently by users from the United States and India. Then, the regression model was modeled based on the multi-logit regression model, which was implemented by using the MATLAB built-in algorithm *(70)*. Four parameters are used for validating the model performance: Log-Likelihood Ratio, X^2, the goodness of fit test, and model significance. As a result, the model shows good performance in modeling sentiment based on the higher significance (sig. = 0.03), lower error (Log-Likelihood Ratio = 177.24, X^2 = 12.31), and high model fitness (goodness of fit test = 0.623). **Table 3** shows the result of the sentiment regression model; the pandemic and country are the main factors influencing the sentiment. Amid-pandemic, the sentiment is more positive (OR = 1.05, P<0.05), and the USA's passengers are more positive than India's (OR = 1.01, P<0.05).

TABLE 3 RESULT OF THE RIDESHARING REGRESSION BASED ON THE MULTI-LOGIT MODEL

Step	Items	B	Stad. E.	Wald	Freedom	Sig.	Exp(B)	95% CI
1	Intercept	-0.48	0.15	10.19	1.00	0.00		

	Pandemic	0.22	0.11	3.67	1.00	0.03	1.01	0.89	1.13
	Gender	0.04	0.06	0.55	1.00	0.89	1.04	0.93	1.18
	Age	-0.77	0.09	0.75	1.00	0.79	0.92	0.77	1.10
	Country	0.33	0.12	4.45	1.00	0.02	1.12	1.01	1.23
2	Intercept	-0.49	0.16	9.65	1.00	0.02	-	-	-
	Pandemic	0.12	0.13	4.66	1.00	0.03	1.05	1.01	1.15
	Gender	0.01	0.12	0.01	1.00	0.97	1.00	0.79	1.27
	Age	0.09	0.67	2.12	1.00	0.12	1.11	0.96	1.26
	Country	-0.18	0.09	3.48	1.00	0.02	1.01	0.99	1.22

CONCLUSIONS

This paper describes how to use an LDA to model the ridesharing trip time topic, as well as how to extract and analyze sentiment using the BERT and multi-logit models. Four topics including wait time, time cost, trip happening time, and a pandemic, are extracted from the original tweets. The paper compares the difference in ridesharing time topic distribution between gender, age, and country group before and after the pandemic. Meanwhile, the sentiment of each tweet in each group based on time series is extracted using the BERT model. The result shows the model has good performance. Then, the significance and correlation of sentiment with each variable are analyzed; the sentiment regression is modeled based on the multi-logit model, and the main factors influencing sentiment are found. The major findings are as follows:

1. The distribution of topics in all groups shows that topics of trip happening time raise high concerns. Meanwhile, the topic of waiting time has gained more attention amid-pandemic. It indicated that passengers more care about the trip time happening, such as in the morning and other times. Meanwhile, the waiting time may increase with the pandemic, therefore gaining more concern from passengers.
2. The USA and India have the same topics concerning pre-and amid-pandemic. The male is more concerned about the waiting time than the female. Males' and females' hottest frequency of topics is waiting and happening time of the trip, before the pandemic. Females are more concerned about the time cost amid-pandemic. The younger groups are more concerned about trips happening time before the pandemic and more concerned about the time cost after the pandemic. When it comes to the older, waiting time is a hot topic before the pandemic but not during the pandemic; they are more concerned with time cost and the happening time of the trip amid-pandemic.
3. The sentiment of the pre-and amid-pandemic and countries' groups showed a significant difference. Customers show more positivity amid-pandemic, and in the USA, customers have a more positive attitude toward ridesharing trip time. Then, the sentiment of gender and country groups has a significant difference, which is that females and American passengers are more positive about the ridesharing trip time in pre-pandemic. But no significant difference in each group amid-pandemic. When comparing the differences between pre-and amid-pandemic in each group, the sentiments of older and American passengers are shown to have significant differences. Positive sentiment is increased in those groups amid-pandemic, and the pandemic has different degrees of impact on individuals.
4. As a result of the regression model, the pandemic and country are the main factors influencing the sentiment. In the amid-pandemic, the sentiment is more positive, and the USA's passengers are more positive than India's.

This study provides a method for ridesharing trip time topic modeling and sentiment analysis. The framework of this paper considers topic occurrence and trend change, sentiment time series variables, which has been hardly done before. Results from this paper can be used for the ridesharing industries' in-service topic modeling and sentiment analysis.

In future studies, additional research should be carried out on this topic to validate the modeling approach and the algorithm. More importantly, sentiment analysis should be associated with topic modeling. Emotion analysis can also be used to enhance the detailed analysis of sentiment. More variables such as sadness, happiness, wiled, etc., can be extracted based on the future study.

REFERENCES

1. Gehrke, S. R. (2020, May 25). Uber service area expansion in three major American cities. Journal of Transport Geography, 86.
2. Shaheen, S., Totte, H., & Stocker, A. (2018). Future of Mobility White Paper. UC Berkeley.
3. M. Barrios, J., Hochberg, Y., & Yi, H. (2020). The cost of convenience: Ridehailing and traffic fatalities. NATIONAL BUREAU OF ECONOMIC RESEARCH.
4. Tirachini, A., & del Río, M. (2019, October). Ride-hailing in Santiago de Chile: Users' characterisation and effects on travel behaviour. Transport Policy, 82, 46-57.
5. Tirachini, A. (2020). Ride-hailing, travel behaviour and sustainable mobility: an international review. Transportation, 47, 2011–2047.
6. Clewlow, R., & Mishra , G. S. (2017, October). Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States. UC Davis Institute of Transportation Studies.
7. Botsman, R. (2017). Who Can You Trust?: How Technology Brought Us Together and Why It Might Drive Us Apart. New York: PublicAffairs.
8. Kong, Y., Wang, Y., Hajli, S., & Featherman, M. (2020, July). In Sharing Economy We Trust: Examining the Effect of Social and Technical Enablers on Millennials' Trust in Sharing Commerce. Computers in Human Behavior, 108.
9. Raut, A., Bhosale, R., Avhad, K., Awari, M., & Jadhav, S. (2020, March). A Survey on: Real time Smart Car Pooling and Ride Sharing System using Android Application. INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS, 7(1), 593-597.
10. Gleave, S. D. (2007). London Travel Report. London: Transport for London.
11. Li, Z., Hong, Y., & Zhang, Z. (2016). An empirical analysis of on-demand ride sharing and traffic congestion. Thirty Seventh International Conference on Information Systems.
12. Du, J., & A. Rakha, H. (2020, October 30). COVID-19 Impact on Ride-hailing: The Chicago Case Study. Transport Findings.
13. Beck, M. J., & Hensher, D. A. (2020, December). Insights into the impact of COVID-19 on household travel and activities in Australia – The early days of easing restrictions. Transport Policy, 99, 95-119.
14. Smith, C. (2022, April 1). Uber Statistics, User counts, Facts & News (20220. Retrieved from expandedramblings: https://expandedramblings.com/index.php/uber-statistics/
15. Burgueño Salas, E. (2021, October 20). Active riders of Lyft from 2016 to 2021, by quarter. Retrieved from statista: https://www.statista.com/statistics/916456/lyft-number-of-rides/
16. Kadali, B. R., Gadiraju, S. S., & Gaddam, H. K. (2022). Analysis of Travelers' Risk Perceptions About Public Transit Systems During COVID-19 Pandemic. Transportation Research Record, 1-12.
17. Oluyede, L., Cochran, A. L., Wolfe , M., Prunkl, L., & McDonald, N. (2022). Addressing transportation barriers to health care during the COVID-19 pandemic: Perspectives of care coordinators. Transportation Research Part A: Policy and Practice, 159, 157-168.

18. Pan, Y., & He, S. Y. (2022). Analyzing COVID-19's impact on the travel mobility of various social groups in China's Greater Bay Area via mobile phone big data. Transportation Research Part A: Policy and Practice, 159, 263-281.
19. Iio, K., Guo, X., Kong, X., Rees, K., & Wang, X. B. (2021). COVID-19 and social distancing: Disparities in mobility adaptation between income groups. Transportation Research Interdisciplinary Perspectives, 10, 100333.
20. Fatmi, M. R., Thirkell, C., & Hossain, M. S. (2021). COVID-19 and travel: How our out-of-home travel activity, in-home activity, and long-distance travel have changed. Transportation Research Interdisciplinary Perspectives, 10, 100350.
21. Schmidt, K., Sieverding, T., Wallis, H., & Matthies, E. (2021). COVID-19 – A window of opportunity for the transition toward sustainable mobility? Transportation Research Interdisciplinary Perspectives, 10, 100374.
22. Politis, I., Georgiadis, G., Papadopoulos, E., Fyrogenis, I., Nikolaidou, A., Kopsacheilis, A,. Verani, E. (2021). COVID-19 lockdown measures and travel behavior: The case of Thessaloniki, Greece. Transportation Research Interdisciplinary Perspectives, 10, 100345.
23. Liu, S., & Yamamoto, T. (2022). Role of stay-at-home requests and travel restrictions in preventing the spread of COVID-19 in Japan. Transportation Research Part A: Policy and Practice, 159, 1-16.
24. Tiikkaja, H., & Viri, R. (2021). The effects of COVID-19 epidemic on public transport ridership and frequencies. A case study from Tampere, Finland. Transportation Research Interdisciplinary Perspectives, 10, 100348.
25. de Palma, A., Vosough, S., & Liao, F. (2022). An overview of effects of COVID-19 on mobility and lifestyle: 18 months since the outbreak. Transportation Research Part A: Policy and Practice, 159, 372-397.
26. Rafiq, R., McNally, M. G., Uddin, Y. S., & Ahmed, T. (2022). Impact of working from home on activity-travel behavior during the COVID-19 Pandemic: An aggregate structural analysis. Transportation Research Part A: Policy and Practice, 159, 35-54.
27. Ye, Q., Ozbay, K., Zuo, F., & Chen, X. (2021). Impact of Social Media on Travel Behaviors during the COVID-19 Pandemic: Evidence from New York City. Transportation Research Record, 1-20.
28. Habib, M. A., & Anik, M. H. (2021). Impacts of COVID-19 on Transport Modes and Mobility Behavior: Analysis of Public Discourse in Twitter. Transportation Research Record, 1-14.
29. Nian , G., Peng , B., Sun, D. J., Ma , W., Peng , B., & Huang, T. (2020). Impact of COVID-19 on Urban Mobility during Post-Epidemic Period in Megacities: From the Perspectives of Taxi Travel and Social Vitality. Sustainability, 12(19), 7954.
30. Manley, E., Ross, S., & Zhuang, M. (2021). Changing Demand for New York Yellow Cabs during the COVID-19 Pandemic. Finding.
31. Xu, A., Chen , J., & Liu, Z. (2021). Exploring the Effects of Carpooling on Travelers' Behavior during the COVID-19 Pandemic: A Case Study of Metropolitan City. Sustainability, 13(20), 11136.
32. Basnak, P., Giesen, R., & Muñoz, J. C. (2022). Estimation of crowding factors for public transport during the COVID-19 pandemic in Santiago, Chile. Transportation Research Part A: Policy and Practice, 159, 140-156.
33. Halvorsen, A., Wood, D., Jefferson, D., Stasko, T., Hui, J., & Reddy, A. (2021). Examination of New York City Transit's Bus and Subway Ridership Trends During the COVID-19 Pandemic. Transportation Research Record, 1-14.
34. Teixeira, J. F., Silva, C., & e Sá, F. M. (2022). The role of bike sharing during the coronavirus pandemic: An analysis of the mobility patterns and perceptions of Lisbon's GIRA users. Transportation Research Part A: Policy and Practice, 159, 17-34.
35. Tirachini, A., & Cats, O. (2020). COVID-19 and public transportation: Current assessment, prospects, and research needs. Journal of Public Transportation, 22(1), 1-21.

36. Zhang, W., & Liu, L. (2022). Exploring non-users' intention to adopt ride-sharing services: Taking into account increased risks due to the COVID-19 pandemic among other factors. Transportation Research Part A: Policy and Practice, 180-195.
37. Young, M., & Farber, S. (2019). The who, why, and when of Uber and other ride-hailing trips: An examination of a large sample household travel survey. Transportation Research Part A: Policy and Practice, 119, 383-392.
38. Kooti, F., Grbovic, M., Aiello, L. M., Djuric, N., Radosavljevic, V., & Lerman, K. (2017). Analyzing Uber's Ride-sharing Economy. WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion, (pp. 574–582).
39. Susilawati, & Lim, T. S. (2021). A study of the scheduling effect on shared autonomous vehicles adoption. Transportation Research Interdisciplinary Perspectives, 10, 100394.
40. Dinhobl, M., Fasanando, S., Dudensing, R., & Mjelde, J. (2022). Impact of COVID-19 on Texas Rural Transit Districts with Emphasis on their Older Adult Riders. Transportation Research Record, 1-12.
41. Shah, P., Varghese, V., Jana, A., & Mathew, T. (2020). Analysing the ride sharing behaviour in ICT based cab services: A case of Mumbai, India. Transportation Research Procedia, 48, 233-246.
42. Middleton, S. R., Schroeckenthaler, K. A., Gopalakrishna, D., & Greenberg, A. (2021). Effect of Price and Time on Private and Shared Transportation Network Company Trips. Greenberg, 2675(8), 458–467.
43. Monchambert, G. (2020). Why do (or don't) people carpool for long distance trips? A discrete choice experiment in France. Transportation Research Part A: Policy and Practice, 132, 911-931.
44. CIARI , F., & AXHAUSEN, K. W. (2012). Choosing carpooling or car sharing as a mode: Swiss stated choice experiments. TRB 2012 Annual Meeting.
45. Jiancheng, L., Weimin, T., Y, S. W., & Yao, L. (2018). Ride-sharing with travel time uncertainty. Transportation Research Part B: Methodological, 118, 143-171.
46. Li, Y., & Chung, S. H. (2020). Ride-sharing under travel time uncertainty: Robust optimization and clustering approaches. Computers & Industrial Engineering, 149, 106601.
47. Agatz, N., Erera, A., Savelsbergh, M., & Wang, X. (2012). Optimization for dynamic ride-sharing: A review. European Journal of Operational Research, 223(2), 295-303.
48. Shah, D., Kumaran, A., Sen, R., & Kumaraguru, P. (2019). Travel Time Estimation Accuracy in Developing Regions: An Empirical Case Study with Uber Data in Delhi-NCR. WWW '19: Companion Proceedings of The 2019 World Wide Web Conference, (pp. 130–136).
49. Morris, E. A., Zhou, Y., Brown, A. E., Khan, S. M., Derochers, J. L., Campbell, H., . . . Chowdhury, M. (2020). Are drivers cool with pool? Driver attitudes towards the shared TNC services UberPool and Lyft Shared. Transport Policy, 94, 123-138.
50. Chen, J. M., de Groote, J., Petrick, J. F., Lu , T., & Nijkamp, P. (2020). Travellers' willingness to pay and perceived value of time in ride-sharing: an experiment on China. Current Issues in Tourism, 23(23), 2972-2985. doi:
51. WANG, B., SHAO, Y., & MIAO, M. (2022). A SOCIAL WELFARE ESTIMATION OF RIDE-SHARING IN CHINA: EVIDENCE FROM TRANSACTION DATA ANALYSIS. Technological and Economic Development of Economy, 28(2), 419–441.
52. Mcauliffe, J., & Blei, D. (2007). Supervised topic models. Advances in neural information processing systems, 20.
53. Tufts, C., Polsky, D., Volpp, K. G., Groeneveld, P. W., Ungar, L., Merchant, R. M., & Pelullo, A. P. (2018). Characterizing tweet volume and content about common health conditions across Pennsylvania: retrospective analysis. JMIR Public Health and Surveillance, 4(4), e10834.
54. Karami, A., Webb, F., & Kitzie, V. L. (2018). Characterizing transgender health issues in twitter. Proceedings of the Association for Information Science and Technology, 55(1), 207-215.

55. Karami, A., & Shaw, G. (2019). An exploratory study of (#) exercise in the Twittersphere. iConference 2019 Proceedings.
56. Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? Information Processing & Management, 54(6), 1292-1307.
57. Karami, A., Bennett, L. S., & He, X. (2018). Mining public opinion about economic issues: Twitter and the us presidential election. International Journal of Strategic Decision Sciences (IJSDS), 9(1), 18-28.
58. Pournarakis, D. E., Sotiropoulos, D. N., & Giaglis, G. M. (2017). A computational model for mining consumer perceptions in social media. Decision Support Systems, 93, 98-110.
59. Karami, A., & Collins, M. (2018). Social Media Analysis for Organizations: US Northeastern Public and State Libraries Case Study.
60. Sun, C., Huang, L., & Qiu, X, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence", NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference (2019).
61. Adoma, A. F., Henry, N. M., & Chen, W. (2020, December). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). (pp. 117-121). IEEE.
62. G. Henry, "Improved sentiment analysis using a customized distilbert NLP configuration", Advances in Engineering: An International Journal (ADEIJ), Vol.3, No.2
63. Twitter Intelligence Tool (TWINT). Available online: https://github.com/twintproject/twint (accessed on 21 June 2021).
64. Gheewala, S., & Patel, R. (2018, February). Machine learning based Twitter Spam account detection: a review. In 2018 Second International Conference on Computing Methodologies and Communication (ICCMC). (pp. 79-84). IEEE.
65. Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks.
66. Moreno, A., & Iglesias, C. A. (2021). Understanding Customers' Transport Services with Topic Clustering and Sentiment Analysis. Applied Sciences, 11(21), 10169.
67. Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).
68. G. Henry, "Improved sentiment analysis using a customized distilbert NLP configuration", Advances in Engineering: An International Journal (ADEIJ), Vol.3, No.2
69. Agresti, A., An Introduction to Categorical Data Analysis. 2007. John Wiley and Sons Inc.
70. Mathworks. Multinomial logistic regression. (2022) Available online: https://nl.mathworks.com/help/stats/mnrfit.html.