AN ANALYSIS OF RIDESHARING COST PRE- AND AMID-COVID-19 PANDEMIC -- THE APPLICATION OF ADVANCED TEXT MINING TECHNIQUE -- THE USA VS INDIA 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 Khabushani, 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

Amid-pandemic, ridesharing companies provide a lot of discount policies to face the influence of the pandemic, but those policies have different effects in various countries and user groups. The Latent Dirichlet Allocation is used for topic modeling, then the distribution of price topics is extracted and labeled in topic graphs, each graph represents the difference of specific groups` (divided based on users` characteristics and countries) topics between pre-and amid-pandemic eras. The result of topic modeling shows that topics have a significant difference between the USA and India. Results from this paper can be used for ridesharing industries in enhancing the service and providing a suitable price standard amid-pandemic.

Keywords: ridesharing, travel cost, topic modeling, sentiment analysis, Twitter data, COVID-19

INTRODUCTION

Overview

With the emergence of ride-hailing and ridesharing systems, a huge transformation occurred in transportation and travel behaviors. A ridesharing system can reduce travel time, contribute to air quality and reduce pollution, lead to a reduction in Vehicle Miles Traveled (VMT), and support economic growth *(1)*. In fact, people reduce travel costs by sharing their vehicles with others and becoming travel companions *(2)*. In the United States, the Uber and Lyft platforms are among the largest programs of this type of service, of which Uber has the largest share of users *(3)*. If these systems cannot meet the demands and preferences of users, such as low cost, short travel time, and boarding and disembarking time, no one may use them in the future *(4)*. Yan et al. consider ride matching and dynamic pricing as two problems in ridesharing systems and believe that these two problems cause the system to lose its efficiency *(5)*.

WHO reported that until mid-July 2022, COVID-19 caused the deaths of more than 6.3 million people and nearly 558 million infected people worldwide *(6)*. The disease has affected the global economy as well as transportation, causing changes in departure times, mode selection, travel destinations, and route selection, and VMT has also decreased due to the spread of telecommuting during the COVID-19 era, but the total number of trips has increased. Cost and convenience, which played a key role in modality selection decisions, are replaced by a reduced risk of infection. Such behavior could return travel demand to private cars, cycling, and even walking amid-COVID-19, as seen during the pandemic. Therefore, the public transportation sector, ridesharing, and other emerging platforms are expected to face serious financial problems as a result of the loss of revenue during the quarantine period and the subsequent reduction in demand *(7)*. As mentioned, cost reduction is one of the goals of ridesharing, which forces users and drivers to use these systems. If this demand is not met, they will switch to other modes of travel. For example, a study in Spain showed that the highest percentage of people after the quarantine period of

COVID-19 tend to use public transportation because it is cheaper, and they believe that taxis and ride-hailing services are luxury services and only when those companies pay for their work trips or use these services at night when there is no public transportation *(8)*. Many studies proposed algorithms to reduce travel costs and match users' wishes for platforms, which we discuss in the literature review section. However, few pieces of research have been conducted on the opinions of users on the cost of traveling with ridesharing and ride-hailing platforms, which seems to be a gap in current research, especially in the pre-and amid-COVID-19 eras in America and India. In this research, our main question is: what are the opinions of American and Indian users about the cost of traveling with ridesharing services, especially in the era pre-and amid-COVID-19?

Research context

In order to answer this question in the present study, the reduction of users and trips in ridesharing systems in different age and gender groups in terms of travel costs in the Twitter application was investigated in order to determine what changes in users' opinions on travel costs and willingness to pay were made. The purpose of this study is to investigate the views of a large community of ridesharing users on the Twitter application for the United States and India, which can cause differences in cultural exchange activities due to unstable economic conditions, cultural norms, lack of infrastructure, and movement habits in these countries. The results of this study can help governments better plan for transportation systems, especially ride-hailing and ridesharing platforms. To do this research, 63,800 tweets were collected from Twitter using text mining methods, specifying age, gender, and country in the pre- and amid-COVID-19 pandemic era. 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.
To organize this article, the research conducted in the field of travel cost for different travel modes and the proposed solutions to these studies are given in the Literature Review section, and then the methodology and data collection methods are examined. Finally, we will discuss the results of this research.

LITERATURE REVIEW

Analysis of the ridesharing service

As we mentioned, COVID-19 has had a great impact on access, mobility, travel mode choices, and reduced use of public transportation, ridesharing, and ride-hailing, which has hit the economy of these services. Research results in China show that COVID-19 has a greater impact on reduced mobility and travel for disadvantaged and low-income groups than others *(9)*. Meanwhile, in two studies from Texas, America, and Greece during the pandemic, unlike China, low-income travelers are likely to make more daily trips because they have no choice but to leave home *(10)*. In a study from America, it was found that disabled people in low-income neighborhoods of Seattle, who have less access to private vehicles, are more dependent on transportation services *(11)*. COVID-19 has also further exposed structural disparities in

access to transportation services for low-income Americans of color *(12)*. At the same time, according to an online survey of 398 Uber users in Egypt, it was found that they used the system during the pandemic because of its usefulness (saving time and money), its speed, efficiency, and convenience, and they think that this service is more valuable than other modes of transportation. In a study about public transportation in three large regions of Sweden, it was found that public transport users decreased by between 40 and 60% *(13),* and in another study in Budapest, this number was 80% *(14)*, and also in Italy, this reduction was seen in public transportation *(15)*. Meanwhile, in a survey of an article from Spain in the spring of 2020, 89.7% of people want to return to public transportation after the quarantine period, and this is only because of the low cost of this type of service. Also, based on a survey in a study from Chile, it was found that young and low-income people are relatively less sensitive to public transport congestion *(16)*. In another paper, the reduction in bus and subway use in areas of New York such as the Bronx and East Brooklyn (lower income areas) was much less than in the CBD *(17)*.

Currently, according to two studies, one of them was from India, it was found that people with higher income tend to use ride-hailing and people with lower income choose ridesharing *(18, 19)*. In a study conducted with a survey of 4365 users in the United States, it was found that a price difference of $1 per mile can increase the probability of sharing by more than 8% *(1)*. But in the case of ride-hailing in the three cities of Boston, San Francisco, and Washington DC, it was found that middle-income households do not want to use these services *(20)*. Passengers who use the Jetti platform in New Mexico City have moderate to high incomes, and twice the national average own a private vehicle *(21)*. Based on data collected from 532 people in China, it was found that tourists as well as people over 40 years old tend to pay more for ridesharing than those under 30 years old to reduce waiting time and travel time *(22)*. In another study from China, it was found that with the increase in the price of ride-hailing or the travel distance, the possibility of choosing ride-hailing decreases, and more passengers turn to taxis and ridesharing, while the opposite is also true *(23)*. In some studies, suggestions have been made to increase users' trust in platforms and increase their satisfaction. For example, Morris et al. mention two cases of reducing the cost of carpooling, especially the cost of ride-hailing, and reforming the pricing system by giving discounts to passengers when their travel takes longer than estimated *(24)*. On the other hand, it suggests an appropriate cost-sharing mechanism that is understandable for users to ensure users balance their interests with the system and increase demand *(25)*. Financial reward, receiving a lower price from a more reliable platform than a less reliable one *(26)*, payment security, driver certification, feedback mechanism, and incremental pricing are among the things that make customers trust the system *(27)* and, ultimately, demand can be increased with appropriate pricing *(28)*.

It is very important to consider the needs of all users of ridesharing and ride-hailing services and to plan properly for the future of these systems. In general, incentives, financial benefits, and sharing of travel expenses may encourage people to share their trips *(4)*. In a research, Hasanpour Jesri and Akbarpour Shirazi consider the costs of ridesharing services to include fixed costs related to the number of vehicles and variable costs related to the distance traveled by vehicles *(29)*. On the other hand, Yan et al. believe that the existing pricing methods in ridesharing systems may lead to the loss of efficiency of these systems *(5).* To solve this problem, it can be said that the synergy of Dynamic Pricing and Dynamic Waiting can reduce price fluctuation, and by improving the pricing programs, it is possible to control and choose the best price *(30)*. In order to improve the pricing, research has been done, among which we can mention the

ADAPT-Pricing method from Asghari and Shahabi in New York, in which the revenue created for the platform in each period is up to 5% and 15%, respectively, increase, while reducing the travel cost by 5% *(31)*.

Method of Ridesharing Service

Among topic models, LDA *(32)* is a valid and widely used model, which assumes that there is an exchange between words and documents in a corpus represented by bag-of-words. LDA has been used in both long-length (e.g., abstracts) and short-length (e.g., tweets) corpora for different applications such as health *(33-34)*, e-petitions *(35)*, politics *(36)*, and investigation of social media strategy *(37)*. For example, Pournarakis *(38)* has carried out the topic modeling for transportation services based on LDA. For the task of clustering tweets into the different topics, this study designed and implemented a Genetic Algorithm based on LDA which improved the K-means clustering approach. Another relevant study has been carried out to analyze ridesharing services based on the Twitter data, the result shows that LDA topic modeling could provide the capacity to extract the most discussed topics in a large dataset in a short period of computing time *(39)*.

When it comes to sentiment analysis, BERT has been popular recently, which is designed to help computers understand the sentiment of ambiguous language in the text by using surrounding text to establish context. For example, Sun et. al. *(40)* created an auxiliary sentence to convert (T)ABSA from a single sentence classification task to a sentence pair classification task based on BERT. The result shows that BERT-pair beats other models on aspect detection and sentiment analysis by a significant margin on the SentiHood dataset. Historically, language models could only read text input sequentially, but couldn't do both at the same time *(41)*. BERT is different because it is designed to read in both directions at once. This capability, enabled by the introduction of Transformers, is known as directionality *(42)*. Meanwhile, the BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with a question and answer datasets. This method achieves the drawback of the supervised method in dataset transfer and a limited amount of data. The BERT was utilized as a reference model in this study, after the sentiment is extracted, the sensitive and correlation analysis are based on logistic regression.

METHODOLOGY FRAMEWORK

Figure 1 shows the methodology's flowchart.. First, this paper collects text data consisting of all characterization indicators from Twitter. Then text data is filtered based on error deletion, and the noise of text is reduced based on the text's meaning. Secondly, the keyword of all texts is extracted and sorted based on the word's frequency using the flash-text method, then the price-related texts are selected from those keywords. Then, the clusters are modeled based on the price-related dataset using LDA, 5 clusters are modeled, each cluster has 20 keywords, and the topics are extracted based on those keywords manually. Furthermore, the topics' differences between each group and pre-and amid-pandemic are compared. Third, this paper uses VADER and BERT to model the sentiment of each sentence of Twitter text, and a more suitable model is chosen for further analysis. Sensitivity and significance analyses are used to examine pre- and amid-pandemic gender, age, and country differences, as well as pre- and amid-pandemic gender

groups, age groups, and pre- and amid-pandemic USA/India. The multi-logit regression model is then used to assess the correlation and regression between each variable.

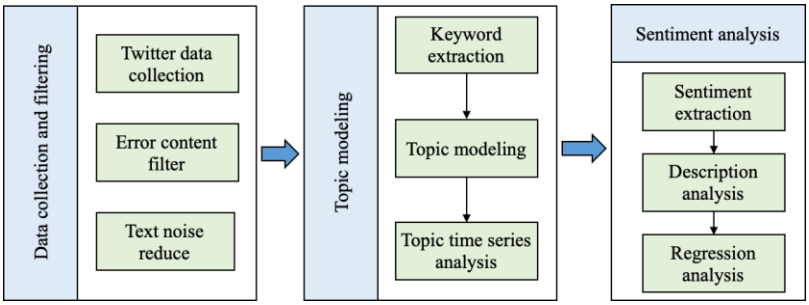


FIGURE 1 THE METHODOLOGY FRAMEWORK'S FLOWCHART

Data Collection and Filtering

The tweets were gathered using Twitter Intelligence Tool (43), a sophisticated Twitter scraping tool. The text data is gathered from 1 January 2019 to 4 May 2022 using the terms "Uber Pool," "Uber Black," "Uber Comfort," "Uber X," "Lyft XL," "Lyft Lux," "Lyft Black," "Lyft Line," "Lyft Shared," "Ola KaaliPeeli," "BlaBlaCar Carpool," "Sride," "Ibibo Ryde," "Meru". The texts were chosen based on the following criteria: they were written in English, they were located in the United States and India, duplicated texts were eliminated, and spam content was removed using a method for identifying spammers (44).The Twitter database has data problems, such as missing data, characters that make no sense, and noise in the data. The phrases are cleaned up of any nonsensical characters, such as emojis, emoticons, URL paths, digits, punctuation marks, symbols, English stop words, non-alphabetical words, and tokens with less than one character.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; the empty text whose length was less than five characters are deleted from the dataset since this paper considers an English word must have at least five characters to provide any signification information *(44)*. 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 filtering	Before filter: @Uber### I like ☺ and miss # uberpool much, prices are odeeeeeeeeer cheaper #uber.

	https://t.co/OOLOYLexyC After filter: I like and miss uber pool, prices are cheaper.

TOPIC MODELING OF RIDESHARING MONEY COST

Time-related tweet extraction

The main content of this paper concentrated on the ridesharing money cost, therefore the tweets related to this topic are kept for further analysis. To extract the related tweets about ridesharing money costs, the words of the original dataset are extracted and sorted based on the word frequency based on the flash-text model in Python. 2000 words are found from the dataset, then, the words related to ridesharing money cost, such as “pay”, “expansive”, “cheaper”, etc., are selected from those keywords, then used for selecting the ridesharing money cost-related tweets. After the data collection and cleaning, 2406 texts related to ridesharing money costs based on 2219 Twitter users are kept in the dataset. In this paper, the happen threshold of a pandemic is chosen as March 2020, the data pre-pandemic includes 1286 texts, and amid-pandemic has 1120 texts. When it comes to gender, 1440 text belongs to males, and 966 to females. The mean of users age = 32.03, S.D. = 5.32, this paper chooses age =37 (only using this threshold, the data show the significance) to divide the users as younger and older (number of tweets belonging to younger = 1858, older = 548). Note that, the countries in this paper include USA and India, 2094 tweets belong to the US users, and 312 to Indian users.

Keywords Extraction :This paper extract keywords based on LDA. A generative statistical model called the LDA enables unobserved groups to explain sets of observations, which explains why some portions of the data are similar.For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics.

To obtain the underlying structure of latent topics in our dataset based on LDA, Python’s Gensim library *(45)* is used, which allows the execution of the algorithm with multi-threads, resulting in effectiveness and fast calculation. This paper concentrates on exploring the difference between the countries, gender, and ages pre- and amid-pandemic, the data is divided into 7 groups (all tweets, USA tweets, India tweets, male tweets, female tweets, younger tweets, and older 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 5 clusters' LDA model *(46)*. Finally, 35 clusters, each cluster has 20 keywords, are collected.

Topic labeling based on keywords combination:As discussed in the previous section, the keywords of 5 clusters are created in each group. However, the result of LDA does not provide the documents’ topic but only a distribution of probabilities to the different topics. Some studies considered the use of several clustering techniques to group keywords of clusters into predefined topics, such as Moreno *(47)* use method involves the use of a K-mean clustering algorithm and Genetic Algorithm combined with a local convergence algorithm to integrate the topics based on LDA result. 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 information of text and the hidden variables of topics.

To deal with those drawbacks, this paper proposes a topic label method for topic labeling. At first, the cluster ordering step is re-order the clusters based on the coherence measures 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 a good performance in the model, therefore, 5 clusters are sorted based on the enhancement of the coherence measure score. Then, this paper selected the three most discussed topics in the latest research including price, expansive, and cheaper *(2)* as references for topic labeling. Each keyword in the clusters is scored based on the correlation with four topics manually. A 10 score means the keyword has a high probability of belonging to a certain topic. Once the keyword has got a 1 score in all topics, the new topic would be created based on the meaning of the keyword, for example, the word pandemic does not belong to any topics, therefore, the pandemic is created as the fourth topic. Finally, after all of the keywords are scored, topics are generated and labeled as topic-1, …, topic-4. Then, each cluster is changed to a related topic based on the keywords labeled. Take the amid-pandemic, male's cluster 1 as an example, as can be seen in **Figure 2**, each keyword in the clusters is transferred to the topic label, and the clusters are relabeled to topic 2 based on the keywords' frequency in each cluster. 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.

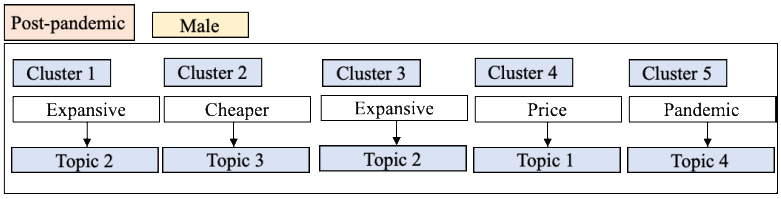


FIGURE 2 EXAMPLE OF TOPIC LABELING

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 models are used for extracting the sentiment of each text, then, the sensitivity and significance pre- and amid-pandemic are analyzed. The logistic regression model is used for interpreting the correlation between passengers' characteristics, countries, pandemic, and sentiment. Then, the regression model is modeled 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 specifically attuned to sentiments expressed in social media. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) that are generally labeled according to their semantic orientation as either

positive or negative. VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is. The VADER lexicon performs exceptionally well in the social media domain. The correlation coefficient shows that VADER (r = 0.881) performs as well as individual human raters (r = 0.888) at matching ground truth (aggregated group mean from 20 human raters for sentiment intensity of each tweet). In this paper, the VADER model is constructed based on the Vader module which is 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 example, the word "okay" has a positive valence of 0.9, "good" is 1.9, and "great" is 3.1, whereas "horrible" is –2.5, the frowning emoticon ":(" is –2.2, and "sucks" and "sux" are both –1.5. More details and is available for download can be found in *(48)*.

BERT Model:The sentiment model attains a greater accuracy of 92% for sentiment when utilizing the cased version of BERT analysis *(41)*. The model is composed of one or more input sequences, added with an initial token "CLS" and a token "SEP" to separate segments. All tokens are represented by word embeddings, concatenated with position embeddings and segment embeddings. Each model is made 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. Each sublayer output is normalized and added to the sublayer input. we define two vectors S and E (which will be learned during fine-tuning) both having shapes (1x768). We then take a dot product of these vectors with the second sentence's output vectors from BERT, giving us some scores. We then apply SoftMax over these scores to get probabilities. The training objective is the sum of the log-likelihoods of the correct start and end positions.
In this paper, the BERT base model is 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 *(49)*.

RESULTS

Topic Modeling Performances and Results

Keywords distribution and results:Tweets were collected from ridesharing customers addressing the @ridesharing Twitter platform from 2019.01 to 2022.05. Then, each tweet related to ridesharing money cost is collected based on the citation of the methodology section. After data collection and cleaning, the frequency of tweets in each month is shown in **Figure 3**. The results show that the number of tweets has a significant decrease in 2020.04. This decrease occurred with a high probability due to the ridesharing users' decrease during the pandemic. Then, the number of tweets increased after 2020.04, and the trend

shows a stable wave, which means that people share their problems and opinion on other transport platforms or other topics amid-pandemic at first, then come back to the ridesharing money cost topic nowadays.

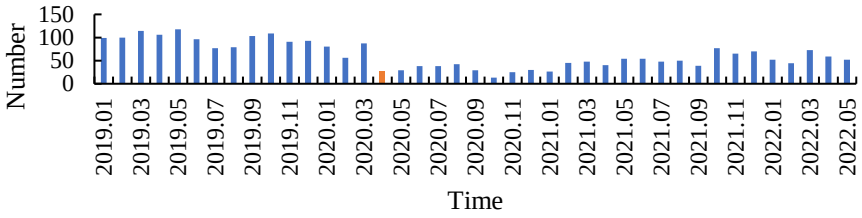
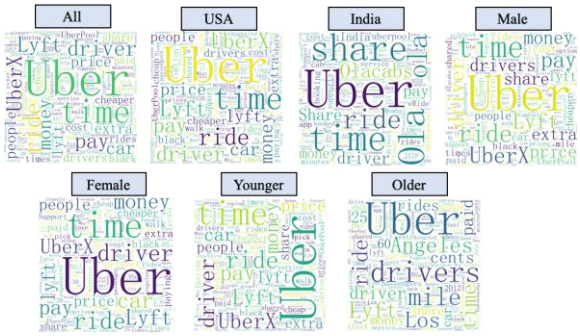
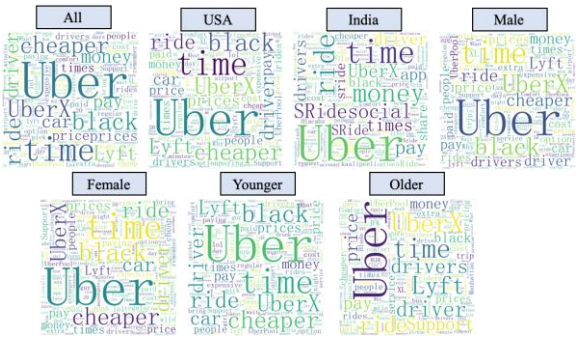


FIGURE 3 THE TWEET FREQUENCY EACH MONTH

The frequency of keywords in each group (country, gender, and age) pre- and amid-pandemic are collected for checking the most common and relevant issues posted on the platform in each time period. As shown in **Figure 4**, the word cloud shows each word with different sizes according to the frequency of words on the dataset. Before the pandemic, words like "Uber", "price", "driver", and "cheaper" are the most frequent words in all groups. These results show that people commonly tweet about ridesharing to discuss problems or recommendations related to ridesharing companies, the ride money cost, and the feeling of price change. In the USA, people care more about the ridesharing price than in India. When it comes to gender and age groups, males and females have the same opinion, their hot topics are both focused on ridesharing companies and drivers' services. During the pandemic, "cheaper" become more popular in all groups, this result shows that more people care about the discount on ridesharing amid-pandemic. In the country group, the Indians more care about the money cost than the USA amid-pandemic. In the gender and age group, the difference is small, they both care about the difference between ridesharing companies' prices, and price changes amid-pandemic.



THE WORD CLOUD OF RIDESHARING PRE-PANDEMIC



THE WORD CLOUD OF RIDESHARING AMID-PANDEMIC

FIGURE 4 WORD CLOUD OF RIDESHARING PRE- AND AMID-PANDEMIC

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 pre- and amid-pandemic are sorted based on the coherence score, then change to the topics based on the proposed method. As a result, 4 topics are labeled, the details of the topics and their content can be seen in **Table 2.**

TABLE 2 DESCRIPTION OF LABELED TOPICS AND THEIR CONTENTS

Item	Label	Content	Description
Ridesharing money cost	1	Price	Money cost of the ride
	2	Expensive	The feeling of the price rise
	3	Cheaper	The feeling of the price discount
	4	Pandemic	Topic related to pandemic

The topic of each cluster is marked in the graph, as shown in **Figure 5**, the left of the graph shows the group and ridesharing trip time label, then, the center of the graph shows the distribution of the topics. In the graph, the F means the frequency of the hottest topic in all periods, pre-pandemic and amid-pandemic. As a result, the discount of ridesharing price – cheaper (3) is the hottest topic in all periods, and hot topic has difference between pre- and amid-pandemic, the topic of price discount – cheaper (2) gains more attention amid-pandemic. It indicates that passengers more care about the discount on prices amid-pandemic, and the companies put some discount policies to deal with the influence of the pandemic.

In the country groups, both USA and India paid attention to the price of ridesharing (1). In both groups, they more care about topic "1" pre-pandemic, then the American passengers are more sensitive about the price discount amid-pandemic than India. In the gender groups, both males and females paid attention to the topic "1". Males' hottest frequency of topic is "1" pre-pandemic and changer to "3" amid-pandemic. Different from males, the hottest topic of females is "1" pre-pandemic, which is changed to "1 & 3" amid-pandemic. It indicated that both of them care about the discount of ridesharing amid-pandemic, and the females are more concerned about how much the price decreases than males amid-pandemic. In the age

groups, both groups have the hot topic “1”. When it comes to the difference between pre- and amid-pandemic, the younger group has a significant difference, topic “2 & 3” gained more attention amid-pandemic instead of topic “1” pre-pandemic. It indicates that younger users are more caring about price changes amid-pandemic.

Group	Topics	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
All	1										
	2										
	3										
	4										
F: 3		Pre-pandemic. F: 2 & 3					Post-pandemic. F: 3				

THE TREND GRAPH OF TOPICS BASED ON ALL TWEET DATA

Group	Topics	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
USA	1										
	2										
	3										
	4										
F: 1		Pre-pandemic. F: 1					Post-pandemic. F: 3				
India	1										
	2										
	3										
	4										
F: 1		Pre-pandemic. F: 1					Post-pandemic. F: 1				

THE TREND GRAPH OF TOPICS IN THE COUNTRY GROUP

Group	Topics	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Male	1										
	2										
	3										
	4										
F: 1		Pre-pandemic. F: 1					Post-pandemic. F: 3				
Female	1										
	2										
	3										
	4										
F: 1		Pre-pandemic. F: 1					Post-pandemic. F: 1 & 3				

THE TREND GRAPH OF TOPIC IN GENDER GROUP

Group	Topics	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Younger	1										
	2										
	3										
	4										
F: 1		Pre-pandemic. F: 1					Post-pandemic. F: 2 & 3				
Older	1										
	2										
	3										
	4										
F: 1		Pre-pandemic. F: 1					Post-pandemic. F: 1				

THE TREND GRAPH OF TOPICS IN THE AGE GROUP

Figure 5 The trend graph of ridesharing service topic modeling

Sentiment analysis of ridesharing service

Sensitive and significant analysis:In this section, the VADER and BERT models have been employed to model the sentiment of each tweet. The proposed models have been implemented on Intel (R) Xeon (R) Silver 4110 CPU @ 2.10 GHz with 64 GB RAM and IDE disk under Centos 7.6 operating system. The Anaconda 2021.03, open-source software is used for developing the algorithm in Python. Then, the NVIDIA V100 GPUs are used to fine-tune the models. To further verify the applicability of the model, 400 tweets’ sentiment is checked manually, the result shows that the BERT model still has high accuracy (81.3%) and less Mean Absolute Error (0.12) than VADER (accuracy = 62%, Mean Absolute Error = 0.25). 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 results of BERT associated with the time series are shown in **Figure 6**. Based on the sentiment values, it is observed that 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 a positive attitude increases at the beginning of the pandemic, then, decreasing trend amid-pandemic. This result may be that the ridesharing company has put some discount policy for customers to respond to the pandemic at the beginning, which enhances the customer's positive sentiment.

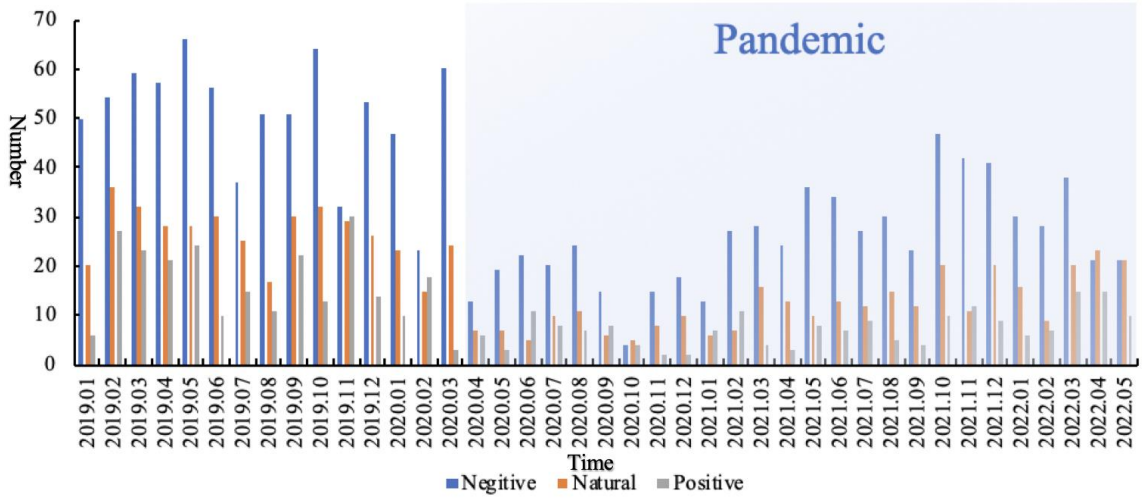
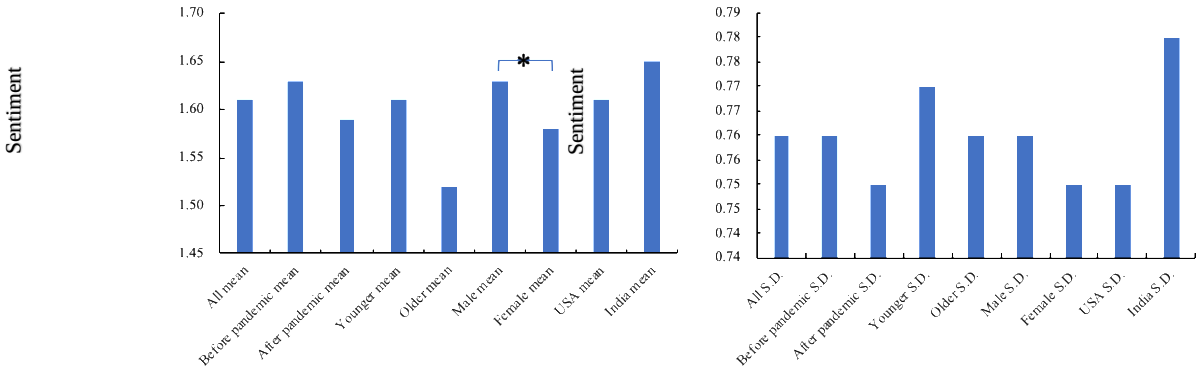


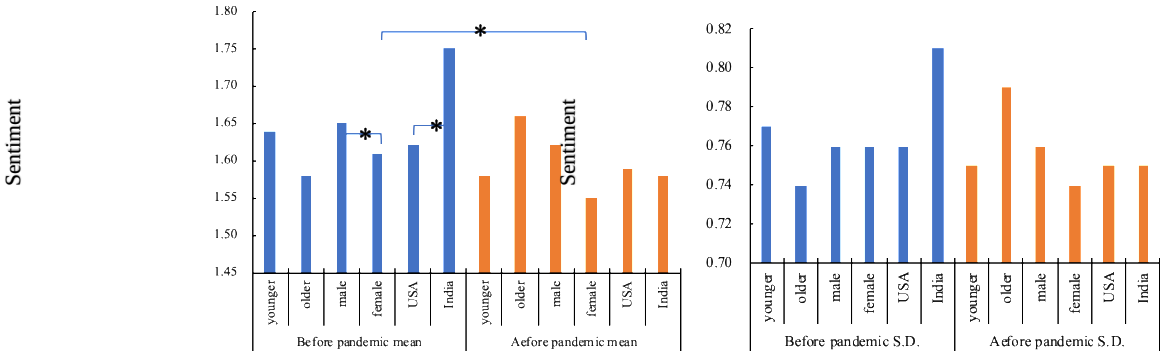
FIGURE 6 THE VOLUME OF RIDESHARING SENTIMENTS ASSOCIATED WITH THE TIME SERIES

To further analyze the sentiment, 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 difference between groups (pandemic, gender, age, and country), the mean and standard deviation (S.D.) of sentiment value are compared. As can be seen in **Figure 7 (a)**, the sentiment in the different gender groups shows a significant difference. When comparing the mean of each group, customers show more positively amid-pandemic, and the younger, male, Indian customers have a more positive attitude toward ridesharing money cost. When it comes to S.D., the customer always keeps a more stable sentiment pre-pandemic. Similarly, older, female, and American customers also have a more stable sentiment about ridesharing money costs than others. Then, the difference between each group pre- and amid-pandemic also be analyzed, as can be seen in **Figure 7 (b)**, the result shows that the sentiment of gender and country groups have significantly different pre-pandemic, but no significant differences in each group amid-pandemic. Meanwhile, younger, females and Indian passengers are more positive about the ridesharing money cost pre-pandemic. The older and the USA passengers enhance a positive attitude amid-pandemic, it may be those groups have more sensitive to the price change, and change their attitude with the company putting some discount policy. Meanwhile, only females enhance their negative thought significantly amid-pandemic, it's an indicator that they may think the discount cannot cover the negative influence of the pandemic. As a result

of S.D., older and Indian passengers keep a more stable attitude than others pre-pandemic, and females enhance the 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).*



THE DIFFERENCE PRE- AND AMID-PANDEMIC, AND CUSTOMS' CHARACTERISTICS



THE DIFFERENCE BETWEEN CUSTOMS' CHARACTERISTICS PRE- AND AMID-PANDEMIC

FIGURE 7 THE DESCRIPTION AND SIGNIFICANT ANALYSIS OF SENTIMENT IN EACH GROUP

Ridesharing service multi-logit regression model:To model the relationship between sentiment and customers' characteristics, country, and pandemic, this paper analyzed the correlation between each related variable and sentiment. The sentiment has a significant correlation with country (0.021*), but there has no significant correlation with other variables. It indicated that the sentiment has different performances in USA and India. Then, the regression model has be modeled based on the multi-logit regression model, which was implemented by using the MATLAB built-in algorithm *(50)*. Four parameters are used for validating the model performance as Log-Likelihood Ratio, X^2, the goodness of fit test, and model significance. As a result, the model shows a good performance in modeling sentiment based on the higher significance (sig.=0.05), and lower error (Log-Likelihood Ratio=213.47, X^2=19.34). **Table 3** shows the result of the sentiment regression model; country and gender are the main factors to

influence the sentiment. In India, the sentiment is more positive (OR=1.17, P<0.05), and the female passengers are more positive than the male (OR=0.99, P<0.05).

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

Step	Items	B	Stad. E.	Wald	Freedom	Sig.	Exp(B)	95% CI	
1	Intercept	1.26	0.26	22.54	1.00	0.00			
	Pandemic	0.05	0.19	0.79	1.00	0.13	1.05	0.87	1.21
	Gender	-0.05	0.19	0.08	1.00	0.05	0.98	0.65	1.34
	Age	-1.55	0.12	0.08	1.00	0.77	0.95	0.64	1.37
	Country	0.33	0.16	1.07	1.00	0.03	1.18	0.86	1.64
2	Intercept	0.35	0.29	1.44	1.00	0.00	-	-	-
	Pandemic	0.02	0.12	0.02	1.00	0.13	1.02	0.98	1.23
	Gender	-0.09	0.21	0.00	1.00	0.05	0.99	0.79	1.27
	Age	0.09	0.67	2.12	1.00	0.97	0.99	0.66	1.49
	Country	0.15	0.18	0.74	1.00	0.04	1.17	0.81	1.68

CONCLUSIONS

This essay explores the issue of ridesharing's financial costs both before and after the pandemic. It also takes into account factors like passenger profiles and regional disparities. The ridesharing money cost subject is modeled using the LDA model, and the sentiment is examined using BERT and multi-logit model. From the original tweets, four topics—price, costly, affordable, and pandemic—were taken out. A graph is used to examine the subject distribution differences across gender, age, and country groups before and after the epidemic. The BERT model is used to extract the emotion of each tweet in each group based on time series. The outcome demonstrates that the BERT model outperforms the VADER model in terms of sentiment extraction. The primary elements impacting sentiment are identified after significance and correlation analysis, regression modeling using the multi-logit model, and sentiment analysis. The following are the main conclusions:

- The discount on ridesharing pricing is always a hot issue, but there are distinctions between hot topics pre- and post-pandemic. Post-pandemic, passengers care more about the price discount, and businesses have implemented several discount rules to deal with the pandemic's effects.
- Both the USA and India included the cost of ridesharing while comparing the differences across categories. American travelers are more sensitive than Indian travelers to the discount during the epidemic. Both men and women in the study's gender-based groups were interested in ridesharing's cost. Males worry more about the price discount than females do. The heated subject of pricing is present in both age groups at all times, with the younger generation worrying more about discounts and price increases due to the pandemic.
- Prior to the epidemic, younger, female, and Indian passengers had a more optimistic attitude on the cost of ridesharing; yet, after the pandemic, older, USA-based passengers had an even stronger sense

of optimism. It's possible that such groups are more sensitive to pricing changes and that the company's decision to implement a discount strategy causes them to alter their behavior. Only women, however, believed that the pandemic's effects could not be offset by the discount. Prior to the epidemic, older and Indian travelers had more steady attitudes than other passengers, and female passengers now further the stable tendency.

- According to the regression model's findings, gender and nation are the primary variables that affect sentiment. Users are more optimistic in India, while female passengers are more optimistic than male travelers.

This study's major contribution is a technique for sentiment analysis and trip money cost modeling for ridesharing. Topic occurrence and trend shift, sentiment time series variables that have seldom been used before, are taken into account in the paper's approach. Results from this study can be used for sentiment analysis and topic modeling in the ridesharing area. On the basis of this article, those industries may improve their services and offer a reasonable pricing design standard, which will increase affiliated firm rivalry.

Future studies should investigate this issue further to validate the modeling strategy and the algorithm. More crucially, subject modeling and sentiment analysis ought to be combined. Additional variables, such as sad, joyful, wild, etc., may be retrieved based on further research and utilized to enhance the deep analysis of sentiment using the emotion analysis.

REFERENCES

1. Middleton, S. R., Schroeckenthaler, K. A., Gopalakrishna, D., & Greenberg, A. (2021, August 1). Effect of Price and Time on Private and Shared Transportation Network Company Trips. Transportation Research Record, 2675(8), 458-467.
2. Xu , A., Chen , J., & Liu, Z. (2021, October 9). Exploring the Effects of Carpooling on Travelers' Behavior during the COVID-19 Pandemic: A Case Study of Metropolitan City. Sustainability, 13(20), 11136.
3. Shaheen, S., Totte, H., & Stocker, A. (2018). Future of Mobility White Paper. UC Berkeley.
4. 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.
5. Yan, P., Lee, C.-Y., Chu, C., ChendZhi, C., & Luo, Z. (2021, July). Matching and pricing in ride-sharing: Optimality, stability, and financial sustainability. Omega, 102, 102351.
6. WHO. (2022, July 15). WHO Coronavirus (COVID-19) Dashboard. Retrieved from World Health Organization: https://covid19.who.int/
7. 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.
8. Awad-Núñez, S., Julio, R., Gomez, J., Moya-Gómez , B., & González , J. S. (2021, March 10). Post-COVID-19 travel behaviour patterns: impact on the willingness to pay of users of public transport and shared mobility services in Spain. European Transport Research Review, 13(20).
9. Pan, Y., & He, S. Y. (2022, May). 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.
10. Iio, K., Guo, X., & Kong, X. (2021, June). COVID-19 and social distancing: Disparities in mobility adaptation between income groups. Transportation Research Interdisciplinary Perspectives, 10, 100333.
11. Wang, Y., Shen, Q., Abu Ashour, L., & Dannenberg, A. L. (2022, May). Ensuring equitable transportation for the disadvantaged: Paratransit usage by persons with disabilities during the COVID-19 pandemic. Transportation Research Part A: Policy and Practice, 159, 84-95.
12. Chen, K. L., Brozen, M., Rollman, J. E., Ward, T., Norris, K. C., Gregory, K. D., & Zimmerman, F. J. (2021, June). How is the COVID-19 pandemic shaping transportation access to health care? Transportation Research Interdisciplinary Perspectives, 10, 100338.
13. Jenelius, E., & Cebecauer, M. (2020, November). Impacts of COVID-19 on public transport ridership in Sweden: Analysis of ticket validations, sales and passenger counts. Transportation Research Interdisciplinary Perspectives, 8, 100242.
14. Bucsky, P. (2020, November). Modal share changes due to COVID-19: The case of Budapest. Transportation Research Interdisciplinary Perspectives, 8, 100141.
15. Esposti , P. D., Mortara , A., & Roberti , G. (2021, February 10). Sharing and Sustainable Consumption in the Era of COVID-19. Sustainability, 13(4), 1903.
16. Basnak, P., Giesen, R., & Muñoz, J. C. (2022, May). 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.
17. Halvorsen, A., Wood, D., Jefferson, D., Stasko, T., Hui, J., & Reddy, A. (2021, September 8). Examination of New York City Transit's Bus and Subway Ridership Trends During the COVID-19 Pandemic. Transportation Research Record, 1–14.
18. 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.
19. 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. doi:https://doi.org/10.1016/j.tra.2018.11.018
20. Gehrke, S. R. (2020, June). Uber service area expansion in three major American cities. Journal of Transport Geography, 86, 102752.
21. Tirachini, A., Chaniotakis, E., Abouelela, M., & Antoniou, C. (2020). The sustainability of shared mobility: Can a platform for shared rides reduce motorized traffic in cities? Transportation Research Part C: Emerging Technologies, 117, 102707.
22. Chen, J. M., de Groote, J., & Petrick, J. F. (2020, June 18). Travellers' willingness to pay and perceived value of time in ride-sharing: an experiment on China. Current Issues in Tourism, 23(23), 2972-2985.
23. Wu, Y., Chen, X., & Ma, J. (2018). Modeling Passengers' Choice in Ride-Hailing Service with Dedicated-Ride Option and Ride-Sharing Option. Proceedings of the 4th International Conference on Industrial and Business Engineering (pp. 94–98). ICIBE' 18.
24. 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.
25. Fielbaum, A., Kucharski, R., Cats, O., & Alonso-Mora, J. (2022, September 16). How to split the costs and charge the travellers sharing a ride? aligning system's optimum with users' equilibrium. European Journal of Operational Research, 201(3), 956-973.
26. Jang, S., Farajallah, M., & Fung So, K. (2020, January 20). The Effect of Quality Cues on Travelers' Demand for Peer-to-Peer Ridesharing: A Neglected Area of the Sharing Economy. Journal of Travel Research, 60(2), 446-461.
27. Shao, Z., & Yin , H. (2019, September 19). Building customers' trust in the ridesharing platform with institutional mechanisms: An empirical study in China. Internet Research, 29(5), 1066-2243.
28. Harding, S., Kandlikar, M., & Gulati, S. (2016, June). Taxi apps, regulation, and the market for taxi journeys. Transportation Research Part A: Policy and Practice, 88, 15-25.
29. Hasanpour Jesri, S., & Akbarpour Shirazi , M. (2022, June 17). Bi Objective Peer-to-Peer Ridesharing Model for Balancing Passengers Time and Costs. Sustainability, 14(12), 7443.
30. Yan, C., Zhu, H., Korolko, N., & Woodard, D. (2019, November 15). Dynamic pricing and matching in ride-hailing platforms. Naval Research Logistics, 67(8), 705-724.
31. Asghari, M., & Shahabi, C. (2018). ADAPT-pricing: a dynamic and predictive technique for pricing to maximize revenue in ridesharing platforms. Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information (pp. 189–198). SIGSPATIAL '18.
32. Mcauliffe, J., & Blei, D. (2007). Supervised topic models. Advances in neural information processing systems, 20.
33. 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.
34. 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.
35. Karami, A., & Shaw, G. (2019). An exploratory study of (#) exercise in the Twittersphere. iConference 2019 Proceedings.
36. 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.
37. 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.
38. 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.
39. Karami, A., & Collins, M. (2018). Social Media Analysis for Organizations: US Northeastern Public and State Libraries Case Study.
40. 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).
41. 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.
42. G. Henry, "Improved sentiment analysis using a customized distilbert NLP configuration", Advances in Engineering: An International Journal (ADEIJ), Vol.3, No.2
43. Twitter Intelligence Tool (TWINT). Available online: https://github.com/twintproject/twint (accessed on 21 June 2021).
44. 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.
45. 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.
46. Moreno, A., & Iglesias, C. A. (2021). Understanding Customers' Transport Services with Topic Clustering and Sentiment Analysis. Applied Sciences, 11(21), 10169.
47. 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).
48. G. Henry, "Improved sentiment analysis using a customized distilbert NLP configuration", Advances in Engineering: An International Journal (ADEIJ), Vol.3, No.2
49. Agresti, A., An Introduction to Categorical Data Analysis. 2007. John Wiley and Sons Inc.
50. Mathworks. Multinomial logistic regression. (2022) Available online: https://nl.mathworks.com/help/stats/mnrfit.html.