A NEW METHOD FOR THE RIDESHARING SERVICE ANALYSIS BASED ON GRAPH CONSTRUCTION — THE USA CASE STUDY ACROSS DIFFERENT AGE AND GENDER GROUPS PRE- AND POST-COVID-19 PANDEMIC

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

Junhua Wang, Ph.D., Professor, School of Transportation Engineering, Tongji University, 4800 Cao'an Highway, Shanghai, 201804, China, (+86)188-0196-2395, Fax: (+86)021-69585717, benwjh@163.com

Hongzhang Mu, Research Assistant, Institute of Information Engineering, Chinese Academy of Sciences School of Cyber Security, University of Chinese Academy of Sciences, 89 Minzhuang Road, Haidian District, Beijing, 100093, China, (+86)152-1003-3919, Fax: (+86)010-82546751, muhongzhang@iie.ac.cn

ABSTRACT

The proposed methodology framework considers the change in the occurrence and trend of topics, and sentiment time series feature. A total of 13411 Tweets from 1 January 2019 to 30 April 2022 are used for method validation. The sentiment of each tweet in each group (in time series) is extracted using BERT, with a high accuracy of 92%. The positivity of sentiment increases with time, and females are found to be significantly more positive compared to males.

Keywords: ridesharing, topic modeling, sentiment analysis, graph construction, Twitter data, COVID-19

INTRODUCTION

Overview

Ridesharing is a transportation rent service that provides the a connection platform between passengers who want to ride an individual automobile and drivers who want to give rent based on internet applications. This service has been popular in recent years based on its' cheaper cost and effective services. In 2021, more than half-billion users were active in using ridesharing smartphone applications to rent an individual car, and more than 540 million individuals are signing in the ridesharing APPs worldwide *(1)*. The competition between ridesharing and taxi companies, and among ridesharing companies such as Lyft, Uber, Didi, etc., is very strong. Passengers' comfort, cost, time-saving, and other factors directly influence the users' choice of the rent service and ridesharing app (in other words, the ridesharing company). Topic modeling and sentiment toward ridesharing services can help the ridesharing platform to change the service model and price standard thus enhancing the competitive ability.

Topic modeling is wildly used for extracting the main influence factor of ridesharing services, which is an efficient and systematic method, to extract the main factors from thousands of documents in a minute *(2)*. As one of the most popular topic modeling methods, Latent Dirichlet Allocation (LDA) has advantages in representing bag-of-words of documents based on statistical distributions of words in the documents *(3)*. However, researchers are modeling the hot topic of ridesharing based on LDA while ignoring the time-varying and multi-variables features of topics *(4)*. The occurrence of ridesharing topics is a random process among time-varying transversion, and it is also influenced by several users' characteristics such as age, gender, etc. *(5-6)*.

Sentiment analysis can automate the mining of sentimental data, e.g. attitudes, opinions, views, and emotions, from a text. It involves understanding and classifying opinions in text into categories as “positive”, “negative” or “neutral”. Many researchers have attempted to model ridesharing sentiment from varied methods such as *(7-8)*. Due to the advantage of feature generation automation, deep learning methods have become widely used in text analysis *(9)*. In the beginning, previous-applied deep learning models do not consider the temporal variation of data and thus provide unsatisfying modeling performance *(10-11)*. To solve the shortcomings of existing models, some recently proposed algorithms are considered as being promising. For example, studies have employed neural networks including stacked autoencoder, Long Short-Term Memory, or Convolutional Neural Networks. These methods, however, cannot fully capture the passenger characteristics and temporal features *(12-14)*. Besides, the deep learning process is a black box, researchers can hardly capture variables' features from the result of the model. Therefore, there requires a multivariable time series deep learning-based method (MTDL) for characterizing the interaction of multi-variables and mining the time series feature of data, then improving the accuracy and interpretability of the ridesharing services analysis.

Research context

To solve the shortcomings of topic modeling and sentiment analysis, this paper uses Twitter data to propose a MTDL for modeling ridesharing services analysis. The MTDL is constructed with LDA, graph construction, Bidirectional Encoder Representations from Transformers (BERT), and the Logistic regression model. The LDA and graph construction are used for topic modeling, in detail, LDA is used for keyword extraction, and the drivers' characteristics and time-series features are

presented based on graph construction. The BERT and Logistic regression model are used for sentiment analysis, meanwhile, the sentiment of ridesharing is extracted using BERT, and the sensitive and correlation analysis is using Logistic regression model. In summary, the objective of the method is to propose a new method, that can characterize the feature (e.g. pandemic, passenger characteristics) of multi-variables and time attributes, for ridesharing topic modeling and sentiment analysis. As the main contribution, results from this paper can be used for ridesharing industries in enhancing the service, and provide, for instance, the suitable price standard.

LITERATURE REVIEW

This section focuses on two aspects: dataset of the ridesharing service analysis, and topic modeling and the sentiment analysis method. The literature on these aspects is discussed below.

Dataset of the ridesharing service analysis

Different data sources have been used previously in analyzing ridesharing services. The studies have been conducted by collecting text review data from Google App Store or interviewing users *(2)*. Recent developments in social interaction platforms have fostered independent data-logging systems with big data. Some social media, such as Facebook, Twitter, Instagram, Weibo, etc. provide an internet-based platform for users to exchange information with others *(15)*. When it comes to ridesharing service topics, the information that users share on social media includes their personal opinion, interesting topic, and sentiment. That information is very effective and complete for the ridesharing service analysis, which makes social media become an important data collection platform *(15)*.
Twitter, which has 330 million monthly active users with over 1.3 billion accounts, is one of the largest media platforms. Each Twitter is less than 240 characters, which may include users' opinions about some topic, service, etc. *(16)*. Twitter creates a chance for extracting the specific data for a particular event such as ridesharing, pandemic, etc., based on the hashtags, which is a label that provides users update threads with trending topics *(17)*. In this paper, to reveal the multivariate-temporal features of ridesharing services, the large-scale Twitter data is considered as data sources. Meanwhile, more than 90% of Twitter users are from the USA, and 50 million USA users are daily activities, which accounts for about 24% of USA adults *(18)*. To keep the data normalization and study the impact of the COVID-19 on ridesharing services, this paper collects Twitter data of USA users pre and post the global pandemic.

Topic Modeling and Sentiment Analysis Method of Ridesharing Service

Among topic models, LDA *(19)* 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 *(20-21)*, e-petitions *(22)*, politics *(23)*, and investigation of social media strategy *(24)*. For example, Pournarakis *(25)* 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 *(26)*.

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. *(27)* 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 *(28)*. 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 *(29)*. 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.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary and multi dependent variable, although many more complex extensions exist *(30)*. Lots of studies on regression have obtained reasonable results based on this model. For example, Ali *(31)* investigated the relationship between weather conditions and driver lane-keeping performance using the logistic regression model. The result shows that heavy rain can significantly increase the standard deviation of lane position, which is a very widely used method for analyzing lane-keeping ability. More detail about the logistic regression model can be seen in Agresti *(32)*.

METHODOLOGY FRAMEWORK

The flowchart of the methodology is presented in Figure 1. The methodology has three parts. i) Data collection and filtering. This part consists of characterizing the indicators which are extracted from Twitter data, filtering text data and deleting the text error, and reducing the noise of text based on the Part-Of-Speech analysis method. ii) Ridesharing service topic modeling. This part cluster the text data based on LDA, each cluster has 20 keywords, and the topics are extracted based on those keywords manually. Then, the time series feature and the changing trend between topics are built using the graph method. iii) Ridesharing services sentiment analysis. This part uses BERT to model the sentiment of each Twitter text. Then, the difference between pre- and post-pandemic, gender and age, as well as gender group pre- and post-pandemic, and age groups pre- and post-pandemic, are compared based on sensitivity and significant analysis. Then, the correlation and regression between each variable are analyzed based on the multi-logit regression model.

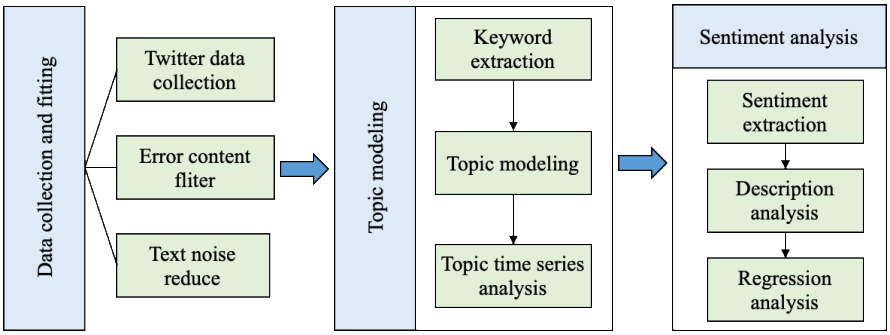


Figure 1 Flowchart of the Methodology Framework

Data Collection and Filtering

The collection of tweets was done through an advanced Twitter scraping tool called Twitter

Intelligence Tool *(33)*. The keywords “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 Carpool”, “Ola Share”, “Ola Carpool” and their derivatives are used to collect the text data from 1 January 2019 to 30 April 2022. The data were selected through the following criteria: the location is the USA, and the text is English written; Duplicated texts were eliminated; spam text was deleted using the detecting spammers method as *(34)*. After the data collection and cleaning, 13411 texts based on 5593 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 include 9411 texts, and post-pandemic is 4000 texts. When it comes to the comparison of gender, 7155 text belongs to males, 6256 for females. Then, the mean of age = 30.20, S.D. = 8.07, this paper chooses age =44 (only using this threshold, the data show the significance) to divide the users as younger and older (number of tweets belonging to younger = 7760, older = 5651).

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, the empty text whose length was less than five characters are deleted from the dataset since this paper considers an English word must have least five characters to provide any signification information *(34)*. 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 tweet (rewrite or not).
Sample of the tweet before and after 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 uberpull, this prices are cheaper.

Topic Modeling of Ridesharing Service

Keywords Extraction*:* This paper extract keywords based on LDA. The LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts 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. Normally, LDA identifies the correlation between topics and documents *(P_(T|D))*, words and topics *(P_(W|T))*, this correlation can be presented as *formula (1)*:

$$LDA \rightarrow Words \begin{pmatrix} P(W_1|T_1) & \cdots & P(W_1|T_t) \\ \vdots & \ddots & \vdots \\ P(W_m|T_1) & \cdots & P(W_m|T_t) \end{pmatrix} \& Topics \begin{pmatrix} P(T_1|D_1) & \cdots & P(T_1|D_t) \\ \vdots & \ddots & \vdots \\ P(T_t|D_1) & \cdots & P(T_t|D_t) \end{pmatrix} \quad (1)$$

Where the $P(W_i|T_k)$ is the probability of each of the words *i* was given a topic *k*, and the $P(T_k|D_j)$ is the probability of each of the topic *k* given a document *j*. The keywords of each topic are calculated based on the descending order of $P(W_i|T_k)$, the $P(T_k|D_j)$ is used for extracting the weight of the topic in the documents.

To obtain the underlying structure of latent topics in our dataset based on LDA, Python's Gensim library *(35)* is used, which allows the execution of the algorithm with multi-threads, resulting in effectiveness and fast calculation. This paper concentrate on exploring the difference between gender and age pre- and post-pandemic based on time series, the data is divided into 5 groups (all tweet, male tweet, female tweet, younger tweet, and older tweet). 200 documents, which include 5 groups, 40 months (2019.01-2022.04), are used for topic modeling based on the 5 topics' LDA model *(36)*. Finally, 1000 clusters, each cluster has 20 keywords, are collected. The structure of the keyword's dataset can be seen in **Figure 2.**

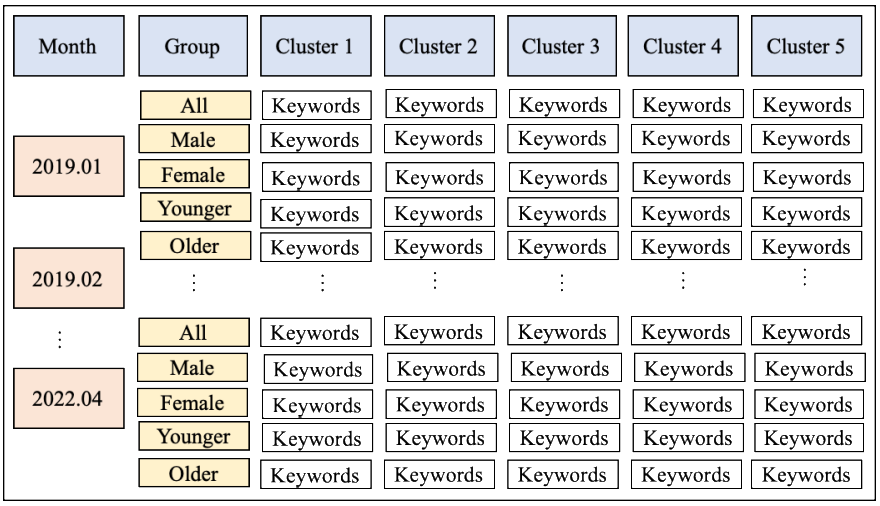


Figure 2 The structure of the keyword dataset

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. The same topics are always exited in the different clusters, for example, cluster 1 may has 3 topics which are extracted based on 20 keywords, as ridesharing cost, ridesharing service feels, and people in the car, the cluster 2 also has the same topics such as ridesharing cost and people in the car. Some studies considered the use of several clustering techniques to group keywords of clusters into predefined topics, such as Moreno *(36)* use method involves the 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. Meanwhile, those methods 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 modeling method that includes 3 steps: cluster ordering, topic generating and dataset labeling. The cluster ordering step is re-order 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 a good performance in the model, therefore, 5 clusters are sorted based on the enhancement of the coherence measure score. The topic generating step is the main step, before defining the name of each topic, previous research on ridesharing platforms and passengers' services were reviewed *(2)*. The six most discussed topics in the latest research include *ridesharing felling, time cost, money and payments, the people in the car, pick up location, and ridesharing company & model* are used as reference to this paper. In this paper, each keyword in the clusters is scored manually based on the correlation of words with six topics, using a 10-Likert scale (1-10). In the detail, 10 score means the keyword has a high probability of belonging to a certain topic. Once the keyword is got 1 score in all six topics, the new topic would be created based on the meaning of the keyword, for example, the word pandemic does not belong to mentioned topics, therefore, the pandemic is created as the seventh topic. In the dataset labeling step, after all of the keywords are scored, topics are generated and labeled as *topic-1, …, topic-n*. Then, the keyword is further sorted and labeled as secondary index based on the 10-Likert scale. For example, the keyword Uber and Uberpull are both belong to *topic-1* (ridesharing company), the Uber has 9 scores, and the Uberpull have 7 scores, therefore, the Uber is labeled as *1-1*, and the Uberpull is *1-2*, once the score is same, the label would random arrangement. Take the 2019.01, male's cluster 1 as an example, as can be seen in **Figure 3**, each keyword in the clusters is transferred to the topic label, such as in the left of figure, the "money" is changed to *3-2*, the first label means the topic, and the second label means the importance of keywords in the topic (keyword sort). Note that, only meaningful keywords are kept for topic modeling, words such as can, much, etc. which have no sense are deleted. After topic labeling, the dataset is changed to the statistic version.

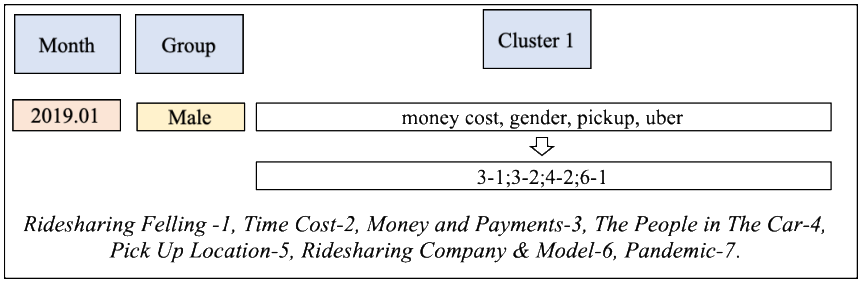


Figure 3 Example of topic labeling

Topic development trend analysis based on graph

To analyze the topic development trend based on time series, this paper put forward the graph method based on the labeled dataset. Let $(X-Y)_1$, ..., $(X-Y)_{40}$ donate the sequence of the topic monthly from 2019.01 to 2022.04. Two dynamic thresholds T_{X} and T_{Y} are calculated for determining the hot topic and content of each month. To improve the expressiveness of the graph, only the hot topic and content be created as a piece in graph. The two dynamic thresholds can be defined as:

$$T_{\mathrm{X}} = \mathrm{X} \mid \max\ frequence\ of\ (X) \qquad (1)$$

$$T_{\mathrm{Y}} = \mathrm{Y} \mid \max\ frequence\ in\ (T_{\mathrm{X}}) \qquad (2)$$

Where X is the topic labeling, and Y is the content labeling. In the graph construction, only the hot topic which has the max frequency present in each month will be created as such topic indicate a significant change in the ridesharing service.

Take an example as 2019.01, male's cluster 1-5 (**Figure 4**), the bottom of the graph is the time series of the topics, on the left of the graph, is the label of the topic and its content, the yellow piece means that in the 2019.01, the hot topic is 4, and the main content in this topic is 2. Note that, once the

frequency of X is equal, both topics will be presented in the graph, when it comes to Y, it is the same.

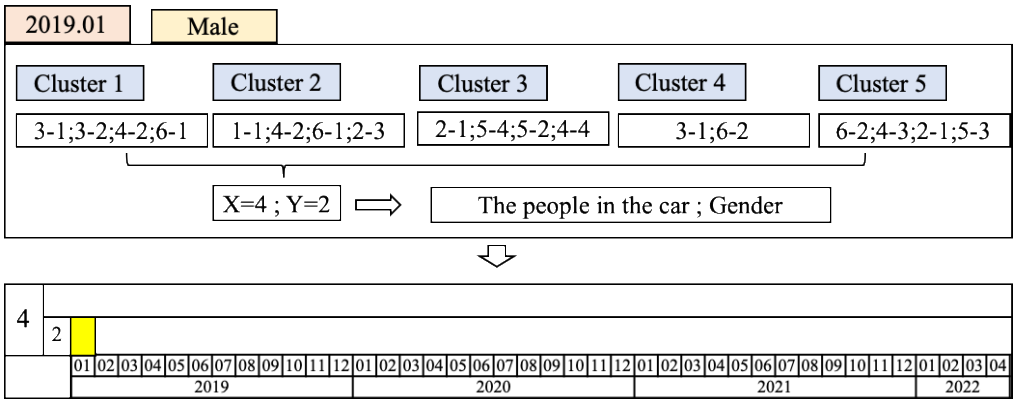
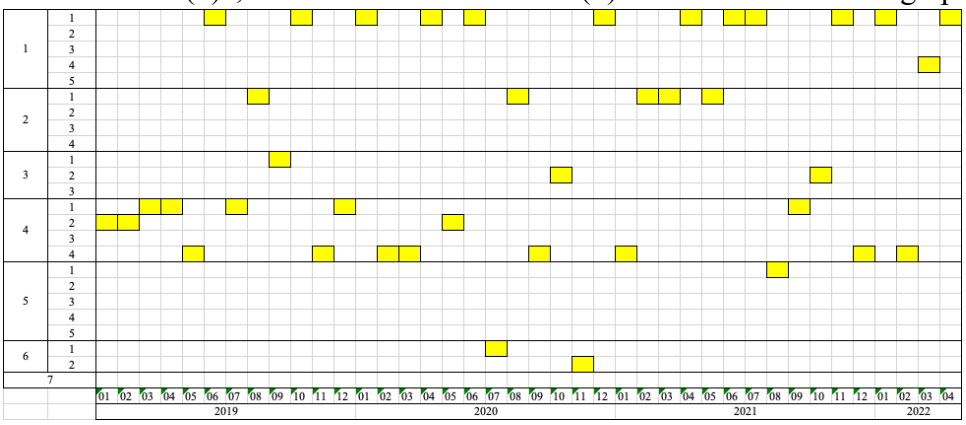
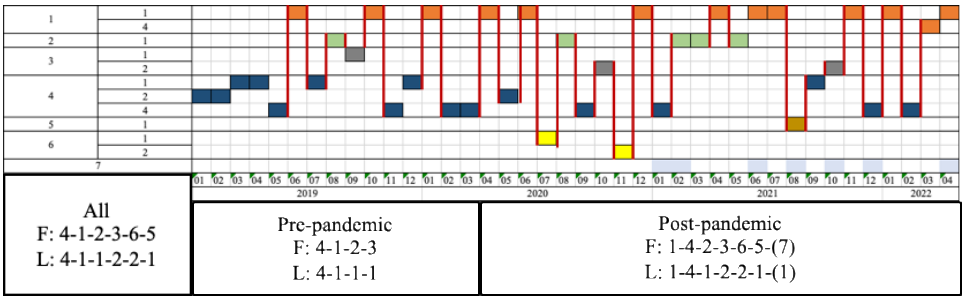


Figure 4 The piece of the topic analysis graph

The graph construction is built based on pieces as shown in **Figure 5 (a)**. The left side of the graph lists the topic and its content labels. The center of the graph shows the pieces for each month. To improve the visually of graph, the empty topic or content is removed. As can be seen in **Figure 5 (b)**, the red line means the transfer of topics in this month. At the bottom of the graph, it shows the passengers' characteristics such as all, male, female, younger, and older. The F means the frequency of the topic in the current group, the L means the hottest content of this topic. Then, both F and L of pre- and post-pandemic are collected and shown in the graph. The sort of the F and L is determined by the occurring frequency of contents. Note that, this paper takes the pandemic as a core variable in the analysis, therefore, once the pandemic topic is shown in the clusters, it would be marked in the graph, and the topic is labeled as "(7)", its' content is marked as "(1)" in the bottom of the graph.



a) The piece of the ridesharing topic graph



b) The construction of the ridesharing topic graph

Figure 5 Example of ridesharing topic graph

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 BERT model is used for extracting the sentiment of each text, then, the sensitivity and significance pre- and post-pandemic is analyzed. The logistic regression model is used for interpreting the correlation between passengers' characteristics, time series, and sentiment, and construing the regression model based on these variables. Therefore, the BERT model and logistic regression, which are related to this paper, are introduced as follows.

BERT Model

The sentiment model attains a greater accuracy of 92% for sentiment when utilizing the cased version of BERT analysis *(29)*. 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). This paper takes a dot product of these vectors with the second sentence's output vectors from BERT. Then this paper applies 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 *(32)*.

RESULTS

Topic Modeling Performances and Results

Keywords distribution and results: Tweets were collected from ridesharing customers addressing the @ridesharing Twitter platform from Jan, 2019 to April, 2022. The frequency of tweets in each month is shown in **Figure 6**. The results show that the number of tweets has a significant decrease during the period from Jan, 2022 to April, 2022. This decrease occurred with a high probability due to the ridesharing users' decrease during the pandemic. Then, the number of tweets is not increased, the trend shows a stable wave, this means that people share their problems and opinion on the other transport platforms or other topics post-pandemic *(17)*.

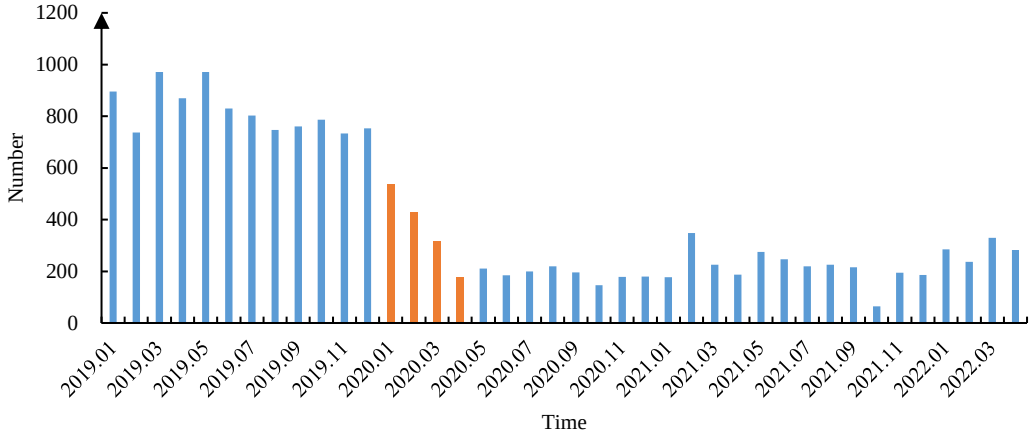


Figure 6 The tweet frequent each month

The frequency of keywords in each group (gender and age) pre- and post-pandemic are also collected for checking the most common and relevant issues posted on the platform. As shown in **Figure 7**, the word cloud shows each word with different size according to the frequency on the dataset. Before the pandemic, words like “Uber”, “time”, “driver”, and “girl” are the most frequent words in all groups. This result shows that people commonly tweet to @ridesharing to discuss problems or recommendations related to ridesharing company, the time cost, drivers' service, and passengers' gender. In the age group, youngers more care about the difference in ridesharing service between ridesharing companies such as Uber and Lyft, and the older are more concentrated on the trip time and the drivers' service. When it comes to gender groups, the male and females have the same opinion on ridesharing, their hot topics are both focused on ridesharing companies and drivers' services. When compare the difference pre- and post- pandemic, “money” and “safe” become a new common tweet topic post- pandemic. This result shows that more people are being absorbed by the ridesharing company's discounts, and they began to care about safety during the trip post-pandemic. Then, in the age group post- pandemic, younger still care about the difference between ridesharing companies, as well as they paid more interested in the time cost than pre-pandemic; The older is more care about the driver's service than pre-pandemic. In the gender group post- pandemic, the difference is still small, and they both care about the difference in the service between ridesharing companies pre- and post-pandemic.

The word cloud of ridesharing pre-pandemic (all-younger-older-male-female)

The word cloud of ridesharing post-pandemic (all-younger-older-male-female)

Figure 7 Word cloud of ridesharing pre- and post-pandemic

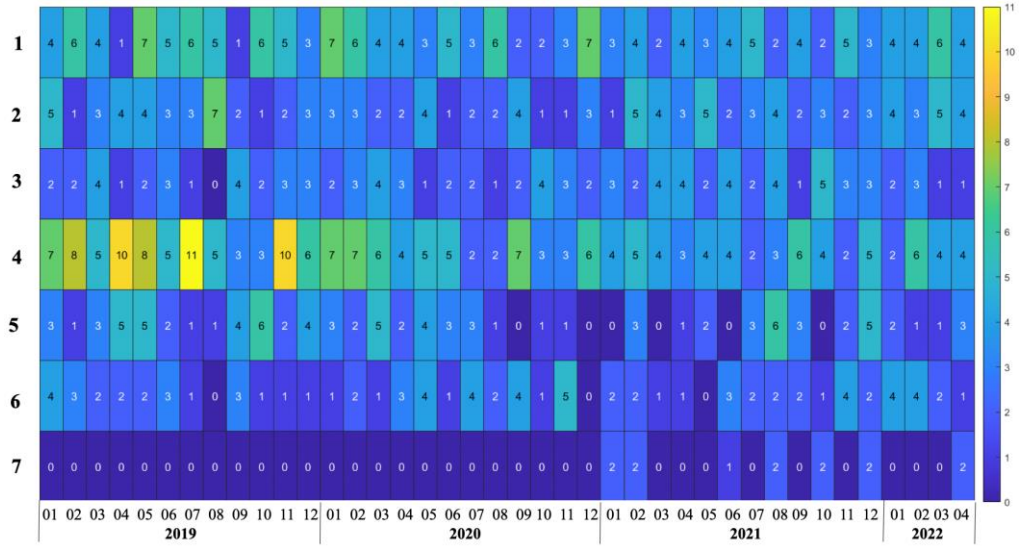
Topic labeling and development trend analysis

As mentioned in the methodology section, to perform the task of extracting the most discussed topics in each group pre- and post-pandemic, the clusters of five groups (all-younger-older-male-female) are extracted based on the LDA. Then, the clusters in each month are sorted based on the coherence score, and the topics and contents of each topic are labeled based on the proposed method. As a result, 7 topics are labeled, and 24 contents are assigned for those topics. The details of the topic and its content can be seen in **Table 2.**

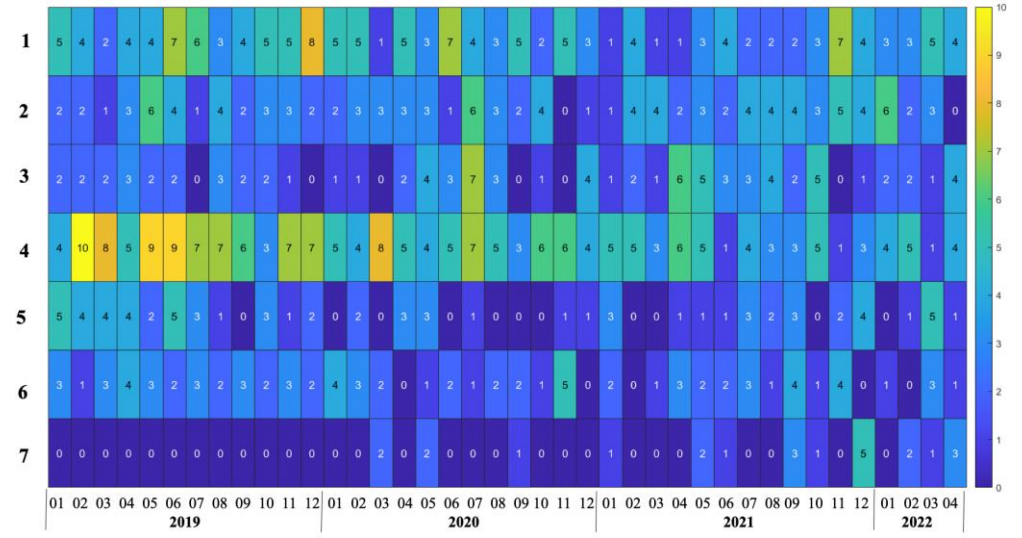
Table 2 Description of labeled topics and their contents

Label	Topic	Label	Content	Description
1	The feeling during a ridesharing trip	1	Feel	Feeling about the whole trip
		2	Car environment	Feeling about car environment
		3	Talk	Feeling about others
		4	Service	Feeling about service
		5	Safety	Feeling about trip safety
2	Trip time cost	1	The time	The time cost of the trip
		2	Wait	Wait time
		3	Morning	Trip happens time
		4	Night	Trip happens time
3	Money cost	1	Money	Ridesharing money
		2	Cost	Cost of the trip
		3	Pay	Pay after trip
4	The people in the car	1	People	People meet during the trip
		2	Gender	Another passengers' gender
		3	Passenger	Another passenger
		4	Driver	Drivers' service
5	Pick up location	1	Location	Depart location
		2	Family	Depart location
		3	Airport	Depart location
		4	Destination	Arrive location
		5	Pickup	Depart location
6	Ridesharing company/car	1	Car type	The car model and Uber-type
		2	Ridesharing company	Ridesharing company
7	Pandemic	1	Pandemic	The influence of pandemic

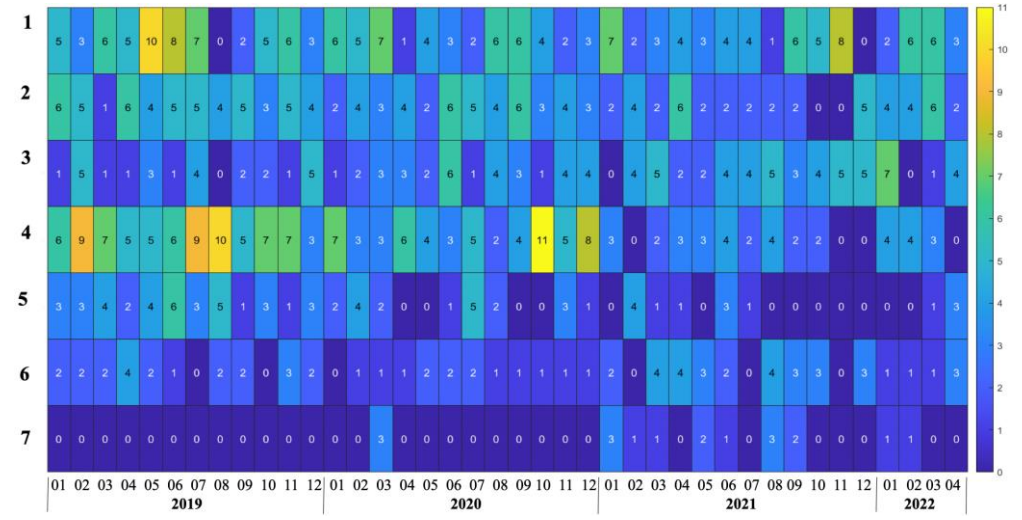
To analyze the transfer trend of the ridesharing topic, this paper constructed the topic modeling graph for each group based on time series. 5 topic trend graphs are built after topic modeling, main content extraction, topic piece construction, and graph optimization. After content extraction and topic piece construction, the distribution of each topic is shown in **Figure 8**, the numbers in the heatmap represent the frequency of each topic discussed in the month. In the all-tweet group, the “the people in the car-driver (4)”, “the feeling during ridesharing trip (1)” and “trip time cost (2)” are the popular topic. The same distribution has existed in the gender group, but male is more care about the topic of “pandemic (7)” and “pick up location (5)”. Moreover, the older have a more complex distribution than the younger, who pay more attention to “money cost (3)” and “ridesharing company (6)” than the younger.



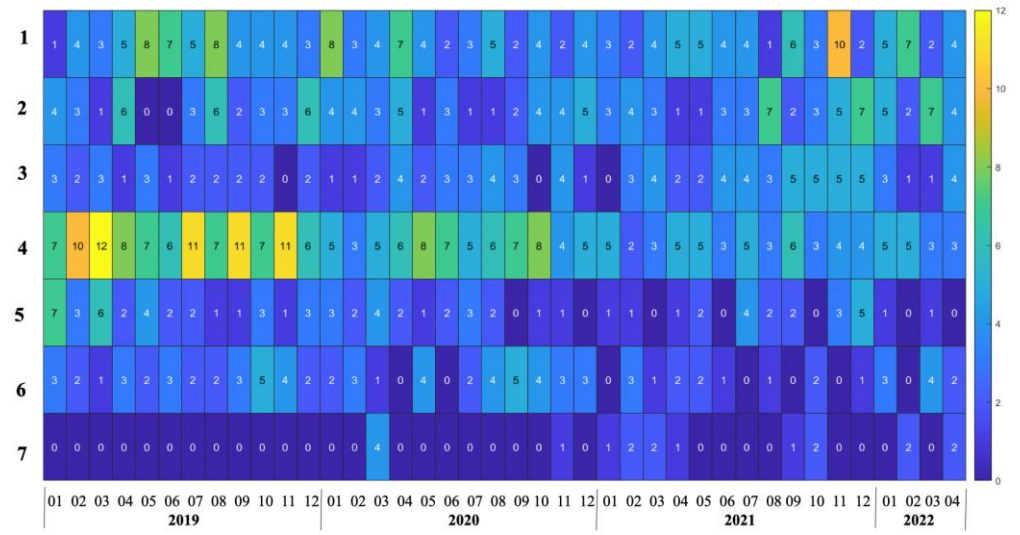
The distribution of ridesharing services in all group



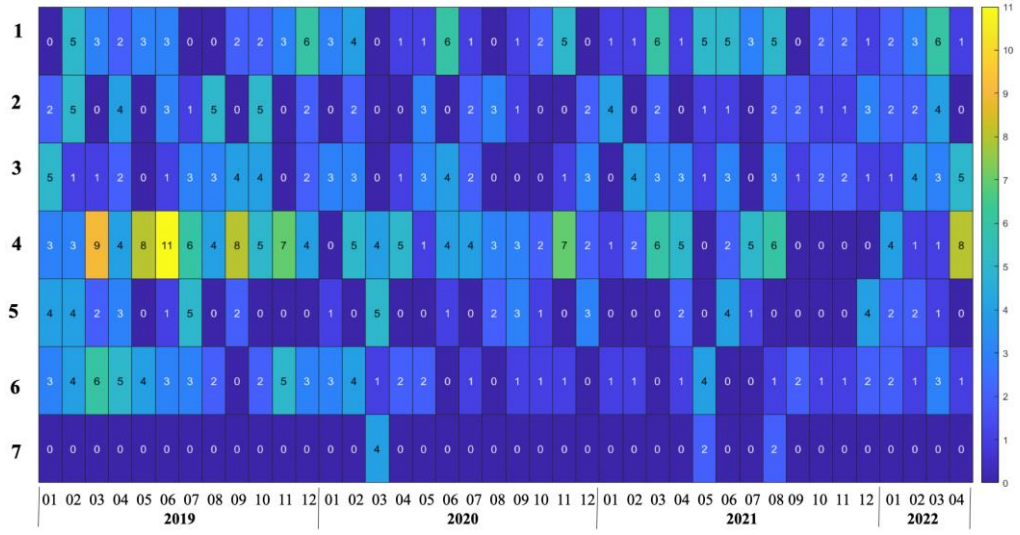
The distribution of ridesharing services in the male group



The distribution of ridesharing services in the female group



The distribution of ridesharing services in the younger group



The distribution of ridesharing services in the older group

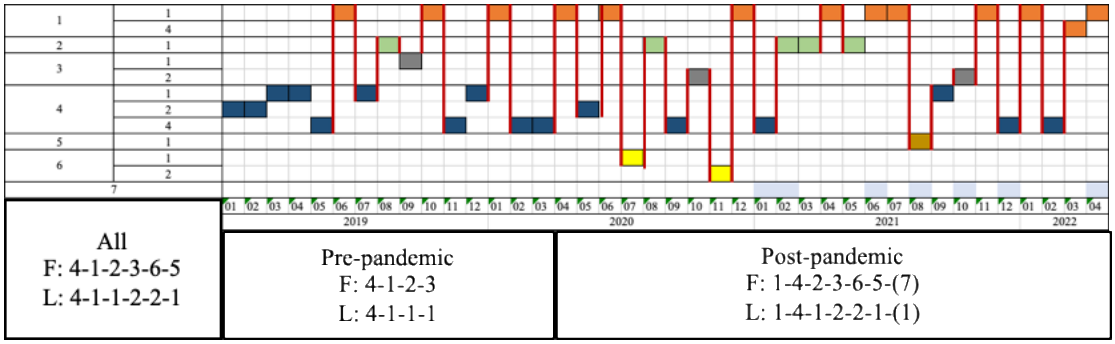
Figure 8 The distribution of ridesharing service topic modeling

The graph is constructed based on the above mention method, the results are shown in **Figure 9**, the trend of the topic in all groups shows that the mutual transfer phenomenon exists between topics. The combination with the greatest transfer frequency is from topics 1 to 4 (1-4). Topics such as “4”, “1” and “2” gain a high concern over time. The topic of the pandemic has been a hot topic since January 2021, the topic distribution has a significant difference between pre- and post-pandemic. More topics have present post-pandemic, such as topics “6” and “5” gained more attention post-pandemic. This could be caused by ridesharing companies putting forwards a lot of policies for customers to cope with the impact of the pandemic, the policy of each company has differences, therefore, bringing more discussion on Twitter. Then, the pickup location also has some changes to the pandemic, which cause some inconvenience to customers and thus get a hot topic during this period. Meanwhile, the sort of hot topics also has some change post-pandemic, the topic of “1” instead of “4” be the most popular topic, it indicated that customers more care about the service of drivers (pandemic prevention in the car, pickup speed, etc.) post-pandemic.

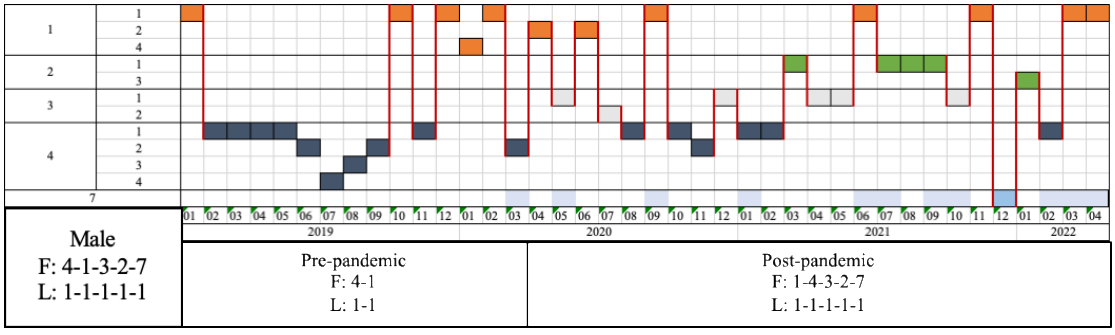
In the gender groups, both males and females paid attention to the topics “1”, “2”, “3”, and “4”, and

they began to focus on the pandemic from March 2020. The male more care about the pandemic and the driver' service than the feeling of a ridesharing trip, opposite with females. Males' transfer of topics is more frequently post-pandemic, the highest frequency of combination is “1-4” pre-pandemic, which is changed to “2-3” post-pandemic. Different from male, the high frequency of combination is “2-3” in the female group, which keep stable in the time series. In the male group, “2”, “3” and “7” gained more attention post-pandemic. It indicated that male is more care about the time and money cost post-pandemic. When it comes to females, they have more topics than male's pre-pandemic, and the hot topics are not changed post- pandemic, which shows that female can keep more stable concerns than males over time. Meanwhile, the sort of hot topic also has some change post-pandemic in both gender groups, the topic of “1” is the most popular topic, it indicated that customers more care about the service of driver post-pandemic. In the female group, the “2” and “3” become more popular, which indicated that they more care about time and money costs post-pandemic.

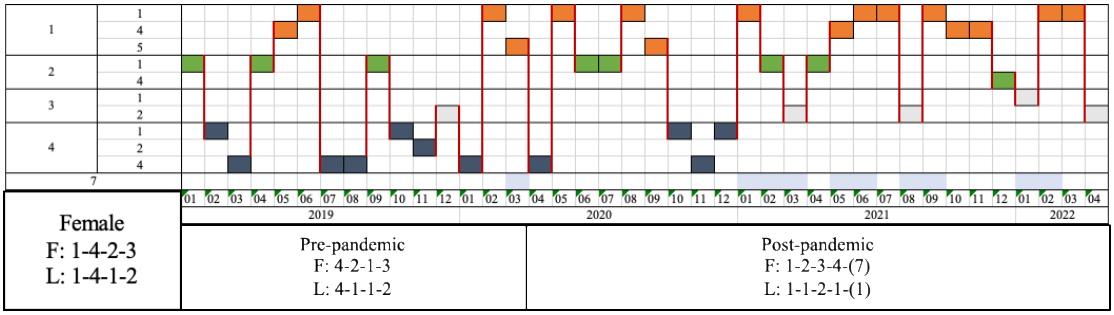
In the age groups, both groups have the sequence “4-1-2-3” of hot topics, and they began to focus on the pandemic in March 2020. The older have more hot topics than the younger (“5” and “6”). Youngers' transfer of topics is more stable than older, and the highest frequency of combination is “1-4”. Different from younger, the high frequency of combinations is “2-3” in the older group. In the younger group, “3” gained more attention post-pandemic. It indicated that younger is more care about the money cost post-pandemic. When comes to the older, they have more topics than the younger pre-pandemic, and the topic “6” is not popular during the pandemic, which shows that the older pay less attention to ridesharing companies post-pandemic. Meanwhile, the sort of hot topic also has some change post-pandemic in younger groups, the topic of “1” be the most popular topic, it indicated that customers more care about the service of driver post-pandemic. However, the order of hot topic in the older group keeps stable as time pass.



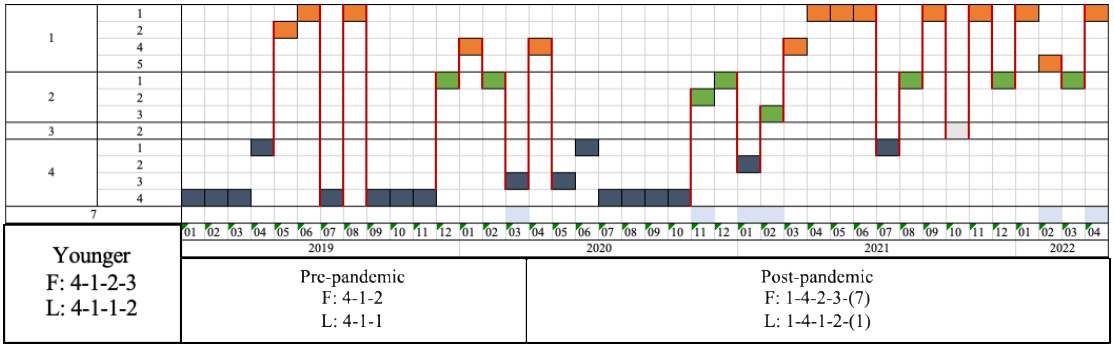
The trend graph of topics based on all tweet data



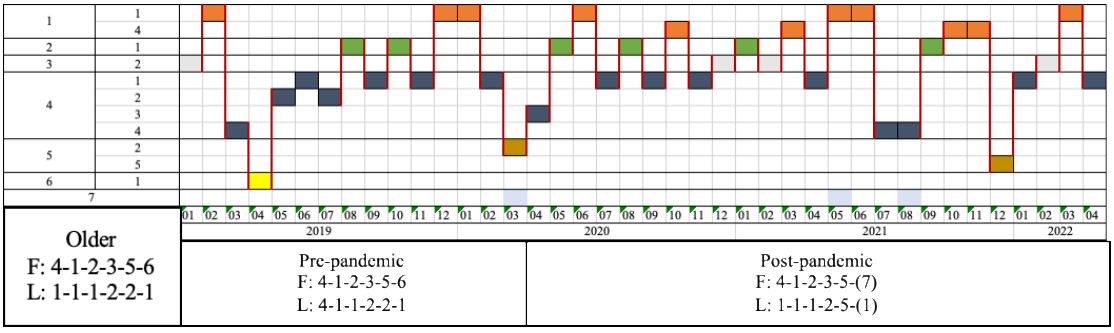
The trend graph of topics in the male group



The trend graph of topic in female group



The trend graph of topics in the younger group



The trend graph of topics in the older group

Figure 9 The trend graph of ridesharing service topic modeling

Sentiment analysis of ridesharing service

Sensitive and significant analysis: In this section, the BERT model has been employed to model the sentiment of each tweet. The proposed model has 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. The model is approximated to the ground truth with high accuracy (92%). To further verify the applicability of the model, 400 tweets' sentiment is checked manually, the result shows that the model still has high accuracy (76.4%) and Mean Absolute Error (0.12), which means this model has good performance in dealing with sentiment analysis problem of Twitter data.

The sentiments associated with the time series are shown in **Figure 10**. 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) 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, keeping the stable trend post-pandemic. This result may be that the ridesharing company has put some discount policy for customers to respond to the pandemic, which enhances the customer's positive sentiment.

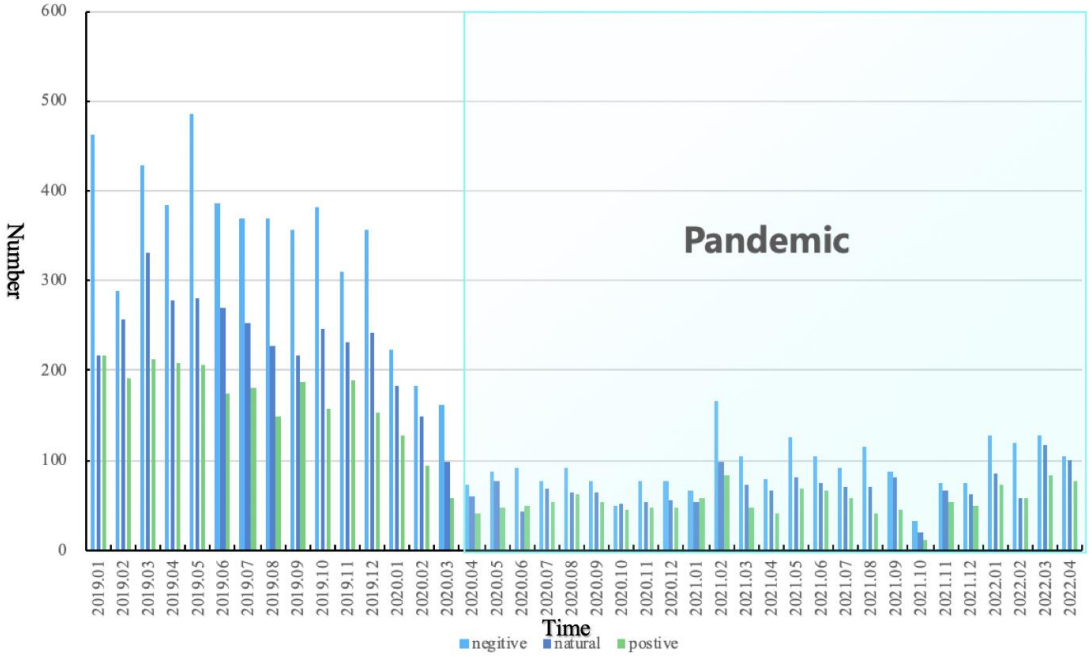
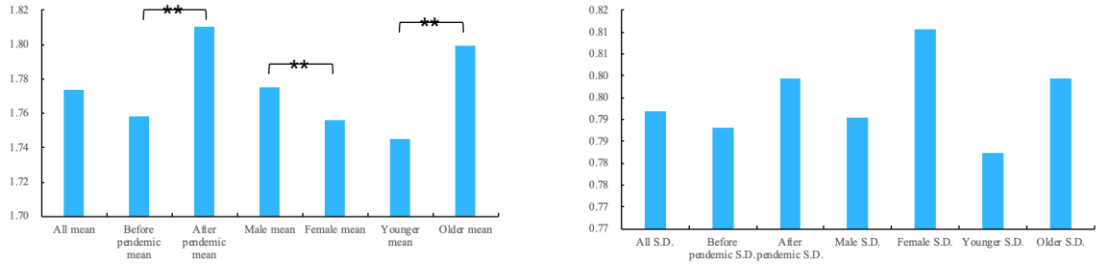
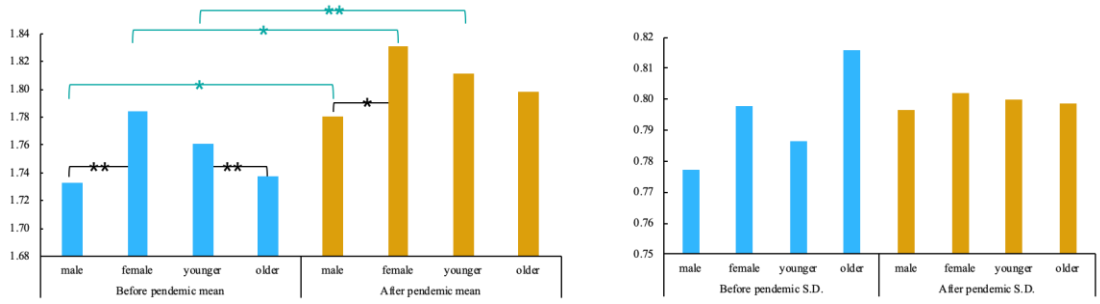


Figure 10 the volume of ridesharing sentiments associated with the time series

To further analyze the sentiment, the sentiment is assigned with values based on the positive enhancement, as the negative is 1, the natural is 2, and the positive is 3. To explore the sentiment difference between groups (pandemic, gender, and age), the mean and standard deviation (S.D.) of each group are compared. As can be seen in **Figure 11 (a)**, the sentiment of the time, gender, and age groups are all shown the significant difference. When comparing the mean of each group, customers show more positively post-pandemic, and the male, older customers have a more positive attitude toward ridesharing service. Meanwhile, the customer always keeps a more stable sentiment pre-pandemic. Similarly, the male and younger customers also have a more stable sentiment about ridesharing. Then, the difference between each group pre- and post-pandemic also be analyzed, as can be seen in **Figure 11 (b)**, the result shows that the sentiment of gender and age groups have significantly different pre-pandemic, but only gender groups have significant differences post-pandemic. Meanwhile, females and younger are more positive about the ridesharing service in all periods, and they keep a more stable attitude than males and older. When comparing the sentiment performance pre- and post-pandemic in each group, the result shows that the positive sentiment is increased in gender and age groups post-pandemic, it may be the customers are got more discounts from ridesharing companies such as Uber. Meanwhile, the attitude stability is decreased in the male, female and younger groups post-pandemic, which indicated that the pandemic has different degrees of impact on individuals. *Note that: * means that the difference was statistically significant at the significance level of 5% (0.01< p-value <0.05); ** means that the difference was statistically significant at the significance level of 1% (p-value <0.01).*



The difference between time series and customs' characteristics



The difference between customs' characteristics pre- and post-pandemic

Figure 11 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 & time series, this paper analyzed the correlation between each characteristic variable, time, and sentiment. There has a significant correlation between sentiment with time (0.065**) and gender (0.028**), but there has no significant correlation between sentiment and age (0.009). It indicated that the sentiment would change as time passed, and gender would also influence the sentiment. Then, the regression model has be modeled based on the multi-logit regression model, which was implemented by using the MATLAB built-in algorithm *(37)*. 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.=0.000), lower error (Log-Likelihood Ratio=143.24, X^2=34.30), and high model fitness (goodness of fit test=0.856). **Table 3** shows the result of the sentiment regression model. The result shows that time and gender are the main factors to influence the sentiment. As the time passed, the sentiment is more positive (OR=1.08, P<0.001), and the females is more positive than the males (OR=1.12, P<0.001).

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	0.58	0.07	79.55	1.00	0.00			
	Time series	0.16	0.03	20.44	1.00	0.00	1.17	1.09	1.25
	Gender	0.18	0.03	31.47	1.00	0.00	1.20	0.99	1.28
	Age	-0.09	0.06	1.75	1.00	0.19	0.92	0.81	1.04
2	Intercept	0.12	0.07	2.69	1.00	1.10	-	-	-
	Time series	0.08	0.04	4.62	1.00	0.03	1.08	1.02	1.16
	Gender	0.12	0.03	11.43	1.00	0.00	0.98	0.87	1.12
	Age	0.14	0.07	3.84	1.00	0.05	1.15	1.00	1.32

CONCLUSIONS

This paper presents a multivariable time series deep learning-based method for ridesharing service analysis which integrates the use of LDA and graph for modeling the ridesharing service topic, and the use of the BERT and multi-logit model for extracting the ridesharing service sentiment. The graphs for describing topics (from the text data to the label) are built through a data abstraction process. Each graph represents a specific group (all, gender, and age groups), and the content of graph construction includes the topic occurrence and trend through time and the topic distribution of different pre and post-pandemic in each group. Then, the sentiment of each tweet in each group based on time series is extracted using the BERT model for evaluating emotions of passengers towards ridesharing services. The methodology is applied in a case study with twitter data of ridesharing pre- and post-COVID pandemic. The result shows the model has good performance in sentiment analysis. The significance and correlation are analyzed, the regression is modeled using the multi-logit model, and main factors in influencing sentiment are determined. Key findings can be summarized:

1) The topic of ridesharing services is analyzed based on the LDA and graph. the trend of topics in all groups shows that topics such as “the people in the car”, “the feeling during ridesharing trip”, “trip time cost”, and “money cost”, are gain the high concerns. The topic of pandemic has been a hot topic since March 2021. More topics have present post-pandemic. The “money cost”, “ridesharing company” and “pick up location” gained more attention post-pandemic as ridesharing companies put forwards a lot of policies for customers to cope with the impact of the pandemic. Meanwhile, the sort of the hot topic also has some changes post-pandemic, the topic of “the feeling during ridesharing trip” instead of “the people in the car” be the most popular topic, it indicated that customers more care about the service of driver post-pandemic.

2) The topics have a significant difference between gender and age groups. In the gender group, the male more care about the pandemic and the drivers' service than the feeling of a ridesharing trip pre-pandemic, opposite with females. During the pandemic, the male is more care about the time and money cost than female. Females can keep more stable concerns than males over time. In the age group, the older has more hot topics than the younger. Youngers' transfer of topics is more stable than older, and more care about the money cost post-pandemic. The older more care about the pickup location than younger.

3) The sentiment of the ridesharing service is extracted based on BERT and analyzed based on the multi-logit regression model. As a result, the sentiment of the time, gender, and age groups are all shown a significant difference. Customers show more positively after the pandemic, and the male, older customers have a more positive attitude toward ridesharing service. Then, the sentiment of gender and age group have significantly different pre-pandemic, females and younger are more positive about the ridesharing service. The sentiments of male, female, and younger are show significant differences between pre- and post-pandemic, the positive sentiment is increased in all gender and age groups post-pandemic, and the pandemic has different degrees of impact on individuals. As the result of the regression model, time series and gender are the main factors to influence the sentiment. As time passed, the sentiment is more positive, and the females is more positive than the males.

As the main contribution, this study provides a new method for ridesharing service topic modeling and sentiment analysis. The framework of this paper considers topic occurrence and trend change, sentiment time series variables which have been hardly done before. The proposed

methodology framework considers the change in occurrence and trend of topics, and sentiment time series feature, which have been hardly explored before. Work of this paper can be used for ridesharing industries in enhancing the service, and provide, for instance, the suitable price standard. Meanwhile, the method proposed in this paper can be used for prediction and correlation analysis of topics, which helps relevant enterprises to take relevant measures in advance to deal with the occurrence of special events, such as pandemic.

In future studies, additional research should be carried out on this topic to validate the modeling approach and the algorithm. More importantly, the sentiment analysis should be associated with the topic modeling. The emotion analysis also can be used to enhance the detailed analysis of sentiment, more variables such as sad, happy, wiled, etc., can be extracted based on the future study. Finally, the developed method in this paper can be integrated with ridesharing company analysis software for determining and predicting the most suitable service and price design to test its performance and robustness in practical uses.

ACKNOWLEDGMENTS

This research was supported by the Shanghai Sailing Program (20YF1451800), the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), the Shanghai Municipal Commission of Science and Technology Project (19511132101).

We thank Amir Mahdi Khabushani for help through the data collection.

AUTHOR CONTRIBUTIONS

The authors confirm contributions paper as follows: study conception and design: Wenxiang Xu, Anae Sobhani, Ting Fu; data collection: Wenxiang Xu, Hongzhang Mu; analysis and interpretation of results: Wenxiang Xu, Ting Fu; draft manuscript preparation: Wenxiang Xu, Ting Fu, Anae Sobhani, Junhua Wang. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Mahmud, M. S., Bonny, A. J., Saha, U., Jahan, M., Tuna, Z. F., & Al Marouf, A. (2022, March). Sentiment Analysis from User-Generated Reviews of Ride-Sharing Mobile Applications. In 2022 6th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 738-744). IEEE.
2. Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and research: a systematic literature review through text mining. IEEE Access, 8, 67698-67717.
3. Karami, A., & Gangopadhyay, A. (2014). Fftm: A fuzzy feature transformation method for medical documents. UMBC Faculty Collection.
4. Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks.
5. Shaheen, S., Chan, N., Bansal, A., & Cohen, A. (2015). Shared mobility: A sustainability & technologies workshop: definitions, industry developments, and early understanding.
6. Morshed, S. A., Khan, S. S., Tanvir, R. B., & Nur, S. (2021). Impact of COVID-19 pandemic on ride-hailing services based on large-scale Twitter data analysis. Journal of Urban Management, 10(2), 155-165.
7. 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.
8. Karimi, A., Rossi, L., & Prati, A. (2021, January). Adversarial training for aspect-based sentiment analysis with BERT. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 8797-8803). IEEE.
9. Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. Mathematical and Computational Applications, 23(1), 11.
10. W.-C., Hong, Dong, Y., Zheng, F., and Wei, S. Y., Hybrid evolutionary algorithms in a SVR traffic flow forecasting model. Appl. Math. Comput., 2011. 15: 6733-6747.
11. X., Feng, Ling, X., Zheng, H., Chen, Z., and Xu, Y., Adaptive multi-kernel SVM with spatial-temporal correlation for short-term traffic flow pre- diction. IEEE Trans. Intel. Transp. Syst., 2019. 6: 2001–2013.
12. Yanamandra, V. H., Pant, K., & Mamidi, R. (2021, September). Towards Sentiment Analysis of Tobacco Products' Usage in Social Media. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) (pp. 1545-1552).
13. Liao, S., Wang, J., Yu, R., Sato, K., & Cheng, Z. (2017). CNN for situations understanding based on sentiment analysis of twitter data. Procedia computer science, 111, 376-381.
14. Zheng, H., Lin, F., Feng, X., and Chen, Y., A Hybrid Deep Learning Model With Attention-Based Conv-LSTM Networks for Short-Term Traffic Flow Prediction. IEEE Transactions on Intelligent Transportation Systems,·2020. 8: 1-10.
15. Collins, C., Hasan, S., & Ukkusuri, S. V. (2013). A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. Journal of Public Transportation, 16(2), 2.
16. 60 Incredible and Interesting Twitter Stats and Statistics. (2020). Brandwatch. https://www.brandwatch.com/blog/twitter-stats-and-statistics/#:%7E:text¼Twitter%20user%20statistics,users%20write%2080%25%20of%20tweets.
17. Beck, M. J., & Hensher, D. A. (2020). 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.
18. Poushter, J. (2016). Smartphone ownership and internet usage continues to climb in emerging economies. Pew research center, 22(1), 1-44.
19. Mcauliffe, J., & Blei, D. (2007). Supervised topic models. Advances in neural information processing systems, 20.
20. 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.
21. 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.
22. Karami, A., & Shaw, G. (2019). An exploratory study of (#) exercise in the Twittersphere. iConference 2019 Proceedings.
23. 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.
24. 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.
25. 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.
26. Karami, A., & Collins, M. (2018). Social Media Analysis for Organizations: US Northeastern Public and State Libraries Case Study.
27. 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).
28. 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.
29. G. Henry, "Improved sentiment analysis using a customized distilbert NLP configuration", Advances in Engineering: An International Journal (ADEIJ), Vol.3, No.2
30. J., Tolles, and Meurer, W. J., Logistic Regression Relating Patient Characteristics to Outcomes. JAMA, 2016. 5: 533-540.
31. G., Ali, and M.A., M., Utilizing naturalistic driving data for in-depth analysis of driver lane-keeping behavior in rain: Non-parametric MARS and parametric logistic regression modeling approaches. Transportation Research Part C: Emerging Technologies, 2018. 379-392.
32. Agresti, A., An Introduction to Categorical Data Analysis. 2007. John Wiley and Sons Inc.
33. Twitter Intelligence Tool (TWINT). Available online: https://github.com/twintproject/twint (accessed on 21 June 2021).
34. 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.
35. 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.
36. Moreno, A., & Iglesias, C. A. (2021). Understanding Customers' Transport Services with Topic

Clustering and Sentiment Analysis. Applied Sciences, 11(21), 10169.
37. Mathworks. Multinomial logistic regression. (2022) Available online: https://nl.mathworks.com/help/stats/mnrfit.html.