Augmenting Ridership Data with Social Media Data to Analyse the Long-term Effect of COVID-19 on Public Transport

Xu, Yanyan; Krishnakumari, Panchamy; Yorke-Smith, Neil; Hoogendoorn, Serge

DOI
10.1109/MT-ITS56129.2023.10241480

Publication date
2023

Document Version
Final published version

Published in
2023 8th International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2023

2023 8th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) | 978-1-6654-5530-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/MT-ITS56129.2023.10241480

Augmenting Ridership Data with Social Media Data to Analyse the Long-term Effect of COVID-19 on Public Transport

Yanyan Xu[1,*], Panchamy Krishnakumari[1], Neil Yorke-Smith[2], and Serge Hoogendoorn[1]

Abstract—**COVID-19 significantly influenced travel behaviours and public attitudes towards public transport. Various studies have illustrated complicated factors related to long-term travel behaviour, indicating difficulty in understanding and predicting post-pandemic long-term travel behaviour via traditional methods. In these complex circumstances, it is valuable to take advantage of social media data to obtain real-time public opinions to understand dynamic travel behaviour changes from the passenger perspective. The present study provides a means – leveraging Twitter data and state-of-art Natural Language Processing (NLP) technologies – to interpret the underlying associations among public attitude, COVID-19 trends and public travel behaviour. Concretely, New York City has been selected due to its dependence on public transit for daily commuting. More than 500K tweets have been collected from January 2019 to June 2022. Automated text mining, topic modelling, and sentiment analysis have been implemented in these contexts to identify dynamic public reactions. A consistently negative attitude to public transit is detected and five main topics, including derivative topics from COVID-19, are discovered within the COVID-19 duration. Policy makers and transit managers can use these topics to take onboard the public's concerns. The paper thus exemplifies how social media data and NLP technologies can support policy-making progress and can benefit other tasks in the transportation domain.**

Index Terms—**public transport travel behaviour; social media; topic modelling; sentiment analysis; COVID-19; natural language processing**

I. Introduction

The COVID-19 pandemic dramatically influenced travel behaviour and urban mobility system worldwide. The fast spread of the virus in closed environments raised the public's risk awareness of public transit [1]. With lockdown policies, remote working modalities, and unpredictable variant viruses, public transport travel demand became much more complex. Although extensive research has illustrated the impact of COVID-19 on travel behaviours in both short-term and long-term – and pointed out the gradual recovery trend of mobility patterns [2] – at the time of writing, public transit patterns have not returned to the pre-COVID-19 norms in many countries. Further, the long-term impact involves socio-economic and cultural factors, which complicate attempts to explain and predict the dynamic public transit pattern in the post-COVID-19 period (i.e., the time after the end of pandemic phase). Despite increasing interest in the literature on travel behaviour change with such factors, there remains insufficient attention on incorporating the passenger perspective to manifest the underlying individual travel demand and policy requests [3]. Therefore, studying the long-term impact of COVID-19 on public transportation systems from user perspectives will support efficient transit management and provide rational policy options for policy-makers. Specifically, this paper addresses three questions:

1) How has COVID-19 influenced attitudes towards public transportation and how is it associated with public transport travel behaviour in the long term?
2) How can NLP technologies improve the understanding of the change in public transport travel behaviour with respect to COVID-19?
3) What are the potential advantages of integrating social media data with other data resources (mobility data, socio-economic data, etc.) in the transportation domain for future applications?

Researchers have explored related reasons for the travel behaviour changes during the pandemic. Zuo et al. [4] illustrated the sharp decline of transit ridership in New York City after the effectiveness of the stay-at-home policy at the early stage of COVID-19. Policy interventions, socio-economic and culture factors have also been discussed as how their complicated relations influencing mobility patterns in the long-term. Liu et al. [5] implemented logistic regression analysis to explore the transit ridership decline ratio with socio-economic reason in the US. Van Wee et al. [6] analyzed the possible long-term effects of COVID-19 on activity-travel behaviour based on the first and second waves from economic, psychology, sociology, and geography perspectives. Maleki et al. [7] revealed the importance levels of various socio-economic and policy variables in affecting human behaviours in the US. These studies illustrated how different parameters integrated together influencing mobility trend, but most of them analyzed from the top-to-bottom perspective which lack of the understanding from individual considerations in travel decision-making progress in the long-term.

A recent trend is to study the pandemic's influence on transportation systems from the passenger perspective. Shelat et

[1] Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2628CN Delft, Netherlands

[2] STAR Lab, Delft University of Technology, 2600GA Delft, Netherlands

*Corresponding author: Y.Xu-10@tudelft.nl

This work received funding from the European Union's Horizon 2020 programme under project TULIPS, grant agreement 101036996.

978-1-6654-5530-5/23/$31.00 ©2023 IEEE

al. [8] utilized a stated choice experiment from train passengers in the Netherlands to explore the relationship between public attitude – 'COVID conscious' and 'infection indifferent' with travel behaviour within the first wave of pandemic. Bert et al. [9] pointed out that general consideration of 'social distancing' and 'cleanliness' has a significant influence on travel decisions, which leads to a large increase in private vehicle usage in both the US and China. This study also showed that public transit usage declined after COVID-19 ended in China, indicating that public concerns and negative attitudes towards public transit are inconsistent with COVID-19 cases. Anke et al. [10] stated that shifts from public transit mainly because of the risk perception instead of the effect of lockdown policy, based on a German micro-mobility online survey. All these studies show the necessity of studying public perception of travel behaviour change during the pandemic. However, most of them are based on costly-to-acquire, non-real-time survey data. It motivates the need for a comprehensive, user-orientated analysis of public emotion and opinion via real-time social media data associated with individual travel behaviour choices.

There is precedent of taking advantage of natural language processing (NLP) technologies via social media data to comprehend public reactions to COVID-19 and track their changes in time series. Xue et al. [11] used social media data to analyze public sentiment trends by identifying people's emotions in both US and China to help policy-makers better understand people's needs and make optimal policies. Ye et al. [12] pointed out a significant time lag between travel-related tweets and urban mobility trends. However, such research has focused on the general sentiment analysis with travel behaviour in the first wave of COVID-19: there is a lack of topic modelling application to figure out how sentiments change based on state-of-art NLP technologies in relation to long-term effect on public transport travel behaviour.

Based on the research gaps identified above, this paper proposes an interpretive analysis framework using social media and NLP methods for sentiment and topic analysis to comprehend the long-term public transport travel behaviour change before, within, and after pandemic.

II. Methodology

We propose three analytical methods to understand the dynamic public transit patterns: (1) time-series comparison of public transit ridership with COVID-19 cases and related policy; (2) sentiment analysis on public attitudes towards public transit and COVID-19; and (3) topic modelling for specific context mining. Fig. 1 overviews the paper's framework.

A. Time Series Comparison

In order to observe relations among COVID-19 cases, public transit ridership, sentiment of public transit ('PT sentiment') and COVID-19 ('COVID-19 sentiment'), we implement a descriptive analysis via comparing these factors in time series. Firstly, the ridership ratio is calculated by comparing per day after COVID-19 with the same date before COVID-19 in 2019 to obtain the per-day change ratio. Then we applied descriptive analysis in comparing multiple pairs of time series data:

1) ridership volume/ratio with COVID-19 positive cases;
2) PT sentiment with COVID-19 sentiment;
3) PT sentiment and COVID-19 sentiment with ridership volume/ratio;
4) PT sentiment and COVID-19 sentiment with COVID-19 positive cases.

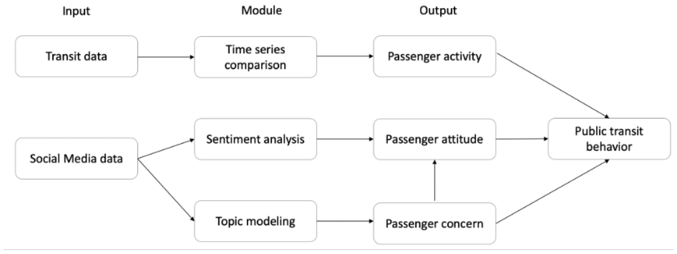

Fig. 1. Flowchart of analysis framework

We also aligned COVID-19-related events and policy with the factors mentioned above to explore the policy impact on public transit in different phases of COVID-19.

B. Sentiment Analysis

In order to acquire passenger perception on public transit and explore how it relates with public ridership change among different phases of COVID-19, we perform a sentiment analysis of Twitter data. In this study, since our data has a specific focus on COVID-19-related topics, we fine-tuned the COVID-Twitter-BERT (CT-BERT) model [13] for sentiment classification into three labels (positive, neutral, negative). CT-BERT was developed from the BERT model [14], [15] and has been proven to have a higher performance than baseline BERT (around 10% to 30%) on COVID-19-related tasks [13].

C. Topic Modelling

We implement the BERTopic model to get a more comprehensive understanding of public concerns among various COVID-19 phases [16]. First, we embed the pre-processed tweets into numerical vectors based on the pre-trained language model – a deep bidirectional transformer 'paraphrase-MiniLM-L12-v2' [17]. In this embedding process, each tweet is considered as a single document and the semantic relations among each word could be efficiently captured. To reduce the dimension of the embedding, we used Uniform Manifold Approximation and Projection (UMAP) algorithm [16] with cosine-similarity to measure the distance between each embedding. Then, we clustered the embedding into various topics via the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm [18] to find the similarity of each content. Meanwhile, cosine similarity has been applied to figure out the best numbers of clusters by iteratively integrating topics in the BERTopic model. Finally, dynamic topic modelling has been implemented via Class-based term frequency-inverse document frequency (c-TF-IDF) representations of topics to explore each topic trend in time series. The model will first generate the main topics based on the whole period of time and then calculate each topic's local representation by multiplying the term frequency of documents at different timestep [16].

III. Experimental Setup

We used the three analytical methods above for dynamic public transit pattern analysis in the case of New York City (NYC) using Twitter data. First, we explain the data sources and how we pre-process the data. Then we detail how the three analytical methods are applied to the specific scenario.

A. Data

The three datasets are: social media data (i.e., Twitter), public transit ridership data, and COVID-19-related data. The study period reflects our interest in the long-term impact of COVID-19 from January 2019 to June 2022, which includes the pre-COVID-19 term, COVID-19 term, and post-COVID-19 term based on the COVID-19 positive cases and policies in NYC. Since NYC officially experienced COVID-19 on 1 March 2020, we collected COVID-19-related data from that date.

- **Social media data.** Two types of Twitter datasets have been collected via Twitter streaming API and Python library *Snscrape*: public transit-related tweets, and COVID-19-related tweets in NYC. Then, six queries ('subway', 'metro', 'trains', 'MTA', 'public transit', 'public transportation') for public transit-related tweets and eight queries ('Covid19', 'coronavirus', 'virus', 'pandemic', 'disease', 'SARS-CoV-2', 'Delta variant', 'Omicron variant') for COVID-19-related tweets have been used respectively to filter the raw tweets from 1 Jan 2019 to 30 June 2022 and 1 February 2020 to 30 June 2022 respectively. Finally, 253,940 public transit-related tweets and 368,520 COVID-19-related tweets were extracted for the next step.
- **Public mobility trend data.** We collected the subway ridership data from MTA Turnstile in NYC to represent the public transit-related mobility trend because the majority of public transit happened in the subway/metro in NYC [19]. This dataset contains 432 metro stations in New York City with Entries and Exit ridership from 1 Jan 2019 to 30 June 2022. The daily amount of ridership with COVID-19 cases in time series is shown in Fig. 2.
- **COVID-19 cases and related policy data.** The COVID-19 dataset comes from the New York Department of Health, which contains the number of positive cases by diagnosis date from 1 March 2020 to 30 June 2022 in NYC. The COVID-19-related policy and events are collected from online resources [20] . During the study period, there were four COVID-19 waves in NYC related to lockdown policy, variant viruses, vaccines, etc.

B. Ridership Analysis

We first performed the descriptive analysis to explore how public transport travel behaviour change at different stages of time by clustering subway ridership per day via its volume and ratio and comparing it with COVID-19 positive cases. Meanwhile, we aligned the COVID-19-related events and policy in time series with the ridership change to figure out the relations between policy and public transit.

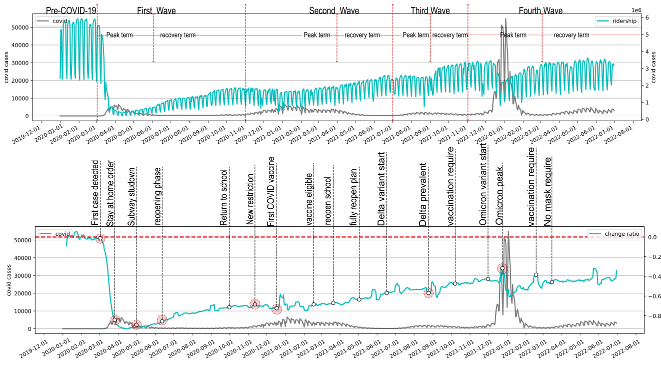

Fig. 2. Ridership change with COVID-19 related policy

C. Sentiment Analysis

We implemented sentiment analysis on the two Twitter datasets which sentiment on COVID-19 news (COVID-19 sentiment) is considered as a baseline to provide a comparison reference with the sentiment on public transit (PT sentiment). In order to get a more accurate performance on this domain-specific task, we fine-tuned the CT-BERT model to classify the sentiment of each tweet (positive as 1, negative as -1, neutral as 0). Specifically, we used *NTKL* library in Python for text cleaning. Then, we fine-tuned CT-Bert model by 'COVIDSenti' data which contains 9000 tweets with 6,280 positive, 16,335 negative, and 67,835 neutral for the classification task [21]. Then we used the fine-tuned model on the two pre-processed tweets data to obtain the sentiment labels for each tweet. Third, we compared and visualized PT sentiment and COVID-19 sentiment in time series. The changing points have been detected compared with public transit ridership, COVID-19 cases and related policies.

D. Text-Mining and Topic Modelling

After obtaining the PT sentiment and COVID-19 sentiment, we first applied text-mining to get a general overview of public perceptions before and after COVID-19 by aggregating the pre-processed texts monthly. Then, we applied topic modelling via BERTopic to get a deeper insight into public concerns on public transit and how various dominant topics changed among different time stages. Specifically the steps in implementing topic modelling are: first, contents from PT-related tweets from Jan 2019 to June 2022 have been clustered to get the main topics with keywords and relations of each topic by visualizing the distance among them. Then, dynamic topic modelling was applied to these topics in time series to observe how public perception changes over a longer time. Finally, we calculate the percentage of each topic in the five stages to yield a more quantitative interpretative explanation of the possible public transport travel decision change from the passenger perspective.

IV. Result and Discussion

A. Ridership Pattern explanation

In Fig. 2 we divide the overall trend of public transport travel behaviour into five stages: pre-COVID-19 and the following four waves. More specifically, it started with a dramatic decrease in the first wave of COVID-19 (March 2020) and was followed by a gradual recovery stage, including three slight drops in each COVID-19 wave. Among each wave, a peaking term and recovery term has been observed. After aligning the policy in time series, we found the related events and policy had a strong and immediate influence (red circle) during the outbreak of COVID-19, like ‘stay-at-home-order’ or ‘reopening’ but gradually diminished its impact in the third and fourth waves. Thus, it is necessary to explore other possible factors influencing travel behaviour as public transit patterns still did not recover to the pre-COVID-19 period even when the COVID-19 restrictions were lifted.

B. Sentiment Analysis

Based on CT-BERT model, we classified COVID-19-related data and public-transit related data into three sentiment types (negative as -1, neutral as 0, positive as 1). Then we aggregated to obtain the average sentiment score per day and visualized in Fig. 4. We notice that the average PT sentiment (blue line) is always in negative mode even before COVID-19, which leads to difficulty in identifying specific factors that causes this frustrating condition. On the other hand, it underlines the essential role of topic modelling for its powerful ability to efficiently extract aspects that provide deeper insights of understanding dynamic travel behavior. Meanwhile, we conducted COVID-19 sentiment in New York City(yellow line) as a baseline compared with PT sentiment. In our original hypothesis, PT sentiment will be largely influenced by COVID-19 sentiment during the pandemic. However, it shows an opposite trend in the first wave and gradually turns to follow the similar trend in subsequent waves. Especially, we detected the changing points in dark blue dash squares(April 2020, Feb 2021, May 2021 and March 2022) to identify the ‘abnormal periods’ when COVID-19 sentiment and PT sentiment have opposite trends. To get a deeper understanding of how these sentiment trends related to public transport travel behaviour and COVID-19 polices, we compared these factors in time series in Fig. 3. In the upper sub-figure, we pointed out the time lag between PT sentiment and ridership decreasing at each wave began in the dark blue squares. In the bottom sub-figure, we aligned COVID-19-related policy to these changing points and found it could only explain part of the apparent changes and for the rest, we still need to consider other factors. Thus, our next step is to implement topic modelling to obtain a comprehensive understanding of all the changing patterns.

C. Topic Modelling

Based on the pre-processed tweets data, we first executed text-mining with *Wordcloud* library in Python to extract the top keywords per month from Jan 2019 to June 2022 to get an overview on public concerns changing trend before and after COVID-19 Fig. 5. It can be clearly seen that public focused more on services issues (time, stop, delay, etc.) before COVID-19. As COVID-19 started to spread in March 2020, the keywords immediately turned into COVID-19 related topics (coronavirus, cleaning, mask, etc.), which reflected a close relationship between COVID-19 and public ridership at that outbreak stage. However, after the first wave, the keywords gradually turned back to transit services issues, weakly focusing on COVID-19. This trend also reflected ridership changing in the long-term with a meaningful connection between public perception and public transport travel behaviour.

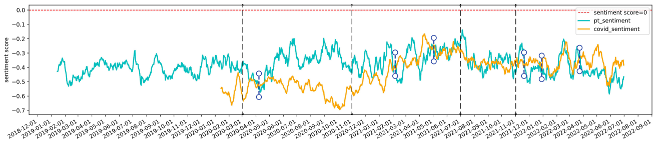

Fig. 3. PT sentiment vs. COVID-19 sentiment

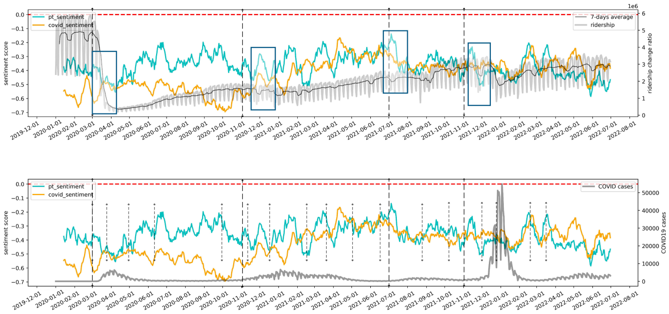

Fig. 4. Sentiment trend with COVID-19 cases and public ridership

To acquire a more quantitative measurement of public concerns, we implemented BERTopic to cluster the active topics. In Fig. 6 and Fig. 7, five main topics with their keywords and distance among each topic have been extracted from Jan 2019 to June 2022. Specifically, the topic of service (time, stop, delay) accounts for the domain concern in the whole period. The topic of COVID-19 (coronavirus, virus, workers) and topic of mask (mask, cleaning) occupied the second and third amounts which close to each other. While the fourth cluster is the topic of safety (shooting, police, NYPD) and the last is the topic of homeless (homeless, home, ride).

The dynamic topic modelling result from Fig. 8 demonstrates topics change in time series. In the pre-COVID-19 period, the topic of transit service dominated, while with the outbreak of COVID-19, the topic of COVID-19 and topic of mask abruptly peaked with the decrease of public concerns on disappointing transit services. When the first wave turned into a recovery term, the transit service topic started to recover and gradually re-dominate the public attentions. It is worth noting the topic of homeless started to rise at the outbreak of pandemic and keep a relatively active level in following time, which indicate the COVID-19-related homeless issue that policy-makers cannot ignore. Moreover, from calculating

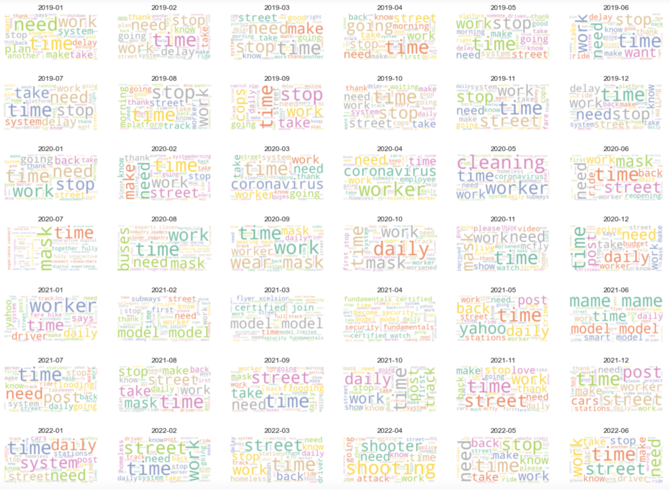

Fig. 5. Top keywords in each month

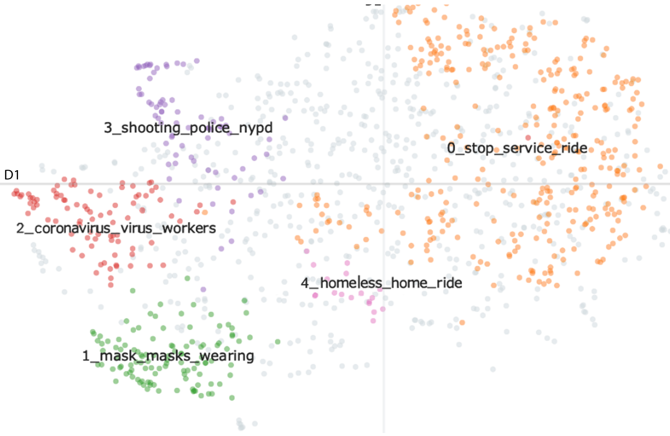

Fig. 6. Overall topic clustering

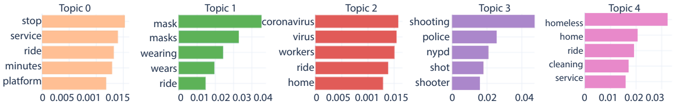

Fig. 7. Keyword of each topic

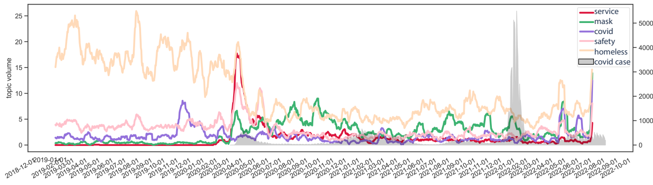

Fig. 8. Dynamic topic modelling

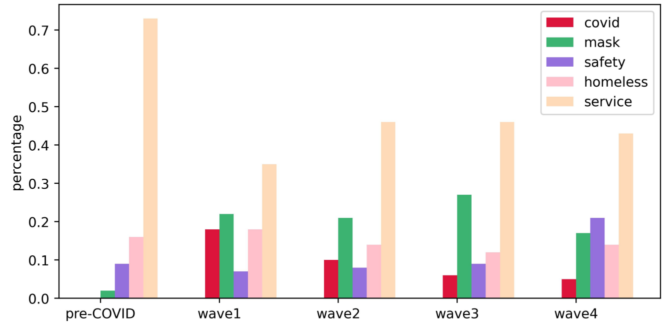

Fig. 9. Topic frequency in each wave of COVID-19

the percentage of each topic in the five phases (Fig. 9), we notice although the transit service topic dominated in all five phases, it decreased from wave1 and did not recovery to the pre-COVID-19 phase till now. Since public transport travel behaviour has a similar trend to the transit service topic, the relationship between transit service concern and public ridership is worth considering. For pandemic-related topics, the mask topic acquired more attention than the COVID-19 topic, illustrating the risk perception for the fast virus spread in narrow spaces. During third and fourth waves, although the concern about COVID-19 dropped significantly, the focus on mask is still at a similar level. Moreover, safety concerns and homeless issues are consecutive throughout the whole period, which required the attention of transit management and policymakers. In general, concerns on public transit became more complex in the long-term COVID-19 instead of the single focus on the negative transit services like delays or stops in pre-COVID-19 period, which implied the COVID-19 derivative issues in public transit that the management agencies needed to pay attention.

The sentiment analysis and topic modelling results call up certain policy implements in the post-COVID-19 period: (1) Public transit service accounts for a large number of concerns even in the post-COVID time to point out the necessity of transit system optimization and improvement to avoid delays, unexpected stops or increase transit frequency in the peak hour to solve crowding issues. (2) COVID-19 has amplified ‘homeless’ and ‘safety’ issues, requiring more policy efforts to address the homeless gathering problem in public transit, like providing them more living services. The government should also provide more police resources to protect passenger safety inside the metro to avoid murder and violent acts. (3) Since ‘mask’ concerns last longer, policy like encouraging keeping distance or wearing masks should be implemented to eliminate passenger risk concerns and prevent future virus spread.

V. Conclusion

This paper was motivated to understand the underlying reasons for the unpredictable changes in public transit patterns due to COVID-19 in the long-term. We implemented state-of-art natural processing language methods to classify and cluster the public transit-related tweets and correlate sentiment with ridership and policy data. The methodology was employed for the case study of New York City, revealing the following findings:

1) COVID-19 cases and related policies strongly impacted public transport travel behaviour in the outbreak of pandemic, while gradually losing effectiveness in the following waves. As policy factor only has a hysteresis impact, it is valuable to explore interpretability of observed patterns from the passenger perspective via social media data, which provides a more comprehensive explanation of long-term public transport travel behaviour.
2) Sentiment analysis pointed out the lasting negative attitudes towards public transit, calling for the attention of transit management agencies and policy-makers to urgent improvement of pubic transit system. It also shows that public attitudes have a time lag with actual travel behaviour change in the beginning of each wave. This benefits understanding and predicting the long-term public transport travel behaviour for future epidemics.
3) Topic modelling method identified five main topics that cause the continuously negative sentiment towards public transit in New York City with different domain variables. More importantly, it revealed the inconspicuous aspects (safety and homeless issues) brought by COVID-19 that will indeed influence passenger travel mode decisions in the longer term. In addition, topic modelling shows consistent concern over mask-wearing and the reduced focus on the transit service in the long-term, which indicate possible reasons why public ridership is still below the pre-COVID-19 level.

In the bigger picture, this paper provides insights into using multiple data sources focusing on user-generated content from social media, to analyze the dynamic public transport travel behaviour. It emphasizes the advantages of social media data and NLP technology in gathering real-time user information, extracting underlying public opinions, and explaining dynamic travel patterns; and the potential of using social media in forecasting travel behaviour by identifying passengers' willingness and attitudes towards different travel modes. Our methodology provides meaningful evidence for transit authorities to estimate the potential travel demands to allocate transit resources in advance and support policy-makers to formulate more efficient policies on transportation systems based on public concerns in future pandemic waves. It also provides feasible evidence of using context-mining via topic modelling for other tasks in the transportation domain to support efficient transit management and policy-making processes.

Future work will include the spatial and temporal analysis in neighbourhood level to gain more insight into dynamic travel behaviour change. Meanwhile, we will study the private vehicle sector to obtain a more comprehensive understanding of various travel behaviour change. Additionally, we would like to implement prediction models based on the multi-modal data sources to forecast the travel demand in future waves. Finally, we will generalize this methodology framework for other tasks in transportation domain like obtain passenger satisfaction or identify congestion patterns.

References

[1] P. Zhao and Y. Gao, "Public transit travel choice in the post covid-19 pandemic era: An application of the extended theory of planned behavior," *Travel Behaviour and Society*, vol. 28, pp. 181–195, 2022.

[2] T. Jain, G. Currie, and L. Aston, "Covid and working from home: Long-term impacts and psycho-social determinants," *Transportation Research Part A: Policy and Practice*, vol. 156, pp. 52–68, 2022.

[3] W. Yao, J. Yu, Y. Yang, N. Chen, S. Jin, Y. Hu, and C. Bai, "Understanding travel behavior adjustment under covid-19," *Communications in Transportation Research*, vol. 2, p. 100068, 2022.

[4] F. Zuo, J. Wang, J. Gao, K. Ozbay, X. J. Ban, Y. Shen, H. Yang, and S. Iyer, "An interactive data visualization and analytics tool to evaluate mobility and sociability trends during COVID-19," *arXiv preprint arXiv:2006.14882*, 2020.

[5] L. Liu, H. J. Miller, and J. Scheff, "The impacts of COVID-19 pandemic on public transit demand in the United States," *Plos one*, vol. 15, no. 11, p. e0242476, 2020.

[6] B. Van Wee and F. Witlox, "Covid-19 and its long-term effects on activity participation and travel behaviour: A multiperspective view," *Journal of transport geography*, vol. 95, p. 103144, 2021.

[7] M. Maleki, M. Bahrami, M. Menendez, and J. Balsa-Barreiro, "Social behavior and covid-19: Analysis of the social factors behind compliance with interventions across the united states," *International Journal of Environmental Research and Public Health*, vol. 19, no. 23, p. 15716, 2022.

[8] S. Shelat, O. Cats, and S. van Cranenburgh, "Traveller behaviour in public transport in the early stages of the covid-19 pandemic in the netherlands," *Transportation Research Part A: Policy and Practice*, vol. 159, pp. 357–371, 2022.

[9] J. Bert, D. Schellong, M. Hagenmaier, D. Hornstein, A. K. Wegscheider, T. Palme *et al.*, "How COVID-19 will shape urban mobility," *City*, vol. 25, pp. 28–1, 2020.

[10] J. Anke, A. Francke, L.-M. Schaefer, and T. Petzoldt, "Impact of sars-cov-2 on the mobility behaviour in germany," *European Transport Research Review*, vol. 13, no. 1, pp. 1–13, 2021.

[11] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, Y. Su, T. Zhu *et al.*, "Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach," *Journal of medical Internet research*, vol. 22, no. 11, p. e20550, 2020.

[12] Q. Ye, K. Ozbay, F. Zuo, and X. Chen, "impact of social media on travel behaviors during the covid-19 pandemic: evidence from new york city," *Transportation Research Record*, p. 03611981211033857, 2021.

[13] M. Müller, M. Salathé, and P. E. Kummervold, "Covid-twitter-bert: A natural language processing model to analyse COVID-19 content on Twitter," *arXiv preprint arXiv:2005.07503*, 2020.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English tweets," *arXiv preprint arXiv:2005.10200*, 2020.

[16] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[17] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese BERT-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[18] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2013, pp. 160–172.

[19] MTA, "Turnstile data," 2022, http://web.mta.info/developers/turnstile.html.

[20] Wikipedia, "Timeline of the COVID-19 pandemic in New York City," 2022, TimelineoftheCOVID-19pandemicinNewYorkCity.

[21] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 1003–1015, 2021.