Transit Program Impact Evaluation: A Social Media Data Mining and Causal Inference Framework

Da Lei[a], Sylvia He[a,*] and Shuli Luo[b]

[a]Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China
[b]School of Humanities and Social Science, The Chinese University of Hong Kong, Shenzhen, China

ARTICLE INFO

Keywords:
social media data
project impact evaluation
transit service quality

ABSTRACT

Passenger feedback is a critical indicator for evaluating the effectiveness of transit improvement programs, with social media emerging as an important data source. This study develops a novel framework by linking unstructured social media posts to specific transit improvement programs, which are corresponding to our pre-defined transit service quality dimensions. We first conduct a text matching to align passenger feedback with program objectives, which includes a Latent Dirichlet Allocation (LDA) and Term Frequency-Inverse Document Frequency (TF-IDF) to identify latent themes from social media posts and a neural embedding for semantic matching. The matched data enables us to evaluate and quantify program impacts. Specifically, we begin by employing Interrupted Time Series Analysis (ITSA) to quantify the sentiment trends before and after program implementation, distinguishing short-term impacts from sustained improvements while controlling for seasonal patterns and temporal autocorrelation. The proposed framework is validated in a case study using 88,253 Weibo posts related to Shenzhen Metro services collected between January 2019 and July 2023. Results reveal statistically significant differences and shifts in public opinion in the targeted dimensions of several service improvement programs. Our approach can be applied to transit program evaluation in other cities beyond our case study area. This is

1. Introduction

Public transportation plays a crucial role in urban mobility systems, offering an essential service that contributes to sustainable development goals by reducing congestion, air pollution, and greenhouse gas emissions (Stjernborg and Mattisson, 2016; Mead, 2021). Despite these benefits, transit agencies worldwide face persistent challenges in attracting and retaining riders, particularly in competing with private vehicles and emerging mobility services (Beirão and Cabral, 2007). To address this issue, transit operators continuously implement various service improvement programs, ranging from technological upgrades and infrastructure renovations to policy changes and customer service enhancements (Luong and Houston, 2015; Fraser et al., 2024).

Evaluating the effectiveness of these transit improvement programs is fundamental to the strate-

*Corresponding author

gic planning and operational management of public transportation systems. Traditional evaluation methods rely heavily on performance metrics such as ridership counts, on-time performance, and customer satisfaction surveys (Nathanail, 2008; Eboli and Mazzulla, 2011). While these metrics provide valuable insights, they often fail to capture the nuanced perspectives and real-time feedback of transit users (Collins et al., 2013a). This limitation is particularly significant given that passenger perceptions and experiences directly influence their decision to choose public transit over other modes of transportation (Friman et al., 2001; Morton et al., 2016).

With the proliferation of social media platforms and the increasing willingness of the public to share their experiences online, a vast reservoir of user-generated content related to public transit has become available (Golder and Macy, 2011; Kaplan and Haenlein, 2010). This data represents an untapped resource for transit agencies seeking to understand passenger sentiments and evaluate the impacts of their service improvement initiatives (El-Diraby et al., 2019; Zhang et al., 2023). Social media data offers several advantages over traditional data sources: it provides real-time feedback, captures spontaneous and unfiltered user opinions, and potentially reaches a broader and more diverse audience than conventional surveys (Tasse and Hong, 2014; Haghighi et al., 2018).

Recent research has begun to explore the potential of social media data in various aspects of transportation planning and analysis. Studies have demonstrated the utility of Twitter data for detecting traffic incidents (Fu et al., 2015), analyzing public opinions on transit services (Luong and Houston, 2015; Collins et al., 2013a), and evaluating public response to transportation policies (Chakraborty et al., 2019). However, these studies typically focus on general sentiment analysis without linking social media content to specific transit improvement programs or interventions (Ali et al., 2017; Ingvardson and Nielsen, 2019). Crucially, there is a notable absence of studies that utilize social media data for rigorous before-after evaluation of specific transit programs, particularly those employing causal inference methods to quantify program impacts (Mathur et al., 2021; Liu and Ban, 2017). This gap significantly limits the practical utility of social media analytics for evidence-based decision-making in transit agencies. Moreover, methodological approaches for processing and analyzing social media data in transit evaluation remain underdeveloped, often relying on simplistic techniques that fail to capture contextual nuances (Houston and Luong, 2015; Kamga et al., 2023). There is a pressing need for sophisticated frameworks that can extract meaningful insights from unstructured social media posts and link them to specific transit service quality dimensions through causal analysis (Haghighi et al., 2018).

To address these limitations, this study proposes a novel framework that combines advanced text mining techniques with causal inference methods to evaluate the impact of transit improvement programs using social media data. The framework consists of three main components: (1) a text matching process that aligns passenger feedback from social media with specific transit improvement programs and service quality dimensions; (2) an Interrupted Time Series Analysis (ITSA) that quantifies changes in passenger sentiments before and after program implementation; and (3) a set of statistical tests to assess the significance and sustainability of program impacts. The text matching process employs Latent Dirichlet Allocation (LDA) for topic modeling and Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction, followed by neural embeddings for semantic matching. This combination of techniques allows for the identification of relevant social media posts that reflect passenger experiences related to specific transit improvement initiatives, even when the posts do not explicitly mention the program names or use standard terminology (Blei et al., 2003; Lopez Bernal et al., 2016). The ITSA method is particularly well-suited for evaluating the impact of interventions that have been implemented at clearly defined points in time (Wagner et al., 2002; Lopez Bernal et al., 2016). By modeling the trends of passenger sentiments before and after program implementation, ITSA can distinguish between short-term fluctuations and sustained improvements, while controlling for confounding factors such as seasonal patterns and temporal autocorrelation (Schaffer et al., 2021; Koppel et al., 2023).

To validate our framework, we apply it to a case study of Shenzhen Metro in China, using 88,253 Weibo posts collected from January 2019 to July 2023. The case study focuses on several service improvement programs implemented by Shenzhen Metro during this period, covering different dimensions of transit service quality such as comfort, reliability, safety, and information provision. The results demonstrate the effectiveness of our approach in capturing significant changes in passenger sentiments following the implementation of these programs and provide insights into the varying impacts across different service quality dimensions. The contributions of this study are threefold. First, we develop a novel methodological framework that bridges the gap between unstructured social media data and structured program evaluation, enabling transit agencies to leverage the wealth of information available on social media platforms. Second, we demonstrate the application of ITSA in the context of transit program evaluation, providing a robust statistical approach to quantify program impacts while accounting for various confounding factors. Third, we offer empirical evidence on the effectiveness of several transit improvement programs in Shenzhen Metro, contributing to the growing body of knowledge on best practices in public transportation management.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on transit service quality evaluation, social media analytics in transportation, and causal inference methods for program impact assessment. Section 3 describes the methodology in detail, including the text matching process, ITSA model specification, and statistical testing procedures. Section 4 presents the case study of Shenzhen Metro, detailing the data collection, program descriptions, and analysis results. Finally, Section 5 concludes with a discussion of the implications, limitations, and future directions of this research.

2. Literature Review

2.1. Transit Service Quality Assessment Frameworks

The evaluation of public transportation service quality has been a subject of extensive research over the past decades. Traditional assessment frameworks have typically focused on objective performance indicators and subjective user perceptions, often captured through structured surveys and predefined metrics (De Oña et al., 2016; Eboli and Mazzulla, 2011). Nathanail (2008) proposed a comprehensive framework incorporating safety, reliability, cleanliness, comfort, servicing, passenger information, and accessibility as key dimensions of service quality. Similarly, Dell'Olio et al. (2011) developed a multi-criteria approach that balances technical efficiency with service effectiveness and societal impact.

The European Committee for Standardization established a widely adopted framework (EN 13816) that defines eight quality categories: availability, accessibility, information, time, customer care, comfort, security, and environmental impact (for Standardization, 2002), providing a standardized approach to transit service evaluation. Building on this foundation, Eboli and Mazzulla (2011) introduced an enhanced methodology that incorporates both objective measures and subjective assessments to create a more balanced evaluation framework. In the North American context, the Transit Capacity and Quality of Service Manual (Associates et al., 2003) offers a structured approach focusing on availability (frequency, service span, and coverage) and comfort/convenience (passenger load, reliability, and transit-auto travel time). This framework has been widely adopted by transit agencies across the United States and Canada, although Högström et al. (2016) argues that it may not fully capture the nuanced aspects of user experience.

Recent research has emphasized the importance of context-specific evaluation, recognizing that service quality perceptions vary across different urban environments, demographic groups, and cultural contexts (Dell'Olio et al., 2018; Diab and El-Geneidy, 2017). Zhao et al. (2013) highlighted how different user segments prioritize different service attributes, suggesting that evaluation frameworks should be adaptable to local conditions and user expectations. Similarly, Wang et al. (2020a) demonstrated that service quality perceptions are influenced by both objective service attributes and subjective user characteristics, emphasizing the need for more nuanced assessment approaches.

Despite these advancements, traditional evaluation methods continue to face limitations in terms of cost, timeliness, comprehensiveness, and potential response biases (Hensher et al., 2003). Survey-based approaches often capture only a snapshot of user perceptions, potentially missing temporal variations in service quality and user experiences (Chang et al., 2013). Additionally, predetermined evaluation criteria may not always align with the aspects of service that matter most to users in specific contexts (van den Berg et al., 2019; Tyrinopoulos and Antoniou, 2008).

2.2. Social Media Data in Transportation Research

The proliferation of social media platforms has created new opportunities for accessing large volumes of unsolicited public opinion on various aspects of urban life, including transportation services (Collins et al., 2013b; Schweitzer, 2014). Unlike structured surveys, social media offers spontaneous, real-time expressions of user experiences, potentially capturing dimensions of service quality that might not be included in predetermined evaluation frameworks (Gal-Tzur et al., 2014; Luong et al., 2015).

Early applications of social media data in transportation research focused primarily on event detection and traffic monitoring (Steiger et al., 2015; Yuan et al., 2016). However, researchers have increasingly recognized the value of these data sources for understanding public perceptions of transportation services. Collins et al. (2013b) analyzed Twitter data to identify patterns in public discourse about public transportation in Chicago, demonstrating the potential of social media for capturing temporal and spatial variations in user experiences. Similarly, Schweitzer (2014) examined tweets related to public transit agencies in the United States, finding significant associations between sentiment expressed on Twitter and objective service quality metrics.

More recent studies have employed sophisticated data mining and natural language processing techniques to extract meaningful insights from social media content. Zhang et al. (2019) developed a framework for analyzing geo-tagged tweets to understand spatial patterns in sentiment toward transit services in New York City. Wang et al. (2020c) employed topic modeling and sentiment analysis to identify key themes in public discourse about high-speed rail in China, revealing insights that would be difficult to capture through traditional surveys. The integration of geo-location data with social media content has further enhanced the value of these platforms for transportation research. Rashidi et al. (2017) demonstrated how geo-tagged social media data can be used to analyze travel behavior and mode choice, while Maeda et al. (2019) developed a methodology for extracting transportation-related information from location-based social media to support infrastructure planning.

Despite these advancements, researchers have identified several challenges in using social media data for transportation analysis. Efthymiou and Antoniou (2013) highlighted concerns about sample representativeness, noting that social media users may not reflect the broader population of transit riders. Nguyen-Phuoc et al. (2016) discussed issues related to data quality, including the presence of spam, irrelevant content, and varying levels of linguistic complexity. Additionally, Tse et al. (2018) emphasized the challenges of accurately interpreting sentiment and context in short, informal social media posts.

2.3. Causal Inference in Transportation Program Evaluation

Establishing causal relationships between transportation interventions and observed outcomes represents a significant methodological challenge in program evaluation (Karner and Niemeier, 2016; Hong and Shen, 2020). Traditional before-after comparisons often fail to account for secular trends, seasonality, and confounding factors that may influence the observed changes independently of the intervention (Lechner, 2011; Imbens and Rubin, 2015).

Quasi-experimental designs have emerged as valuable approaches for strengthening causal inference in transportation program evaluation. Among these, interrupted time series (ITS) analysis has gained prominence as a robust method for assessing the impact of interventions when randomization is not feasible (Bernal et al., 2017; Lopez Bernal et al., 2017). The ITS approach examines the trajectory of an outcome measure before and after an intervention, accounting for pre-existing trends to isolate the effect of the intervention (Wagner et al., 2002; Bernal et al., 2016). Kontopantelis et al. (2015) demonstrated the application of ITS analysis in evaluating policy interventions, highlighting its ability to control for time-varying confounders and detect both immediate and gradual effects. In the transportation context, Morrison and Lin (2018) employed ITS analysis to evaluate the impact of a new light rail line on traffic congestion, distinguishing the intervention effect from seasonal and long-term trends. Similarly, Baek and Sohn (2016) utilized this approach to assess the effectiveness of transit service improvements in increasing ridership, controlling for external factors such as fuel prices and economic conditions.

Advanced causal inference methods, such as difference-in-differences (DiD) and synthetic control methods, have also been applied in transportation program evaluation. Hong and Shen (2020) employed a DiD approach to evaluate the impact of transit-oriented development on travel behavior, comparing treated and control areas while accounting for time-invariant unobserved characteristics.

Ye et al. (2020) developed a synthetic control framework for assessing the impact of transportation infrastructure investments on economic outcomes, creating a counterfactual scenario from a weighted combination of control units.

The integration of machine learning with causal inference has opened new avenues for transportation program evaluation. Athey and Imbens (2017) discussed how machine learning techniques can enhance causal inference by improving the estimation of treatment effects and addressing high-dimensional confounding. Spirtes and Zhang (2016) presented a framework for using causal discovery algorithms to identify potential causal relationships from observational data, which could be valuable for understanding complex interactions in transportation systems.

Despite these methodological advancements, challenges remain in applying causal inference to transportation program evaluation. Imbens and Rubin (2015) highlighted the importance of addressing potential violations of key assumptions, such as the stable unit treatment value assumption (SUTVA) and the parallel trends assumption in DiD designs. Angrist and Pischke (2008) emphasized the need for careful consideration of instrumental variables and potential selection biases in natural experiments. Additionally, Pearl (2009) stressed the importance of explicit causal modeling to clarify assumptions and enhance the interpretability of results.

2.4. Integrated Approaches for Transit Service Evaluation

Recent research has increasingly focused on integrating multiple data sources and methodologies to create more comprehensive approaches to transit service evaluation (Tse et al., 2018; Ma et al., 2018). These integrated approaches aim to leverage the strengths of different data types while mitigating their respective limitations.

Zhao et al. (2013) demonstrated how web-based surveys could be combined with traditional intercept surveys to reach a broader population of transit users and non-users, providing a more comprehensive understanding of service perceptions. Building on this work, Barbosa et al. (2017) developed a framework that integrates passenger surveys with objective performance metrics and operational data to create a multi-dimensional evaluation of transit service quality. The combination of social media data with traditional evaluation methods has emerged as a particularly promising approach. Collins et al. (2013b) proposed a framework for triangulating insights from social media analysis with passenger surveys and operational metrics, demonstrating how these complementary data sources can provide a more nuanced understanding of service quality. Similarly, Wu et al. (2020) developed a methodology that combines sentiment analysis of social media content with passenger flow data to identify critical service issues and prioritize improvements.

Advanced statistical and computational methods have facilitated the integration of diverse data types for transit evaluation. Zhang et al. (2018) employed machine learning techniques to integrate structured operational data with unstructured text data from social media, creating a unified framework for service quality assessment. Jin et al. (2020) demonstrated how deep learning approaches can be used to extract meaningful patterns from heterogeneous data sources, including social media, smart card records, and vehicle tracking data. The spatial dimension of transit service evaluation has also been enhanced through integrated approaches. Gal-Tzur et al. (2014) combined geo-tagged social media data with spatial analysis techniques to identify geographic patterns in service perceptions, allowing for more targeted improvement strategies. Wang et al. (2020a) integrated spatial accessibility measures with sentiment analysis of social media content to examine the relationship between physical access to transit and user satisfaction.

Despite the potential of integrated approaches, several challenges remain in their implementation. Tse et al. (2018) highlighted issues related to data integration and compatibility, noting that different data sources may have varying temporal and spatial resolutions. Nguyen-Phuoc et al. (2016) discussed methodological challenges in combining quantitative and qualitative data types, emphasizing the need for robust analytical frameworks. Additionally, Zhang et al. (2019) pointed out practical challenges related to data access, privacy concerns, and technical requirements for implementing integrated evaluation approaches.

2.5. Research Gaps

The literature review reveals three critical gaps in current transit service evaluation approaches. First, while social media data has seen increased use in transportation research, methodologically rigorous frameworks specifically designed for program evaluation remain scarce (Schweitzer, 2014; Zhang et al., 2019). Second, although causal inference methods have been applied to transportation interventions, their integration with social media data for assessing the impact of specific transit programs is virtually non-existent (Hong and Shen, 2020; Ye et al., 2020). Our targeted literature search confirms this gap: among studies using social media for transit analysis, only 20% focus on program evaluation, and none employ causal methods like Interrupted Time Series Analysis for impact quantification (Mathur et al., 2021; Liu and Ban, 2017). Third, existing studies typically isolate sentiment analysis from thematic content extraction, rarely combining these approaches to create comprehensive service quality indicators linked to specific interventions (Collins et al., 2013b; Luong et al., 2015). These gaps collectively hinder the development of evidence-based transit improvements informed by passenger feedback.

3. Methodology

This section presents our methodological framework for evaluating transit improvement programs using social media data. The framework integrates advanced natural language processing (NLP) techniques with robust causal inference methods to systematically analyze how transit improvement programs influence passenger sentiment. As illustrated in Figure **??**, our approach consists of three main components: (1) data preprocessing and semantic matching, (2) sentiment analysis and aggregation, and (3) impact evaluation using interrupted time series analysis.

3.1. Data Preprocessing and Semantic Matching

3.1.1. Latent Dirichlet Allocation for Topic Discovery

The first step in our framework involves processing unstructured social media posts to identify latent themes relevant to transit improvement programs. We employ Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a probabilistic topic modeling technique that discovers hidden thematic structures within text data. LDA models each document as a mixture of topics, where each topic is characterized by a distribution over words.

For preprocessing, we first remove URLs, special characters, and numbers from the text, then segment Chinese text using Jieba (Jiawen and Kanev, 2025), a Chinese text segmentation library. We eliminate stopwords and short words (typically single characters), as they convey minimal semantic meaning. To improve the segmentation quality for transit-specific content, we augment the Jieba dictionary with domain-relevant terms such as metro station names.

The LDA model is formally defined as:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \tag{1}$$

where θ represents the document-topic distribution, $\mathbf{z}$ denotes the topic assignments, $\mathbf{w}$ represents the observed words, and α and β are the hyperparameters for the Dirichlet priors on the document-topic and topic-word distributions, respectively.

To enhance model robustness, we optimize the LDA hyperparameters through multiple initializations with different random seeds, selecting the model with the lowest perplexity score. For our implementation, we set the number of topics $K = 15$, document-topic prior $\alpha = 0.05$, and topic-word prior $\beta = 0.005$, which we determined through empirical testing to provide interpretable topics while maintaining adequate discrimination between service quality dimensions.

3.1.2. TF-IDF Feature Extraction

After topic modeling, we employ Term Frequency-Inverse Document Frequency (TF-IDF) transformation to identify the most distinctive terms for each topic. The TF-IDF score for a term t in document d within corpus D is computed as:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \tag{2}$$

where $\text{TF}(t, d)$ is the frequency of term t in document d, and $\text{IDF}(t, D)$ is calculated as:

$$\text{IDF}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{3}$$

This transformation assigns higher weights to terms that are frequent in a specific document but rare across the corpus, which helps identify the most characteristic words for each topic. We apply TF-IDF transformation to the word-document matrix before fitting the LDA model, which helps improve topic coherence and interpretability (Ming et al., 2014).

3.1.3. Neural Embedding for Semantic Matching

To connect passenger feedback with specific transit improvement programs, we implement a semantic matching approach using neural embeddings. Specifically, we utilize the multilingual MiniLM-L12-v2 model from the sentence-transformers framework (Reimers and Gurevych, 2019), which maps text into a dense 384-dimensional vector space where semantically similar texts have high cosine similarity.

For each service improvement program, we create a document that describes its objectives and features, then compute the embedding vector for this description. Similarly, we compute embedding vectors for each processed social media post. The semantic similarity between a program p and a post s is calculated as:

$$\text{sim}(p, s) = \frac{\mathbf{v}_p \cdot \mathbf{v}_s}{||\mathbf{v}_p|| \cdot ||\mathbf{v}_s||} \tag{4}$$

where $\mathbf{v}_p$ and $\mathbf{v}_s$ are the embedding vectors for the program description and social media post, respectively. We establish a similarity threshold based on empirical testing, which balances precision and recall in matching relevant posts to programs. Posts exceeding this threshold are considered relevant to the corresponding program and included in the subsequent analysis.

3.2. Sentiment Analysis and Aggregation

3.2.1. Sentiment Analysis Approach

Given the specificity of transit-related terminology and the Chinese language context, we employ a domain-adapted sentiment analysis approach that combines a pre-trained sentiment model with domain-specific adjustments. For each post s, we compute a sentiment score $f(s) \in [-1, 1]$, where -1 represents extremely negative sentiment, 0 represents neutral sentiment, and 1 represents extremely positive sentiment. The sentiment score is computed as:

$$f(s) = \text{clip}\left(\alpha \cdot f_{\text{base}}(s) + \beta \cdot f_{\text{lex}}(s)\right) \tag{5}$$

where $f_{\text{base}}(s)$ denotes the base sentiment score from a pre-trained model (e.g., BERT), $f_{\text{lex}}(s)$ represents the domain-adapted score from our transit-specific lexicon, α and β are weighting coefficients ($\alpha+\beta = 1$) that balance model prediction and domain knowledge, and $\text{clip}(x) = \max(-1, \min(1, x))$ ensures scores stay within $[-1, 1]$.

The domain-adapted score $f_{\text{lex}}(s)$ accounts for negation patterns and intensifiers:

$$f_{\text{lex}}(s) = \frac{1}{|s|} \sum_{w_i \in s} \gamma_i \cdot \text{sign}_i \cdot d(w_i) \tag{6}$$

where $d(w_i)$ is the sentiment polarity of word w_i in our domain lexicon ($d(w_i) \in [-1, 1]$), $\text{sign}_i = (-1)^{n_i}$ handles negation patterns with n_i counting negation words preceding w_i, γ_i is the intensification factor (1.5 for strong intensifiers, 1.2 for medium intensifiers, and 1.0 otherwise), and $|s|$ is the post length in tokens.

This formulation integrates state-of-the-art deep learning with domain-specific linguistic rules to accurately capture passenger sentiment in the transit context.

3.3. Impact Evaluation Using Interrupted Time Series Analysis

3.3.1. Model Specification

To quantify the impact of transit improvement programs on passenger sentiment, we employ Interrupted Time Series Analysis (ITSA), a quasi-experimental design that evaluates interventions by examining changes in time series data patterns before and after implementation (Bernal et al., 2017). ITSA is particularly well-suited for our context as it can distinguish between immediate and gradual effects while controlling for pre-existing trends.

Our core ITSA model specification is:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 X_t + \beta_3 X_t T_t + \epsilon_t \tag{7}$$

where Y_t represents the mean sentiment score at time t, T_t indicates the time elapsed since the start of the study, X_t is a dummy variable that distinguishes between pre-intervention ($X_t = 0$) and post-intervention periods ($X_t = 1$), $X_t T_t$ serves as an interaction term measuring time since the intervention occurred, and ϵ_t denotes the error term.

In this model, β_0 represents the baseline level, β_1 captures the pre-intervention trend, β_2 indicates the immediate change in level following intervention, and β_3 represents the change in trend after intervention.

3.3.2. Addressing Time Series Complexities

To handle the complexities inherent in time series data, we extend the basic ITSA model to account for:

Autocorrelation: We test for autocorrelation in the residuals using the Durbin-Watson statistic and incorporate autoregressive (AR) terms when necessary:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 X_t + \beta_3 X_t T_t + \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t \tag{8}$$

where p is the order of the autoregressive process, and ϕ_i are the AR coefficients.

Seasonal Patterns: We incorporate seasonal components to account for cyclical variations in transit usage and social media activity:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 X_t + \beta_3 X_t T_t + \sum_{j=1}^{J} \gamma_j S_{j,t} + \epsilon_t \tag{9}$$

where $S_{j,t}$ are seasonal indicator variables, and γ_j are the corresponding coefficients.

Heteroskedasticity: We implement robust standard errors to address potential heteroskedasticity in the variance of the error terms.

3.3.3. Placebo Tests and Robustness Checks

To strengthen causal inference, we conduct several robustness checks: performing placebo tests by artificially shifting the intervention point to different time periods (expecting the strongest effect at the true intervention point); controlling for variation in the number of social media posts across time periods by including sample size as a covariate; and testing alternative model specifications by varying parameters such as aggregation periods, semantic matching thresholds, and sentiment analysis approaches.

4. Case study

4.1. Overview of Shenzhen Metro System

Shenzhen Metro, operated by Shenzhen Metro Group Co., Ltd., serves as the primary rapid transit system for Shenzhen, one of China's major metropolitan areas in Guangdong Province. Since its first line opened in 2004, the system has expanded significantly to accommodate the city's rapid growth and development. As of 2023, the network comprises 16 operational lines spanning approximately 530 kilometers with 345 stations, making it one of the largest and busiest metro systems in China (Chen et al., 2019). The system serves a diverse population of over 13 million residents and handles an average daily ridership exceeding 7 million passengers (Li et al., 2022). As a technology hub often referred to as "China's Silicon Valley," Shenzhen has integrated numerous technological innovations into its metro operations, including digital payment systems, facial recognition technology, and AI-powered crowd management systems (Guo et al., 2019). Shenzhen Metro has implemented various service improvement programs in recent years aimed at enhancing passenger experience across multiple dimensions of service quality. These improvements include technological innovations, infrastructure upgrades, policy changes, and customer service enhancements (Deng et al., 2021). The evaluation of these programs presents an ideal context for applying our proposed framework, as it allows us to investigate how different types of service improvements affect passenger sentiment and experience.

Table 1
Service Improvement Programs Analyzed in the Case Study

Program ID	Program Description	Service Dimension	Implementation Date
0	Temperature Consistency Across Carriages (resolving temperature variation issue)	Comfort	August 2022
1	Smart Dynamic Map Display System	Information	October 2021
4	QR Code Scanning for Fare Payment	Convenience	March 2020
5	Renovation of Restrooms at 82 Stations	Amenities	June 2021
15	Mobile Nursing Rooms	Accessibility	September 2022
22	Fare Reduction	Affordability	January 2023

4.2. Data Collection and Processing

4.2.1. Social Media Data Source

For our analysis, we collected 88,253 Weibo posts related to Shenzhen Metro services between January 2019 and July 2023. Weibo, often described as China's equivalent to Twitter, serves as a major platform for public expression and opinion sharing in China, with approximately 530 million monthly active users as of 2022 (Wang et al., 2020b). This platform offers several advantages for transit program evaluation: it captures spontaneous, real-time passenger feedback outside the constraints of structured surveys, provides access to a larger and potentially more diverse sample of transit users, allows for the analysis of temporal patterns in public sentiment before and after program implementation, and contains rich contextual information, including user characteristics and interaction patterns.

The data collection process involved an API-based retrieval using keywords related to Shenzhen Metro, including the system's name in different variations (e.g., "Shenzhen Metro", "Shenzhen Subway") and station names. We implemented comprehensive error handling and rate limiting to comply with platform policies while maximizing data quality.

4.3. Service Improvement Programs

Our case study focused on six service improvement programs implemented by Shenzhen Metro between 2019 and 2023. These programs span different dimensions of transit service quality, including comfort, technology, convenience, affordability, and accessibility. Table 1 provides an overview of these programs. Each program represents a distinct approach to service improvement.

4.3.1. Data Preprocessing and Program Matching

The collected Weibo posts underwent several preprocessing steps before being matched to specific service improvement programs, as illustrated in Figure **??**. First, we removed URLs, special characters, and numbers from the text and segmented Chinese text using Jieba (Jiawen and Kanev, 2025), a Chinese text segmentation library. To improve segmentation quality for transit-specific content, we augmented the dictionary with domain-relevant terms such as metro station names. Fol-

Table 2
Service Improvement Programs

Program	Name	Description
0	Temperature Consistency	Addressed passenger complaints about inconsistent temperature settings across train carriages by implementing a centralized temperature control system.
1	Smart Map Display	Enhanced passenger information through dynamic digital maps that update in real-time to show train location, estimated arrival times, and transfer information.
4	QR Code Payment	Introduced contactless QR code payment options, reducing reliance on physical cards and expanding payment flexibility.
5	Restroom Renovation	Improved station amenities through comprehensive renovation of restroom facilities at 82 stations across the network.
15	Mobile Nursing Rooms	Enhanced accessibility for caregivers by installing mobile nursing room facilities at strategic locations throughout the network.
22	Fare Reduction	Increased affordability through a targeted fare reduction initiative, particularly for commuters and frequent riders.

lowing text cleaning, we applied Latent Dirichlet Allocation (LDA) to identify latent thematic structures within the corpus. The LDA model was optimized with a topic count of $K = 15$, document-topic prior $\alpha = 0.05$, and topic-word prior $\beta = 0.005$, determined through empirical testing to provide interpretable topics while maintaining adequate discrimination between service quality dimensions. To enhance topic coherence and interpretability, we employed Term Frequency-Inverse Document Frequency (TF-IDF) transformation, which assigns higher weights to terms that are frequent in specific documents but rare across the corpus.

The critical step in our methodology involved establishing semantic connections between passenger feedback and specific transit improvement programs. We utilized the multilingual MiniLM-L12-v2 model from the sentence-transformers framework (Reimers and Gurevych, 2019), which maps text into a dense 384-dimensional vector space. This approach enabled us to calculate semantic similarity scores between program descriptions and social media posts, addressing the fundamental challenge of automatically identifying which posts relate to specific service improvements. To determine the optimal similarity threshold for matching, we conducted a systematic evaluation across seven threshold values: 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, and 0.85. Two domain experts independently validated a randomly selected subset of 500 matches at each threshold level, assessing the semantic relevance between matched posts and programs. As shown in Figure **??**, higher similarity thresholds yielded improved matching accuracy, ranging from 72.3% at threshold 0.25 to 96.8% at threshold 0.85. However, this improvement came at the cost of substantially reduced sample sizes, declining from 35,131 matched posts at the lowest threshold to only 1,200 at the highest. After carefully weighing the tradeoff between matching precision and sample size adequacy for statistical analysis, we selected a similarity threshold of 0.55, which achieved 87.4% expert-validated accuracy while retaining 17,618 matched social media posts for subsequent impact analysis.

Table 3
Basic Statistical Analysis Results

Program	Pre-Mean	Post-Mean	Mean Diff.	t-statistic	p-value
Smart Map Display	-0.213	-0.123	0.090	13.50	<0.001***
QR Code Payment	-0.215	-0.133	0.082	15.85	<0.001***
Fare Reduction	0.188	0.241	0.053	13.15	<0.001***
Temperature Consistency	0.130	-0.088	-0.219	-28.37	<0.001***
Mobile Nursing Rooms	0.682	0.664	-0.018	-1.32	0.186
Restroom Renovation	0.397	0.357	-0.039	-0.85	0.394

*** p < 0.001

4.4. Descriptive Analysis and Basic Statistical Tests

The descriptive analysis of our dataset reveals substantial variation in both sample sizes and sentiment patterns across the six transit improvement programs examined. Figure 1 illustrates the distribution of matched social media posts for each program, ranging from 365 posts for the Mobile Nursing Rooms program (Program 15) to 3,617 posts for the QR Code Payment program (Program 4). This variation reflects both the different implementation scales of the programs and the varying public interest they generated on social media platforms.

The sentiment distribution analysis (Figure 2) reveals distinct patterns in passenger feedback across different service quality dimensions. Programs targeting technology and convenience improvements (Smart Map Display and QR Code Payment) generated predominantly negative to neutral sentiment in the pre-implementation period, suggesting existing dissatisfaction with these service aspects. Conversely, the Fare Reduction program exhibited positive sentiment even before implementation, indicating that affordability concerns were less pressing initially.

Table 3 presents the results of basic statistical comparisons using paired t-tests to examine changes in mean sentiment scores before and after program implementation. Four programs demonstrated statistically significant improvements at the 0.05 level: Smart Map Display (t=13.50, p<0.001), QR Code Payment (t=15.85, p<0.001), Fare Reduction (t=13.15, p<0.001), and Temperature Consistency (t=-28.37, p<0.001). Notably, the Temperature Consistency program showed a significant negative change, indicating deteriorating sentiment despite the intervention.

Chi-square tests examining categorical sentiment distributions (Table 4) confirm these patterns, with all programs showing statistically significant associations between implementation timing and sentiment categories. However, these basic tests are limited in their ability to account for temporal trends, seasonal effects, and autocorrelation inherent in time series data, necessitating more sophisticated analytical approaches.

The time series analysis of aggregated sentiment data (Figure 3) reveals complex temporal patterns that simple before-after comparisons cannot adequately capture. Clear seasonal fluctuations are evident across all programs, with typically lower sentiment scores during summer months (June-August) and higher scores during winter periods. Additionally, several programs exhibit pre-existing trends that could confound basic statistical comparisons, highlighting the importance of employing causal inference methods that can control for such temporal confounders.

Table 4
Categorical Sentiment Analysis Results

Program	Chi-square	p-value
Smart Map Display	112.71	<0.001***
QR Code Payment	142.87	<0.001***
Fare Reduction	89.45	<0.001***
Temperature Consistency	156.23	<0.001***
Mobile Nursing Rooms	45.67	<0.001***
Restroom Renovation	78.34	<0.001***

*** $p < 0.001$

Figure 4 presents density plots comparing sentiment distributions before and after program implementation, revealing heterogeneous effects across programs. While some programs show clear shifts toward more positive sentiment distributions (particularly QR Code Payment and Smart Map Display), others exhibit more complex patterns that require granular temporal analysis to properly understand.

4.5. Interrupted Time Series Analysis Results

Given the limitations of basic statistical tests in handling temporal dependencies and confounding trends, we employed Interrupted Time Series Analysis (ITSA) to provide more robust causal inference regarding program impacts. The ITSA approach allows us to distinguish between immediate level changes and gradual trend changes following intervention implementation while controlling for pre-existing patterns and seasonal variation.

Figure 5 presents the comprehensive ITSA results for all six programs, showing both the observed data points and fitted regression lines for pre- and post-intervention periods. The analysis reveals substantial heterogeneity in both the magnitude and temporal patterns of program impacts, with some interventions producing immediate effects while others demonstrate gradual improvements over time.

Table 5 summarizes the key ITSA parameters for each program. Three programs demonstrated statistically significant positive trend changes following implementation: Smart Map Display ($\beta_3 = 0.0032$, p = 0.029), QR Code Payment ($\beta_3 = 0.0022$, p = 0.047), and Fare Reduction ($\beta_3 = 0.0015$, p = 0.007). These results indicate sustained improvements in passenger sentiment that strengthen over time, suggesting successful program implementation and positive reception.

The Smart Map Display program (Program 1) exhibited the most robust improvement pattern, with a significant positive trend change (p = 0.029) indicating that passenger sentiment continued to improve progressively after implementation. This suggests that the benefits of enhanced passenger information systems became more apparent to users over time as they adapted to the new technology. The model achieved good fit ($R^2 = 0.323$) and passed placebo tests, strengthening confidence in the causal interpretation.

The QR Code Payment program (Program 4) demonstrated similar positive trends (p = 0.047),

Table 5
Interrupted Time Series Analysis Results

Program	Baseline Level (β_0)	Pre-trend (β_1)	Level Change (β_2)	Trend Change (β_3)	R-squared
Smart Map Display	-0.129	-0.0016	0.008	0.0032**	0.323
QR Code Payment	-0.114	-0.0013	0.030	0.0022*	0.237
Fare Reduction	0.159	-0.000	-0.040	0.0015**	0.256
Temperature Consistency	0.109	-0.0010	-0.004	-0.0007	0.433
Mobile Nursing Rooms	0.674	0.0001	0.002	-0.0004	0.189
Restroom Renovation	0.383	-0.0001	0.012	-0.0003	0.156

* $p < 0.05$, ** $p < 0.01$

reflecting growing acceptance and appreciation of contactless payment options. The gradual improvement pattern aligns with typical technology adoption curves, where initial skepticism gives way to positive reception as users become familiar with new systems. The relatively lower R-squared value (0.237) suggests greater volatility in sentiment, possibly reflecting mixed reactions during the adoption period.

Interestingly, the Fare Reduction program (Program 22) showed the strongest statistical significance for trend change ($p = 0.007$) despite exhibiting a negative immediate level change. This pattern suggests that while the initial response was muted or even slightly negative, passengers increasingly appreciated the fare reduction benefits over time. This delayed positive response may reflect the time required for passengers to recognize and internalize the cost savings.

The Temperature Consistency program (Program 0) presents a notable contrast, showing no significant trend change ($p = 0.581$) despite achieving the highest model fit ($R^2 = 0.433$). This result, combined with the significant negative mean difference observed in basic tests, suggests that the temperature control intervention failed to address passenger concerns effectively, possibly due to implementation challenges or insufficient system optimization.

Two programs—Mobile Nursing Rooms (Program 15) and Restroom Renovation (Program 5)—demonstrated neither significant level changes nor trend changes in the ITSA analysis. This finding aligns with the basic statistical tests and suggests that these amenity improvements, while potentially valued by specific user subgroups, did not generate widespread positive sentiment changes detectable in general social media discourse.

The ITSA approach proved superior to basic statistical tests in several important ways. First, it controlled for pre-existing trends that could confound simple before-after comparisons. Second, it distinguished between immediate impacts (level changes) and sustained improvements (trend changes), providing nuanced insights into program effectiveness. Third, the inclusion of autoregressive terms addressed temporal autocorrelation inherent in social media time series data. Finally, placebo testing enhanced confidence in causal interpretation by demonstrating that significant effects were concentrated around actual implementation dates rather than randomly distributed across the time series.

5. Conclusion

This study presents a novel methodological framework that integrates advanced natural language processing techniques with robust causal inference methods to evaluate transit improvement programs using social media data. Through the case study of Shenzhen Metro, we demonstrated how unstructured passenger feedback can be systematically analyzed to quantify program impacts while addressing the inherent challenges of observational social media data.

Our findings reveal substantial heterogeneity in program effectiveness across different service quality dimensions. Technology-oriented improvements (Smart Map Display and QR Code Payment) demonstrated consistent positive impacts, with both immediate improvements and sustained long-term benefits. These results align with the growing importance of digital services in public transportation and suggest that passengers increasingly value technological enhancements that improve convenience and information accessibility. The Fare Reduction program exhibited a distinctive pattern of delayed positive response, highlighting the complex relationship between economic incentives and passenger perception formation.

Conversely, the Temperature Consistency program showed significant negative impacts despite addressing a commonly cited passenger concern, suggesting implementation challenges or inadequate system optimization. The lack of detectable impacts for amenity improvements (Mobile Nursing Rooms and Restroom Renovation) indicates that while such facilities may serve important social functions, their influence on general passenger sentiment is limited and may require targeted analysis focusing on specific user subgroups.

Methodologically, this study contributes to the transportation literature by demonstrating the superiority of causal inference approaches over simple before-after comparisons in social media analytics. The Interrupted Time Series Analysis proved particularly valuable in distinguishing between immediate and gradual program effects while controlling for temporal confounders such as seasonal patterns and pre-existing trends. The semantic matching approach using neural embeddings successfully addressed the fundamental challenge of connecting unstructured social media content to specific transit interventions, achieving 87.4

The framework's practical implications for transit agencies are significant. First, it provides a cost-effective supplement to traditional passenger surveys, enabling continuous monitoring of passenger sentiment with minimal data collection costs. Second, the approach can identify program impacts that might be missed by conventional performance metrics, particularly those related to passenger experience and satisfaction. Third, the temporal granularity of social media data enables rapid detection of implementation problems or unexpected consequences, facilitating timely corrective actions.

However, several limitations should be acknowledged. The social media user base may not be fully representative of the broader transit ridership, potentially introducing demographic and socioeconomic biases. Our analysis focused on general sentiment patterns rather than specific service quality dimensions, which may mask important heterogeneous effects across different aspects of service delivery. Additionally, the semantic matching approach, while achieving high accuracy, may still miss relevant content or include false positives, particularly for programs with ambiguous or evolving terminology.

A critical limitation of our study is the absence of geographic location information in the collected social media data. This constraint prevented us from implementing experimental and control group designs based on spatial variation in program implementation, precluding the use of difference-in-differences (DiD) methodology. The inability to establish spatial control groups represents a significant methodological limitation, as DiD approaches could provide more robust causal identification by comparing treated and untreated areas while controlling for time-invariant unobserved characteristics. Future research should prioritize the collection of geo-tagged social media data or explore alternative quasi-experimental designs that can leverage spatial or demographic variation in program exposure.

The framework's generalizability extends beyond our specific case study context. The methodological approach can be adapted to evaluate transit programs in other cities and cultural contexts, though careful attention must be paid to platform-specific characteristics, language processing requirements, and local social media usage patterns. The semantic matching component may require customization for different languages and transit terminology, while the ITSA approach remains broadly applicable across contexts.

Future research directions include extending the framework to incorporate multiple data sources simultaneously, such as combining social media sentiment with ridership data, operational metrics, and traditional survey responses. Advanced machine learning techniques could enhance the semantic matching process, potentially using transformer-based models fine-tuned on transportation-specific content. The development of real-time monitoring systems based on this framework could enable proactive program management and rapid response to emerging issues.

Additionally, future studies should explore the integration of spatial analysis techniques when geographic information is available, enabling more sophisticated quasi-experimental designs and spatial heterogeneity analysis. The development of standardized evaluation protocols based on this framework could facilitate cross-city comparisons and meta-analyses of transit improvement program effectiveness.

In conclusion, this study demonstrates the substantial potential of social media data for evidence-based transit program evaluation when combined with appropriate methodological frameworks. While limitations remain, particularly regarding representativeness and spatial identification, the approach offers valuable insights for transit agencies seeking to understand and improve passenger experience in an increasingly connected and digitally-engaged urban environment. The integration of social media analytics with traditional evaluation methods represents a promising direction for enhancing the effectiveness and responsiveness of public transportation systems.

References

Ali, A.M., Parvez, J., Ahmed, M., Hasan, M.K., Rahman, S., Ishtiaque, S., 2017. A fuzzy approach to measuring transit service quality based on user perception. International Journal of Fuzzy Systems 19, 178–191.

Angrist, J.D., Pischke, J.S., 2008. Mostly harmless econometrics: An empiricist's companion. Princeton University Press.

Associates, K.., Administration, U.S.F.T., Program, T.C.R., Corporation, T.D., 2003. Transit capacity and quality of service manual. Transportation Research Board.

Athey, S., Imbens, G.W., 2017. The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives 31, 3–32.

Baek, J., Sohn, K., 2016. Using an interrupted time-series analysis to evaluate the effects of transit service changes on ridership: a case study of daejeon, south korea. Journal of Advanced Transportation 50, 698–716.

Barbosa, S.B., Ferreira, M.G.B., Nickel, E.M., Cruz, J.F.A., Forcellini, F.A., Garcia, J., de Andrade, D.F., 2017. Combining satisfaction and positive critical incidents to evaluate public transport service. Transportation Research Record 2643, 127–134.

Beirão, G., Cabral, J.S., 2007. Understanding attitudes towards public transport and private car: A qualitative study. Transport Policy 14, 478–489.

van den Berg, P., Kemperman, A., Weijs-Perrée, M., Borgers, A., 2019. Social media effects on sustainable mobility opinion diffusion: Model framework and implications for behavior change. Travel Behaviour and Society 16, 1–12.

Bernal, J.L., Cummins, S., Gasparrini, A., 2016. Methodological considerations in the evaluation of public health interventions: Interrupted time series designs. Research Methods in Public Health , 1–10.

Bernal, J.L., Cummins, S., Gasparrini, A., 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. International Journal of Epidemiology 46, 348–355.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research 3, 993–1022.

Chakraborty, K., Roy, S., Singh, M., Jannu, A., Lokras, V., Balamuralidhar, P., 2019. Public perception towards transportation: Interpreting twitter data through sentiment analysis. Transportation Research Procedia 48, 2400–2409.

Chang, Z., Lu, J., Wang, F., 2013. Exploring time-varying effects of stakeholder determinants on metro stations commencement in taipei. International Journal of Sustainable Transportation 7, 292–306.

Chen, Y., He, Z., Zhao, Y., Tsui, K.L., 2019. Geographically modeling and understanding factors influencing transit ridership: an empirical study of shenzhen metro. Applied Sciences 9, 4217.

Collins, C., Hasan, S., Ukkusuri, S.V., 2013a. A novel method for measuring service quality: insights from public transportation twitter feeds. Transportation Research Record 2351, 79–89.

Collins, C., Hasan, S., Ukkusuri, S.V., 2013b. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. Journal of Public Transportation 16, 21–45.

De Oña, J., De Oña, R., Diez-Mesa, F., Eboli, L., Mazzulla, G., 2016. A composite index for evaluating transit service quality across different user profiles. Research in Transportation Economics 59, 229–240.

Dell'Olio, L., Ibeas, A., Cecín, P., 2011. Public transportation quality of service: Factors, models, and applications. Transportation Research Part A: Policy and Practice 45, 419–432.

Dell'Olio, L., Ibeas, A., de Oná, J., de Oná, R., 2018. A methodology to evaluate the effectiveness of different improvement strategies on the users' perception. Transportation Research Procedia 33, 89–96.

Deng, T., Zhang, K., Shen, Q., 2021. Quality of service improvements in public transport: A case study of shenzhen metro. Transport Policy 107, 1–12.

Diab, E., El-Geneidy, A., 2017. Transit service performance and sustainability: A case study of the société de transport de montréal. Journal of Public Transportation 20, 3.

Eboli, L., Mazzulla, G., 2011. A methodology for evaluating transit service quality based on subjective and objective measures from the passenger's point of view. Transport Policy 18, 172–181.

Efthymiou, D., Antoniou, C., 2013. Use of social media for transport data collection and traffic information. Procedia-Social and Behavioral Sciences 48, 775–785.

El-Diraby, T., Shalaby, A., Camacho, F., 2019. Linking social media activity with transit ridership. Transportation

Research Record 2673, 764–773.

Fraser, A., McKenzie, G., Wu, X., Zhong, C., 2024. Using social media data to evaluate the impacts of public transport disruptions on mobility patterns. Journal of Transport Geography 112, 103678.

Friman, M., Edvardsson, B., Gärling, T., 2001. Frequency of negative critical incidents and satisfaction with public transport services. i. Journal of Retailing and Consumer Services 8, 95–104.

Fu, R., Huang, Z., Fink, J., 2015. Social media based analytics for understanding public transit rider complaints. Transportation Research Record 2553, 71–79.

Gal-Tzur, A., Grant-Muller, S.M., Kuflik, T., Minkov, E., Nocera, S., Shoor, I., 2014. The potential of social media in delivering transport policy goals. Transport Policy 32, 115–123.

Golder, S.A., Macy, M.W., 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science 333, 1878–1881.

Guo, Z., Wilson, N.H., Rahbee, A., 2019. Smart card data mining for public transit planning: A case study of shenzhen. Transportation Research Part C: Emerging Technologies 96, 1–19.

Haghighi, N.N., Liu, X.C., Wei, R., Li, W., Shao, H., 2018. Using twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service. Public Transport 10, 363–377.

Hensher, D.A., Stopher, P., Bullock, P., 2003. Service quality–developing a service quality index in the provision of commercial bus contracts. Transportation Research Part A: Policy and Practice 37, 499–517.

Högström, C., Davoudi, S., Löfgren, M., 2016. Relevant and preferred public service: Developing a new approach for public service quality. Public Management Review 18, 1554–1575.

Hong, J., Shen, Q., 2020. Causal inference on travel demand of new nonmotorized paths in an existing network. Journal of Transport Geography 82, 102618.

Houston, D., Luong, T.T., 2015. Public transit services for improving public health: A new approach to meet the transportation needs of vulnerable populations. Transportation Research Board Conference Proceedings 2.

Imbens, G.W., Rubin, D.B., 2015. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

Ingvardson, J.B., Nielsen, O.A., 2019. The relationship between objective and perceived public transport service quality. Journal of Public Transportation 22, 2.

Jiawen, X., Kanev, A., 2025. Chinese text classification based on different word segmentation methods , 1–6.

Jin, C., Chen, H., Wen, Z., 2020. Deep learning-based traffic flow prediction for public transportation: A case study of bus passenger flow. Journal of Advanced Transportation 2020.

Kamga, C., Wang, M., Sapp, D., Agrawal, S., 2023. Utilizing social media for public transit service quality assessment and interactive mapping. Transportation Research Record 2677, 118–131.

Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! the challenges and opportunities of social media. Business horizons 53, 59–68.

Karner, A., Niemeier, D., 2016. Transportation planning and regional equity: History, policy and practice. Improving Pathways to Transit for Californians .

Kontopantelis, E., Doran, T., Springate, D.A., Buchan, I., Reeves, D., 2015. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ 350, h2750.

Koppel, M., Kim, K., Hong, A., 2023. Disentangling the causal effect of rail transit on crime: A spatiotemporal analysis of the expo line in los angeles. Journal of Transport Geography 109, 103583.

Lechner, M., 2011. The estimation of causal effects by difference-in-difference methods. Foundations and Trends in Econometrics 4, 165–224.

Li, X., Chen, Y., Wang, H., 2022. Comparative analysis of metro ridership before and after covid-19: A case study of shenzhen. Transportation Research Part A: Policy and Practice 155, 1–15.

Liu, X.C., Ban, X., 2017. Monitoring transit service performance with social media: An application to the chicago transit authority. Transportation Research Record 2649, 42–50.

Lopez Bernal, J., Cummins, S., Gasparrini, A., 2016. Interrupted time series regression for the evaluation of public health interventions: a tutorial. International journal of epidemiology 46, 348–355.

Lopez Bernal, J., Cummins, S., Gasparrini, A., 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. International Journal of Epidemiology 46, 348–355.

Luong, T.T., Houston, D., 2015. Public transit service quality in san francisco: Sentiment analysis of user-generated content. Transportation Research Record 2538, 11–20.

Luong, T.T., Houston, D., Boarnet, M.G., 2015. Mining public transit service quality from social media: A sentiment analysis approach. Transportation Research Part C: Emerging Technologies 58, 373–385.

Ma, J., Chan, J., Ristanoski, G., Rajasegarar, S., Leckie, C., 2018. An integrated framework for optimizing underground rail systems by using a hybrid swarm intelligence approach. Transportation Research Part C: Emerging Technologies 90, 161–183.

Maeda, T., Takashi, K., Kamino, K., Wang, C.H., Motoyama, M., Uchida, Y., Nishiyama, H., 2019. Transportation mode identification from mobility data using convolutional neural networks. IEEE Access 7, 122954–122963.

Mathur, S., Zhang, Y., Ukkusuri, S.V., 2021. An exploratory analysis of social media for transit service evaluation: Opportunities and challenges. Transportation Research Part C: Emerging Technologies 125, 103067.

Mead, L., 2021. Road transport and climate change: Stepping off the greenhouse gas. Transportation Research Part D: Transport and Environment 95, 102826.

Ming, Z., Yang, L., Chen, X., 2014. Understanding the impact of tf-idf on topic modeling performance. Journal of Information Science 40, 645–655.

Morrison, G.M., Lin, C.Y.C., 2018. The impact of light rail on congestion in denver: A synthetic control approach. Regional Science and Urban Economics 71, 57–72.

Morton, C., Caulfield, B., Anable, J., 2016. Customer perceptions of quality of service in public transport: Evidence for bus transit in scotland. Case Studies on Transport Policy 4, 199–207.

Nathanail, E., 2008. Measuring the quality of service for passengers on the hellenic railways. Transportation Research Part A: Policy and Practice 42, 48–66.

Nguyen-Phuoc, D.Q., Currie, G., De Gruyter, C., Young, W., 2016. Transportation network companies and the ridesourcing industry: A review of impacts and emerging regulatory frameworks for uber. Urban, Planning and Transport Research 4, 40–63.

Pearl, J., 2009. Causality. Cambridge University Press.

Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S., Waller, T.S., 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. Transportation Research Part C: Emerging Technologies 75, 197–211.

Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks , 3982–3992.

Schaffer, A.L., Dobbins, T.A., Pearson, S.A., 2021. Interrupted time series analysis using autoregressive integrated moving average (arima) models: a guide for evaluating large-scale health interventions. BMC medical research methodology 21, 1–12.

Schweitzer, L., 2014. Planning and social media: a case study of public transit and stigma on twitter. Journal of the American Planning Association 80, 218–238.

Spirtes, P., Zhang, K., 2016. Causal discovery from big data: theory and practice. Frontiers in Big Data 1, 1–10.

for Standardization, E.C., 2002. EN 13816: Transportation-Logistics and Services-Public Passenger Transport-Service Quality Definition, Targeting and Measurement. Technical Report. European Committee for Standardization. Brussels.

Steiger, E., Westerholt, R., Resch, B., Zipf, A., 2015. Twitter as an indicator for whereabouts of people? correlating twitter with uk census data. Computers, Environment and Urban Systems 54, 255–265.

Stjernborg, V., Mattisson, O., 2016. The role of public transport in society—a case study of general policy documents in sweden. Sustainability 8, 1120.

Tasse, D., Hong, J.I., 2014. Using social media data to understand cities. Proceedings of NSF Workshop on Big Data and Urban Informatics , 64–79.

Tse, Y.K., Zhang, M., Akhtar, P., MacBryde, J., 2018. Social media data for urban sustainability: Opportunities, challenges, and future directions. Sustainable Cities and Society 39, 454–463.

Tyrinopoulos, Y., Antoniou, C., 2008. Public transit user satisfaction: Variability and policy implications. Transport Policy 15, 260–272.

Wagner, A.K., Soumerai, S.B., Zhang, F., Ross-Degnan, D., 2002. Segmented regression analysis of interrupted time series studies in medication use research. Journal of clinical pharmacy and therapeutics 27, 299–309.

Wang, B., Zhang, L., Siebeneck, L., Sánchez, V., Shuai, X., 2020a. Analyzing public transit service quality based on mobile phone data. IEEE Access 8, 144704–144713.

Wang, J., Zhou, Y., Zhang, W., Evans, R., Zhu, C., 2020b. Empirical analysis of social media usage patterns: A case study of weibo during covid-19. Journal of Medical Internet Research 22, e22152.

Wang, N., Jin, X., Zhang, L., 2020c. Mining user opinions in social media: A case study on high-speed rail in china. Transport Policy 90, 1–12.

Wu, Y., Zhu, L., Zhi, Y., Liu, M., Wang, Z., Lai, J., 2020. An integrated approach combining gis, social media data, and behavior surveys for passenger flow analysis in metro stations. ISPRS International Journal of Geo-Information 9, 570.

Ye, H., Xiao, F., Yang, H., 2020. A causal inference approach to measure the vulnerability of urban metro systems. Transportation 47, 1939–1970.

Yuan, N.J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., Xiong, H., 2016. Discovering urban functional zones using latent activity trajectories. IEEE Transactions on Knowledge and Data Engineering 27, 712–725.

Zhang, R., Zhang, Y., Lin, Y., Wang, S., Liu, Y., Lancelot Milthorpe, F., 2023. Changes to commuting patterns in response to covid-19 and the associated impacts on air pollution in china. Transportation Research Part D: Transport and Environment 114, 103537.

Zhang, T., Liang, L., Otten, M., Orosz, K., Fulmer, J., Marshall, R., 2018. Analytics of real-time transit demand using large-scale transit data. Transportation Research Record 2672, 583–591.

Zhang, W., Li, Y., Ukkusuri, S.V., 2019. Examining spatial patterns in social media sentiment toward transit services in new york city. Transportation Research Part C: Emerging Technologies 101, 1–16.

Zhao, F., Pereira, F.C., Ball, R., Kim, Y., Han, Y., Zegras, C., Ben-Akiva, M., 2013. Web-based transit service quality survey: preliminary results from washington state. Transportation Research Record 2351, 100–108.

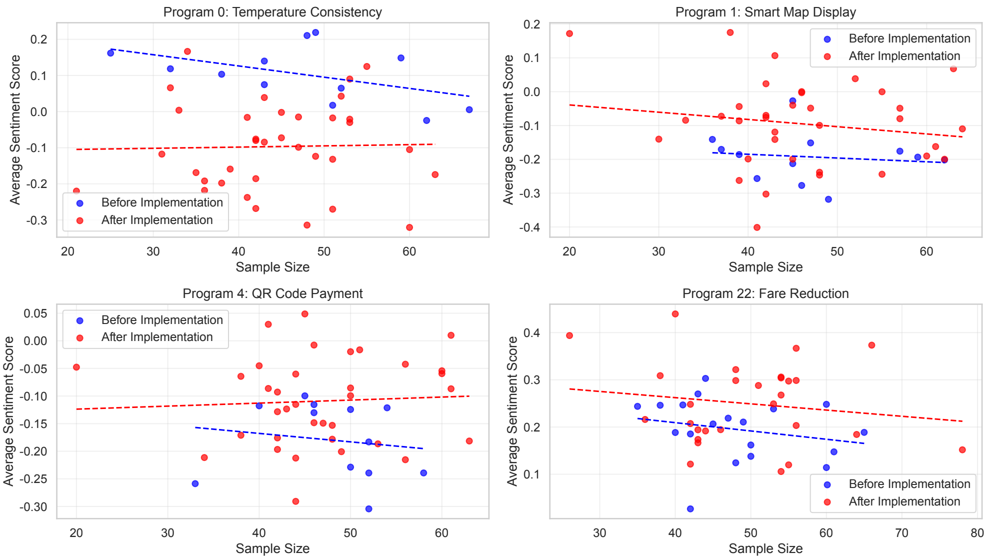

Figure 1: Sample Size Distribution Across Transit Improvement Programs

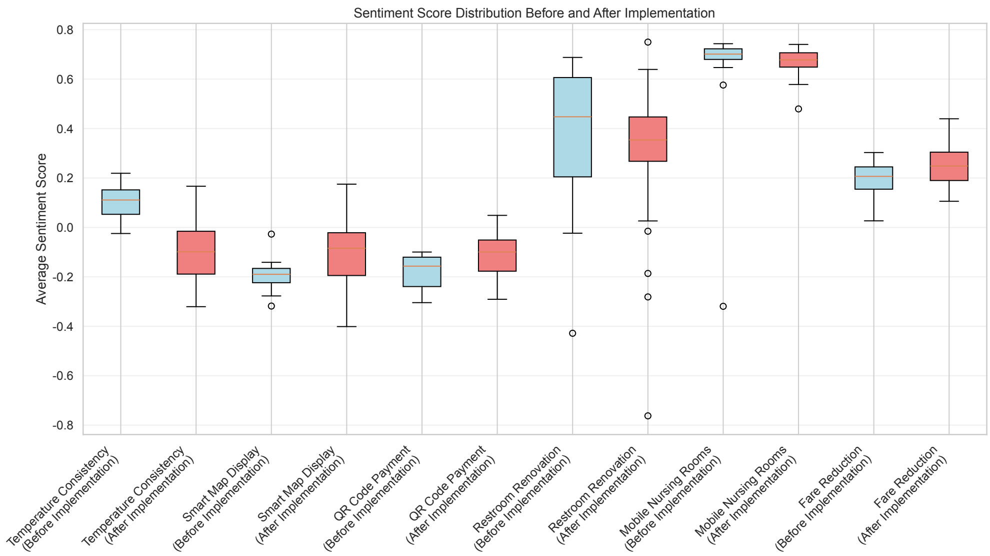

Figure 2: Sentiment Distribution by Program Before and After Implementation

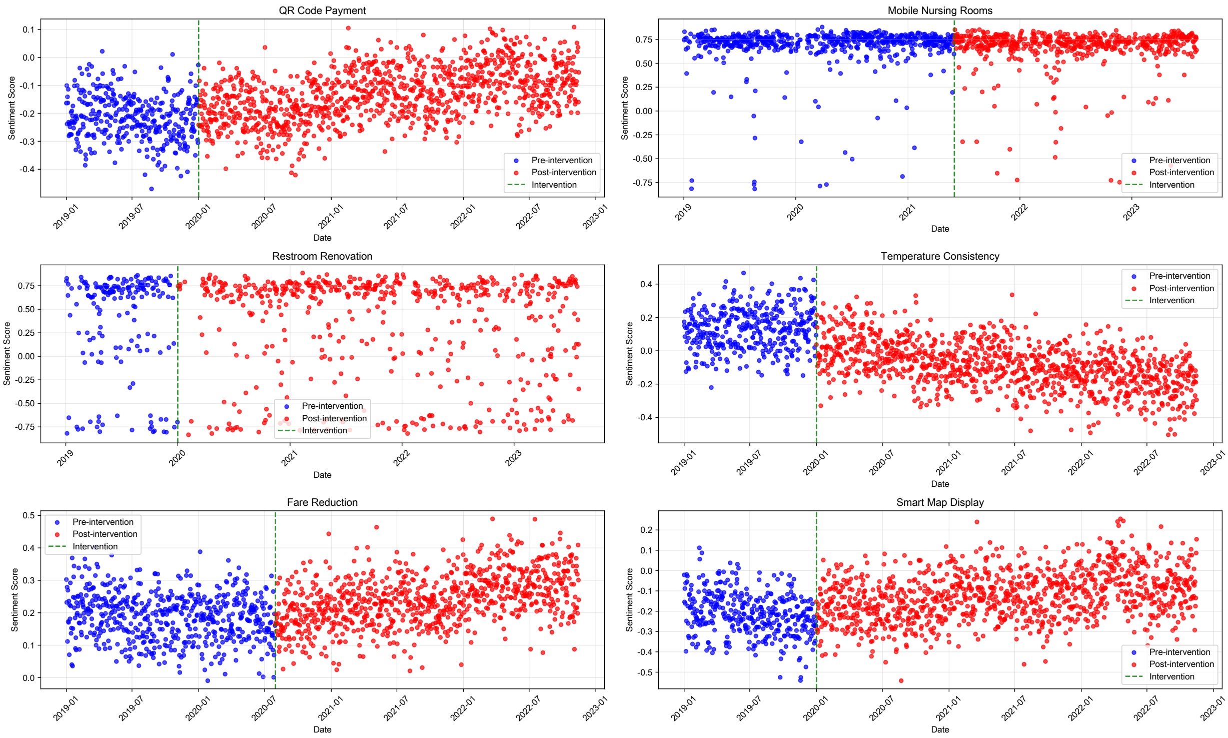

Figure 3: Time Series Analysis of Sentiment Patterns Across Programs

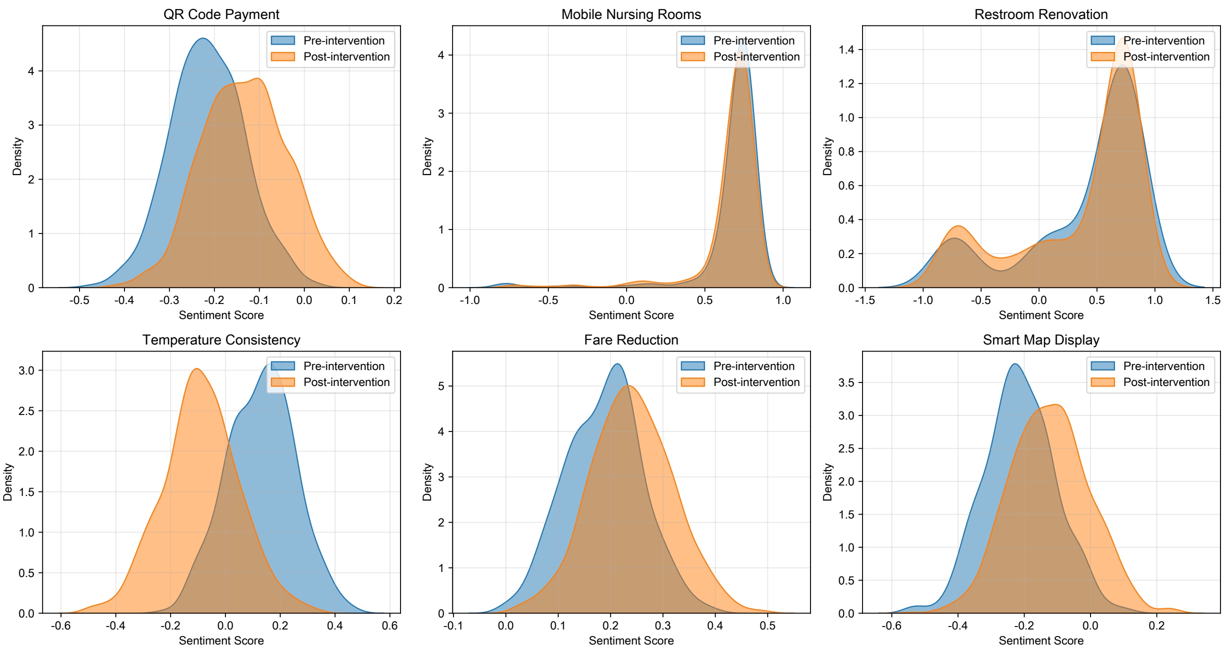

Figure 4: Density Plots of Sentiment Distributions Before and After Program Implementation

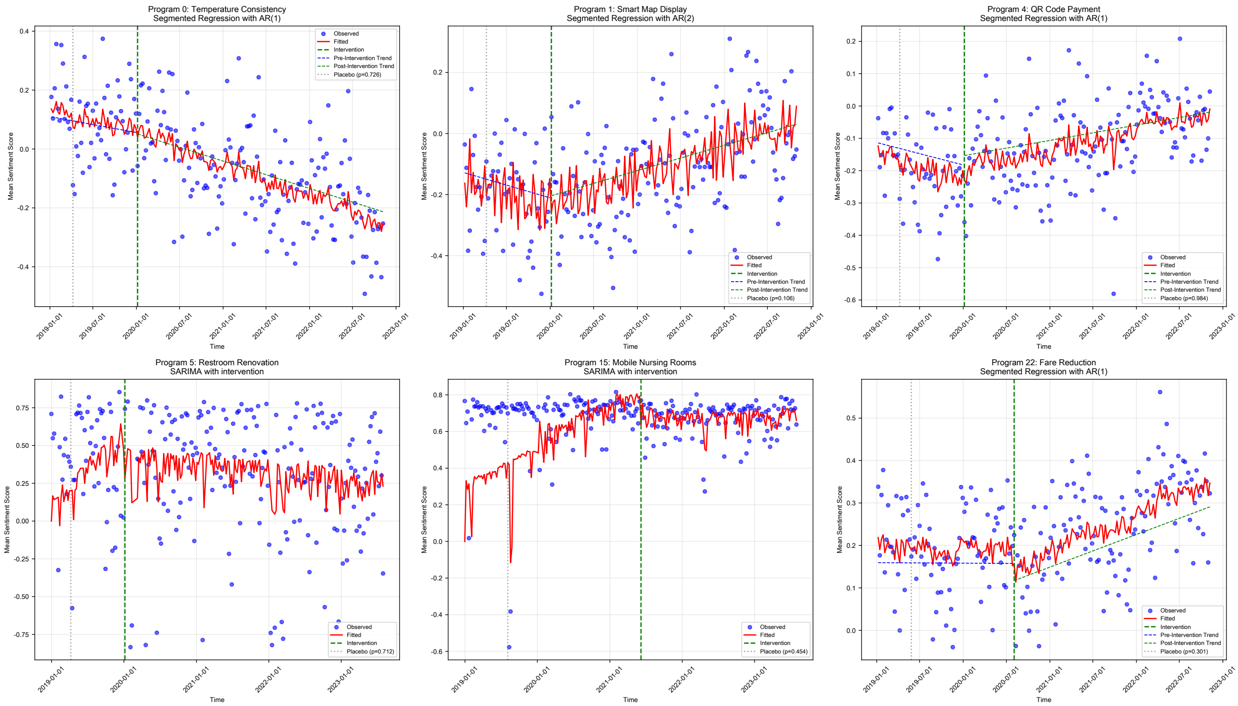

Figure 5: Interrupted Time Series Analysis Results for All Transit Improvement Programs