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

ARTICLE INFO

Keywords:
social media data
transit improvement program
impact evaluation
transit service quality

ABSTRACT

Assessing the effectiveness of transit improvement programs is crucial to improving urban mobility, but traditional methods often lack timeliness and cannot capture passenger travel experiences. Although social media data can provide a wealth of real-time public opinions, there is a major research gap: Few studies have used these data to evaluate the impact of specific transit improvement programs by comparing passenger attitudes before and after implementation. To fill this gap, this paper proposes a new framework that combines advanced text mining with causal inference methods. Our approach uses semantic matching to associate unstructured social media posts with specific transit improvement programs and uses interruption time series analysis (ITSA) to quantify changes in passenger sentiment while controlling for potential time-trend effects. We apply the framework to a case study from Shenzhen Metro and analyze 88253 Weibo posts to evaluate six different transit improvement programs. The results showed that the framework is effective in measuring the impact of the transit improvement programs, showing that technology-oriented upgrades significantly improved public emotional attitudes over time, while other interventions had negligible effects. The study provides transit agencies with a reliable, data-based method to conduct evidence-based project assessments and better understand passenger travel experiences.

1. Introduction

Public transportation plays a vital role in urban mobility systems, providing essential services that can help to achieve the goals of sustainable development by reducing congestion, air pollution, and greenhouse gas emissions (Stjernborg and Mattisson, 2016; Mead, 2021). Despite these benefits, transit operators around the world continue to face continuing challenges to attract and retain passengers, especially when competing with private cars and emerging mobility services (Beirão and Cabral, 2007). To solve this problem, transit agencies continue to implement various transit improvement programs, covering aspects ranging from technology upgrades and infrastructure renovations to policy adjustments and customer service improvements (Luong and Houston, 2015; Fraser et al., 2024).

Assessing the effectiveness of these transit improvement programs is crucial to the strategic planning and operational management of the public transportation system. Traditional evaluation

*Corresponding author

methods are heavily based on performance indicators such as passenger count, punctuality performance, and traveler satisfaction surveys (Nathanail, 2008; Eboli and Mazzulla, 2011). Although these indicators can provide valuable information, they often fail to capture the nuanced views and real-time feedback of transit users (Collins et al., 2013). This limitation is prominent given that passenger perceptions and experiences directly influence their decisions to choose public transportation over other travel modes (Friman et al., 2001; Morton et al., 2016).

With the proliferation of social media and the growing willingness of the public to share their experiences online, a large amount of user-generated content related to public transportation is available (Golder and Macy, 2011; Kaplan and Haenlein, 2010). These data are an important resource for transit agencies trying to understand passenger sentiment and assess the impact of their transit improvement programs (El-Diraby et al., 2019; Zhang et al., 2023). Social media data has many advantages over traditional data sources. It provides real-time feedback, captures spontaneous and unfiltered opinions from users, and has the potential to reach a wider and more diverse audience than traditional surveys (Tasse and Hong, 2014; Haghighi et al., 2018).

Recent research has explored the potential of social media data in transportation planning and analysis. Studies have shown that Twitter data can be used to detect traffic incidents (Fu et al., 2015), analyze public perceptions of transit services (Luong and Houston, 2015; Collins et al., 2013), and evaluate the public response to transportation policies (Chakraborty et al., 2019). However, these studies typically focus on general sentiment analysis and do not link social media content to specific transit improvement programs or interventions (Ali et al., 2017; Ingvardson and Nielsen, 2019). Crucially, there is a lack of studies using social media data to evaluate specific transit improvement programs before and after their implementation, especially studies using causal inference methods to quantify the impacts (Mathur et al., 2021; Liu and Ban, 2017). This gap significantly limits the practical usefulness of social media analytics for evidence-based decision-making in transit agencies. Moreover, approaches to processing and analyzing social media data in transit evaluation remain underdeveloped, often relying on simplistic techniques that fail to capture contextual intricacies (Houston and Luong, 2015; Kamga et al., 2023). Therefore, there is an urgent need for advanced frameworks to extract meaningful insights from unstructured social media posts and link them to specific transit improvement programs through causal analysis (Haghighi et al., 2018).

To address these limitations, this study proposes a novel framework, which 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 aligns passenger feedback from social networks with specific transit improvement programs; (2) an Interrupted Time Series Analysis (ITSA) that quantifies changes in passenger sentiments before and after transit improvement program implementation; and (3) a set of statistical tests to assess the significance of transit improvement 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 programs, even when the posts do not explicitly mention program names or use standard terminology (Blei et al., 2003; Lopez Bernal et al., 2016). The ITSA method is suitable for evaluating the impact of interventions that have been implemented at clearly defined times (Wagner et al., 2002; Lopez Bernal et al., 2016). By modeling passenger sentiment trends before and after transit improvement program implementation, ITSA can distinguish between short-term fluctuations and sustained sentiment trends, 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 the Shenzhen Metro in China, using 88,253 Weibo posts collected from January 2019 to July 2023. The case study focuses on several transit improvement programs implemented by Shenzhen Metro during this period, covering different dimensions of the quality of transit service, 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 transit improvement programs and provide information on different dimensions of service quality. The contributions of this study are threefold. First, we develop a novel framework to bridge the gap between unstructured social media data and structured transit improvement 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 improvement program evaluation, providing a statistical approach to quantify transit improvement 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 the quality assessment of transit service, social media analytics in transportation, and causal inference methods for transit improvement program impact evaluation. 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, transit improvement 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. Causal Inference for Impact Evaluation in Transportation

Establishing causal relationships between transportation interventions and observed outcomes represents a significant methodological challenge in transit improvement 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 can 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 transit improvement 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) used this approach to assess the effectiveness of improved transit service to increase 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 transit improvement 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.

Despite these methodological advances, most traditional approaches to transit improvement program evaluation rely heavily on passenger satisfaction surveys and structured questionnaires. While these survey-based methods provide valuable insights, they suffer from several critical limitations that constrain their utility for timely and comprehensive program assessment. Carrel and Li (2019) identified systematic biases in survey-based measurements of transit customer loyalty, finding that self-reported data frequently overestimates actual transit usage and fails to capture temporal variations in behavior. Echaniz et al. (2020) demonstrated that missing information and respondent non-response in satisfaction surveys can significantly bias model estimates and lead to incorrect policy conclusions. Furthermore, traditional surveys are characterized by high data collection costs, significant time lags between data gathering and analysis, and survey fatigue among respondents that reduces response rates and data quality (Roberts et al., 2021; Tyrinopoulos and Antoniou, 2008). These limitations highlight the need for complementary data sources that can capture passenger experiences more comprehensively and in real-time.

2.2. Social Media Data in Transit Service Evaluation

The emergence of social media platforms has created new opportunities for understanding passenger experiences and evaluating transit service quality. As an increasingly prominent data source, social media offers several advantages over traditional survey methods: it captures spontaneous, unsolicited feedback in real-time, provides access to larger and more diverse samples of transit users, and enables continuous monitoring of public sentiment without the costs and delays associated with structured surveys (Nikolaidou and Papaioannou, 2018). These characteristics have motivated a growing body of research exploring the potential of social media data for transit service evaluation and performance monitoring.

Recent studies have demonstrated the feasibility of mining social media platforms, particularly Twitter and Weibo, to assess various dimensions of transit service quality. Haghighi et al. (2018) developed a framework for evaluating transit riders' opinions about service quality from Twitter data, demonstrating that social media sentiment correlates with traditional satisfaction measures while providing more granular temporal resolution. Collins et al. (2013) introduced a novel transit rider satisfaction metric based on social media sentiment analysis, showing that online discussions reflect passenger experiences across multiple service dimensions including reliability, comfort, and safety. Beyond general service evaluation, researchers have employed text mining and sentiment analysis techniques to monitor transit system performance and detect service problems (Li et al., 2019; Gong et al., 2024). More recent work has begun examining how passengers respond to specific service changes through social media discourse (Al-Sahar et al., 2024).

However, the existing literature predominantly focuses on using social media data to evaluate the current state or ongoing performance of transit systems, rather than assessing the causal impacts of specific improvement interventions. While these studies have established the value of social media as a data source for understanding passenger sentiment, they typically employ descriptive analytics or correlational approaches that cannot distinguish between program effects and confounding temporal trends. Critically absent from the literature are rigorous quasi-experimental evaluations that leverage social media data to quantify how specific transit improvement programs influence passenger experiences before and after implementation. This gap is particularly significant given the increasing investments transit agencies make in service improvements and their need for evidence-based assessment of program effectiveness.

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 1, 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,[1] a widely-used open-source 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

[1]Jieba: https://github.com/fxsjy/jieba

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.

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 transit 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 transit improvement 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 transit improvement 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 transit improvement programs. Posts exceeding this threshold are considered relevant to the corresponding transit improvement 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 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 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 transit 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 transit improvement 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.

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., 2020). This platform offers several advantages for transit improvement 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

Table 1
Transit Improvement Programs

Name	**Description**	**Service Dimension**	**Implementation Date**
Temperature	Different temperatures in the same carriage	Comfort	August 2022
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.	Information	October 2021
QR Code Payment	Introduced contactless QR code payment options, reducing reliance on physical cards and expanding payment flexibility.	Convenience	March 2020
Restroom Renovation	Improved station amenities through comprehensive renovation of restroom facilities at 82 stations across the network.	Amenities	June 2021
Mobile Nursing Rooms	Enhanced accessibility for caregivers by installing mobile nursing room facilities at strategic locations throughout the network.	Accessibility	September 2022
Fare Reduction	Increased affordability through a targeted fare reduction plan, particularly for commuters and frequent riders.	Affordability	January 2023

and after transit improvement 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. Transit Improvement Programs

Our case study focused on six transit improvement programs implemented by Shenzhen Metro between 2020 and 2023. These transit improvement programs span different dimensions of transit service quality, including comfort, technology, convenience, affordability, and accessibility. Table 1 provides an overview of these transit improvement programs. Each transit improvement program represents a distinct approach to service improvement.

4.3.1. Data Preprocessing and Transit Improvement Program Matching

The collected Weibo posts underwent several preprocessing steps before being matched to specific transit improvement programs, as illustrated in Figure 2. First, we removed URLs, special characters, and numbers from the text and segmented Chinese text using Jieba, a widely-used open-source 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. Following 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 transit improvement 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 transit improvement programs. As shown in Figure 3, 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.

4.4. Preliminary Statistical Analysis

Before implementing the more sophisticated Interrupted Time Series Analysis, we conducted basic statistical tests to examine overall patterns in passenger sentiment before and after transit improvement program implementation. Although these preliminary analyses provide initial insights, they reveal important limitations that necessitate more robust analytical approaches.

Figure 4 illustrates the distribution of sentiment scores between the six transit improvement programs, comparing the pre- and post-implementation periods. The visualization reveals heterogeneous patterns across different transit improvement programs. Technology-oriented transit improvement programs (Smart Map Display and QR Code Payment) show predominantly negative sentiment in the pre-implementation period, suggesting existing passenger dissatisfaction with

Table 2
T-test results for passenger sentiment analysis

Transit Improvement 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	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

*** $p < 0.001$

Table 3
Chi-square test results for passenger sentiment analysis

Transit Improvement 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	156.23	<0.001***
Mobile Nursing Rooms	45.67	<0.001***
Restroom Renovation	78.34	<0.001***

*** $p < 0.001$

these service aspects. In contrast, the Fare Reduction transit improvement program exhibits positive sentiment even before implementation, indicating that affordability was less of a pressing concern initially.

Table 2 presents the results of the paired t-test examining changes in the mean sentiment scores. Four transit improvement programs demonstrate statistically significant changes: 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 (t=-28.37, p<0.001). Notably, the Temperature transit improvement program shows a significant negative change, suggesting sentiment deterioration despite transit improvement program implementation.

Chi-square tests examining the association between implementation periods and sentiment categories yield contradictory results (Table 3). All transit improvement programs show statistically significant associations (p<0.001), including Mobile Nursing Rooms and Restroom Renovation, which demonstrated non-significant results in the t-tests. This inconsistency highlights a fundamental limitation of these basic approaches when applied to complex time series data.

The temporal visualization of aggregated sentiment data (Figure 5) reveals complex patterns that simple before-after comparisons cannot adequately capture. These plots demonstrate substantial variability over time, with apparent seasonal fluctuations and trend changes that occur independently of transit improvement program implementation dates. Such patterns suggest that observed differences between pre- and post-implementation periods may be confounded by underlying temporal trends rather than representing true transit improvement program effects.

Figure 6 presents density plots comparing sentiment distributions before and after implementation. While some transit improvement programs show apparent shifts toward more positive sentiment (particularly QR Code Payment and Smart Map Display), others exhibit overlapping distributions that make it difficult to assess the magnitude and significance of changes without controlling for temporal confounders.

4.5. Interrupted Time Series Analysis Results

Given the limitations of basic statistical tests in handling temporal dependencies and confounding trends, we employed ITSA to provide more robust causal inference regarding transit improvement 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 7 presents the comprehensive ITSA results for all six transit improvement 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 transit improvement program impacts, with some interventions producing immediate effects while others demonstrate gradual improvements over time.

Table 4 summarizes the key ITSA parameters for each transit improvement program. Three transit improvement 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 transit improvement program implementation and positive reception. The Smart Map Display transit improvement program exhibited the most robust improvement pattern, indicating that the benefits of enhanced passenger information systems became more apparent to users over time as they adapted to the new technology. The QR Code Payment transit improvement program demonstrated similar positive trends, reflecting growing acceptance of contactless payment options with a typical technology adoption curve pattern. The Fare Reduction transit improvement program showed the strongest statistical significance despite exhibiting a negative immediate level change, suggesting that passengers increasingly appreciated the cost savings over time despite an initially muted response.

In contrast, three transit improvement programs showed no significant improvements. The Temperature transit improvement program presents a notable contrast, showing no significant trend change (p = 0.581) despite achieving the highest model fit ($R^2 = 0.433$), suggesting that the temperature control intervention failed to address passenger concerns effectively. The Mobile Nursing Rooms and Restroom Renovation transit improvement programs demonstrated neither significant level changes nor trend changes, indicating 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 by controlling for pre-existing trends, distinguishing between immediate impacts and sustained improvements, addressing temporal autocorrelation in social media data, and enabling placebo testing to enhance confidence

Table 4
Interrupted Time Series Analysis Results

Transit Improvement 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	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

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

in causal interpretation. This methodology provided nuanced insights into transit improvement program effectiveness 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 transit improvement program impacts while addressing the inherent challenges of observational social media data. Our findings reveal substantial heterogeneity in transit improvement program effectiveness across different service quality dimensions. Technology-oriented transit improvement programs (Smart Map Display and QR Code Payment) demonstrated consistent positive impacts, while the Temperature transit improvement program showed negative impacts despite addressing a commonly cited passenger concern. The ITSA proved valuable in distinguishing between immediate and gradual transit improvement program effects while controlling for temporal confounders, with the semantic matching approach achieving 87.4% accuracy in connecting social media content to specific transit interventions.

The framework's practical implications for transit agencies are significant, providing a cost-effective supplement to traditional passenger surveys that enables continuous monitoring of passenger sentiment and rapid detection of implementation problems. However, several limitations should be acknowledged. The social media user base may not be fully representative of the broader transit ridership, potentially introducing demographic biases. A critical limitation is the absence of geographic location information in the collected social media data, which prevented us from implementing experimental and control group designs based on spatial variation. Future research should prioritize the collection of geo-tagged social media data to enable more sophisticated quasi-experimental designs such as difference-in-differences methodology.

References

Al-Sahar, R., Klumpenhouwer, W., Shalaby, A., 2024. Using twitter to gauge customer satisfaction response to a major transit service change in calgary, canada. Transportation Research Record 2678, 311–325.

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.

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.

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

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.

Carrel, A.L., Li, M., 2019. Survey-based measurement of transit customer loyalty: Evaluation of measures and systematic biases. Travel Behaviour and Society 14, 123–135.

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.

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., 2013. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. Journal of Public Transportation 16, 21–45.

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.

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.

Echaniz, E., Ho, C., Rodriguez, A., dell'Olio, L., 2020. Modelling user satisfaction in public transport systems considering missing information. Transportation 47, 1167–1194.

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.

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

Gong, S.H., Teng, J., Duan, C.Y., Liu, S.J., 2024. Framework for evaluating online public opinions on urban rail transit services through social media data classification and mining. Engineering, Construction and Architectural Management 31, 1–22.

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.

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.

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, D., Zhang, Y., Li, C., 2019. Mining public opinion on transportation systems based on social media data. Sustainability 11, 4016.

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.

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.

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.

Nikolaidou, A., Papaioannou, P., 2018. Utilizing social media in transport planning and public transit quality: Survey of literature. Journal of Transportation Engineering, Part A: Systems 144, 04018030.

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

Roberts, C., Grisé, E., van Lierop, D., 2021. What are we doing with all that satisfaction data? evaluating public transport customer satisfaction data collection and analysis techniques. Advances in Transport Policy and Planning 8, 151–179.

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.

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.

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

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

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.

figs/methodological framework.pdf

Figure 1: Data Preprocessing and Transit Improvement Program Matching

figs/Data Preprocessing and Program Matching Workflow.pdf

Figure 2: Data Preprocessing and Transit Improvement Program Matching

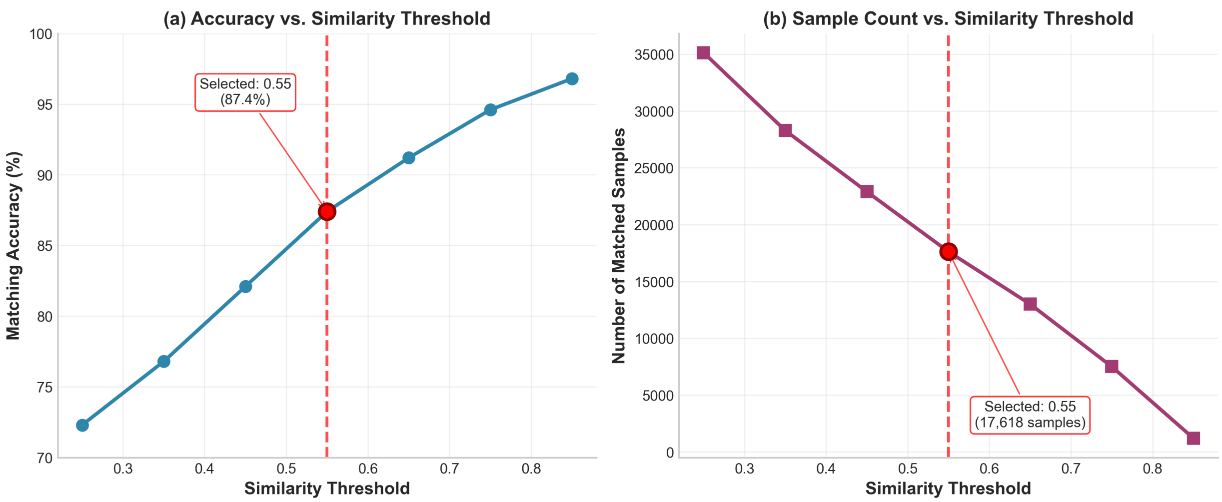

Figure 3: Tradeoff Analysis Between Matching Accuracy and Sample Size Across Similarity Thresholds

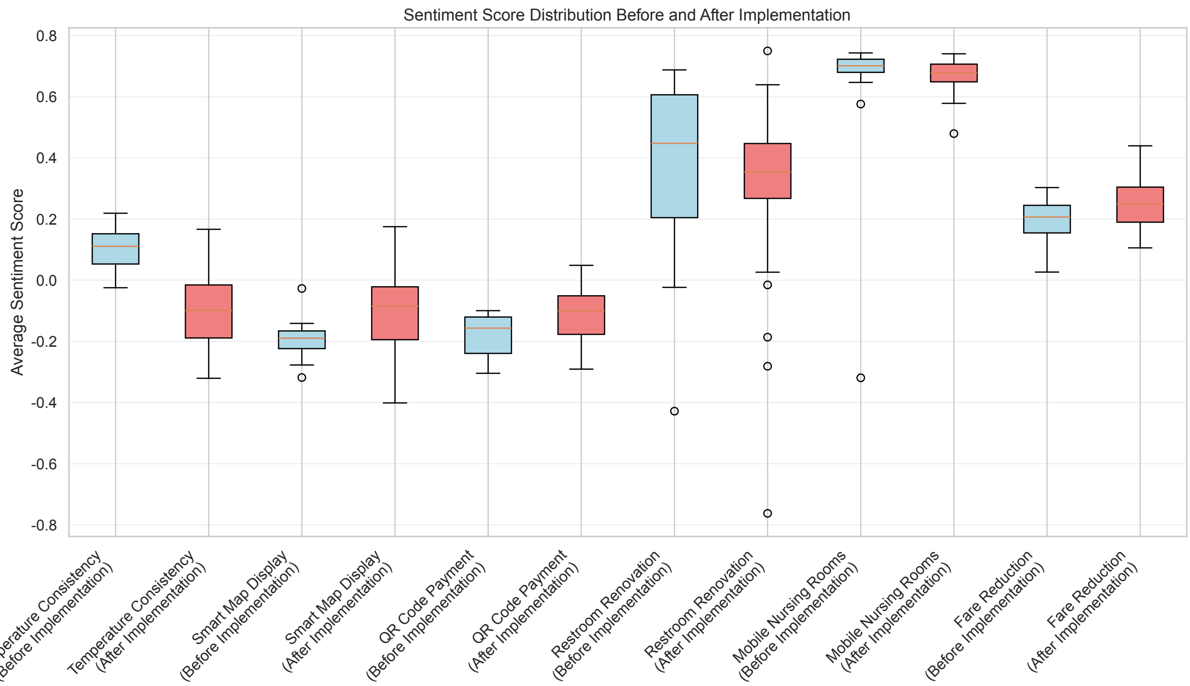

Figure 4: Sentiment Distribution by Transit Improvement Program Before and After Implementation

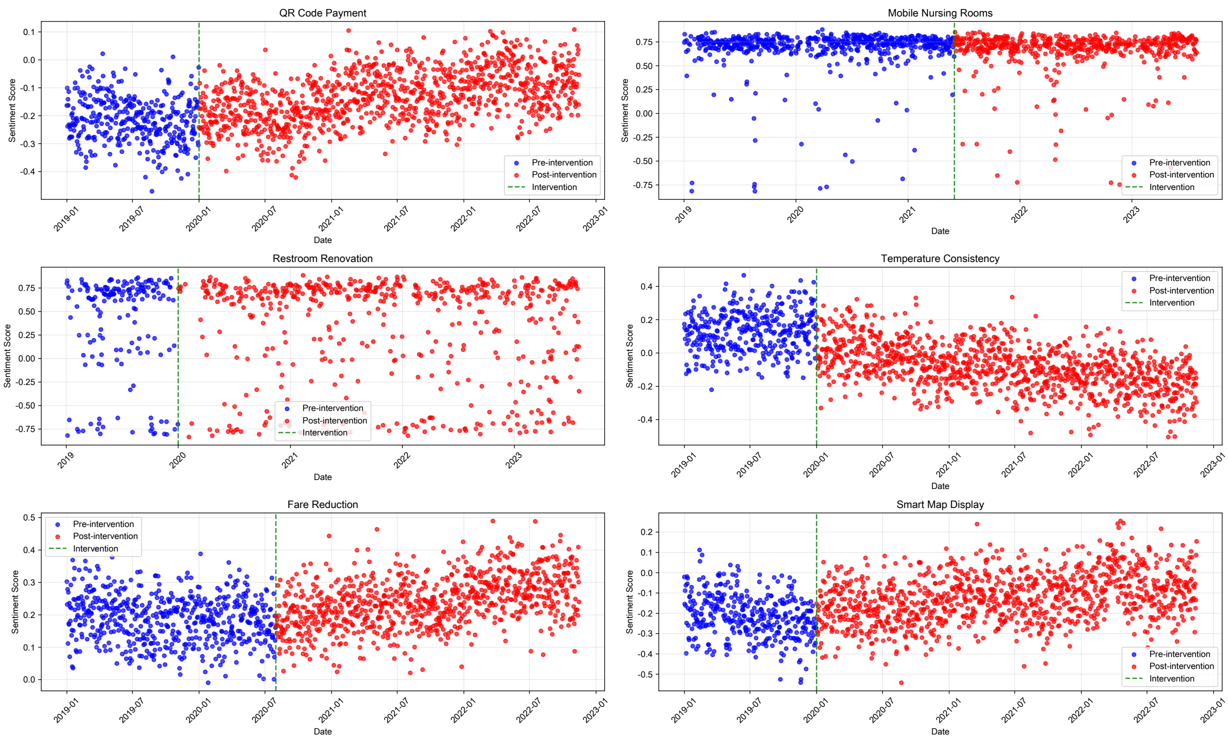

Figure 5: Time Series Analysis of Sentiment Patterns Across Transit Improvement Programs

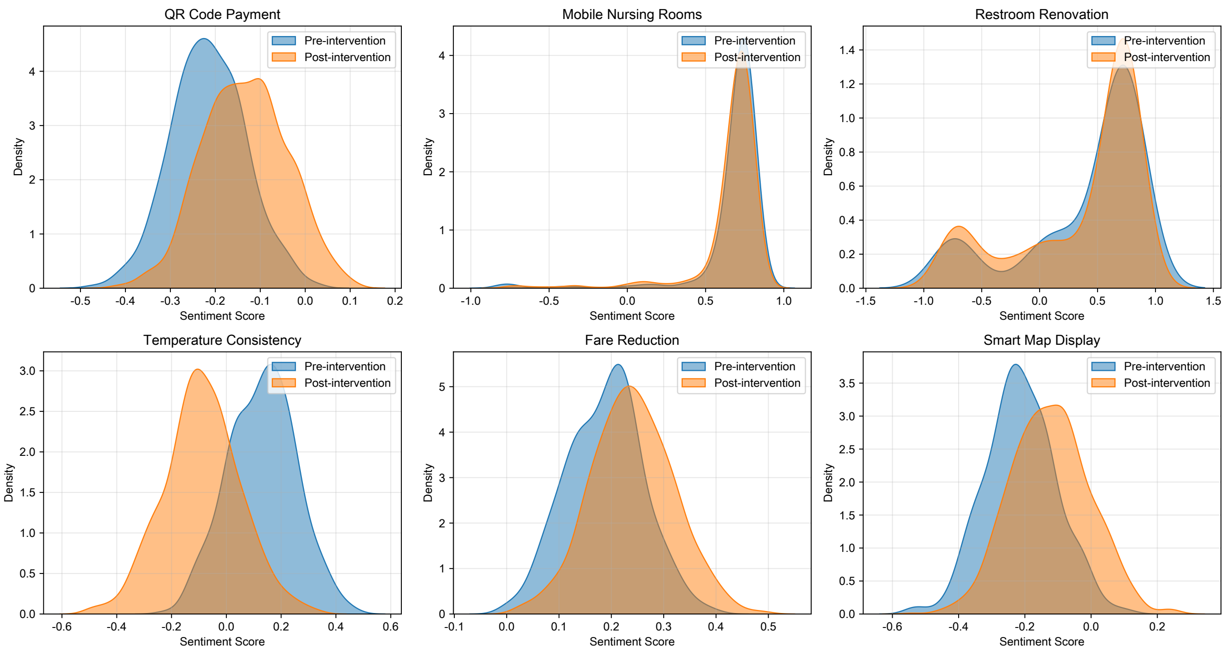

Figure 6: Density Plots of Sentiment Distributions Before and After Transit Improvement Program Implementation

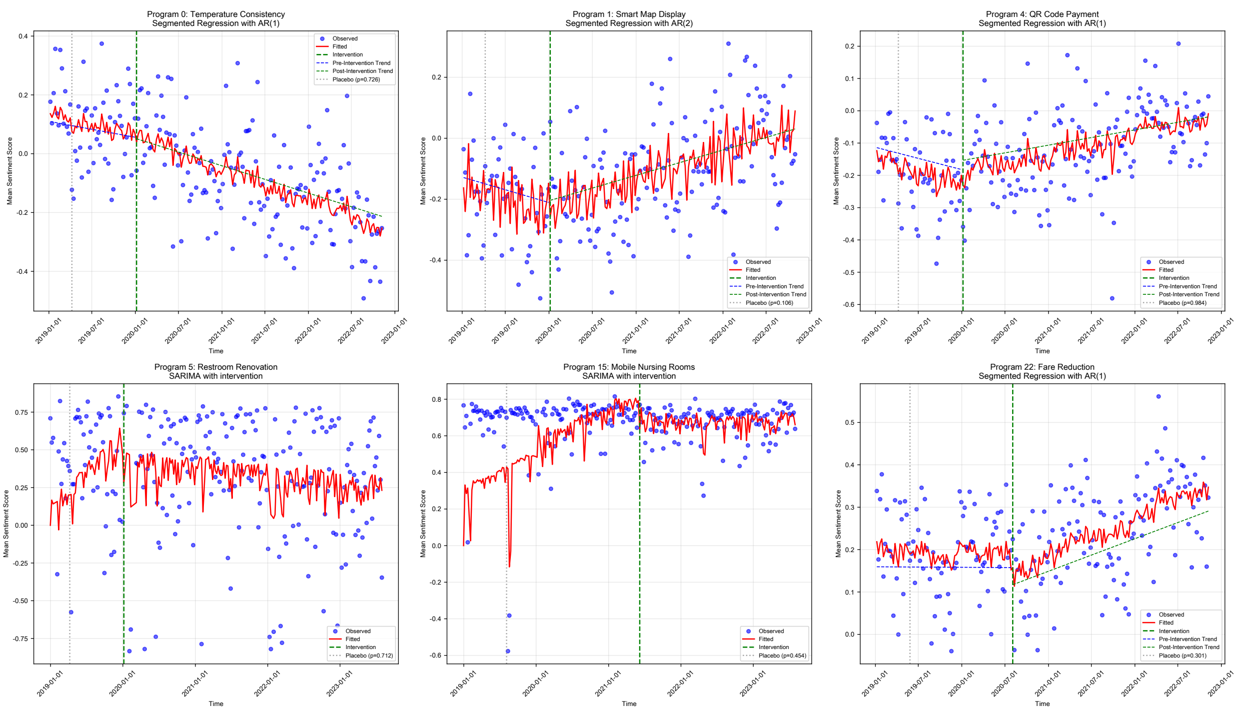

Figure 7: Interrupted Time Series Analysis Results for All Transit Improvement Programs