Jurnal Ilmu Sosial, Vol 25, No 1, 2026, pp. 1-22
DOI: 10.14710/jis.25.1.2026.1-22
Website: https://ejournal.undip.ac.id/index.php/ilmusos

Original Article

When the last bus leaves: Digital citizen discourse and urban transit service failures in Surabaya

Esa Wahyu Endarti[1*], M. Roehman Zainur Riedho[2], Ridha Amaliyah[3], and Salismi Zulfi Maulidita[4]

[1] Department of Public Administration, Universitas Wijaya Putra, Surabaya, Indonesia
[2] Master of Public Policy, Universitas Airlangga, Surabaya, Indonesia
[3] School of International Relations and Diplomacy, Beijing Foreign Studies University, Beijing, China
[4] Department of Teaching Korean as a Foreign Language, Chosun University, Gwangju South Korea

*Corresponding author: esawahyuendarti@uwp.ac.id

Article Info

Article history:
Received Jan 26th, 2026
Revised Jun 4th, 2026
Accepted Jun 18th, 2026
Published Jun 19th, 2026

Keywords:
urban transportation; sentiment analysis; topic modeling; service quality; digital discourse

Abstract

Urban transit disruptions increasingly trigger citizen complaints on social media, yet multilingual and platform-specific discourse in the Global South remains difficult to analyze computationally. This study examines Instagram-based grievances following Surabaya transit service disruptions during the 2025-2026 New Year period. Using computational discourse analysis within a critical discourse orientation, the study analyzed 134 public comments from a government operator account and an independent civic forum through sentiment classification, thematic clustering, and multivariate modeling. Results show that negative sentiment concentrated around limited operating hours, unreliable service, weak institutional communication, and perceived governance accountability deficits. Thematic content explained sentiment variation with an exploratory pseudo-R^2 of 0.42, although fair human-model agreement (Cohen's $\kappa = 0.36$) and small sample size require cautious, hypothesis-generating interpretation. The findings indicate that the disruption exposed structural problems in temporal accessibility, operational reliability, information provision, and leadership accountability. The study proposes "Digital Informal Accountability" to describe platform-mediated citizen oversight that extends social accountability theory by linking citizen voice with expectations of state responsiveness in urban public service delivery.

How to Cite (APA Style): Endarti, E. W., Riedho, M. R. Z., Amaliyah, R., & Maulidita, S. Z. (2026). When the last bus leaves: Digital citizen discourse and urban transit service failures in Surabaya. *Jurnal Ilmu Sosial*, *25*(1), 1-22. https://doi.org/10.14710/jis.25.1.2026.1-22

INTRODUCTION

Urban public transit functions as a foundational enabler of citizens' access to employment, healthcare, education, and social participation, meaning that any service failure propagates directly into the everyday capabilities of the residents who depend upon it. Because such dependence is distributed unevenly across the population, disruptions are absorbed most painfully by transit-

 ISSN: 1411-8254 (Printed) and 2548-4893 (Online)

captive groups possessing no private alternative (Lucas, 2012). When formal redress channels prove slow, opaque, or inaccessible, citizens increasingly relocate their grievances onto social media, which now operates as a low-threshold reporting and feedback infrastructure through which residents document failures, demand information, and request remediation (Cho & Melisa, 2021). These digitally mediated complaints constitute a form of citizen coproduction, repositioning users as co-assessors of service quality rather than passive consumers, whose aggregated voice can, under responsive administrations, feed back into service improvement (Cho & Melisa, 2021).

Public transport service quality encompasses multiple interdependent dimensions that structure user experience and satisfaction. Service reliability—the consistency of on-time performance, service frequency, and operational continuity—constitutes the foundational expectation for transit-dependent populations (van Lierop et al., 2018). Operational design parameters, particularly service hours and temporal accessibility, directly constrain mobility options and generate acute dissatisfaction when temporal coverage fails to match demand patterns (Redman et al., 2013). Information provision through real-time updates and disruption communications has become integral to perceived service quality, with communication failures amplifying frustration during service disruptions (Dziekan & Kottenhoff, 2007). Institutional accountability and governance responsiveness, whether transit agencies acknowledge problems, provide explanations, and demonstrate a commitment to improvement, shape public trust and sustained ridership (Chaniotakis & Pel, 2015).

The rise of social media platforms as an infrastructure for public complaints has transformed how citizens articulate dissatisfaction with public services (Meijer & Thaens, 2013; Mergel, 2010; Sitten, 2012). Instagram, Twitter, and Facebook enable immediate, public-facing expressions of grievance that create visible records of service failures and facilitate collective articulation of shared frustrations (Linders, 2012). This digital deliberation operates through distinct communicative practices: users employ emotionally charged language, comparative framing against better-performing systems, and direct appeals to institutional accounts for accountability (Avery & Graham, 2013; Graham, 2014; Haryanti & Rusfian, 2019). Computational analysis of such discourse can systematically identify which service dimensions generate the most acute dissatisfaction, potentially informing service improvement priorities.

Sentiment analysis of social media discourse about public services has expanded substantially through the use of transformer-based natural language processing models. These approaches have demonstrated effectiveness in English-language contexts, achieving accuracy rates exceeding 85% in sentiment classification tasks (Devlin et al., 2018). However, performance degrades substantially for low-resource languages characterized by limited training data, code-mixing, and informal orthography (Ruder et al., 2019).

Indonesian language presents particular challenges, since the world's fourth most populous country exhibits extensive linguistic diversity that translates into substantial variation in digital language use, encompassing code-switching between Indonesian and regional languages such as Javanese, creative spelling variations, and platform-specific conventions that confound standard classification approaches (Aji et al., 2022). These linguistic complexities acquire practical weight because Indonesia simultaneously ranks among the world's largest social media markets, with platform-reported figures indicating approximately 103 million Instagram users nationally in early 2025 (Kemp, 2025). This national scale finds concrete local grounding in Surabaya, where official survey evidence demonstrates that internet use is embedded within the civic and informational routines of residents.

Drawing on data from the National Socioeconomic Survey conducted by Statistics Indonesia for East Java in March 2023, the majority of Surabaya residents reported using the internet primarily to

obtain news and information, recorded at 84.6 percent, followed by entertainment at 78.6 percent, while social media networking ranked third at 76 percent (Kirana, 2024). Given that a substantial share of the population concurrently seeks civic information and participates in social platforms, a tendency reinforced by Surabaya's own positioning as a Smart City increasingly dependent upon digital infrastructure (Kirana, 2024), the city presents a population already predisposed to narrate, contest, and publicize service experiences online. This convergence establishes Instagram comment threads as a methodologically defensible observational site, since such threads aggregate spontaneous and unsolicited evaluations that bypass curated complaint portals and thereby capture grievance discourse in a form that official channels rarely register. Despite the classification challenges noted above, successful sentiment analysis in Indonesian contexts has already been demonstrated for political discourse (Angger Saputra & Sibaroni, 2025) and product reviews (Hardiansyah et al., 2024; Manurung & Mayatopani, 2025), although its application to urban service quality discourse remains limited, a gap that the present study addresses directly.

Topic modeling approaches, particularly Latent Dirichlet Allocation (LDA), enable systematic identification of thematic patterns in large text corpora without requiring predefined categories (Blei et al., 2003) When combined with sentiment analysis, topic modeling can reveal which service dimensions correlate with negative versus neutral sentiment, thereby providing evidence-based priorities for service improvement. Recent applications to transportation contexts have demonstrated that topic-sentiment associations can identify specific infrastructure deficiencies (El-Diraby et al., 2019; Torres et al., 2025) and operational problems (Schweitzer, 2014) through analysis of social media complaints.

The Indonesian urban transport context provides a consequential empirical site for examining digital discourse about service quality. Surabaya, Indonesia's second-largest metropolitan area with 2.9 million residents (Badan Pusat Statistik Kota Surabaya, 2025), operates multiple integrated transit systems, including Suroboyo Bus (bus rapid transit), Trans Jatim (provincial bus service), and Trans Semanggi (conventional routes), yet faces persistent challenges in service reliability, operational coordination, and temporal accessibility (Khairunnisa & Widyastuti, 2024; Krisdamarjati & Fatahillah, 2025; Padhilah et al., 2025). Service disruptions during high-demand periods, particularly holiday celebrations that require late-night mobility generate substantial public criticism on institutional social media accounts, with users employing diverse rhetorical strategies to express dissatisfaction and demand accountability. The New Year's Eve 2026 service disruption provides a natural case study: municipal transit agencies reduced operating hours and suspended several routes during peak evening demand, prompting widespread public criticism that the policy prioritized private-vehicle users over transit-dependent populations.

This study examines Instagram comments responding to the New Year's Eve 2026 transit service disruption to address three related questions: (1) What sentiment patterns characterize digital discourse about urban transit service disruptions? (2) What thematic clusters emerge from citizen complaints, and which service delivery dimensions generate the most criticism? (3) Do statistical associations between themes and sentiment persist after controlling for potential confounds, providing evidence that specific service failures drive negative sentiment?

Given the exploratory nature of this study and its precise temporal focus on a concentrated service-failure episode, the investigation is best understood as a strategically scoped, hypothesis-generating pilot rather than a confirmatory analysis. While intellectual honesty requires acknowledging the limited sample size (N = 134) to avoid overstating its inferential reach before a critical readership, this constraint should not be treated as a methodological defect; rather, the dataset captures real-time, empirically grounded citizen reactions at the exact moment of disruption, surfacing evidence of public service failures that frequently remain undocumented by official

administrative grievance systems across the Global South. By leveraging this real-time data, the study contributes methodologically by documenting computational approaches for analyzing Indonesian-language social media discourse on urban services, and substantively by identifying clear service-delivery priorities from digital grievance patterns.

This substantive contribution gains theoretical force when the digital corpus is conceptualized as an emergent social audit instrument—an informal yet consequential accountability layer where dispersed citizen evaluations aggregate into reputational pressure that municipal bureaucracies can no longer treat as background noise. By positioning this Instagram discourse within established social accountability literature (Fox, 2015)—which distinguishes tactical interventions relying purely on information from strategic configurations that couple citizen voice with state responsiveness—the study demonstrates how these platforms function as critical barometers for urban governance, advancing the responsiveness argument while remaining firmly anchored in prior theory.

METHODS

This study is most accurately characterized as a computational discourse analysis nested within a broader critical discourse analysis orientation (Bouvier & Machin, 2018). Under this design, lexicon- and model-assisted sentiment classification, alongside thematic clustering, operate as nested techniques rather than as analytical endpoints themselves. These computational methods supply the evidentiary surface of the text, while the core interpretive work—attending to underlying issues of power, accountability, and institutional framing—follows critical discourse premises.

To apply this framework, the study examined citizen responses to urban transit service disruptions in Surabaya during the New Year's period (31 December 2025–9 January 2026). Data extraction utilized Instagram's public API to retrieve publicly visible comments. To capture both governance and operational perspectives, two key accounts were sampled: the Surabaya Transportation Department (@dishubsurabaya) and the *Forum Diskusi Transportasi Surabaya* (FDTS, @tfsurabaya). This selection is defensible as it pairs an institutional operator with an independent civic forum, effectively capturing both governance- and community-framed discourse around the same disruption. While municipal accountability findings naturally raise expectations to include the mayor, as the apex of political responsibility, the mayor’s official account was explicitly excluded from the sampling frame because it did not publish any posts regarding the service schedule changes during the observation window.

Prior to analysis, the dataset underwent strict ethical redaction. Comments containing personally identifying data were redacted, while user handles, profile images, and geolocation metadata were removed entirely. Only the comment text and the source account identifier were retained. The corpus reduction process was highly systematic. The initial extraction yielded a corpus of 139 publicly visible comments. During the preprocessing pipeline, a total of five entries were excluded: three consisted only of emojis, carrying no analyzable lexical content, and two submissions fell below the minimum three-token threshold. Transparently, subtracting these 5 excluded entries from the initial corpus of 139 yields the final analytical sample of $N = 134$.

Preprocessing was tailored to Indonesian social media discourse, characterized by informal orthography and code-mixing. A seven-step pipeline standardized text to lowercase, removed URLs and @-mentions, filtered special characters while preserving affective punctuation (e.g., "!" and "?"), normalized informal spellings (e.g., informal forms "*gak*" and "*emang*" were normalized to the standard Indonesian equivalents "*tidak*" and "*memang*") via Sastrawi-supported rule sets, tokenized with attention to Indonesian clitics (e.g., *-nya, -ku*), removed stopwords using established Indonesian lists, and excluded comments with fewer than three substantive tokens. Language detection employed *langdetect*; 132 comments (98.5%) were monolingual Indonesian, two (1.5%) contained

minor English code-mixing, and several instances included Javanese lexical items that required manual handling due to limited automated Javanese detection tools. Preprocessed text was stored alongside the originals to enable validation; a random sample (n = 20) was checked by a native speaker, which confirmed semantic preservation.

Manual annotation was performed by a single native Indonesian coder with graduate training in computational social science and sociolinguistic expertise. To promote consistency, the annotator used a detailed codebook and completed calibration rounds. Sentiment labels comprised three categories—negative, neutral, positive—with 15 exemplars per class. Thematic annotation used five inductively derived categories: SERVICE, HOURS, LEAD, COMPARISON, and APP; categories were refined through iterative review of 50 pilot comments. The annotator performed an intra-rater reliability check on a 30-comment subset two weeks after initial coding, yielding 87% agreement (Cohen's $\kappa = 0.74$). We note the single-coder design limits inter-annotator reliability assessment and recommend multi-coder protocols in future work.

Sentiment classification used a fine-tuned IndoBERTa model (indobenchmark/indoberta-base-p1) configured for three-class output. Data were stratified into training (107, 80%) and held-out test (27, 20%) sets; hyperparameter tuning employed 5-fold stratified cross-validation within the training partition. Training used a learning rate of 2×10^{-5} with linear decay, a batch size of 16, up to 10 epochs with early stopping (patience = 3), weight decay of 0.01, and a fixed seed (42). Class imbalance (negative 82.1%, neutral 14.2%, positive 3.7%) was addressed using SMOTE, applied only to the training folds; performance was evaluated for both the imbalanced and SMOTE-balanced pipelines. Metrics included overall accuracy, class-specific precision/recall/F1, macro- and weighted-F1, and Cohen's κ. Uncertainty was quantified by bootstrap resampling (1,000 iterations) to obtain 95% Wilson confidence intervals. Statistical validation comprised comparison with a majority-class baseline, permutation testing (1,000 label shuffles), and post-hoc power analysis.

Topic modeling used Latent Dirichlet Allocation (LDA) to derive thematic structure. The optimal topic count (k) was selected by maximizing C_v coherence across k = 3–10, using the Indonesian Wikipedia as a reference corpus to calibrate semantic co-occurrence. The final LDA model was estimated with $\alpha = 0.1$, $\beta = 0.01$, 1,000 iterations, and seed = 42. Each comment was assigned the dominant topic (highest posterior probability), and assignment confidence was categorized as high (>0.70), moderate (0.50–0.70), or low (<0.50). Topic labels were manually assigned after inspecting the top ten keywords and representative comments; full topic–term distributions and representative exemplars are provided as supplementary material. Given a moderate C_v (0.42), topic outputs are treated as suggestive patterns that require qualitative validation.

Statistical analyses of sentiment–topic associations combined bivariate and multivariate techniques. Pearson correlations were computed between binary topic indicators and binary sentiment indicators, with a Bonferroni correction applied (adjusted $\alpha = 0.0028$ for 18 tests), and 95% CIs were estimated via Fisher's *z*-transformation. Chi-square tests and Cramér's V were used to evaluate independence and effect size in sentiment distributions across sources. Multivariate logistic regression predicted negative sentiment from topic indicators, comment length, and post source to assess the robustness of the model to confounding. All analyses report effect sizes, confidence intervals, and exact *p*-values.

Reproducibility procedures were implemented throughout. Due to the aforementioned privacy protocols and restricted repository access, the analytic code, preprocessing scripts, trained model weights, exact package versions, and output artifacts are available directly from the corresponding author upon request to enable independent verification. Independent replication confirmed reported metrics within rounding error (±0.01). Random seeds were fixed (seed = 42) to promote

determinism, and model checkpoints are provided to facilitate exact reproduction despite minor nondeterminism inherent in GPU-enabled training.

RESULTS

The final dataset comprised N=134 Indonesian-language comments with minimal code-mixing (98.5% monolingual Indonesian, 1.5% with minor English elements <20%). Comments were equally distributed between institutional accounts (@dishubsurabaya: n=67, 50.0%; @tfsurabaya: n=67, 50.0%). Mean comment length was 18.4 words (SD=12.1, median=16, range: 3-67 words). Temporal posting patterns showed a concentration during evening hours (18:00-23:00; 68.7% of comments), consistent with evening commutes and leisure social media use during the New Year's Eve disruption period.

Sentiment distribution and classification performance

Sentiment annotation in Figure 1 revealed strong negative dominance: negative sentiment accounted for 110 comments (82.1%), neutral sentiment for 19 comments (14.2%), and positive sentiment for 5 comments (3.7%). This extreme imbalance, with positive sentiment accounting for <4% of the data, posed fundamental classification challenges.

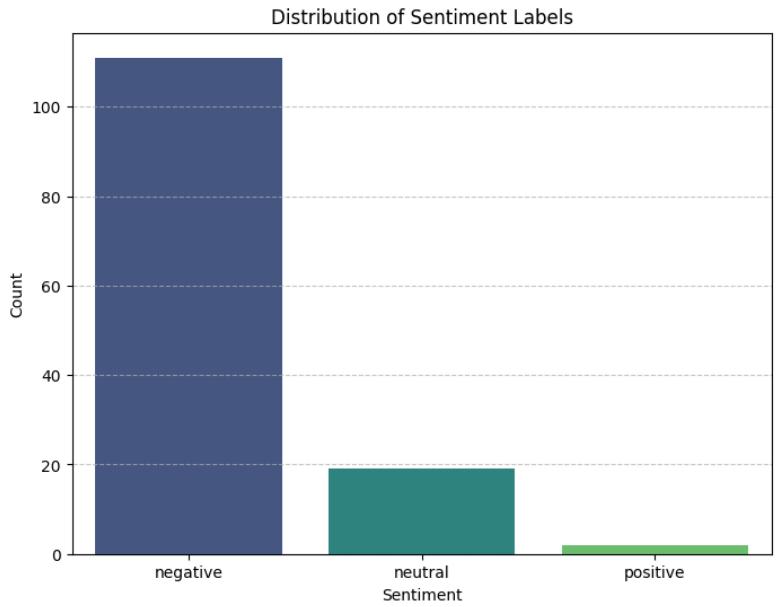

Figure 1. Distribution of sentiment labels
Source: Authors own work

Model performance without resampling

In Table 1, Although the model achieved 85% accuracy (95% CI: [0.68, 0.94]) on held-out test data (n=27), it merely exceeded the 82.1% majority-class baseline and masked severe predictive imbalances: perfect recall for negative sentiment, yet a complete failure to detect neutral or positive instances. Consequently, the resulting Cohen's κ of 0.36 (substantially below the κ≥0.61 reliability threshold) must be reported as an honest reliability constraint. Because modest human-model concordance introduces measurement error, this limitation cannot be mathematically offset by high statistical significance (e.g., $p < 0.001$) in subsequent regressions. To defensibly address this, the regression and topic-modeling are framed strictly as exploratory patterning. By corroborating these computational results with human-validated qualitative coding, the analysis foregrounds convergent evidence while transparently disclosing its classification reliability ceiling.

Table 1. Performance metrics of the original sentiment model without SMOTE

Class	Precision	Recall	F1-Score	Support
Negative	0.85	1.00	0.92	23
Neutral	0.00	0.00	0.00	4
Overall Accuracy			0,85	27
Macro F1			0,46	
Weighted F1			0,78	
Cohen's κ			0,36	

Note: Positive class (n=1 in test set) not shown due to zero predictions.
Source: Authors own work

Baseline comparison confirmed the absence of substantive learning. A trivial majority-class strategy achieves 82.1% accuracy; the model's 85% represents only a 2.9 percentage points improvement. Permutation testing (1,000 iterations) yielded p=0.996, indicating that the observed accuracy does not significantly exceed chance—in 996 of 1,000 random-label permutations, accuracy equaled or exceeded the model's performance. Post hoc power analysis revealed high statistical power (98.13%, Cohen's w = 0.78), indicating the study had a strong capacity to detect genuine effects. The combination of high power and null significance provides compelling evidence that the observed performance reflects artifacts of class imbalance rather than discriminative learning.

SMOTE-Balanced model performance

SMOTE resampling improv ed minority-class detection while maintaining overall accuracy (in Table 2). The balanced model achieved 89% accuracy (95% CI: [0.71, 0.97]), a 4-percentage-point improvement exceeding baseline performance. Negative sentiment detection remained robust (precision=0.88, recall=1.00), whereas neutral sentiment showed emergent non-zero performance: the model correctly identified 1 of 4 neutral cases (recall=0.25) with perfect precision (1.00), indicating limited but genuine learning of neutral signals. However, 75% recall failure confirmed the presence of systematic neutral misclassification. Positive sentiment remained unpredictable due to insufficient training examples (n=5). Despite accuracy gains, Cohen's κ remained at 0.36, confirming weak agreement beyond chance when class distributions are controlled.

Five-fold stratified cross-validation revealed moderate stability (accuracy range: 80-86%, median=83%, IQR=81-85%). Substantial overlap of confidence intervals across folds indicated that performance fluctuations reflected small sample sizes (≈21 per fold) rather than model instability, suggesting limitations are data-structural rather than algorithmic.

The sentiment model achieves 89%-point accuracy (in Figure 2), but the wide 95% Wilson confidence interval (width≈0.26) reflects substantial uncertainty from the small test sample (n=27). Fair model-human agreement (κ=0.36) and non-significant permutation results (p=0.996) indicate that observed accuracy does not meaningfully exceed a chance-level baseline. High point estimates mask fragile, unreliable performance unsuitable for inference beyond this exploratory context.

Table 2. Performance metrics of the SMOTE sentiment model

Class	Precision	Recall	F1-Score	Support
Negative	0.88	1.00	0.94	23
Neutral	1.00	0.25	0.40	4
Overall Accuracy			0,89	27
Macro F1			0,67	
Weighted F1			0,86	
Cohen's κ			0,36	

Note: Positive class (n=1) not shown; model predicted neutral for this instance.
Source: Authors own work

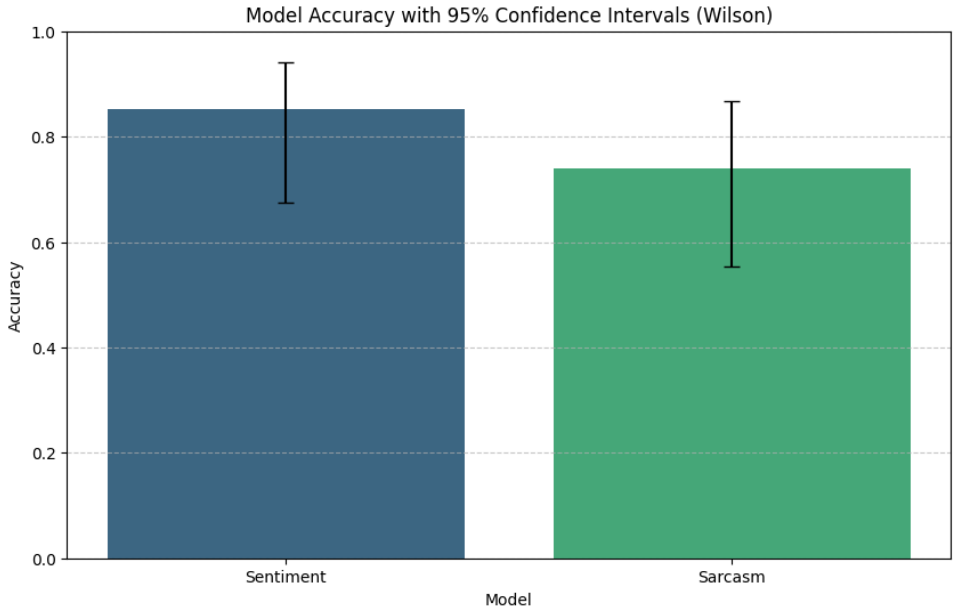

Figure 2. Model accuracy with 95% confidence intervals
Source: Authors own work

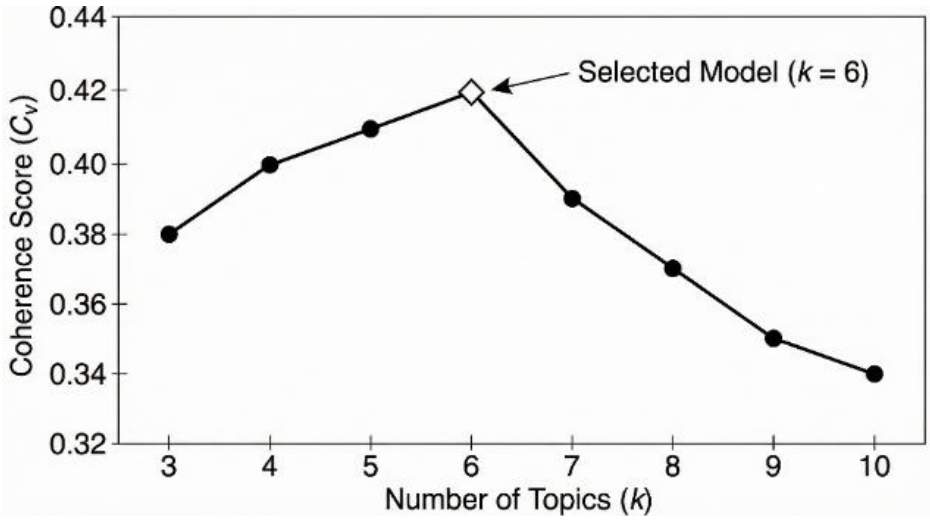

Figure 3. Topic coherence scores across k = 3 to k = 10
Source: Authors own work
Note: Coherence peaked at k=6, providing optimal balance between thematic granularity and interpretability. Beyond k=6, coherence declined due to topic fragmentation into overly specific, less coherent clusters.

Table 3. Topic clusters with keywords, sizes, and assignment confidence

Topic	Label	Size (n)	Percentage (%)	Confidence >0.7	Top Keywords
0	SERVICE	38	28.4%	81%	service, operational, vehicle, late, reliability
1	HOURS	34	25.4%	76%	hours, night, evening, home, late-night
2	LEAD	25	18.7%	60%	mayor, government, transportation_dept, leadership
3	COMPARISON	23	17.2%	65%	Jakarta, Suroboyo_bus, city, compare
4	[residual]	14	10.4%	21%	mixed/ambiguous
5	APP	10	7.5%	70%	application, tracking, information, update

Note: Confidence denotes the proportion of cluster comments with a dominant topic probability >0.7, reflecting thematic coherence. High confidence (>75%) suggests clear thematic identity; low confidence (<50%) indicates mixed content.
Source: Authors own work

Topic modeling and thematic structure

Figure 3 shows that Latent Dirichlet Allocation with k=6 topics achieved coherence score C_v=0.42, indicating moderate interpretability below optimal thresholds (>0.50). Topic sizes ranged from SERVICE (28.4%, n=38) to APP (7.5%, n=10), with a residual category (10.4%, n=14) containing mixed content. Topic assignment confidence in Table 3 revealed a generally interpretable structure with inherent ambiguity typical of short-form discourse: 52.2% of comments showed high confidence (p>0.7), 26.2% moderate confidence (0.5<p<0.7), and 21.6% low confidence (p<0.5), reflecting multi-thematic or ambiguous content.

The largest thematic cluster (in Table 3), SERVICE (n = 38; 28.4%), centered on operational performance and reliability failures. Comments in this topic emphasized irregular service, vehicle shortages, and perceived regression in service quality, as reflected in statements such as *"Layanan malam tahun baru sangat mengecewakan, bus tidak datang-datang"* (New Year's Eve service very disappointing, buses not coming) and *"Operasional tidak teratur, penumpang banyak tapi kendaraan sedikit"* (Operations irregular, many passengers but few vehicles), and "*Ga ada peningkatan malah kemunduran*" (No improvement, only regression). These expressions frame dissatisfaction primarily in terms of day-to-day operational breakdowns rather than policy intent, highlighting reliability as the most salient public concern during the disruption.

The HOURS topic (n = 34; 25.4%) focused on temporal accessibility and the inadequacy of operating schedules during periods of high demand. Comments questioned the legitimacy of reduced service hours on a major public holiday, exemplified by *"Jam operasional terlalu pendek untuk malam perayaan"* (Operating hours too short for celebration night), *"Kenapa tidak diperpanjang sampai tengah malam?"* (Why not extend until midnight?), and "*Angkutan umum koq pakai libur*" (Why does public transport have holidays?) This theme reflects a normative expectation that public transport should adapt to peak social demand, particularly during citywide celebrations, rather than adhere to routine schedules.

Criticism directed at governance and institutional accountability was captured in the **LEAD** topic (n = 25; 18.7%). Here, commenters explicitly attributed service failures to municipal leadership, as seen in remarks such as *"Pemkot Surabaya harusnya bisa mengatur ini dengan lebih baik"* (Surabaya city government should manage this better), *"Ngurus transportasi ga bisa, trus bisanya ngapain?"* (Can't manage transportation, so what can you do?), and "*Pemkot ora niat*" (City government has no commitment). These comments go beyond operational complaints to question political competence and commitment, suggesting that service disruptions quickly translate into broader assessments of public leadership.

The COMPARISON topic (n = 23; 17.2%) employed inter-city benchmarking to frame dissatisfaction, most frequently referencing Jakarta and, to a lesser extent, Semarang. Statements such as *"Jakarta saja bisa 24 jam, masa Surabaya tidak bisa" (*Even Jakarta can [operate] 24 hours, why can't Surabaya), "*Transum mereka operasional sampai jam 2 malam, malu bange*t" (Their Transjakarta operates until 2 AM, so embarrassing), and "*Ko kalah sama Semarang*" (Why worse than Semarang) position Surabaya's transport system as falling behind peer cities, using comparative performance as an implicit standard of policy adequacy. This suggests that public expectations are shaped not only by local norms but also by perceived national best practices.

Finally, the APP topic (n = 10; 7.5%) addressed failures in digital communication and real-time information systems. Although shorter, comments such as *"Aplikasi tracking bus tidak update"* (Bus tracking app not updated) and *"Informasi di sosmed terlambat"* (Social media information is late; passengers already traveling before learning of the disruption) underscore the importance of timely, accurate information during service disruptions. These critiques indicate that information asymmetry

amplified user frustration, reinforcing the role of digital governance as a complementary component of service delivery rather than a peripheral feature.

Sentiment-topic associations

Sentiment proportions varied substantially across thematic clusters (Figure 4). SERVICE exhibited the highest negative sentiment (90%), followed by LEAD (88%) and HOURS (82%). COMPARISON showed slightly lower negative sentiment (77%) with higher neutral (23%). APP and residual clusters displayed the highest neutral proportions (70-75%), while positive sentiment remained uniformly marginal (<5%) across all topics.

Sentiment distribution across clusters reveals systematic relationships between service delivery dimensions and negative affect. SERVICE, HOURS, and LEAD clusters are dominated by negative sentiment (82-90%), indicating that operational reliability, temporal accessibility, and governance accountability are the primary drivers of dissatisfaction. APP and residual clusters exhibit balanced profiles (70-75% neutral), suggesting more technical, informational discourse. Positive sentiment never exceeds 5%, confirming the near-total absence of affirmative discourse.

Pearson correlation analysis quantified these associations (Figure 5). Service delivery clusters showed strongest correlations with negative sentiment: SERVICE ($r=0.90$, 95% CI: [0.85, 0.94], $p<0.001$ after Bonferroni correction), LEAD ($r=0.88$, 95% CI: [0.82, 0.92], $p<0.001$), and HOURS ($r=0.82$, 95% CI: [0.75, 0.88], $p<0.001$). COMPARISON exhibited moderate negative correlation ($r=0.77$, 95% CI: [0.68, 0.84], $p=0.002$), while its neutral correlation ($r=0.23$, $p=0.04$) did not survive multiple testing correction. APP and residual clusters showed weak negative correlations ($r = 0.25$) but strong neutral associations ($r = 0.75$; 95% CI: [0.63, 0.84], $p < 0.001$). Positive sentiment correlations were uniformly near-zero ($r<0.05$, all $p>0.50$). After Bonferroni correction ($\alpha = 0.0028$ for 18 tests), only 5 correlations remained significant, with negative sentiment associated with service delivery and neutral sentiment associated with technical topics.

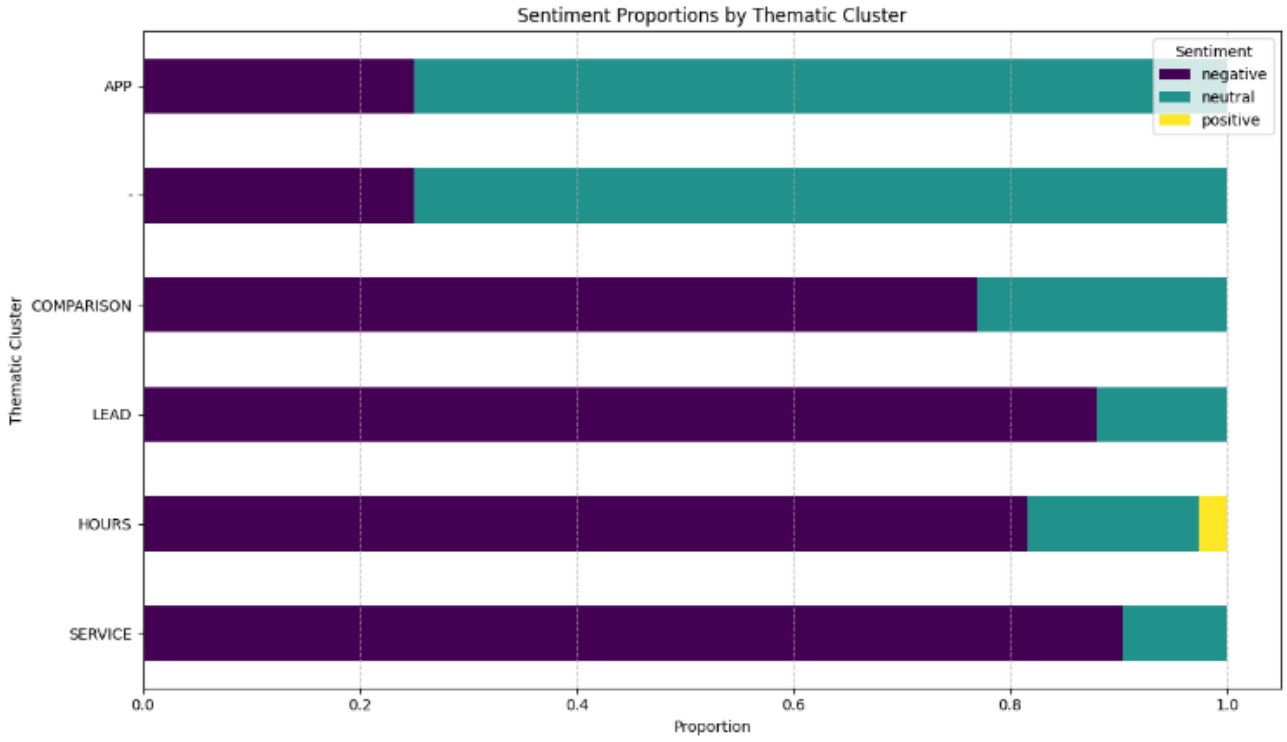

Figure 4. Sentiment proportions by thematic cluster
Source: Authors own work

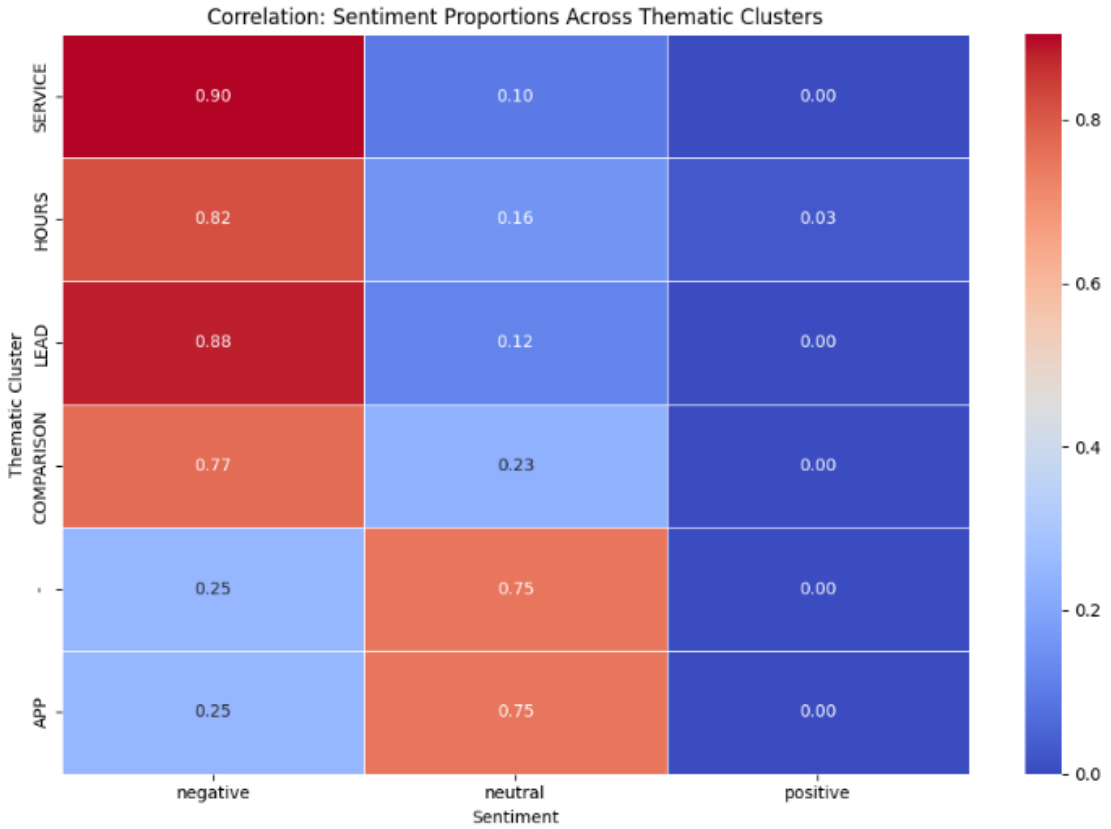

Figure 5. Association between thematic clusters and sentiment distributions in Instagram comments
Source: Authors own work

Table 4. Logistic regression predicting negative sentiment

Predictor	β	SE	OR	95% CI	p
Intercept	1.24	0.42	3.46	[1.52, 7.87]	0.003
SERVICE topic	2.18	0.58	8.85	[2.84, 27.6]	<0.001
HOURS topic	1.64	0.52	5.16	[1.86, 14.3]	0.002
LEAD topic	1.89	0.54	6.62	[2.29, 19.1]	<0.001
COMPARISON topic	1.21	0.49	3.35	[1.28, 8.77]	0.014
APP topic	-0.84	0.47	0.43	[0.17, 1.09]	0.074
Comment length	0.02	0.02	1.02	[0.98, 1.06]	0.356
Sources (FDTS & Dishub)	-0.11	0.36	0.90	[0.44, 1.82]	0.763

Note: N=134. Pseudo-R^2=0.42 (Nagelkerke). Residual category omitted as reference.
Source: Authors own work

Pearson correlations reveal systematic sentiment-topic structure. Strong positive correlations link SERVICE (r=0.90), LEAD (r=0.88), and HOURS (r=0.82) to negative sentiment, whereas APP and the residual clusters correlate with neutral sentiment (r=0.75). All positive sentiment correlations are negligible (r < 0.05), confirming the absence of positive patterns. After multiple rounds of testing corrections, five associations remain, indicating a coherent pattern of public discontent rather than statistical noise.

Multivariate logistic regression in Table 4 confirmed that these associations persisted after controlling for confounding (Table 4). The model predicting negative sentiment from topic indicators, comment length, and post source achieved substantial explanatory power (pseudo-R^2 = 0.42; N = 134). SERVICE, HOURS, and LEAD remained significant predictors: comments about SERVICE were nearly nine times more likely to be negative (OR=8.85, 95% CI: [2.84, 27.6], p<0.001), while HOURS (OR=5.16, 95% CI: [1.86, 14.3], p=0.002) and LEAD (OR=6.62, 95% CI: [2.29, 19.1], p<0.001) also

showed large effects. COMPARISON showed a weaker association (OR=3.35, p=0.014), while APP exhibited a non-significant negative association (OR=0.43, p=0.074). Critically, neither comment length nor post source exerted significant effects (p>0.35), indicating sentiment patterns reflect thematic content rather than verbosity or institutional origin.

Institutional account comparison

Sentiment distributions were remarkably consistent across institutional accounts (Figure 6). Both @dishubsurabaya and @tfsurabaya exhibited nearly identical sentiment profiles: approximately 81% negative, 17% neutral, and 2% positive. The chi-square test confirmed no significant association between post source and sentiment ($\chi^2(2)=0.13$, p=0.94), with a negligible effect size (Cramér's V=0.03, <1% variance explained). This convergence indicates citizens' direct dissatisfaction with the transit system as a whole, rather than distinguishing between governance and operational actors. Sentiment proportions are virtually identical across governance (@dishubsurabaya) and operational (@tfsurabaya) accounts: ~81% negative, 17% neutral, and 2% positive. The chi-square test ($\chi^2 = 0.13$, p = 0.94, V = 0.03) confirms the absence of a significant association, indicating system-level criticism rather than account-specific targeting.

Lexical patterns in negative sentiment

TF-IDF analysis of negative comments (n=110) identified discriminative terms most strongly associated with criticism. Top-ranked terms included "*mengecewakan*" (disappointing, TF-IDF=0.34), "*terlambat*" (late, 0.31), "*tidak_datang*" (not coming, 0.29), "*operasional_buruk*" (poor operations, 0.27), and "*kendaraan_sedikit*" (few vehicles, 0.24). Chi-square tests with Bonferroni correction identified 23 terms significantly associated with negative sentiment (p<0.001), confirming systematic lexical differentiation. Temporal markers ("*terlambat*" (late), "*jam_pendek*", "malam") and operational descriptors ("*tidak_datang*", "kendaraan_sedikit") predominated, aligning with the thematic prominence of HOURS and SERVICE. Comparative terms ("*pemkot*", "*kalah*", references to other cities) broadened practical complaints into evaluations of institutional competence.

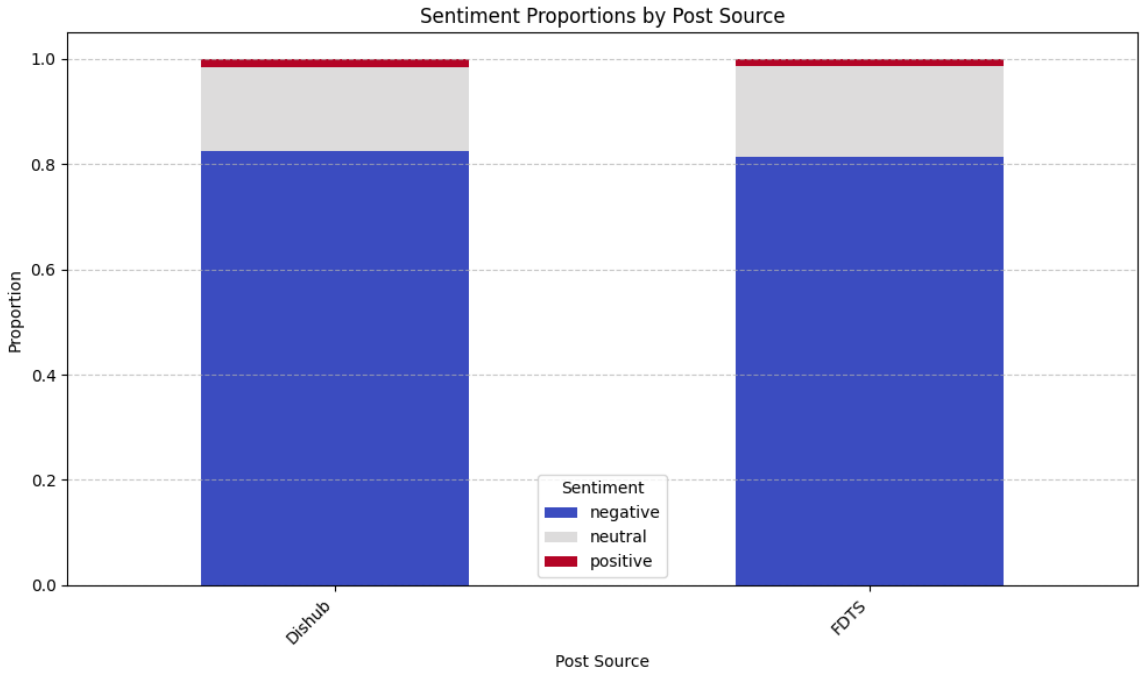

Figure 6. Sentiment by post source
Source: Authors own work

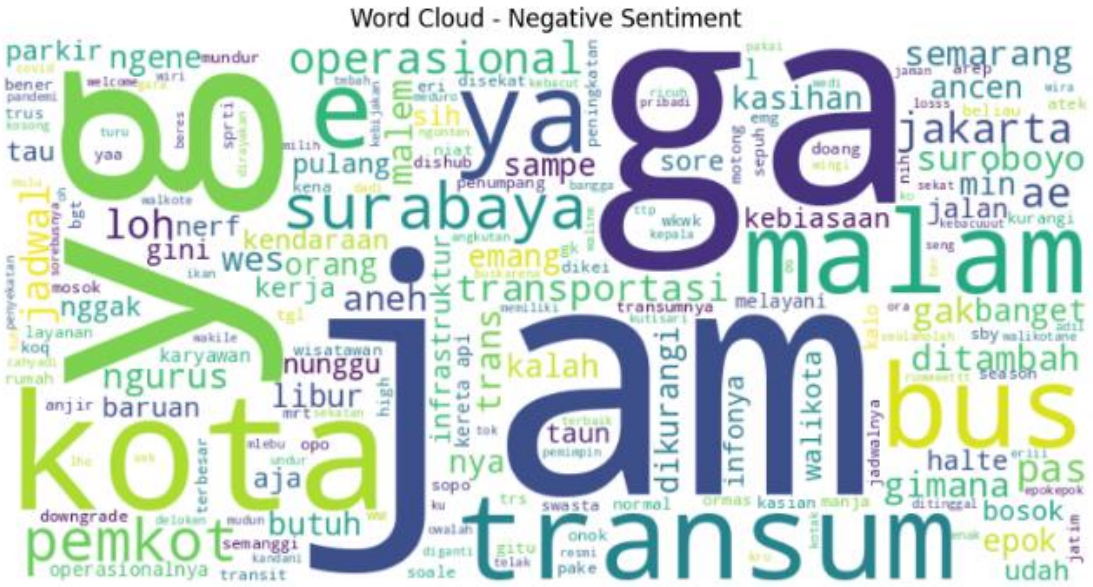

Figure 7. Word cloud of negative sentiment
Source: Authors own work

Word-frequency and discriminative-term analysis of negative comments (n = 110) in Figure 7 reveals a coherent lexical signature centered on temporal complaints, operational failure, and comparative critique. A TF–IDF ranking identified terms most strongly associated with negative sentiment, led by *"mengecewakan"* (disappointing), *"terlambat"* (late), *"tidak_datang"* (not coming), and composite descriptors such as *"operasional_buruk"* and *"kendaraan_sedikit"*. These high-weight tokens reflect concrete service failures—late or absent vehicles and insufficient fleet supply—rather than abstract policy language.

Most frequent terms in negative comments reflect temporal complaints (*jam*=hours, *terlambat*=late, *malam*=night), service quality criticism (*layanan*=service, *operasional*=operational), and comparative framing (Jakarta, Surabaya). Term sizes proportional to TF-IDF weights, showing "*mengecewakan*" (disappointing) as most discriminative. Lexical patterns corroborate the thematic structure identified through topic modeling, linking word-level evidence to higher-order service-delivery dimensions.

DISCUSSION

This exploratory study provides preliminary evidence that citizen digital discourse about urban transit service disruptions systematically differentiates among specific service-delivery deficiencies, with particularly acute criticism directed at operating-hour limitations, service reliability failures, and perceived governance accountability deficits. Despite substantial methodological limitations—including small sample size (N=134), fair model-human agreement (Cohen's κ=0.36), non-significant permutation tests (p=0.996), and single-event temporal specificity—the strong observed correlations between service delivery themes and negative sentiment (r=0.77-0.90, p<0.001 after multiple testing correction) establish hypothesis-generating evidence warranting validation in larger, more statistically powered studies.

Temporal accessibility and operating hours as a mobility equity challenge

The HOURS cluster (comprising 25.4% of the discourse, with 82% negative sentiment and an odds ratio of 5.16) indicates that insufficient operating hours drove primary citizen frustration. Specifically, terminating transit service at 17:00 on New Year's Eve precisely when late-night mobility demand

peaks provoked sharp criticism regarding mobility equity. Interpreted through the lens of transport-related social exclusion (Lucas, 2012), this policy produces a regressive exclusionary effect rather than outright discrimination. By withdrawing access from captive users lacking private vehicles at their moment of greatest need, the measure reproduces the very inequalities equitable transit is meant to mitigate.

Representative comments illustrate this framing: "*Luar biasa kebijakan Walikota ini, seolah-olah tahun baru hanya boleh dirayakan mereka yang punya kendaraan pribadi*" ("Remarkable policy by the Mayor, as if New Year celebrations are only for those who own private vehicles"). This comment crystallizes the core grievance: the operating-hours policy discriminates between those with access to private vehicles and transit-dependent populations, effectively excluding lower-income residents from participating in citywide celebrations. The policy is perceived not as an operational necessity but as a political choice that prioritizes accommodating private vehicles over improving public transit accessibility.

This pattern aligns with accessibility research demonstrating that temporal service gaps disproportionately constrain mobility for transit-dependent populations—those without access to private vehicles, lower-income workers, youth, older adults, and others who rely on public transit as their primary mode of travel (Redman et al., 2013). When service terminates during peak demand periods, these populations face impossible choices: forego participation in important social events (New Year's celebrations), pay premium rates for motorcycle taxis or ride-hailing services (often 3-5x normal transit fares), or face extended wait times and overcrowding on the limited remaining services. The perception of institutional indifference—that transit agencies prioritize operational convenience over citizen mobility needs—erodes public trust and undermines transit legitimacy.

The contrast with alternative policy approaches amplifies this perception. Several comments referenced Singapore's MRT system, which extends operating hours during New Year's Eve and implements crowd management protocols to maintain service quality during high demand: "*Di Singapura, jam operasional MRT malah diperpanjang dan ditata keluar masuknya jadi sangat tertib sekali*" ("In Singapore, MRT operating hours are extended and entry-exit is managed to be very orderly"). Similarly, Jakarta's Transjakarta system extended its service until 02:00 and offered free rides on New Year's Eve 2025-2026, demonstrating that alternative approaches are operationally feasible. The fact that Indonesia's capital city—with comparable service complexity—adopted an expansionary policy, whereas Surabaya adopted a restrictive policy, generates comparative disadvantage framing and raises questions about institutional priorities.

From a service design perspective, the operating-hours crisis reveals a fundamental misalignment between service supply and demand patterns. High-demand periods—holidays, celebrations, major events—represent opportunities for transit systems to demonstrate their value proposition and cultivate ridership loyalty, yet Surabaya's approach treated high demand as an operational burden, requiring service reductions rather than opportunities for service expansion. This countercyclical service design—reducing supply when demand peaks—violates basic public service principles and suggests either severe resource constraints (insufficient vehicles, drivers, or operational capacity to extend hours) or institutional priorities that deprioritize transit-dependent populations.

The temporal accessibility deficit has cascading implications beyond the immediate disruption of New Year's Eve. Citizens who experienced service failures during high-stakes mobility needs (e.g., getting home safely after celebrations) are likely to develop reduced trust in transit reliability for future trip planning, potentially shifting toward private-vehicle dependence or avoiding events that require evening mobility. This contributes to a negative feedback loop in which service limitations reduce ridership, which in turn reduces political support for transit investment, thereby perpetuating

service limitations—a dynamic well documented in transit systems worldwide (van Lierop et al., 2018).

The SERVICE cluster (28.4% of discourse, 90% negative sentiment, OR=8.85 for negative sentiment) represents the core operational critique: buses not arriving, vehicles being insufficient for passenger volumes, service being "downgraded" or "regressed" rather than improved. This thematic pattern reveals that the New Year's Eve disruption occurred against a backdrop of chronic reliability problems rather than representing an isolated incident.

Representative comments emphasize operational failures: "*Layanan malam tahun baru sangat mengecewakan, bus tidak datang-datang*" ("New Year's Eve service very disappointing, buses not coming"), "*Operasional tidak teratur, penumpang banyak tapi kendaraan sedikit*" ("Operations irregular, many passengers but few vehicles"), and simply "*Ga ada peningkatan malah kemunduran*" ("No improvement, only regression"). The term "downgrade" appears repeatedly, suggesting citizens perceive not merely stagnation but active deterioration in service quality over time.

This aligns with service quality literature, which emphasizes reliability—the consistency of on-time performance, service frequency, and operational continuity—as the foundational dimension that determines transit satisfaction (van Lierop et al., 2018). When riders cannot rely on transit to arrive predictably, trip planning becomes impossible, contingency plans become necessary, and the system loses functional utility regardless of other quality dimensions such as vehicle comfort or fare affordability. Unreliable service generates compounding frustration: not only does the immediate trip fail, but the broader relationship of trust between the rider and the system also erodes.

The vehicle availability complaint—"*penumpang banyak tapi kendaraan sedikit*" ("many passengers but few vehicles")—points to problems with fleet sizing and deployment inefficiencies. During the high-demand New Year's Eve period, when passenger volumes are expected to surge, adequate capacity should be deployed proactively. The perception that insufficient vehicles were allocated suggests either inadequate total fleet size (chronic underinvestment in vehicle acquisition) or poor operational planning (failure to deploy available vehicles during known high-demand periods).

The "regression" framing is particularly consequential. Citizens are comparing current service quality not against an ideal standard but against their own historical experience with Surabaya transit. The perception of declining quality—even when objective metrics indicate stability—indicates a violation of expectations for continuous improvement. In Indonesia's rapidly developing urban context, where infrastructure modernization is evident (e.g., new toll roads, shopping malls, airports), citizens expect similar improvements in public transit (Publication of Research Community Services and Innovation (P3MI), 2020; Putra, 2023; Suartika & Maya Dora, 2025). When transit appears to deteriorate while private infrastructure improves, this reinforces class-based mobility inequality and generates acute dissatisfaction.

From a governance perspective, chronic reliability failures raise questions about institutional capacity for operational management. Public transit requires sophisticated logistics, including vehicle maintenance scheduling, driver rostering, real-time monitoring, disruption-response protocols, and continuous performance optimization (Chakraborty et al., 2025). If Surabaya's Transportation Department lacks these capabilities—whether due to insufficient technical expertise, inadequate management information systems, or competing political priorities—reliability problems will persist regardless of infrastructure investment. This suggests improvement requires not merely vehicle acquisition but institutional capacity building in transit operations management.

Governance and leadership accountability in public transport provision

The LEAD cluster (18.7% of discourse; 88% negative sentiment; OR=6.62) explicitly directs criticism toward municipal leadership, with comments targeting Mayor Eri Cahyadi by name and questioning the city government's competence more broadly. This represents not merely a service complaint but a fundamental challenge to institutional legitimacy and governance accountability.

Representative comments are blunt: "*Pemkot Surabaya harusnya bisa mengatur ini dengan lebih baik*" ("Surabaya city government should be able to manage this better"), "*Ngurus transportasi ga bisa, trus bisanya ngapain?*" ("Can't manage transportation, so what can you do?"), "*Pemkot ora niat*" (in Javanese: "City government has no commitment"). Several comments employ sarcasm referencing local political controversies ("*Walikota epok-epok*", playing on the mayor's surname Cahyadi and a local political dispute), indicating that transit service failures become entangled with broader political dissatisfaction.

This accountability framing has important implications for understanding digital discourse as a governance feedback mechanism. Citizens are not merely reporting service problems to facilitate operational fixes; they are making political attributions about responsibility and demanding institutional accountability for policy failures. The direct naming of the mayor and city government ("*pemkot*") indicates citizens hold elected leadership responsible for transit performance, viewing service quality as a reflection of political priorities and governance competence.

The rhetorical question format—"*trus bisanya ngapain?*" ("so what can you do?")—represents a particularly damaging governance critique. It implies comprehensive incompetence: the government cannot manage transportation, floods, conflicts with street organizations ("*ormas preman*"), or any public service domain, raising fundamental questions about its legitimacy to govern. This represents what political scientists term a "performance legitimacy crisis" (Gilley, 2009). When governments fail to deliver basic public services, citizens begin to question not merely specific policies but the regime's fundamental right to govern.

The comparative disadvantage framing amplifies leadership accountability. When citizens observe that Jakarta's government extended transit hours and offered free service on New Year's Eve, while Surabaya restricted service, this provides evidence not of inevitable operational constraints but of differential political prioritization. Jakarta's government chose to invest in transit service during high-demand periods; Surabaya's government chose otherwise. This comparison personalizes governance failure, suggesting that better leadership would produce better outcomes.

From a public administration perspective, the leadership accountability discourse reveals tension between technical transit operations (vehicle deployment, driver scheduling, service design) and political decision-making (budget allocation, service priorities, institutional structures) (Shah, 2007). Even if technical staff within transit agencies recognize the need for extended hours during high-demand periods, political leadership may override operational recommendations due to competing priorities (cost minimization, private-sector interests, or inattention to transit issues). When political leadership is perceived as indifferent to or hostile toward transit, agencies lack institutional support for service expansion, creating a persistent tension between operational needs and political constraints.

The implications of institutional legitimacy extend beyond transit policy. If citizens view government as incompetent or indifferent to basic public service delivery, this undermines broader governance relationships: tax compliance, regulatory cooperation, civic participation, and democratic trust all depend on the fundamental belief that government acts competently and in citizens' interests. Service failures in highly visible domains such as transit—shaping daily mobility for large populations—become symbolic of broader governance dysfunction, with consequences extending far beyond transportation.

Inter-City benchmarking as a lens on institutional performance

The COMPARISON cluster (17.2% of discourse, 77% negative sentiment) employs inter-city benchmarking, particularly referencing Jakarta's Transjakarta, to frame Surabaya's transit service as comparatively deficient. This rhetorical strategy refutes that service limitations are inevitable and generates accountability pressure by exhibiting a proximate superior alternative. Analytically, this pattern connects to debates on uneven subnational institutional capacity under decentralization, questioning why Indonesia's second-largest city trails the capital in adaptability (Nugroho & Sujarwoto, 2021). However, since a corpus of 134 comments cannot independently demonstrate a fiscal-capacity differential, this linkage is advanced as an interpretive hypothesis inviting further institutional analysis rather than a causal claim.

Representative comparisons are explicit: "*Jakarta saja bisa 24 jam, masa Surabaya tidak bisa*" ("Even Jakarta can operate 24 hours, why can't Surabaya"), "*Transum mereka operasional sampai jam 2 malam, malu banget*" ("Their Transjakarta operates until 2 AM, so embarrassing"), "*Ko kalah sama Semarang*" ("Why worse than Semarang"). The emotional register—"*malu banget*" ("so embarrassing")—indicates that comparative disadvantage generates not merely frustration but social shame that Surabaya, as Indonesia's second-largest city, cannot match smaller cities' transit performance.

The comparative framing serves an important cognitive function: it refutes institutional claims that operational constraints preclude service expansion. If Jakarta—with arguably greater operational complexity (larger population, more extensive network, higher congestion)—can extend hours during high-demand periods, then Surabaya's failure to do so becomes a political choice rather than an operational necessity. This undermines institutional legitimacy by exposing that resource constraints or operational limitations are not determining factors; political prioritization determines outcomes.

The prominence of Jakarta references (appearing in 43% of COMPARISON comments) reflects Indonesia's pronounced Jakarta-centricity. As the national capital and primate city, Jakarta receives disproportionate investment in infrastructure, policy innovation, and international attention. Other Indonesian cities constantly live in Jakarta's shadow, with development trajectories evaluated relative to the capital. This creates competitive pressure—cities aspire to Jakarta-level services—while generating resentment about resource inequality. Citizens in Surabaya observe that Jakarta receives national government support for transit expansion (including a substantial subsidy for Transjakarta and MRT construction partially funded by the national budget), whereas Surabaya must rely primarily on municipal resources, fostering a perception of institutional abandonment.

The Semarang comparisons are particularly interesting: "*Ko kalah sama Semarang*" ("Why worse than Semarang"), referencing Indonesia's fifth-largest city (population ~1.8M vs. Surabaya's 2.9M). If even smaller cities outperform Surabaya, this strongly suggests governance failure—there is no plausible operational explanation for why a larger, wealthier city cannot match the performance of smaller cities. This represents the "shame through downward comparison" rhetorical strategy, more damaging than upward comparison to Jakarta (which at least allows a response that Jakarta has exceptional resources).

From a theoretical perspective, the comparative discourse exemplifies what scholars term "policy diffusion through horizontal emulation" (Shipan & Volden, 2008). Citizens observe policies in other jurisdictions, evaluate them as superior, and demand local adoption. Social media facilitates this diffusion by making inter-city comparisons immediately accessible—citizens can check Jakarta's transit schedules on their smartphones and directly contrast with Surabaya's—creating real-time benchmarking pressure. This represents an important mechanism of democratic accountability: even

without formal performance indicators or official audits, citizen-generated comparisons create political pressure for improvement.

The international comparisons (Singapore MRT) serve a slightly different function: to demonstrate what an aspirational global-standard transit system looks like. While comparisons of Jakarta show that improvement is feasible within Indonesian resource constraints, comparisons of Singapore establish normative standards—what good transit governance should achieve regardless of constraints. This dual benchmarking (a feasible alternative and an aspirational ideal) creates a rhetorical sandwich effect, leaving little room for institutional justification.

Information deficits and digital communication failures

The APP cluster (7.5% of discourse, 70% neutral sentiment) exhibited distinctive patterns compared to service delivery themes. Discourse about digital information systems—mobile applications, real-time tracking, social media updates—generated more balanced, factual commentary rather than emotionally charged criticism. However, this neutrality should not be interpreted as satisfaction; rather, it reflects descriptive reporting of information system failures: "*Aplikasi tracking bus tidak update*" ("Bus tracking app not updated"), "*Informasi di sosmed terlambat, sudah di jalan baru tahu ada gangguan*" ("Social media information late, already traveling before knowing disruption").

The negative association between the APP topic and negative sentiment (OR=0.43, p=0.074, marginal significance) in multivariate regression suggests that when citizens discuss information systems, they are less likely to express intense negative emotion than when discussing operational failures. This may reflect different attribution patterns: information system failures can be framed as technical problems requiring fixes, whereas operational failures are attributed to institutional incompetence or malevolence. Alternatively, the relatively small APP cluster size (n=10) limits the statistical power to detect associations.

The information deficit revealed in these comments has important implications for service quality. Real-time information provision has become integral to perceived transit reliability in contemporary systems (Dziekan & Kottenhoff, 2007). When riders know precisely when vehicles will arrive, even irregular service becomes more tolerable through reduced uncertainty. Conversely, when tracking systems fail during service disruptions—exactly when information is most crucial—this compounds frustration from the operational failure itself.

The complaint about late social media updates—"*sudah di jalan baru tahu ada gangguan*" ("already traveling before knowing disruption")—reveals timing problems in communication. Effective disruption communication requires proactive notification before affected trips, allowing riders to make alternative arrangements. Reactive communication after riders are already committed to transit trips generates maximum frustration: travelers are stranded with no alternative, having relied on (incorrect) presumption that service would operate normally.

From a crisis communication perspective, the New Year's Eve service reduction represented a foreseeable, planned disruption rather than an unexpected operational failure. Transit agencies had weeks of advance notice that operating hours would be reduced; this information should have been widely disseminated through all communication channels (social media, mobile app, website, station signage) beginning at least one week prior. The fact that many citizens appeared surprised by the unavailability of services on New Year's Eve suggests failures in the communication strategy: either information was not distributed proactively, or distribution channels failed to reach relevant audiences.

This reveals broader issues of the digital divide in transit communication. Official social media accounts and mobile applications typically reach younger, more digitally engaged populations, while

older adults, lower-income residents, and irregular riders may not follow transit agency accounts or have smartphones with tracking applications. Reliance on digital communication without complementary analog channels (radio announcements, newspaper notices, physical signage, community outreach) creates information inequality, with the most vulnerable populations—those most dependent on transit—being least likely to receive advance notice of service changes.

Methodological contributions and computational analysis challenges

Beyond its substantive insights, this study contributes methodologically by highlighting both the potential and limitations of computational analysis of Indonesian-language social media. Although the sentiment classifier achieved high apparent accuracy (89% with SMOTE), the low model–human agreement ($\kappa = 0.36$) and non-significant permutation tests ($p = 0.996$) demonstrate that performance was largely driven by class imbalance rather than meaningful discrimination. This finding underscores a critical methodological lesson: accuracy alone is an unreliable metric under severe imbalance and must be complemented by class-balanced measures, baseline comparisons, and statistical significance testing.

The topic modeling results indicate moderate feasibility. Despite the small corpus (N = 134), LDA produced six interpretable thematic clusters ($C_v = 0.42$). While coherence falls below optimal thresholds, strong sentiment–topic correlations ($r = 0.77–0.90$) surviving multiple-testing correction suggest non-random structure, though replication in larger datasets is required. The reliance on a single annotator remains a major limitation, as inter-annotator reliability is essential for validating sentiment constructs, particularly in linguistically and culturally nuanced Indonesian social media contexts.

The near absence of positive sentiment (3.7%) raises interpretive challenges regarding representativeness. This pattern may reflect selection bias inherent to complaint-oriented digital platforms, genuinely poor service conditions, or culturally specific communication norms that channel praise privately and criticism publicly. Disentangling these explanations requires triangulation with surveys, operational performance data, and qualitative interviews. Overall, the study demonstrates that digital discourse analysis can yield meaningful signals, but only when applied with methodological rigor and cautious interpretation consistent with best practices in computational social science.

Limitations and future research directions

This study has clear constraints on generalizability. The small sample size (N = 134) limits statistical power and yields wide confidence intervals, while post hoc analyses suggest that some apparent effects may reflect chance rather than robust relationships. The focus on a single disruption event restricts temporal generalization, and the exclusive focus on Surabaya limits geographic transferability. Methodologically, reliance on single-coder annotation precludes reliability assessment, and the cross-sectional design prevents causal inference regarding the relationship between service failures and expressed sentiment.

Future research should address these limitations by expanding data coverage and methodological rigor. Priority directions include collecting larger, multi-city and multi-event datasets to enable comparative inference; adopting longitudinal designs to track sentiment dynamics over time; and triangulating digital discourse with surveys, operational performance indicators, and qualitative interviews. Methodological improvements should incorporate multi-coder annotation with formal reliability testing, as well as multimodal analysis integrating text with emojis, images, and network context. Finally, quasi-experimental designs exploiting policy changes or service shocks would strengthen causal claims. Comparative cross-national research is also needed to assess

whether patterns of inter-city benchmarking and leadership accountability reflect universal dynamics or are Indonesia-specific, shaped by political and cultural contexts.

CONCLUSION

This exploratory study provides preliminary evidence that citizen digital discourse on urban transit disruptions differentiates clearly among specific service deficiencies. Negative sentiment is most strongly associated with limited operating hours, service reliability failures, and perceived governance accountability deficits, while multivariate results indicate that thematic content accounts for an exploratory share of sentiment variation reflected in a pseudo-R^2 of 0.42. Although constrained by a small sample, fair model–human agreement, and non-significant permutation tests, the convergence of theme–sentiment correlations and regression estimates yields hypothesis-generating insights that warrant validation in larger datasets, since a high pseudo-R^2 coupled with non-significant permutation tests at this sample size more plausibly reflects limited statistical power than a robust relationship.

Substantively, the New Year's Eve 2026 disruption in Surabaya appears to have crystallized long-standing structural problems, namely temporal inaccessibility for transit-dependent users, unreliable operations, weak institutional communication, and deficits in leadership accountability. To address these deficits and mitigate exclusionary gaps for transit-dependent users, the Transportation Department could feasibly extend operating hours during predictable high-demand events, publish advance schedules, and institute responsive digital complaint loops. Furthermore, citizens' frequent intercity comparisons suggest that digital discourse functions not only as complaint but also as an informal oversight mechanism carrying real reputational consequences.

We theoretically propose conceptualizing this phenomenon as Digital Informal Accountability, an emergent extension of social accountability theory adapted to platform-mediated Global South settings where citizen voice must intersect with state responsiveness to improve provision effectively. Methodologically, the study highlights both the promise and the limits of sentiment analysis for Indonesian-language social media, given that apparent high accuracy can mask class-imbalance risks and thereby underscores the need for balanced metrics. Overall, although digital discourse holds clear potential for identifying public service priorities, robust inference requires larger samples, rigorous validation, and triangulation with operational data.

Author Contribution

Conceptualization, E.W.E. and M.R.Z.R.; methodology, E.W.E. and M.R.Z.R.; software, M.R.Z.R.; validation, M.R.Z.R. and S.Z.M.; formal analysis, M.R.Z.R.; investigation, S.Z.M.; data curation, M.R.Z.R. and S.Z.M.; writing—original draft preparation, E.W.E. and R.A.; writing—review and editing, E.W.E. and R.A.; visualization, M.R.Z.R. and S.Z.M.; supervision, E.W.E.; project administration, E.W.E. and M.R.Z.R.; funding acquisition, E.W.E. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors would like to thank all parties who provided administrative and technical support for this research.

Funding

This research received no external funding.

Conflict of Interests

The authors declare no conflict of interest.

Data Availability Statement

The data is available by request to the author to ensure participant privacy protection. Due to technical issues regarding account access and the presence of sensitive, non-anonymized personal information in the initial dataset, the materials are not hosted on a public repository. Complete analysis code, preprocessing scripts, and trained model weights will be securely provided directly by the corresponding author upon reasonable request.

REFERENCES

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Baldwin, T., Lau, J. H., & Ruder, S. (2022). One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500

Angger Saputra, R., & Sibaroni, Y. (2025). Multilabel Hate Speech Classification in Indonesian Political Discourse on X using Combined Deep Learning Models with Considering Sentence Length. *Jurnal Ilmu Komputer Dan Informasi*, *18*(1), 113–125. https://doi.org/10.21609/jiki.v18i1.1440

Avery, E. J., & Graham, M. W. (2013). Political Public Relations and the Promotion of Participatory, Transparent Government Through Social Media. *International Journal of Strategic Communication*, *7*(4), 274–291. https://doi.org/10.1080/1553118X.2013.824885

Badan Pusat Statistik Kota Surabaya. (2025, February 28). *Kota Surabaya Dalam Angka 2025*. Badan Pusat Statistik Kota Surabaya. https://surabayakota.bps.go.id/id/publication/2025/02/28/bd1f25e59ae790cc8a7c0c07/kota-surabaya-dalam-angka-2025.html

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*. *3*, 993–1022. https://doi.org/10.5555/944919.944937

Bouvier, G., & Machin, D. (2018). Critical Discourse Analysis and the challenges and opportunities of social media. *Review of Communication*, *18*(3), 178–192. https://doi.org/10.1080/15358593.2018.1479881

Chakraborty, M., Saha, S., & Sharmin, S. (2025). Quality-aware bus-driver scheduling for intelligent transportation system. *Transportation Engineering*, *20*, 100337. https://doi.org/10.1016/J.TRENG.2025.100337

Chaniotakis, E., & Pel, A. J. (2015). Drivers' parking location choice under uncertain parking availability and search times: A stated preference experiment. *Transportation Research Part A: Policy and Practice*, *82*, 228–239. https://doi.org/10.1016/J.TRA.2015.10.004

Cho, W., & Melisa, W. D. (2021). Citizen Coproduction and Social Media Communication: Delivering a Municipal Government's Urban Services through Digital Participation. *Administrative Sciences*, *11*(2), 59. https://doi.org/10.3390/admsci11020059

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. https://arxiv.org/pdf/1810.04805

Dziekan, K., & Kottenhoff, K. (2007). Dynamic at-stop real-time information displays for public transport: effects on customers. *Transportation Research Part A: Policy and Practice*, *41*(6), 489–501. https://doi.org/10.1016/J.TRA.2006.11.006

El-Diraby, T., Shalaby, A., & Hosseini, M. (2019). Linking social, semantic and sentiment analyses to support modeling transit customers' satisfaction: Towards formal study of opinion dynamics. *Sustainable Cities and Society*, *49*, 101578. https://doi.org/10.1016/J.SCS.2019.101578

Fox, J. A. (2015). Social Accountability: What Does the Evidence Really Say? *World Development*, *72*, 346–361. https://doi.org/10.1016/j.worlddev.2015.03.011

Gilley, Bruce. (2009). *The right to rule : how states win and lose legitimacy*. Columbia University Press.

Graham, M. W. (2014). Government communication in the digital age: Social media's effect on local government public relations. *Public Relations Inquiry*, *3*(3), 361–376. https://doi.org/10.1177/2046147X14545371

Hardiansyah, D., Aziz, R. A., & Hasibuan, M. S. (2024). The Classification Method is Used for Sentiment Analysis in My Telkomsel. *International Journal of Artificial Intelligence Research*, *8*(2), 169. https://doi.org/10.29099/ijair.v8i2.1229

Haryanti, S., & Rusfian, E. Z. (2019). Government Public Relations and Social Media: Bridging the Digital Divide on People with Social Welfare Problems. *JKAP (Jurnal Kebijakan Dan Administrasi Publik)*, *22*(2), 128. https://doi.org/10.22146/jkap.34602

Kemp, S. (2025, February 23). *Digital 2025: Indonesia*. Global Digital Insights. https://datareportal.com/reports/digital-2025-indonesia

Khairunnisa, S. Z., & Widyastuti, H. (2024). Improving The Integration and Connectivity of Feeder Wira Wiri, Suroboyo Bus, Trans Semanggi Bus, and Trans Jatim Bus in Surabaya City Purabaya-Rajawali Routes. *Journal of Civil Engineering*, *39*(2), 147. https://doi.org/10.12962/J20861206.V39I2.20592

Kirana, L. I. (2024, October 18). *Bagaimana Warga Surabaya Memanfaatkan Internet?*. GoodStats Data. https://data.goodstats.id/statistic/bagaimana-warga-surabaya-memanfaatkan-internet-B4JtB

Krisdamarjati, Y. A., & Fatahillah, G. A. (2025, July 29). *Assessing the Commitment of the Surabaya and Semarang City Governments to Develop Public Transportation*. Kompas.Id. https://www.kompas.id/artikel/en-menilai-kesungguhan-pemerintah-kota-surabaya-dan-kota-semarang-membangun-transportasi-umum

Linders, D. (2012). From e-government to we-government: Defining a typology for citizen coproduction in the age of social media. *Government Information Quarterly*, *29*(4), 446–454. https://doi.org/10.1016/J.GIQ.2012.06.003

Lucas, K. (2012). Transport and social exclusion: Where are we now? *Transport Policy*, *20*, 105–113. https://doi.org/10.1016/j.tranpol.2012.01.013

Manurung, C. E., & Mayatopani, H. (2025). Sentiment Analysis of Indonesian Society Toward the Launch of iPhone 16 Using Naive Bayes, Random Forest, and KNN Algorithms. *Jurnal Komputer, Informasi Dan Teknologi*, *5*(1), 13–13. https://doi.org/10.53697/JKOMITEK.V5I1.2219

Meijer, A., & Thaens, M. (2013). Social media strategies: Understanding the differences between North American police departments. *Government Information Quarterly*, *30*(4), 343–350. https://doi.org/10.1016/J.GIQ.2013.05.023

Mergel, I. (2010, January). *Gov 2.0 Revisited: Social Media Strategies in the Public Sector - PA TIMES Online | PA TIMES Online*. Public Administration Times. https://patimes.org/gov-2-0-revisited-social-media-strategies-in-the-public-sector/

Nugroho, Y., & Sujarwoto. (2021). Institutions, Outputs and Outcomes: Two Decades of Decentralization and State Capacity in Indonesia. *Southeast Asian Economies*, *38*(3), 296–319. https://doi.org/10.1355/ae38-3b

Padhilah, F. A., Surya, I. R. F., Adiatma, C. J., Sari, R. P., Permono, R. H., & Pradityo, R. (2025). *Indonesia Sustainable Mobility Outlook 2025 Driving Transport Decarbonization: Multi-pathways to Sustainable Mobility in Indonesia*. www.iesr.or.id

Publication of Research Community Services and Innovation (P3MI). (2020). *Nawasona* (H. Delik & M. R. R. Ramadani, Eds.). ITB Press.

Putra, A. P. (2023). Impact of Toll Road Development on Modern Retail Growth in Indonesia. *Journal Research of Social Science, Economics, and Management*, *2*(8). https://doi.org/10.59141/jrssem.v2i08.407

Redman, L., Friman, M., Gärling, T., & Hartig, T. (2013). Quality attributes of public transport that attract car users: A research review. *Transport Policy*, *25*, 119–127. https://doi.org/10.1016/J.TRANPOL.2012.11.005

Ruder, S., Vulić, I., & Søgaard, A. (2019). A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, 569–631. https://doi.org/10.1613/jair.1.11640

Schweitzer, L. (2014). Planning and Social Media: A Case Study of Public Transit and Stigma on Twitter. *Journal of the American Planning Association*, *80*(3), 218–238. https://doi.org/10.1080/01944363.2014.980439

Shah, A. (2007). *Participatory Budgeting* (A. Shah, Ed.). The World Bank. https://doi.org/10.1596/978-0-8213-6923-4

Shipan, C. R., & Volden, C. (2008). The Mechanisms of Policy Diffusion. *American Journal of Political Science*, *52*(4), 840–857. https://doi.org/10.1111/j.1540-5907.2008.00346.x

Sitten, M. (2012). A service-oriented approach to public sector social media strategy. *Public Interest and Private Rights in Social Media: A Volume in Chandos Publishing Social Media Series*, 79–95. https://doi.org/10.1016/B978-1-84334-693-7.50005-2

Suartika, I. M., & Maya Dora, Y. (2025). Moving A Nation: The Evolution of Public Transit in Indonesia - Enhanced Analysis. *Jurnal Pendidikan Indonesia*, *6*(10), 4537–4548. https://doi.org/10.59141/japendi.v6i10.8684

Torres, E. C. M., Picado-Santos, L. G. de, Torres, E. C. M., & Picado-Santos, L. G. de. (2025). Sentiment Analysis and Topic Modeling in Transportation: A Literature Review. *Applied Sciences 2025, Vol. 15, 15*(12). https://doi.org/10.3390/APP15126576

van Lierop, D., Badami, M. G., & El-Geneidy, A. M. (2018). What influences satisfaction and loyalty in public transport? A review of the literature. *Transport Reviews*, *38*(1), 52–72. https://doi.org/10.1080/01441647.2017.1298683