electronics

MDPI

Review

Social Media Sentiment and Service Quality Evidence in Public Transport and Passenger Mobility: A Review

Anastasia Nikolaidou [1,*], Panos Papaioannou [1], Sokrates Basbas [2] and Ioannis Politis [1]

1 Transport Engineering Laboratory, Department of Civil Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; papa@civil.auth.gr (P.P.); pol@civil.auth.gr (I.P.)
2 Department of Transportation & Hydraulic Engineering, School of Rural and Surveying Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; smpasmpa@topo.auth.gr
* Correspondence: nikolaid@civil.auth.gr

Abstract

Social media platforms, online reviews, app store feedback, and complaint records are important sources of user-generated textual data for understanding passenger experience in public transport and related mobility contexts. This paper reviews how social media data mining and natural language processing have been used to transform such data into evidence on service quality and satisfaction. The review focuses on public transport while selectively drawing on adjacent mobility domains, including aviation, ride-hailing, micromobility, and app-based mobility, which offer transferable methodological insights. Public transport studies are treated as the core empirical evidence base, whereas adjacent mobility- and methods-oriented studies support conceptual and methodological interpretation rather than direct claims about public transport. The review examines how textual sources are converted into analytical outputs, including sentiment, topics, aspects, emotions, and composite indicators. It shows that sentiment should not be treated as a direct measure of satisfaction but as an intermediate signal of passengers' expressed perceptions. Key risks include sampling bias, negativity bias, linguistic ambiguity, class imbalance, weak spatial and temporal attribution, platform effects, and limited validation. The paper proposes a validation-oriented framework that maps text-derived signals to service aspects, anchors them in spatial–temporal context, triangulates them with survey, complaint, and operational evidence, and translates them into service-quality-relevant decision-support outputs.

Keywords: social media data mining; sentiment analysis; natural language processing; user-generated content; public transport; service quality; customer satisfaction; text mining

Academic Editor: Ioannis Hatzilygeroudis

Received: 14 June 2026
Revised: 6 July 2026
Accepted: 10 July 2026
Published: 14 July 2026

1. Introduction

Customer satisfaction and perceived service quality have become central concerns in public transport planning, reflecting a broader shift from supply-oriented performance assessment towards a demand-oriented understanding of the passenger experience [1,2]. Reliance solely on objective indicators, such as travel time, frequency, reliability, capacity, or cost, is insufficient, as such indicators often fail to capture passengers' subjective perceptions of service performance [2,3]. The so-called "quality paradox" suggests that service improvements may go unnoticed when they do not align with passengers' expectations or lived experiences [1,3].

Traditional assessment of perceived service quality and customer satisfaction in public transport has relied mainly on structured passenger surveys and stated-preference questionnaires, often analyzed using index-based or multi-criteria models such as the Service Quality (SERVQUAL) model, structural equation modeling, and factor analysis [2,4,5]. These methods remain important because they provide a structured, theoretically grounded, and comparable basis for measuring perceived quality and supporting formal quality management frameworks [1]. However, cross-sectional surveys are typically costly, infrequent, and limited in their ability to capture rapid changes in passenger perceptions during service disruptions [6,7]. Structured questionnaires may also be affected by response biases, including social desirability bias [4,8–10], respondent fatigue, and low response rates [11,12]. In addition, retrospective questions may introduce recall bias and fail to capture immediate emotional reactions to service experiences [13,14]. These limitations have increased interest in using digital tools for continuous monitoring of passenger perceptions [15,16].

Social media platforms, online review sites, mobile application stores, and other digital feedback channels provide a complementary source of passenger-generated information. Unlike traditional surveys, user-generated content provides real-time, unsolicited feedback on passenger experiences [17–21]. Passengers frequently report delays, overcrowding, cancellations, safety concerns, accessibility issues, ticketing problems, staff interactions, and emotional responses through platforms such as Twitter/X, Sina Weibo, TripAdvisor, Skytrax, and app stores [18,22–24]. Since these data are often time-stamped and, in some cases, linked to routes, stations, locations, or service events, they can support near-real-time monitoring of passenger perceptions and emerging service problems [19]. In this respect, social media and related forms of user-generated content may serve as "human sensors" for analyzing passenger experience and monitoring service quality [20,25].

Recent advances in data mining, natural language processing (NLP), and machine learning have enabled the transformation of large volumes of unstructured passenger text into usable evidence for transport analysis. Early studies relied primarily on keyword matching and basic sentiment analysis [21]. More recent research has employed deep learning methods, transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) [26], advanced topic modeling techniques [27], Aspect-Based Sentiment Analysis (ABSA) [28], and generative Large Language Models (LLMs) [29,30]. These outputs are increasingly combined with statistical, spatial, and decision-making methods to develop indicators and monitoring tools [31,32]. This multidisciplinary approach produces sentiment scores, topic clusters, aspect-level insights, and temporal and spatial patterns that can help identify service problems, disruption effects, and recurring passenger concerns [33–35].

However, computational outputs do not necessarily constitute accurate measures of service quality or satisfaction. This issue is particularly evident in sentiment analysis, which is often used as a proxy for user opinions. Customer satisfaction depends on expectations, perceived performance, and the extent to which expectations are fulfilled [36–38]. Sentiment analysis, by contrast, primarily identifies the emotional tone of a message. For example, a negative post may reflect temporary frustration or stress caused by a specific disruption rather than an overall assessment of the transport system [14,39]. Automated sentiment analysis may also fail to capture sarcasm, irony, implicit meanings, and contextual nuance [40,41].

This distinction is important because many studies treat social media sentiment as an indicator of satisfaction without fully addressing the assumptions involved. Sentiment scores are often combined with topic modeling, geographic analysis, and decision-making tools to assess transit performance and passenger complaints [42,43]. However, a simple positive or negative polarity score does not necessarily reveal which service dimension is affected, what caused the response, how severe the issue is, or how it relates to overall satisfaction [44]. Social media data also contains inherent biases. Passengers affected by disruptions may be more likely to share their experiences, while online content is often produced disproportionately by younger and more digitally engaged users. As a result, such data cannot be assumed to represent the entire passenger population [45,46].

The literature, therefore, faces two linked challenges. Technically, social media texts are often noisy, short, multilingual, context-dependent, and difficult to process due to issues such as spam, sarcasm, class imbalance, and weak geospatial attribution. Interpretively, computational outputs must be carefully translated into meaningful evidence of service quality, dissatisfaction, satisfaction, or broader passenger experience, since expressed sentiment in public transport depends strongly on mode, journey purpose, disruption context, user group, expectations, location, and time period.

Recent reviews have synthesized research on sentiment analysis, topic modeling, and social media analytics in transport, with emphasis on data sources, platforms, model performance, applications, and methodological challenges [47–49]. More recent transport studies have further extended this literature by applying social media mining and online review analytics to public transport service quality, rail-system service evaluation, transit service changes, passenger-perception dynamics, and emerging LLM-based interpretation of transit service-quality perceptions [15,22,25,29,50,51]. In parallel, earlier and recent studies have proposed methods for extracting public opinion, service quality perceptions, and satisfaction-related signals from transport-related social media and user-generated textual data [50,52–54]. However, comparatively little attention has been given to the interpretive pathway by which computational outputs are transformed into service-quality-relevant evidence.

Against this background, this paper reviews the use of social media data mining for measuring public transport service quality and satisfaction. It examines how existing studies translate user-generated text into service quality indicators, how computational outputs are interpreted, and how these indicators are validated for reliability and policy relevance. Five research questions guide the review:

1. How have social media and user-generated textual data been used to assess service quality and satisfaction-related evidence in public transport and related passenger mobility contexts?
2. Which data mining and natural language processing methods are commonly applied to extract meaning from transport-related user-generated text?
3. How are sentiment, topics, emotions, and aspects interpreted as service quality or satisfaction-related indicators?
4. What methodological assumptions and limitations affect the reliability of social media-based service quality measurement?
5. What validation approaches are needed to support the use of social media-derived indicators in transport service quality assessment?

This paper makes four main contributions. First, it reframes social media sentiment as an intermediate perception signal rather than as a direct measure of customer satisfaction. Second, it synthesizes the full methodological and interpretive pathway through which user-generated textual data are transformed into evidence of service quality, from data source selection and preprocessing to aspect mapping, indicator construction, and validation. Third, it identifies the key risks that affect the reliability of this transformation, including sampling bias, negativity bias, linguistic ambiguity, class imbalance, weak spatial and temporal attribution, platform effects, and limited construct validation. Fourth, it proposes a validation-oriented framework that clarifies how social media-derived indicators can be interpreted alongside survey-based and operational evidence to support more robust service quality assessment in public transport.

The remainder of the paper is organized as follows. Section 2 presents the scope of the review and the analytical approach. Section 3 examines data sources and text-mining methods. Section 4 discusses the interpretation of sentiment and other computational outputs; Section 5 addresses reliability and validation. Section 6 presents a conceptual framework and future research directions. Section 7 concludes the paper.

2. Review Scope and Analytical Approach

2.1. Review Design and Corpus Rationale

This paper employs a structured conceptual and methodological review to synthesize the ways in which social media data mining and natural language processing convert user-generated textual data into evidence of service quality and customer satisfaction in transport research. The review is centered on core concepts, analyzing studies based on their function within the analytical process from textual data to computational results, interpretation, and validation.

This approach is appropriate because the relevant literature spans transport research, social media analytics, natural language processing, machine learning, service quality theory, and customer satisfaction measurement, and remains fragmented across disciplinary boundaries. The review includes journal articles and peer-reviewed conference papers in transport, computer science, information systems, and data mining, provided they meet the eligibility criteria and make a relevant conceptual or methodological contribution.

Although the review primarily concerns public transport service quality, the corpus also includes studies from adjacent passenger mobility domains, particularly aviation, airports, ride-hailing, micromobility, and app-based mobility. These studies were retained because they provide transferable methodological insights on online reviews, sentiment interpretation, service quality mapping, app feedback, complaint analysis, and validation. They were not treated as direct empirical evidence on public transport service quality unless the transport context explicitly concerned public transport systems.

The final analytical corpus comprised 260 peer-reviewed records. Table 1 summarizes the composition of the corpus by publication type. For reporting purposes, proceedings/series articles and indexed peer-reviewed conference papers are grouped under indexed conference/proceedings records.

Table 1. Composition of the analytical corpus by publication type.

Publication Type	Number of Records	Share of Corpus (%)	Role in the Review
Journal articles	182	70.0	Primary empirical, conceptual and methodological evidence across transport, service quality and computational text analytics
Indexed conference/proceedings records	78	30.0	Retained where peer-reviewed and methodologically relevant, especially for data mining, NLP, machine learning, sentiment analysis and social media analytics
Total	260	100.0	Full analytical corpus retained after eligibility screening

2.2. Literature Identification, Screening, and Eligibility

A structured literature search was conducted across four primary bibliographic databases: Scopus, Web of Science, Institute of Electrical and Electronics Engineers (IEEE) Xplore, and Transport Research International Documentation (TRID). ScienceDirect was used as a supplementary publisher-platform search source due to its overlap with Scopus-indexed records and its differences in search and export functionality. Google Scholar was also used as a supplementary source to cross-check influential records and support backward and forward citation searches. Still, it was not treated as a separately counted primary database due to limitations in export reproducibility and result stability. The search covered publications from 2012 to 2026 and was last updated in June 2026.

The search strategy combined four groups of terms related to: (i) user-generated textual sources, including social media, Twitter/X, online reviews, app reviews, complaints, and open-ended comments; (ii) transport and mobility contexts, including public transport, transit, rail, bus, metro, passenger transport, mobility services, aviation, airports, ride-hailing, and micromobility; (iii) computational text-analytics methods, including text mining, sentiment analysis, topic modeling, natural language processing, aspect extraction, machine learning, deep learning, transformer-based models, and LLMs; and (iv) service-quality and satisfaction constructs, including service quality, customer satisfaction, passenger satisfaction, passenger experience, perception, complaints, dissatisfaction, and service evaluation. Indicative search combinations included "social media" AND "public transport" AND "sentiment analysis"; "Twitter" OR "X" AND "transport" AND "service quality"; "user-generated content" AND "passenger satisfaction"; "online reviews" AND "transport" AND "customer satisfaction"; "topic modeling" AND "public transport"; "NLP" AND "transport service quality"; and "aspect-based sentiment analysis" AND "mobility". The syntax was adapted to the search fields, operators, and export functions supported by each database or platform, and both backward and forward snowballing were used to identify additional influential and transferable studies.

To improve transparency, Supplementary Table S2 provides a summary of search documentation covering the primary bibliographic database searches, supplementary publisher-platform checks, Google Scholar cross-checks, backward and forward citation searches, platform-specific adaptations, and records carried forward to the master screening file. It is intended as a search documentation summary rather than a complete reproducible search-history log, because search interfaces and export functions differed across sources. Because broad and narrow searches within the same database or platform overlapped and relevance refinement was applied before consolidation, query-level counts in Supplementary Table S2 are not additive. After consolidation, 959 records from the primary bibliographic database searches and 155 additional records from supplementary checking were carried forward, resulting in 1114 raw imports before deduplication.

Studies were included if they: (1) used user-generated textual data; (2) focused on transport, mobility, or transport service contexts; (3) applied computational text analytics methods, such as data mining, NLP, sentiment analysis, topic modeling, aspect extraction, or machine learning; and (4) linked textual outputs to service quality, satisfaction, passenger experience, or service evaluation.

Studies were excluded if they were unrelated to transport or service quality, used social media solely for traffic prediction or behavior analysis, relied solely on structured surveys without textual data, or were purely technical and lacked methodological relevance to the review objectives. Duplicates, non-peer-reviewed material, opinion pieces, news articles, commercial reports, short abstracts, and inaccessible full texts were also excluded.

Records retrieved from the primary bibliographic database searches and supplementary checking were exported in the formats supported by each source and consolidated into a single master screening file. Deduplication was performed via bibliographic matching using available metadata, including title, authorship, DOI, publication year, and source, followed by manual review to identify near-duplicates, minor title variations, and multiple versions of the same study. Where both conference and journal versions of the same study were available, the most complete version was retained.

After deduplication, 584 records remained for preliminary screening of title, abstract, publication status, and scope. During this stage, 322 records were excluded based on the criteria described above. Because preliminary screening was used to remove clearly out-of-scope records, exclusion reasons were treated as overlapping scope categories rather than separately counted, mutually exclusive reasons. Final eligibility exclusions were recorded separately during full-text assessment and detailed coding.

The remaining 262 records were retained for detailed eligibility assessment and coding, including final eligibility confirmation, full-text availability checking, and corpus classification. Two records were excluded at this final stage: one because the full text was unavailable, and one because it did not meet the final eligibility criteria after detailed review. The final analytical corpus therefore comprised 260 records.

Peer-reviewed conference papers were included only when they offered original or relevant methodological insight, particularly in data mining, NLP, machine learning, sentiment analysis, topic modeling, aspect extraction, validation, or social media analytics, given that important methodological developments in these fields frequently appear at conferences.

Although this review did not follow the full Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol, its identification and screening process was informed by PRISMA-style reporting principles to enhance transparency in record identification, deduplication, screening, and eligibility assessment. A full PRISMA approach was not adopted because the paper is a structured conceptual and methodological review rather than a systematic effectiveness review, meta-analysis, or quantitative synthesis of comparable empirical findings. Accordingly, the screening process was documented transparently, while the synthesis remained concept-centric rather than intervention- or outcome-centric.

The identification, deduplication, screening, and eligibility process is summarized in Figure 1.

2.3. Coding Framework and Classification Criteria

After the final corpus was defined, each retained record was coded using a structured framework designed to support the conceptual and methodological synthesis. The coding captured bibliographic information, data-source characteristics, transport context, analytical methods, evidence role, service-quality relevance, validation approach, and integration with survey-based or operational evidence. The full record-level coding is provided in Supplementary Table S1.

The coding framework combined descriptive and interpretive classifications. Descriptive classifications captured observable features of each record, such as publication type, transport sector, platform, content type, analysis type, and validation type. Interpretive classifications captured the evidentiary and methodological relevance of each study in relation to the review objectives, including evidence role, service-quality link strength, text-analytics depth, interpretive strength level, validation level, and integration with survey-based or operational evidence. These categories were defined by the authors to support a structured conceptual and methodological synthesis. They should therefore be interpreted as transparent qualitative coding decisions rather than as statistically derived categories or normative rankings of study quality.

The initial coding was performed by the lead author and subsequently reviewed by the co-authors. Ambiguous cases were discussed until consensus was reached, particularly when studies combined public transport with adjacent passenger-mobility contexts, when their contribution was mainly methodological rather than service-quality-oriented, or when interpretive coding required qualitative judgment.

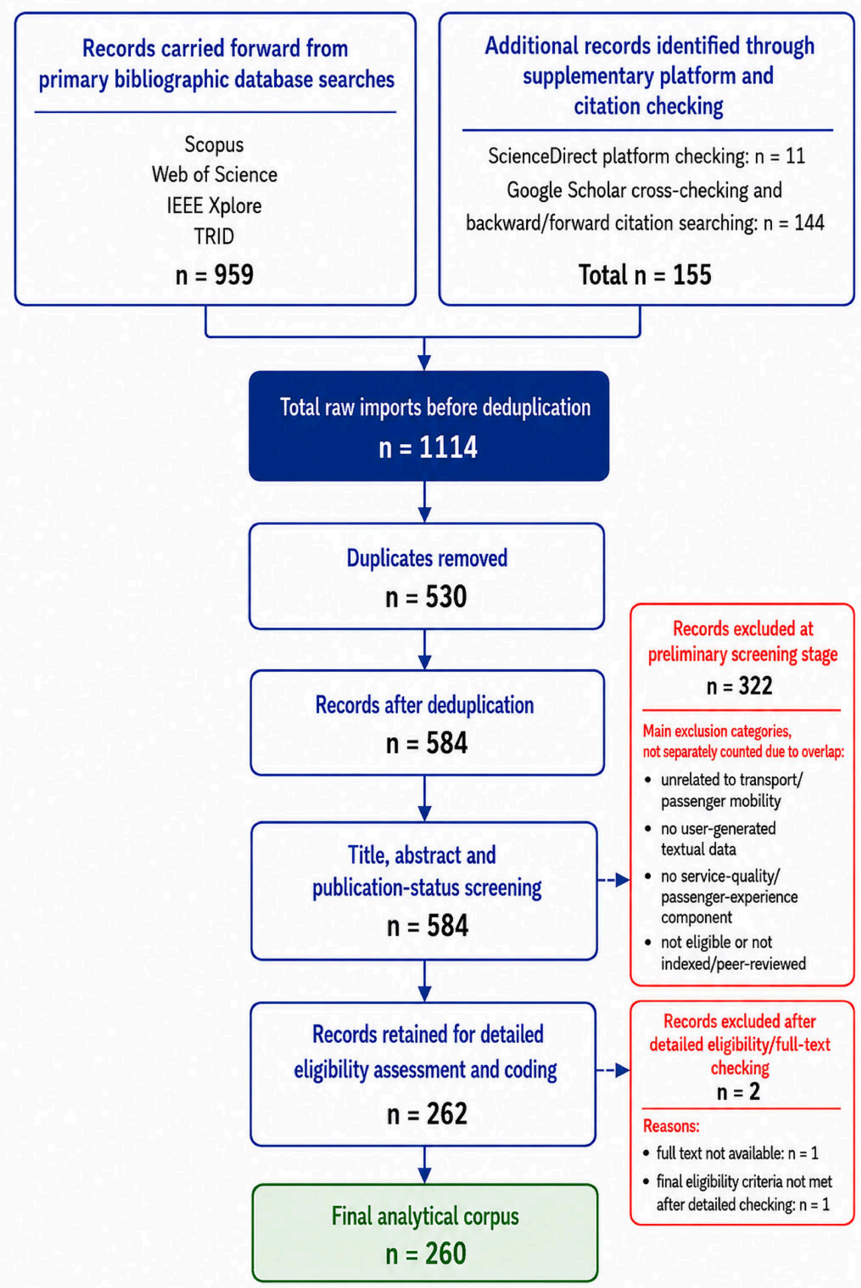

Figure 1. Literature identification and screening process.

No formal inter-coder reliability coefficient was calculated because the review was designed as a structured conceptual and methodological synthesis rather than a formal systematic review with independent duplicate coding. The classification tables should therefore be read as tools for transparent synthesis rather than as statistically validated coding categories.

To improve transparency, Supplementary Table S3 provides the operational definitions and decision rules used for the main coding categories, while Supplementary Table S1 provides the record-level connection between individual studies and the categories summarized in the main text. The coding framework also provides the basis for the analytical role classification presented in the following subsection.

2.4. Analytical Roles and Evidence Weighting

Given that the corpus spans multiple transport domains and methodological traditions, retained records were classified into three analytical roles: core public-transport studies, adjacent passenger-mobility studies, and methods/theory studies (Table 2). This distinction was introduced to avoid assigning equal evidentiary weight to all records. Core studies provide the primary empirical basis for claims about public transport service quality and satisfaction-related evidence. Adjacent mobility studies provide transferable insights from related passenger mobility domains, while methods/theory studies support broader methodological and conceptual synthesis.

Table 2. Analytical role of retained records and use in the synthesis.

Analytical Role	Number of Records	Share of Corpus (%)	Use in Synthesis
Core studies	65	25.0	Direct empirical evidence on public transport service quality, satisfaction, dissatisfaction, or passenger perception using social media or other user-generated textual data.
Adjacent mobility studies	157	60.4	Transferable methodological or interpretive evidence from related passenger-mobility domains, including aviation, airports, ride-hailing, micromobility, shared mobility, MaaS, and app-based mobility.
Methods and theory studies	38	14.6	Methodological, technical, or conceptual support for text analytics, NLP, sentiment analysis, topic modeling, machine learning, LLMs, preprocessing, model evaluation, validation, or satisfaction/service-quality theory.

Records were classified as core studies when they directly addressed public transport service quality, satisfaction, dissatisfaction, or passenger perception using social media or other user-generated textual data. Adjacent mobility studies addressed similar constructs in related domains, such as aviation, airports, ride-hailing, micromobility, shared mobility, Mobility as a Service (MaaS), and app-based mobility, but were not treated as direct evidence for public transport. Methods/theory studies were retained primarily for their computational, methodological, or conceptual contribution to the analytical pipeline from user-generated text to service-quality-relevant interpretation.

This classification was used throughout the synthesis to distinguish direct evidence for public transport from transferable methodological or interpretive support.

2.5. Synthesis Logic and Text-Analytics Depth

The synthesis follows the interpretive pathway by which user-generated textual data are transformed into evidence relevant to service quality. Rather than treating data mining and NLP as isolated technical tools, the review examines how data sources, preprocessing choices, computational outputs, service-aspect interpretation, and validation practices shape the evidentiary value of text-derived signals.

Within this logic, social media sentiment is treated as an intermediate analytical output rather than as a standalone measure of customer satisfaction. Its interpretation depends on data quality, methodological design, service context, and validation against external evidence.

https://doi.org/10.3390/electronics15143086

To document the computational relevance of the corpus, records were also classified according to the depth of text analytics applied or discussed. Table 3 summarizes this classification.

Table 3. Text analytics depth of the retained records.

Text Analytics Depth	Number of Records	Share of Corpus (%)	Definition
High	172	66.2	Studies applying advanced or central computational text analytics, including machine learning, deep learning, transformer-based models, topic modeling, ABSA, emotion detection, clustering or multi-stage NLP pipelines
Medium	79	30.4	Studies applying standard sentiment analysis, keyword-based text mining, descriptive text classification or simpler NLP procedures linked to service evaluation
Low	9	3.5	Studies not centered on computational text analytics but retained as supporting evidence for service quality theory, satisfaction constructs, eWOM (Electronic Word-of-Mouth), review bias, complaint behavior, validation or methodological interpretation

Table 3 shows that the retained corpus is methodologically substantive in terms of text analytics: 66.2% of records were coded as high-depth and 30.4% as medium-depth. Overall, 96.5% of the corpus involves advanced or methodologically explicit forms of computational text analysis, confirming that the review is not dominated by records that rely only on superficial keyword searches or descriptive mentions of social media.

3. Social Media Data Mining in Transport: Data Sources, Methods, and Outputs

3.1. User-Generated Textual Data Sources and Passenger Signals

Transport-related social media data mining uses various user-generated textual sources—microblogging platforms, review sites, app reviews, complaints, Customer Relationship Management (CRM) messages, forums, and open-ended survey comments. These sources differ in format, timing, user motivation, context, and suitability for service quality analysis. Treating these sources as equivalent can be misleading; it is essential to clarify which type of passenger signal each source offers.

Microblogging platforms such as Twitter/X and Sina Weibo are widely used in transport studies for their volume, accessibility, and real-time nature [55,56]. These platforms provide immediate, unsolicited feedback from passengers about delays, disruptions, and staff interactions [57–59], supporting real-time "human sensing" [20,25,60,61]. However, posts are short, noisy, context-dependent, and prone to self-selection and negativity biases [58,62–64]. As such, they are best viewed as real-time signals of perception and disruption, rather than comprehensive measures of overall satisfaction [55,59].

Other sources offer distinct evidentiary value. Online review platforms (TripAdvisor, Yelp, Google Maps, Skytrax) contain reflective, post-use evaluations and structured ratings, enabling quantitative analysis [65–69]. App reviews reveal how digital interfaces shape the transport experience, though they may blur distinctions between physical and digital service quality [70,71]. Complaint records and CRM messages provide structured signals of dissatisfaction [70,71], while open-ended survey comments link to demographic and closed-ended rating data, aiding validation and model training [72,73]. Discussion forums and mobility sites capture more deliberative discourse but often reflect highly engaged or niche communities [74].

This diversity of sources affects interpretation because Twitter/X posts, reviews, complaints, forums, and survey comments capture different moments, emotional intensity, and user motivations in the passenger experience. This distinction is reflected in the corpus structure: core public-transport studies rely more heavily on short-form social media posts, especially Twitter/X and Sina Weibo/Chinese microblogging, whereas online-review evidence is more concentrated in the remaining corpus, particularly in adjacent passenger-mobility domains such as aviation and app-based mobility. Accordingly, conclusions about real-time perception and disruption signals are grounded primarily in the core public-transport subset, while review-based findings are treated mainly as transferable evidence on richer post-use evaluation and rating-linked validation.

The usefulness ratings in Table 4 were assigned as qualitative interpretive judgments rather than quantitative scores. Each source type was assessed according to whether it captures explicit passenger evaluation, provides sufficiently rich text for service-attribute mapping, includes complementary contextual or structured information, supports validation against survey-based, operational, spatial, or temporal evidence, and is affected by known biases such as self-selection, negativity bias, platform effects, and limited representativeness. Sources were classified as highly useful when they satisfied most of these criteria, moderately useful when they provided relevant passenger-experience evidence but had stronger validation or bias limitations, and limited when they mainly supported contextual or issue-specific interpretation.

Table 4. User-generated textual sources and passenger signals.

Source Type	Typical Platforms/Examples	Passenger Signal Captured	Main Analytical Value	Main Limitation	Usefulness for Satisfaction-Related Analysis
Microblogging platforms	Twitter/X, Sina Weibo	Immediate reactions to disruptions, delays, crowding, incidents, staff interactions and service events	Real-time monitoring; event detection; perception tracking; disruption response	Short, noisy and context-dependent texts; self-selection and negativity bias; limited representativeness	Moderate (**)—Useful as near real-time perception and disruption signal, but weak as standalone satisfaction-related evidence
Online review platforms	TripAdvisor, Yelp, Google Maps, Skytrax	Reflective post-consumption evaluations of service attributes	Richer descriptions; often linked to ratings; useful for service attribute analysis	Selective review behavior; overrepresentation of strong positive/negative experiences; tourist or occasional-user bias	High (***)—Especially useful when ratings and textual content are jointly analyzed
Mobile application reviews	Google Play, App Store, MaaS apps, ticketing and journey-planning apps	Evaluation of digital mobility services, app reliability, route planning, payment, ticketing and customer support	Useful for digital service quality and user-interface assessment	Blurs digital service quality and physical transport service quality	Moderate (**)—Useful for digital mobility services, but requires careful separation of app-related and transport-service issues
Complaint records and CRM messages	Formal complaint, customer relationship management databases	Explicit dissatisfaction and reported service failures	Structured dissatisfaction evidence; useful for issue categorization and service recovery	Complaint-driven; captures only users motivated and able to complain	Moderate (**)—Strong for dissatisfaction analysis, but weak for estimating average satisfaction

Table 4. *Cont.*

Source Type	Typical Platforms/Examples	Passenger Signal Captured	Main Analytical Value	Main Limitation	Usefulness for Satisfaction-Related Analysis
Open-ended survey comments	Free-text survey responses	Structured but textual passenger feedback linked to survey variables	Can be linked to demographics, trip characteristics and satisfaction scores	Less spontaneous; shaped by survey design and question wording	High (***)—Strong for validation, supervised model training and linking textual evidence to structured satisfaction measures
Discussion forums and specialized mobility websites	Forums, Q&A platforms, mobility communities	Deliberative discussion of service changes, policy, infrastructure, fares and accessibility	Captures collective interpretation and issue framing	Represents engaged or issue-specific communities	Limited (*)—Useful for contextual understanding, but limited for representative satisfaction measurement

Note: Usefulness levels are indicative qualitative assessments for satisfaction-related analysis, not quantitative scores. *** = high usefulness, ** = moderate usefulness, and * = limited usefulness.

3.2. Analytical Components and Text-Analytics Methods

Despite variation in data sources and analytical methods, the reviewed studies can be organized around a set of recurring analytical components that link user-generated textual data to structured outputs and, in more advanced applications, to service-quality interpretation, indicator-type outputs, and validation. These components are descriptive rather than prescriptive: they summarize methodological functions observed across the literature rather than a fixed sequence implemented by all studies. Individual records combine them selectively and unevenly. Some studies stop at data collection, preprocessing, or sentiment classification, whereas others proceed to service-aspect interpretation, spatial or temporal aggregation, indicator construction, or external validation. Figure 2 summarizes the main literature-derived components of text-based service-quality analysis and indicator development.

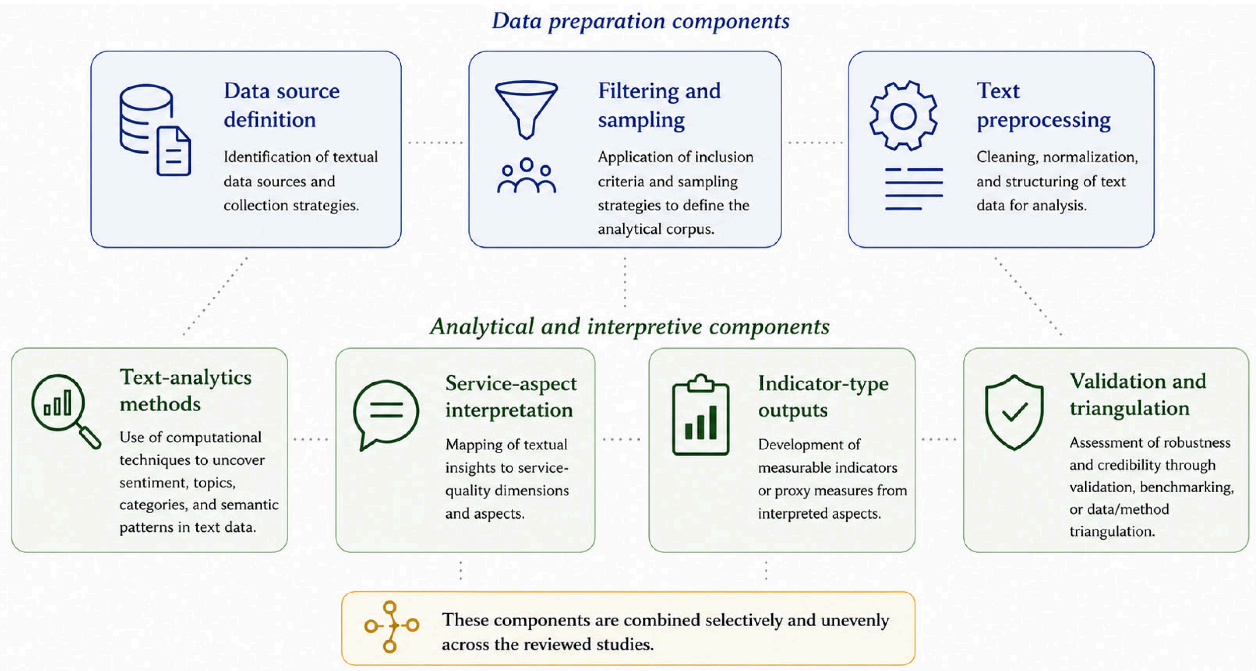

Figure 2. Recurring analytical components identified in social-media-based service-quality and satisfaction studies.

Data source definition, filtering, and preprocessing form the first set of analytical components. Studies gather text via Application Programming Interfaces (APIs), web scraping, exports, app reviews, CRM systems, or survey databases [75]. At this stage, researchers define search terms, platform boundaries, language criteria, temporal scope, geographical coverage, and keywords [62]. These choices matter because posts mentioning official operator accounts may capture actionable complaints, whereas posts collected via station names, route identifiers, hashtags, or generic transport-related terms may capture broader, noisier discussions [19,52,76]. Social media data therefore require relevance filtering to remove false positives, duplicate posts, advertisements, news items, spam, bot-generated content, irrelevant languages, and records outside the study period [19,58,77,78]. It is also important to distinguish official operator messages from passenger-generated content, since service alerts and passenger complaints reflect different perspectives and should not be merged uncritically in satisfaction analysis [39,79].

Text preprocessing prepares raw textual data for computational analysis. Common procedures include lowercasing, tokenization, punctuation and stop-word removal, stemming, lemmatization, spelling correction, emoji handling, and slang normalization [71,80,81]. Although these procedures reduce noise, they may also affect meaning. In transport contexts, route and station names, acronyms, hashtags, and emojis often carry important information. Excessive normalization may therefore remove details needed to identify location, mode, service attributes, or emotional tone [19,42].

Text analytics methods constitute the central computational component of the reviewed studies. These include lexicon-based sentiment analysis, supervised machine learning, deep learning, transformer-based models, topic modeling, and aspect-based methods [82]. Lexicon-based approaches are relatively transparent, but often struggle with domain-specific terminology, sarcasm, and negation [83,84]. Supervised learning methods, such as Naive Bayes, support vector machines, and Random Forests, are used to classify sentiment and service aspects, but require labelled data and may not generalize effectively across contexts [26,77,85–87]. Deep learning and transformer-based models, including BERT and Robustly Optimized BERT Pretraining Approach (RoBERTa), are better at capturing contextual information in short and noisy texts [26,29,30]. However, improved predictive performance does not automatically ensure that outputs are clearly linked to service attributes or customer satisfaction [88].

Topic modeling and aspect-based methods extend the analysis beyond document-level polarity. Topic models, such as Latent Dirichlet Allocation (LDA) and BERTopic, identify recurring themes in reviews, posts, and complaints, such as delays, cleanliness, crowding, fares, and safety [89–91]. ABSA identifies service attributes and links them to sentiment or emotion [28]. This is particularly important because individual comments often contain mixed evaluations [35,45,92]. Aspect-level analysis is therefore more closely aligned with service-aspect interpretation than coarse sentiment polarity, which may overlook distinctions that are operationally relevant for transport providers [44,93].

The more interpretive components involve service-aspect interpretation, indicator-type outputs, and validation or triangulation. Service aspect interpretation translates computational outputs into categories such as safety, information, reliability, comfort, and personnel [9]. For instance, negative sentiment about delays may indicate reliability problems, while negative comments about overcrowding or unclear announcements may relate to comfort or information provision [53]. This stage is essential because it connects text-mining outputs to formal service quality assessment.

Indicator-type outputs are produced by aggregating mapped textual outputs into structured measures that support monitoring, benchmarking, or decision support. Signals may be grouped by station, route, mode, or event to identify spatial or temporal variation in perceived service quality [43]. Aspect-level outputs may also be transformed into dashboards, heat maps, scores, or rankings. Some studies align these outputs with established service quality frameworks, such as SERVQUAL, Airline Service Quality (AIRQUAL), Railway Service Quality (RAILQUAL), or Importance-Performance Analysis [27,53,66,85], or incorporate them into multi-criteria decision-making models [94].

Validation and triangulation provide a methodological safeguard for assessing the robustness and credibility of text-derived outputs. Sentiment scores can be interpreted as satisfaction-related evidence only when they are explicitly linked to service experience, service attributes, and passenger evaluation [95]. Validation approaches include manual annotation, algorithm comparison [96], benchmarking against review ratings, comparison with survey-based satisfaction scores [41], and linkage with operational records, such as Automatic Passenger Count (APC) and General Transit Feed Specification (GTFS) data [51]. However, most studies place greater emphasis on technical accuracy than on validating whether textual outputs operate as robust indicators of service quality or satisfaction [51,97].

Overall, Figure 2 should not be read as a mandatory sequence of transformations. Rather, it identifies the main analytical components through which user-generated text can be transformed into evidence relevant to service quality. The reviewed studies combine these components to varying degrees. Some remain primarily descriptive, focusing on sentiment classification, topic detection, or complaint categorization. Others move further towards service-aspect interpretation, indicator-type outputs, spatial or temporal aggregation, and external validation. The credibility of any resulting indicator therefore depends not only on algorithmic accuracy, but also on the suitability of the data source, the transparency of filtering and preprocessing, the validity of service-aspect mapping, the logic of aggregation, and the strength of validation or triangulation.

The distribution of these components also differs across the review corpus. Core public-transport studies more often combine sentiment or NLP with temporal, event-based, spatial, or operational context, reflecting the need to interpret passenger signals by route, station, disruption, and time window. The remaining corpus contains a stronger concentration of model-comparison, rating-linked, and review-based studies, which are useful for methodological transfer but less directly informative about public-transport operating conditions.

3.3. Analytical Outputs and Interpretive Boundaries

Computational text analysis produces various outputs, including sentiment polarity, topics, aspects, complaint categories, emotion labels, spatiotemporal patterns, and composite indicators. Because these outputs differ in their proximity to service-quality and satisfaction constructs, they should not be treated as equally strong evidence of satisfaction or service performance. The main output types and their interpretive limits are summarized in Table 5.

Sentiment polarity is the most common output, classified as positive, negative, neutral, or as a continuous score [98,99]. It is often aggregated by time, location, mode, or attribute to track shifts in passenger perceptions [19,34]. While useful for mapping expressed tone, sentiment metrics risk being overinterpreted as satisfaction scores and do not fully account for negativity bias, context, or the difference between momentary dissatisfaction and overall system evaluation [14,23,97].

Topic and theme extraction identify recurring operational concerns, such as delays, crowding, cleanliness, fares, and safety [78,91,100]. These outputs indicate which issues dominate passenger discourse, but high topic frequency does not necessarily correspond to importance for satisfaction. Some topics may be discussed more frequently because they are problematic or newsworthy, whereas important service attributes may remain underrepresented when they function adequately [17,53,101].

Table 5. Analytical outputs and interpretive limits.

Output Type	Typical Methods	Managerial Use	Main Interpretive Limitation	Validation Requirement
Sentiment polarity	Lexicon-based sentiment analysis; supervised classification; transformer-based sentiment models	Monitoring positive, negative or neutral user tone across time, platforms, routes or events	Polarity does not equal satisfaction; vulnerable to sarcasm, negativity bias and context loss	Manual annotation; comparison with ratings, surveys or service events
Topic/theme extraction	LDA, BERTopic, clustering, keyword-based topic extraction	Identifying dominant issues such as delays, crowding, fares, safety or accessibility	Topic frequency does not equal importance or dissatisfaction	Human topic validation; comparison with complaint categories or service quality dimensions
Aspect-level evaluation	ABSA, aspect extraction, aspect-level sentiment classification	Identifying which service attributes are positively or negatively evaluated	Requires reliable aspect definitions and domain-specific mapping	Manual aspect annotation; comparison with SERVQUAL-type dimensions or survey attributes
Complaint classification	Supervised classifiers, rule-based complaint detection, CRM classification	Service recovery, issue prioritization, complaint monitoring	Captures dissatisfaction signals, not average satisfaction	Comparison with formal complaint logs and operational incidents
Emotion detection	Emotion lexicons, supervised emotion classification, transformer-based emotion models	Distinguishing anger, anxiety, frustration, trust or joy in passenger discourse	Emotion is culturally and contextually dependent; sarcasm and exaggeration are difficult to detect	Manual emotion annotation; qualitative validation; event-based comparison
Spatial-temporal patterns	Geocoding, station/route extraction, temporal aggregation, spatial analysis	Mapping service problems across routes, stations, events or time periods	Depends on reliable location and time attribution; geotags are sparse	Comparison with GTFS, disruption logs, APC, ridership or operational records
Composite indicators	Weighted sentiment/topic scores, Importance-Performance Analysis (IPA), TOPSIS, VIKOR, Multicriteria Satisfaction Analysis (MUSA), dashboards	Ranking services, prioritizing interventions, decision support	Risk of hidden weighting, construct mixing and false precision	Transparent weighting; sensitivity analysis; survey and operational validation

Aspect-level evaluation is more directly relevant to service-quality measurement because it links specific service attributes to sentiment or emotion [28]. For example, aspect-based models can identify mixed evaluations of different attributes within the same review [102]. This level of granularity is more actionable than aggregate sentiment, which may obscure operationally relevant distinctions [77,103].

Complaint categories, emotion labels, spatial-temporal patterns, and composite indicators provide additional analytical outputs, but each has specific interpretive limits. Complaint classification can support service recovery but mainly reflects the views of users motivated to report problems [104–106]. Emotion detection distinguishes affective states but remains sensitive to sarcasm and cultural variation [40,107,108]. Spatial and temporal analyses can inform planning, but their reliability depends on accurate location and time attribution [6,17,19,101]. Composite indicators, including Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR), Multicriteria Satisfaction Analysis, and Importance-Performance Analysis, condense textual evidence into structured measures but require clear justification and validation [31,109,110].

The central distinction is that analytical outputs such as sentiment, topics, complaints, emotions, spatial patterns, and composite indices are not inherently measures of satisfaction. They become satisfaction-relevant only when they are linked to service attributes, expectations, context, and validation. This distinction provides the basis for the following section, which examines how text-derived outputs can be interpreted as indicators of service quality and satisfaction.

Table 5 shows that the evidentiary value of text-mining outputs depends on how far the analysis moves from descriptive signal detection towards service-quality interpretation. Outputs that remain at the level of general polarity, affect, or topic detection provide weaker or more indirect evidence, whereas outputs mapped to explicit service dimensions, passenger expectations, satisfaction constructs, or external validation sources provide stronger evidence of service quality. Composite indicators offer stronger decision-support potential, but only when their constructs, weighting logic, temporal scope, and validation assumptions are explicit. This distinction underpins the review's separation between direct service-quality evidence, satisfaction-related perception signals, and primarily methodological or contextual contributions.

4. From Sentiment to Service Quality: Interpretation, Mapping, and Conceptual Tensions

4.1. Sentiment Polarity and the Limits of Satisfaction Measurement

A central methodological issue in parts of the literature is the tendency to interpret sentiment polarity as a proxy for customer satisfaction [111]. While sentiment analysis offers scalable outputs that can be aggregated across time, space, platforms, or service attributes [20,98], this simplicity can obscure the fact that sentiment polarity only reflects the affective tone of a text, whereas satisfaction is a broader, holistic judgment shaped by expectations, performance, and overall experience [95,108].

Customer satisfaction is commonly explained, among other approaches, by Expectancy-Disconfirmation Theory (EDT), which conceptualizes satisfaction as an evaluation of the gap between prior expectations and perceived performance [37,91]. In transport contexts, this evaluation is highly variable and is influenced by journey purpose, reliability, familiarity with the service, and user characteristics [34,112]. A negative post about a delay may therefore indicate incident-specific dissatisfaction, accumulated frustration, or a temporary emotional reaction [23]. Such expressions are useful for diagnosing passenger experiences, but they should not be interpreted as equivalent to overall service satisfaction [113].

This distinction matters as social media often captures heightened emotions. Dissatisfied passengers are more likely to post complaints due to negativity bias [38,114,115]. A negative post may indicate a service problem but not overall dissatisfaction [33]. The absence of complaints should not be interpreted as evidence of satisfaction, since routine or satisfactory journeys may not prompt users to post [116]. Neutral and positive sentiment can also be ambiguous, sometimes reflecting irony, factual reporting, or relief rather than genuine satisfaction [33,41,117].

This does not make sentiment analysis irrelevant. Its value lies in identifying real-time affective orientations and highlighting service areas that require further interpretation. The methodological problem arises when a macro-level sentiment score is treated as a final measure of satisfaction rather than as an intermediate analytical output.

4.2. Mapping Textual Signals to Service Quality Dimensions and Interpretive Strength

To be operationally useful, sentiment analysis should move beyond aggregate polarity towards granular, aspect-level evaluation [77,103]. Mapping sentiment and other textual signals to specific service attributes bridges the gap between unstructured textual data and formal evaluation frameworks, enabling more actionable insights for transport operators [27,28,44,118].

Key service-quality dimensions in transport include reliability, comfort, safety, accessibility, information provision, affordability, and customer service [9,10,110]. These dimensions are commonly associated with established frameworks such as SERVQUAL, AIRQUAL, and RAILQUAL [36,85,119]. Social media content can be mapped to such dimensions using topic modeling, lexicon-based or deep-learning classifiers, and ABSA [17,26,28,29,96].

This mapping links textual cues to specific service attributes. References to delays and cancellations may indicate reliability problems, while comments on overcrowding or poor cleanliness may reflect comfort-related concerns [15,19]. Mentions of harassment or unsafe driving relate to safety, whereas ticket prices and refund issues concern affordability [113,120]. Similarly, references to information provision or staff behavior can be assigned to the relevant service-quality dimensions [50,110].

Rigorous mapping can transform sentiment outputs into actionable evidence. Negative sentiment linked to specific aspects, such as reliability or staff behavior, helps identify the underlying sources of dissatisfaction [93,121]. ABSA is therefore more suitable for service-quality evaluation than coarse document-level polarity classification, because it separates posts or reviews into specific targets and attributes [122]. This level of granularity is particularly important because a single service encounter may contain mixed evaluations, such as positive comments about staff assistance alongside negative comments about waiting time or disruption management [45,88].

However, service quality mapping remains methodologically challenging because user-generated transport text is often noisy, ambiguous, and mode-specific [77,123]. Generic terms, overlapping place or station names, and informal expressions can reduce algorithmic accuracy [42,94,124,125]. Robust domain knowledge, context-specific dictionaries, transparent coding procedures, and validation are therefore necessary to support reliable interpretation [26,126].

Traditional service quality frameworks may also fail to capture emerging issues revealed through user-generated text [27]. Contemporary passengers increasingly discuss digital ticketing, health protocols, crowd management, security, accessibility, and sustainability [127–129]. This creates a methodological tension between deductive mapping to established service quality categories and inductive identification of new themes. A robust analytical approach should therefore combine both strategies, preserving comparability while remaining sensitive to emerging passenger concerns [17,130].

Within this broader mapping process, the reviewed literature differs in the strength with which sentiment outputs are interpreted. At the weakest level, sentiment is treated as a direct proxy for satisfaction, with positive sentiment interpreted as satisfaction and negative as dissatisfaction [98,111]. Although simple and scalable, this approach is conceptually fragile because it conflates momentary affective expressions with broader evaluative judgments and struggles with mixed, indirect, or ironic expressions [23,40,131,132].

A stronger approach interprets sentiment as a perception signal linked to specific service aspects, commonly through ABSA or by associating sentiment outputs with service topics, dimensions, routes, stations, disruption events, or contexts [28,77]. For example, clusters of negative sentiment associated with terms such as "crowded" or "delay" may indicate comfort or reliability problems [6,19]. This approach provides actionable diagnostic evidence, but it does not constitute a comprehensive measure of satisfaction [68]. At the strongest level, sentiment is treated as one component of a broader service quality or satisfaction-related indicator, combining sentiment extraction with service quality mapping, contextual interpretation, and external evidence [50,95]. Table 6 summarizes these levels of interpretive strength.

Table 6. Conceptual levels of interpretive strength in the use of sentiment outputs.

Level	Interpretation of Sentiment	Typical Analytical Design	Main Limitation	What This Level Can Support
Level 1: Sentiment as direct proxy	Positive sentiment is treated as satisfaction and negative sentiment as dissatisfaction	Aggregate polarity scores by platform, route, time period or operator	Conflates momentary affect with broader satisfaction; weak construct validity	Descriptive monitoring of online tone, but not robust satisfaction measurement
Level 2: Sentiment as perception signal	Sentiment is interpreted as evidence of user perception toward specific service aspects	Sentiment linked to topics, aspects, routes, stations, disruption events or service dimensions	Provides diagnostic evidence but not a full satisfaction measure	Identification of service issues, disruption reactions and aspect-level perception patterns
Level 3: Sentiment as satisfaction-related evidence	Sentiment is one component of a broader service quality or satisfaction-related indicator	Sentiment combined with service quality mapping, contextual interpretation and external evidence	More data-intensive; requires alignment across sources and explicit assumptions	Stronger service quality interpretation and cautious satisfaction-related inference

The record-level coding further supports this distinction. In the retained corpus, 114 records (43.8%) interpreted sentiment as satisfaction-related evidence within a broader service-quality or evaluative framework, while 67 records (25.8%) treated sentiment primarily as a perception signal. A further 62 records (23.8%) provided non-sentiment-based service-quality mapping, for example through topics, aspects, complaints, or review dimensions. Only 17 records (6.5%) treated sentiment mainly as a direct proxy for satisfaction. This distribution reinforces the need to distinguish between sentiment polarity, perception signals, and satisfaction-related evidence rather than treating all sentiment outputs as equivalent measures of customer satisfaction.

This distinction is particularly important for the core public-transport subset, where direct satisfaction-proxy interpretations are rare, and sentiment is more commonly interpreted as a perception signal, satisfaction-related evidence, or part of broader service-quality mapping. In the remaining corpus, direct-proxy interpretations are more visible, partly reflecting adjacent mobility and review-based studies where ratings, reviews, and satisfaction labels are more frequently combined. The review therefore treats the caution against equating sentiment with satisfaction as especially well supported for public-transport applications.

Context is central to this progression. Identical sentiment scores may have different meanings depending on transport mode, journey stage, location, time period, and operational events [6,18,33,34,110]. A sudden increase in negative posts during a major disruption, safety incident, or extreme weather event should not be interpreted in the same way as persistent negative sentiment under normal operating conditions. Sentiment outputs should therefore be interpreted in relation to mode-specific, temporal, spatial, and platform-specific conditions rather than applied uniformly across contexts [44,77,120]. External validation further strengthens the reliability of this interpretation.

5. Reliability, Validation, and Methodological Risks

5.1. Sampling, Representativeness, and Negativity Bias

A major limitation of social media-based service quality measurement is that user-generated data are rarely representative of the broader passenger population. Unlike structured surveys with defined sampling frames and demographic controls, social media data come from users who self-select to post. This creates a self-selection problem, whereby the data may disproportionately reflect highly visible, digitally active, and engaged segments rather than the full socio-demographic spectrum of transport users [45,46].

This limitation is particularly important in public transport, where passenger populations are socially diverse. Older passengers, low-income users, people with limited digital access, non-native speakers, disabled passengers, and users who depend heavily on public transport may be underrepresented in social media datasets [35,92]. Consequently, concerns expressed online may not fully correspond to those of the broader passenger base, potentially leading to biased assessments of service quality [90].

Platform architecture also shapes representativeness, as different platforms elicit different types of passenger expression. Microblogging platforms like Twitter/X capture short, event-driven comments about disruptions [18,21], while review platforms allow more reflective evaluations [17,66]. App-store reviews often blur the line between physical service quality and digital usability [70,71]. Structured complaint datasets from CRM systems are skewed toward dissatisfied users [26]. Each platform thus provides only a partial, medium-specific view of the passenger experience and must be contextualized before aggregation.

A related issue is negativity bias: the tendency for negative experiences to be more salient and more likely to be expressed than routine or satisfactory ones [114,115]. Transport-related posts typically follow operational failures, such as delays or cancellations [107,133]. Satisfactory journeys are rarely posted about. Thus, high negative sentiment may signal real problems, but it may also partly reflect the higher propensity of dissatisfied users to post [15,51,107].

These biases affect satisfaction measurement. Interpreting negative sentiment as a direct measure of dissatisfaction can exaggerate overall dissatisfaction, while limited positive content may reflect the underreporting of routine journeys rather than low satisfaction [116]. To address this, social media indicators should be normalized, for example, as "complaints per million passengers" using ridership or passenger count data [134]. Sentiment trends must always be considered in an operational context: a spike during disruptions differs from persistent negativity in normal conditions [135].

Negativity bias also affects machine learning. Datasets with mostly negative posts can cause classifiers to learn skewed patterns and perform poorly on minority classes. Techniques like resampling, class weighting, manual annotation, and balanced metrics help address this [136–138]. However, these adjustments improve technical performance but do not solve the deeper issue that social media data are inherently complaint-rich.

5.2. Linguistic Ambiguity and Preprocessing Risks

User-generated transport text is often noisy and context-dependent, with abbreviations, slang, emojis, hashtags, informal spelling, sarcasm, irony, humor, anger, multilingual expressions, and local references [19,42]. These features make automated text analysis difficult and can distort the interpretation of service quality if not properly managed [40].

Sarcasm and irony are major challenges. Passengers may use positive words to describe negative experiences, such as delays or overcrowding, causing lexicon-based models to misclassify sentiment [40]. Sarcasm may become more frequent during disruptions or chronic service problems, potentially masking the severity of frustration if posts are not interpreted in context [24,41].

Linguistic ambiguity is a specific preprocessing risk in transport-related social media mining. Transport keywords, station names, line names, and route identifiers may overlap with commercial brands, neighborhood names, events, venues, or everyday language. For example, "subway" may refer to either an urban rail system or a fast-food brand, while station names may also denote streets, districts, or non-transport entities [19,42,77]. Such ambiguity can introduce false positives into the dataset and affect downstream tasks such as sentiment analysis, topic modeling, aspect extraction, and service quality mapping. Preprocessing should therefore include contextual filtering, entity disambiguation, transit-specific lexicons or ontologies, and manual or semi-automated checks, particularly when keywords have multiple meanings [26,81].

Hashtags and emojis add a further layer of complexity to passenger feedback analysis. Hashtags may compress location, judgment, and sarcasm into a single expression, while emojis can convey frustration, humor, or affective nuance and may alter the meaning of a post [139]. Removing these elements can reduce interpretive value, whereas retaining them without appropriate processing may introduce additional noise. Custom dictionaries for emojis, abbreviations, route identifiers, station names, and local expressions can help minimize information loss and improve sentiment or aspect classification [105].

Multilingual and code-switched posts present an additional challenge. Public transport systems serve diverse user populations, yet sentiment analysis is often restricted to English-language content or to a single dominant language [19]. Excluding non-English content may exclude important user groups from the analysis and reduce representativeness [92]. Incorporating multilingual data requires language-specific models, bilingual evaluation, or translation tools in order to maintain classification quality [92,140].

These challenges demonstrate that preprocessing and model selection are not neutral technical decisions; they shape which meanings are preserved or lost before analysis begins [26,77]. For valid interpretation, it is essential to preserve transport-relevant terms, locations, line names, accessibility references, emotional markers, and negations. Aggressive stop-word removal or excessive normalization may remove complex names, spatial information, or affective cues that are necessary for transport analysis [19]. Over-normalized models may produce cleaner datasets, but they can also generate outputs with limited practical value for transport agencies [107].

5.3. Spatial and Temporal Attribution

Transport service quality is inherently spatial and temporal [57]. Complaints about crowding, delay, safety, or accessibility are more actionable when linked to specific routes, stations, vehicles, times, or events [19,20,22]. However, social media data often lack strong spatial and temporal attribution.

Although posts are time-stamped, the posting time may not match the actual service experience. Users may post during or after journeys, or much later [15,92]. This asynchronous behavior creates uncertainty when matching posts to operational data [141]. A negative comment may refer to a current disruption, a past incident, or a general opinion about unreliability [33,92].

Spatial attribution is challenging because few social media posts have precise Global Positioning System (GPS) geotags; geotagged tweets make up only 1–3% of samples [55,142]. Relying solely on geotags reduces the dataset and excludes relevant feedback. Many posts mention stations, lines, or routes informally, but these references are often incomplete or ambiguous. Without careful extraction and validation, spatial mapping can create a false sense of operational precision [19,92].

Because transport interventions are spatially targeted, operators need to identify where problems occur in order to respond effectively. Network-level sentiment scores are generally less useful than route- or station-level signals, but only when location inference is reliable [17,34]. Temporal aggregation also requires caution: broad time intervals may obscure short-term disruptions, whereas very fine intervals may amplify noise. The appropriate temporal scale depends on the analytical objective. Near-real-time disruption detection requires short intervals, whereas satisfaction monitoring and trend analysis require longer baselines and segmentation between normal operating days and disruption periods [19,143].

Structured operational datasets can help mitigate spatial and temporal uncertainty. Sentiment signals should be compared with passenger counts, GTFS indicators [144], disruption logs, delay records [57,59], and ridership data [134] to assess whether reported problems correspond to actual service conditions. Such methodological triangulation can help agencies distinguish isolated online complaints from systemic and verifiable service failures [51,59].

5.4. Validation, Construct Validity, and Indicator Reliability

Validation remains a central challenge in social media-based service quality measurement. The literature often validates model performance, but less frequently examines whether the resulting outputs reflect actual service quality or customer satisfaction. Technical validation assesses whether a model classifies text accurately, typically using metrics such as accuracy, precision, recall, and F1-score [50,51,145]. However, such metrics assess classification performance rather than whether sentiment scores serve as valid proxies for overall customer satisfaction [146,147].

Substantive validation assesses whether text-derived indicators align with actual transport service quality or satisfaction patterns. This requires triangulation with external evidence. For example, negative sentiment can be compared with delays or cancellations from GTFS or APC data [107,134], sentiment trends can be benchmarked against survey scores [41,79], complaints can be cross-referenced with CRM logs [26], and spatial clusters can be matched to known bottlenecks using spatial analytics [19,22].

Validation strength in the literature ranges from no external validation, where algorithmic outputs are interpreted without external checking, to manual annotation, cross-model comparison, survey or rating benchmarking, and multi-source triangulation. Cross-model comparison, such as BERT versus Support Vector Machine (SVM), is useful for assessing technical performance [82,88], but does not establish construct validity. The strongest validation approaches combine social media, surveys, ratings, complaints, operational data, or spatial-temporal evidence to assess whether text-derived signals correspond to service conditions or passenger evaluations [23,24,144]. Table 7 summarizes these validation levels and their evidentiary strength.

The record-level coding shows that validation remains uneven across the retained corpus. Cross-model comparison was the most common validation level, appearing in 71 records (27.3%), followed by survey or rating benchmarking in 59 records (22.7%) and manual annotation in 57 records (21.9%). Multi-source triangulation was identified in 36 records (13.8%), while 37 records (14.2%) provided no external validation. This distribution confirms that technical validation is more common than substantive validation, and that relatively few studies triangulate social media-derived indicators with both passenger-evaluation data and operational evidence. The literature provides evidence for the feasibility of several validation practices, but not yet for their systematic adoption as a standard decision-support workflow.

Table 7. Validation levels and evidentiary strength of social media-derived service-quality indicators.

Validation Level	What is Validated	Typical Evidence	Main Contribution	Main Limitation
No external validation	Algorithmic outputs are interpreted without external checking	Sentiment scores, topics, complaint labels	Fast and simple descriptive output	High risk of overinterpretation; does not establish construct validity
Manual annotation	Whether labels, topics, aspects, or sentiment assignments are plausible	Human-coded samples, inter-annotator agreement	Improves credibility of classification and interpretation	Does not by itself prove correspondence with satisfaction or service performance
Cross-model comparison	Whether one algorithm outperforms another on a classification task	Accuracy, precision, recall, F1-score, model benchmarking	Useful for technical performance assessment	Validates classification performance, not construct meaning
Survey or rating benchmarking	Whether text-derived indicators align with structured passenger evaluations	Satisfaction surveys, numerical review ratings, star ratings	Provides stronger satisfaction-related evidence	Requires comparable constructs, temporal alignment, and compatible sampling frames
Multi-source triangulation	Whether text-derived signals align with multiple external evidence sources	Social media text, surveys, ratings, CRM records, complaints, GTFS, APC, disruption logs	Strongest basis for triangulated service-quality indicators	Data-intensive; requires transparent assumptions, weighting, uncertainty treatment, and careful interpretation; does not prove causality or representativeness

This unevenness also differs across the analytical roles of the retained records. Technical validation, including model performance assessment and cross-model comparison, is more common in adjacent mobility- and methods-oriented studies. By contrast, operational data comparison, empirical operational linkage, and multi-source triangulation are more prevalent in the core public-transport subset, although they remain far from universal. Direct survey integration is limited in both groups. These patterns support the need for the evidence-weighting and transferability discussion in Section 6.1 before proposing a validation-oriented decision-support framework.

Composite service quality indicators introduce further validation challenges. Recent frameworks combine topic weights and sentiment scores with multi-criteria decision-making methods, including TOPSIS, VIKOR, and related approaches [28,31,148]. Although such indices may be useful for benchmarking, they also risk hidden weighting, construct mixing, and false precision. Hidden weighting occurs when topic frequency is implicitly treated as importance and average sentiment is interpreted as objective performance, as in some IPA-based applications [149]. This assumption is problematic because frequently mentioned topics are not necessarily the most important service attributes. Some core dimensions, such as safety or accessibility, may be highly important but are rarely discussed in relation to passenger expectations [93,101].

Construct mixing arises when complaints, sentiment, topics, operational data, and surveys are combined as though they measure the same phenomenon, even though they represent distinct forms of evidence [23,26,95,114]. False precision is another risk: composite indicators may produce exact numerical scores that appear objective, even when they are based on uncertain algorithms, biased samples, or weak spatial attribution [19,31]. Composite indicators are therefore valuable only when their construction is transparent, and each component is clearly defined, weighted, and benchmarked against relevant survey-based or operational evidence [149].

The reliability of social media-based service-quality measurement depends on rigorous execution across the full analytical process, rather than on model sophistication alone. Data-source selection, sampling, preprocessing, model performance, aspect mapping, spatiotemporal attribution, and external validation all shape the credibility of the final output [15,42]. A reliable approach must define the user signal being captured, distinguish between emotion, sentiment, complaints, and satisfaction, connect outputs to clearly specified service attributes, account for platform bias, and validate interpretations through expert annotation, surveys, or operational data [23,26,33,41,51,150].

Overall, social media data are valuable for real-time transport quality monitoring, but they are not self-validating [33]. User-generated content can reveal what users express, where dissatisfaction emerges, which attributes trigger strong reactions, and how perceptions evolve over time [18–20]. However, its use as a satisfaction indicator requires methodological restraint [95]. Social media analytics should therefore complement, rather than replace, traditional surveys and operational data [33,51,95].

6. Discussion

6.1. Evidence Weighting and Transferability Across the Review Corpus

Before proposing the validation-oriented framework, it is necessary to clarify how evidence is weighted across the review corpus. The synthesis does not assign equal evidentiary status to all retained records. Claims about public-transport service quality and satisfaction-related evidence are grounded primarily in the core public-transport subset, whereas adjacent passenger-mobility and methods/theory studies are used selectively for transferable methodological, conceptual, or validation-related insights. Table 8 summarizes selected coded variables from Supplementary Table S1 to distinguish the direct public-transport evidence base from the remainder of the corpus.

Table 8. Evidence map of corpus characteristics, analytical approaches, and validation/integration patterns.

Evidence-Map Axis	Core Public-Transport Studies, n = 65	Remaining Corpus, n = 195	Interpretation for the Review
Mode/sector	General or multimodal public transport: 28 records (43.1%); metro/light rail: 16 (24.6%); rail: 11 (16.9%); bus: 10 (15.4%).	Adjacent passenger-mobility domains, including aviation, airports, ride-hailing, micromobility, shared mobility, MaaS, app-based mobility, and methods/theory records.	The core evidence base is multimodal but more concentrated in urban public transport contexts. The remaining corpus supports methodological transfer rather than direct public-transport claims.
Platform	Twitter/X: 39 records (60.0%); Sina Weibo/Chinese microblogging: 10 (15.4%).	Twitter/X: 68 records (34.9%); Sina Weibo/Chinese microblogging: 4 (2.1%).	Short-form social media is especially dominant in public-transport service-quality research.
Content type	Social media posts: 52 records (80.0%); online reviews: 7 (10.8%).	Social media posts: 89 records (45.6%); online reviews: 83 (42.6%).	Public-transport evidence is more social-media-based, while online reviews are more prominent in adjacent mobility domains.

Table 8. *Cont.*

Evidence-Map Axis	Core Public-Transport Studies, n = 65	Remaining Corpus, n = 195	Interpretation for the Review
Analytical method	Keyword/frequency analysis: 57 records (87.7%); text mining/NLP: 54 (83.1%); sentiment analysis: 50 (76.9%); content/thematic/manual analysis: 43 (66.2%); temporal/event analysis: 38 (58.5%); model comparison/evaluation: 35 (53.8%); machine-learning classification: 33 (50.8%).	Sentiment analysis: 152 records (77.9%); keyword/frequency analysis: 132 (67.7%); text mining/NLP: 114 (58.5%); model comparison/evaluation: 123 (63.1%); machine-learning classification: 113 (57.9%); content/thematic/manual analysis: 102 (52.3%).	Sentiment analysis is dominant in both groups, but the remaining corpus has a stronger model-evaluation orientation.
Topic, aspect, spatial and temporal analysis	Topic modeling: 21 records (32.3%); aspect/service-attribute analysis: 12 (18.5%); spatial/geospatial analysis: 18 (27.7%); temporal/event analysis: 38 (58.5%).	Topic modeling: 51 records (26.2%); aspect/service-attribute analysis: 43 (22.1%); spatial/geospatial analysis: 12 (6.2%); temporal/event analysis: 47 (24.1%).	Spatial and temporal contextualisation is proportionally stronger in core public-transport studies.
Service-quality link strength	Strong link: 52 records (80.0%); moderate link: 13 (20.0%).	Strong link: 115 records (59.0%); moderate link: 64 (32.8%); weak link: 16 (8.2%).	The core subset provides the strongest direct basis for public-transport service-quality synthesis.
Interpretive strength	Satisfaction-related evidence: 33 records (50.8%); perception signal: 15 (23.1%); non-sentiment service-quality mapping: 16 (24.6%); direct satisfaction proxy: 1 (1.5%).	Satisfaction-related evidence: 81 records (41.5%); perception signal: 52 (26.7%); non-sentiment service-quality mapping: 46 (23.6%); direct satisfaction proxy: 16 (8.2%).	The core subset strongly supports the argument that sentiment should not be treated as a direct satisfaction measure.
Primary validation level	Multi-source triangulation: 23 records (35.4%); manual annotation: 20 (30.8%); no external validation: 9 (13.8%); cross-model comparison: 7 (10.8%); survey/rating benchmarking: 6 (9.2%).	Cross-model comparison: 64 records (32.8%); survey/rating benchmarking: 53 (27.2%); manual annotation: 37 (19.0%); no external validation: 28 (14.4%); multi-source triangulation: 13 (6.7%).	Public-transport studies show proportionally stronger multi-source triangulation, while the remaining corpus relies more on cross-model and rating-based validation.
Specific validation practices	Manual annotation: 32 records (49.2%); model-performance validation: 31 (47.7%); operational-data comparison: 22 (33.8%); survey comparison: 7 (10.8%); cross-model comparison: 7 (10.8%); ratings comparison: 3 (4.6%).	Model-performance validation: 113 records (57.9%); manual annotation: 48 (24.6%); cross-model comparison: 65 (33.3%); ratings comparison: 50 (25.6%); operational-data comparison: 10 (5.1%); survey comparison: 4 (2.1%).	Technical validation is common in both groups, but operational-data comparison is much more visible in public-transport studies.
Survey and operational integration	Survey linkage: 8 records (12.3%); operational linkage: 23 (35.4%).	Survey linkage: 4 records (2.1%); operational linkage: 18 (9.2%).	Survey integration remains limited, while operational-data linkage is more common in the core public-transport subset.
Multi-source triangulation and no external validation	Multi-source triangulation: 23 records (35.4%); no external validation: 9 (13.8%).	Multi-source triangulation: 13 records (6.7%); no external validation: 28 (14.4%).	Multi-source triangulation is more visible in core studies but remains far from universal.

Note: Primary validation level records the main validation approach assigned to each study. Specific validation practices are not mutually exclusive and may be applied to multiple records.

Table 8 shows that the direct public-transport evidence base differs from the remaining corpus in several important ways. Core public-transport studies are more strongly dominated by Twitter/X and short-form social media posts, and they show a stronger spatial, temporal, event-based, and operational orientation. Operational data comparison and multi-source triangulation are also more evident in the core subset. These patterns support the review's public-transport-specific conclusions regarding social media as a signal of perception and disruption, the need for spatiotemporal anchoring, and the importance of operational data for interpreting service-quality evidence.

By contrast, the remaining corpus contributes primarily transferable methodological and interpretive evidence. Adjacent mobility studies, especially aviation, app-based mobility, and online-review studies, provide useful examples of review-based analysis, rating comparison, and service-attribute interpretation, but they are not treated as direct evidence on public-transport service quality. Methods/theory records inform the methodological framing of the review, including text mining, model evaluation, preprocessing, class imbalance, transformer-based modeling, and validation, but do not carry the same evidentiary weight as the core public-transport studies for claims about passenger satisfaction or service quality.

This distinction also clarifies the status of the proposed framework. The framework is not presented as a standardized workflow consistently implemented throughout the literature. Rather, it synthesizes practices that appear unevenly across the reviewed studies, including manual annotation, model performance validation, service-aspect mapping, survey or rating comparison, operational data linkage, and multi-source triangulation. It then extends these observed practices into a more explicit validation-oriented evidence-to-decision logic, including contextual anchoring, uncertainty assessment, confidence thresholds, and conflict-handling rules. These latter elements should therefore be read as a proposed direction for future research and applied decision support, rather than as established practice across the reviewed corpus.

6.2. Proposed Validation-Oriented Framework for Service-Quality Evidence and Decision Support

The literature indicates that social media data mining is most valuable for assessing transport service quality when computational outputs are embedded in a broader measurement and decision-support logic [51,95]. The central question is therefore not whether user-generated content is useful, but under what conditions unstructured textual information can be interpreted as service-quality-relevant evidence and used to support monitoring, planning, or policy evaluation [40,52,142].

Social media, surveys, complaints, and operational data capture different but complementary dimensions of the public transport experience. Social media provide immediacy and sensitivity to emerging issues [6,20]; surveys offer structured and more representative passenger evaluations [45]; complaints capture formal escalation; and operational data provide evidence of service conditions, disruptions, demand, and performance [51]. No single source is sufficient on its own. Social media signals may reveal visible or frustrating events, but they do not automatically establish causality, representativeness, or population-level satisfaction [33,142].

Building on the literature-derived components summarized in Figure 2, the proposed framework in Figure 3 does not introduce a new technical text-mining pipeline. Its contribution is to reposition existing analytical components within a validation-oriented evidence-to-decision logic. Unlike Figure 2, which summarizes components observed unevenly across the reviewed studies, Figure 3 is prescriptive: it identifies the interpretive, contextual, and validation conditions required before social media outputs can inform service-quality monitoring or policy action. In this framework, sentiment scores, topics, aspects, emotions, and complaint labels become decision-support evidence only when they are mapped to explicit service aspects, anchored in spatial and temporal context, assessed against external evidence, and interpreted with uncertainty and source limitations in mind.

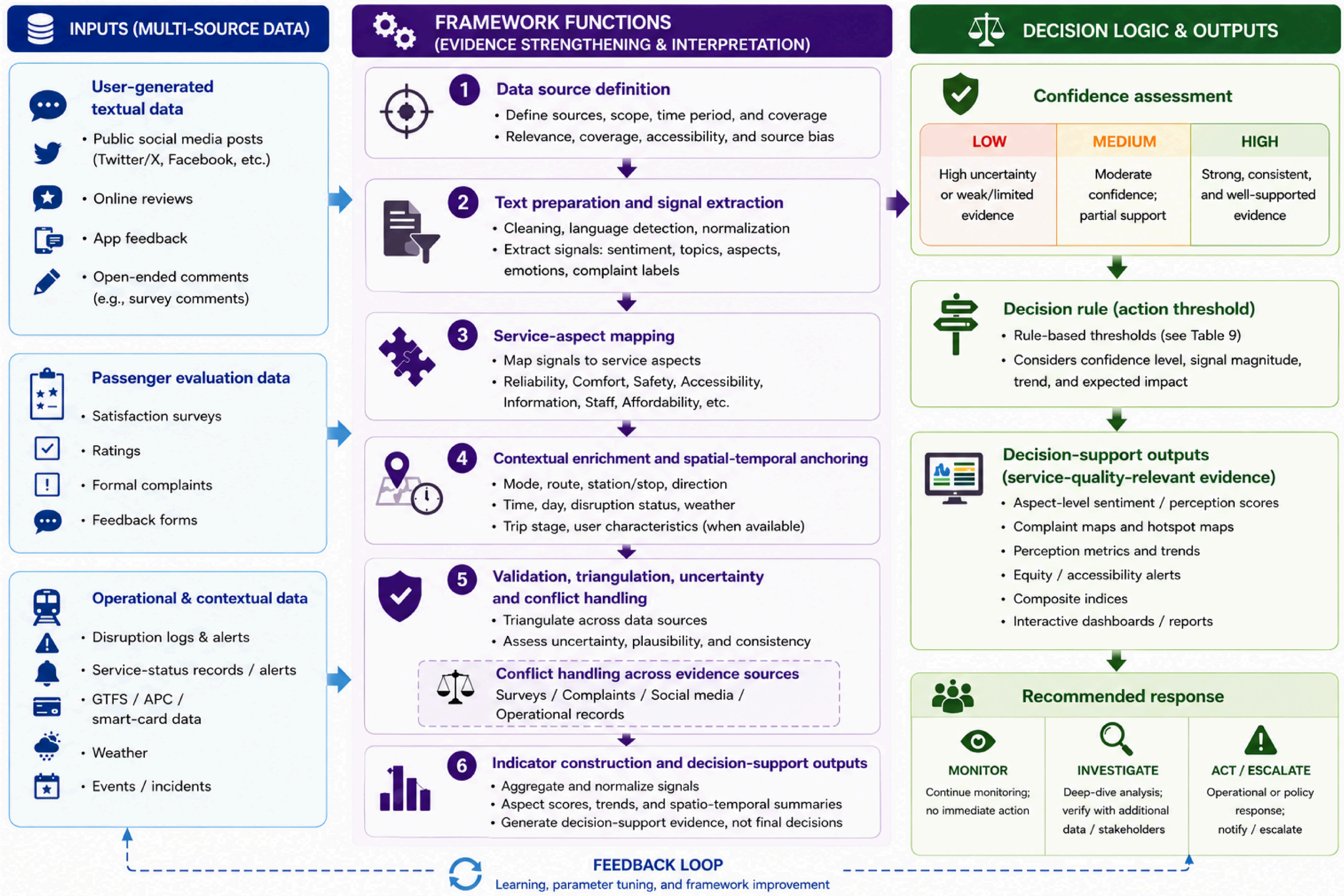

Figure 3. Structure and functions of the proposed validation-oriented framework.

Several reviewed studies approximate parts of this logic by linking social-media-derived outputs with survey data, operational indicators, GTFS or APC data, smart-card validations, complaint records, spatial patterns, or official service indices [41,79]. Other studies move towards indicator construction, urgency classification, or operational decision-support outputs, including complaint maps, performance metrics, topic-based indicators, and response-oriented classifications [106,107]. These examples show that many components of a validation-oriented approach already exist in the literature. However, they remain unevenly developed: explicit confidence thresholds, conflict-handling rules, and actionability criteria are still limited.

Consistent with the evidence-weighting logic outlined above, the proposed framework synthesizes these unevenly observed practices into a more explicit validation-oriented evidence-to-decision logic.

1. Data source definition. This function defines the type of textual evidence used, such as social media posts, online reviews, app feedback, complaints, or open-ended survey comments. Different sources capture different forms of passenger experience; a disruption-related tweet is not equivalent to a formal complaint or a reflective post-use review.
2. Text preparation and signal extraction. This function involves filtering, preprocessing, and extracting computational signals, such as sentiment, topics, emotions, aspects, and complaint labels. Transport-relevant information, including line and station

names, route references, accessibility terms, hashtags, emojis, and negation, should be preserved where analytically meaningful.

3. Service-aspect mapping. This function maps text-derived outputs to specific transport service dimensions, such as reliability, comfort, safety, accessibility, affordability, information provision, staff interaction, and disruption management. This step links computational outputs to meaningful service-quality constructs.
4. Contextual enrichment and spatial-temporal anchoring. This function links textual outputs to contextual information, including mode, route, station, time, disruption status, weather conditions, trip stage, and user characteristics where available. The meaning of user sentiment depends on when, where, and under what operating conditions it is expressed.
5. Validation, triangulation, and uncertainty assessment. This function compares social media-derived signals with external evidence, such as surveys, operational data, disruption logs, smart card data, complaint records, and ratings. The purpose is not to prove that social media data are representative of all passengers, but to assess whether the signal is credible, contextually plausible, and consistent with other forms of evidence.
6. Indicator construction and decision support. This function translates sufficiently validated and contextually interpreted signals into decision-support outputs, such as aspect-level sentiment scores, complaint maps, perception metrics, equity alerts, hotspot maps, or composite indices. These outputs should support monitoring, investigation, policy evaluation, or targeted operational response, while remaining conditional on transparent documentation of data limitations, assumptions, validation evidence, uncertainty, and decision context.

This framework positions social media data mining as part of a broader service-quality measurement and decision-support system. It does not reject sentiment analysis; rather, it defines its appropriate role. Sentiment should be used as a dynamic signal of expressed evaluation, not as a standalone measure of satisfaction. When combined with aspect extraction, contextual enrichment, spatial–temporal anchoring, triangulation, and uncertainty assessment, it can support more credible service-quality monitoring and decision support.

To make the proposed framework more concrete, consider an illustrative scenario for monitoring reliability-related dissatisfaction in a metropolitan public transport network. The example is not presented as a new empirical case study, but as an implementation scenario showing how the framework could be operationalized in practice.

The textual source would consist of public social media posts mentioning the transport authority, operator accounts, line names, route numbers, station names, stops, disruption terms, and mode-specific keywords during a defined monitoring period. Retweets, official service announcements, advertisements, spam, and non-passenger posts would be removed or coded separately so that the analytical corpus focuses on passenger-generated evaluative content. The service aspect measured would be reliability, including references to delays, cancellations, long waiting times, missed connections, unexpected closures, irregular headways, and disruption-related crowding.

Text-derived outputs would include reliability-related post volume, aspect-level sentiment, complaint categories, topic clusters, and, where possible, emotion or urgency labels. These outputs would not be treated as direct satisfaction scores, but as reliability-related perception signals requiring contextual anchoring and validation. Contextual data would include service-status records, disruption logs, planned works, timetable or headway data, ridership or smart-card validations, complaint records, and relevant passenger survey indicators.

Validation would proceed through three checks: manual annotation of a sample, model performance assessment, and external comparison with disruption logs, complaints, ridership anomalies, service status records, or survey trends. Agreement with external evidence would strengthen construct validity, but it would not prove causality, representativeness, or population-level satisfaction.

Conflicting information would be handled through triangulation. Operational records indicate whether a measurable service event occurred; social media posts indicate the visibility and perceived passenger impact of the event among digitally active users; complaints capture formal escalation; and surveys provide broader but less immediate passenger evaluation. The final decision-support output could be a reliability perception dashboard by line, route, station, time period, and disruption event. Table 9 presents a simple decision rule for determining when a social-media-derived indicator is sufficiently reliable to support action.

Table 9. Decision rule for using social-media-derived service-quality indicators.

Signal Confidence	Minimum Evidence Criteria	Interpretation	Recommended Response
Low	Isolated or low-volume posts; uncertain service aspect; weak spatial/temporal match; no external corroboration	Possible perception signal only	Monitor; no direct operational action
Medium	Repeated posts above baseline; acceptable classification confidence; partial spatial/temporal match; partial support from complaints, operational data, or contextual evidence	Plausible emerging service-quality issue	Investigate; check operational logs, complaints, staff reports, passenger information, and local context
High	Clear increase above baseline; validated service-aspect classification; specific line/route/station and time window; corroboration by at least one independent source, such as disruption logs, complaints, survey trend, ridership anomaly, or service-status record	Credible service-quality signal with decision-support value	Act or escalate; adjust operations, improve passenger information, inspect affected service, or trigger targeted follow-up survey

This worked example illustrates the added value of the proposed framework. Existing studies already construct social-media-derived indicators, spatial complaint maps, perception metrics, urgency classifications, and decision-support outputs. However, the proposed framework integrates these elements into a more explicit evidence-to-decision logic by requiring service-aspect mapping, contextual anchoring, triangulation, uncertainty assessment, and a transparent decision rule before social-media indicators are used to support operational or policy decisions.

6.3. Future Research Priorities and Implications

Five research priorities emerge from the review.

First, future studies should develop hybrid indicators that combine social media sentiment, service aspects, surveys, complaints, and operational data [33,51]. Research should move beyond algorithmic accuracy and test whether computational indicators correspond to meaningful service-quality patterns, thereby strengthening construct validity [41,95]. Survey validation can reveal whether online expressions reflect broader passenger opinions or mainly capture extreme dissatisfaction due to negativity bias and self-selection [95,114]. Operational validation can assess whether negative sentiment related to reliability, safety, crowding, or information provision aligns with observed service conditions [51]. Future work should also define confidence criteria for when hybrid indicators are suitable for monitoring, investigation, or intervention, rather than treating any statistically significant association between social media and external data as sufficient evidence for action.

Second, research should focus on aspect-based, emotion-aware, and context-sensitive models. Service quality is multidimensional, and passengers often evaluate conflicting aspects in a single post [28]. Document-level sentiment methods may miss these nuances [88,151], whereas ABSA links sentiment to specific service attributes and provides more actionable diagnostics [44]. Future work should also integrate emotion analysis, since discrete emotions such as anger, fear, frustration, anxiety, trust, and relief may reveal distinct passenger experiences and service-quality concerns [23,108]. These approaches require domain-specific adaptation, including custom ontologies, service-quality dictionaries, transport-aware language models, and validation against manually annotated or externally benchmarked data [19,26,77].

Within this methodological progression, the application of LLMs represents an important emerging direction for interpreting unstructured passenger-generated text [29,30]. LLMs may offer greater flexibility in aspect extraction, contextual sentiment interpretation, and the semantic mapping of informal passenger feedback to service-quality dimensions, particularly in short, noisy texts that contain implicit meanings, mixed sentiment, sarcasm, or references to operational events [29,30]. However, their use introduces methodological risks, including prompt sensitivity, limited reproducibility, model opacity, hallucinated explanations, contextual reasoning errors, ambiguity misclassification, privacy concerns, and the need for domain-specific validation [29,30]. LLMs should therefore be treated as supplementary components within a validation-oriented analytical framework, not as autonomous shortcuts for satisfaction measurement. Their outputs should be anchored to expert-defined service-quality taxonomies, manual annotation, survey-based evidence, and operational data [29,152].

Third, bias mitigation and equity analysis are critical. Social media users are not representative of all passengers, affecting equity-sensitive planning [6]. Online data may overrepresent younger, more digitally active, and more affluent users, while underrepresenting older passengers, low-income users, disabled passengers, non-native speakers, or passengers with limited digital access [35,45,46]. Studies should therefore assess platform, demographic, and language bias and avoid prioritizing interventions solely on the basis of topic volume. Small clusters related to harassment, accessibility, discrimination, safety, or exclusion may signal severe issues even when they are infrequent [22]. Equity-sensitive indicators should therefore combine signal volume with issue severity, affected user group, spatial concentration, and corroborating contextual evidence.

Fourth, research should develop multilingual, multimodal, and platform-sensitive methods. Restricting analysis to English-language data or to a single dominant language may exclude relevant passenger groups and introduce bias [92]. Multilingual models, language-specific preprocessing, validated translation workflows, and multilingual lexicons can help reduce errors across diverse passenger populations [92,140,153]. Platform sensitivity is necessary because platforms such as Twitter/X [18,21,50], app store reviews [70,71], and CRM complaints [26] capture different forms of passenger expression and different moments of the service experience. Multimodal analysis using images, emojis, or visual cues can supplement text, but it presents technical, ethical, privacy, and interpretive challenges and must remain aligned with service-quality goals [88,139].

Fifth, decision-support systems must be transparent, interpretable, and confidence-based. Social media mining can help operators identify problem locations, affected service dimensions, and emerging perception signals, but these outputs should not automatically trigger operational or policy responses. Future systems should make explicit how confidence levels are assigned, how conflicting evidence from social media, complaints, surveys, and operational records is handled, and what type of response is appropriate under low-, medium-, or high-confidence conditions. Explainable Artificial Intelligence (XAI) and transparent conceptual design can help address black-box issues by providing managers with the reasoning behind specific sentiment, aspect, or classification outputs [29,30,154]. A dashboard displaying spatial clusters of negative sentiment is operationally useful only if the underlying data source, classification method, spatial inference logic, validation process, uncertainty level, and recommended response are transparent and methodologically justified to stakeholders [15,41].

The field should therefore move from exploratory social media mining towards validated, context-sensitive, and theory-informed service-quality assessment. Social media-derived indicators can support timely monitoring, disruption detection, complaint categorization, and public perception tracking [19]. Their analytical value increases when sentiment is linked to specific service aspects, spatial and temporal contexts, and corroborated by survey-based or operational evidence [28,51]. Advanced computational models, including transformers, domain-adapted variants, and LLMs, can improve aspect classification and semantic interpretation [26,30,155], but they must be integrated with service-quality theory, domain expertise, transparent assumptions, and rigorous external validation [26,29]. These tools should therefore be understood as complementary diagnostics rather than definitive measures of customer satisfaction [95]. The most credible approach is a triangulated service-quality and decision-support framework, in which social media provides timely perception signals [6], surveys offer structured and demographic passenger evaluations [41], complaints capture formal escalation, and operational datasets provide empirical evidence of service conditions [51]. Integrating these layers offers a more robust and actionable basis for assessing transport service quality than relying on any single source [33,51], provided that uncertainty, confidence levels, and source conflicts are made explicit.

7. Conclusions

This paper reviewed how social media data mining and natural language processing have been used to extract service-quality indicators, satisfaction-related evidence, and decision-support signals from user-generated textual data in transport research. It examined the analytical pathway through which social media posts, online reviews, app-store reviews, and complaint records are transformed into computational outputs, including sentiment scores, topics, aspects, emotions, complaint categories, and composite service-quality indicators.

The review shows that social media and other forms of user-generated text provide valuable but partial evidence of passenger experience. These data capture timely and context-specific expressions of user perception, revealing service problems, dissatisfaction, safety and accessibility concerns, and digital service failures that may be less visible in traditional survey instruments. Their main contribution lies in supporting the monitoring of public reactions, the identification of service pain points, and the detection of shifts in user discourse across operational, spatial, and temporal contexts.

At the same time, the findings confirm that sentiment analysis alone cannot be considered a valid measure of customer satisfaction. Sentiment captures the affective tone of a text, whereas satisfaction is a broader evaluative judgment shaped by expectations, perceived performance, service attributes, and context. Interpreting positive sentiment as satisfaction and negative sentiment as dissatisfaction is therefore conceptually weak unless supported by aspect-level interpretation, contextual information, and external validation. Sentiment should be understood as an intermediate analytical signal rather than as a standalone satisfaction measure.

The reviewed literature demonstrates increasing methodological sophistication, moving from lexicon-based sentiment classification towards machine learning, deep learning, topic modeling, ABSA, emotion detection, and hybrid analytical frameworks. These methods improve the extraction of meaning from large and noisy textual datasets. However, technical sophistication does not automatically produce valid evidence of service quality. Even highly accurate classifiers may generate weak or misleading conclusions if the underlying data are biased, the operational context is unclear, or the outputs are not validated against external evidence.

A central limitation identified in the review is the gap between technical and substantive validation. Many studies assess classification performance using technical metrics, but fewer examine whether text-derived outputs correspond to actual passenger satisfaction or service-quality conditions. In transport research, validation must therefore extend beyond model accuracy to include comparison with surveys, ratings, manual annotations, operational records, complaints, and other external evidence. The strongest approaches are those that triangulate social-media-derived indicators with structured satisfaction measures and operational performance data. Such triangulation should not be understood as a simple search for agreement across sources; it also requires explicit handling of disagreement, uncertainty, source bias, and confidence thresholds before indicators are used to support action.

Several methodological risks remain. Social media users are not representative of the full passenger population, and online discourse is shaped by platform characteristics, self-selection, and negativity bias. Complaint-rich data may overrepresent dissatisfaction, while routine or positive experiences are often underreported. Linguistic ambiguity, sarcasm, multilingual content, emojis, bots, and weak spatial attribution further complicate analysis. These limitations require cautious interpretation, transparent preprocessing, explicit mapping to service-quality dimensions, and rigorous validation.

Overall, this paper positions social media data mining as one component of a broader transport service-quality measurement system. Sentiment provides a dynamic signal of passengers' expressed perceptions; aspect extraction identifies the service dimensions being evaluated; contextual enrichment links these signals to mode, location, time, and disruption conditions; and validation assesses whether the resulting indicators align with broader satisfaction or operational evidence. This approach avoids both dismissing social media data as too biased to be useful and treating sentiment as a direct substitute for traditional survey-based assessment. It also clarifies that social-media-derived indicators should inform monitoring, investigation, or response only when supported by transparent service-aspect mapping, contextual anchoring, validation evidence, and confidence assessment.

Future research should prioritize hybrid, validated, and context-sensitive frameworks for measuring transport service quality. Key priorities include developing aspect-based models tailored to transport service quality, integrating social media analytics with survey, complaint, and operational datasets, addressing representativeness and equity concerns, improving multilingual and platform-specific analysis, and designing transparent decision-support tools with explicit confidence thresholds and conflict-handling rules. The objective should not be to replace traditional passenger surveys with social media analytics, but to integrate both within a more responsive, multi-source, and interpretable assessment framework.

In conclusion, social media data mining has considerable potential to support the evaluation of transport service quality, provided that its outputs are interpreted with conceptual and methodological rigor. Sentiment is not equivalent to satisfaction; it becomes satisfaction-relevant only when linked to service attributes, contextual conditions, and validation evidence. Even then, it should be interpreted as decision-support evidence rather than as proof of population-level satisfaction. This distinction is essential for moving from exploratory opinion mining towards robust, actionable, and theoretically grounded service-quality measurement in transport.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/electronics15143086/s1: Supplementary Table S1: Full analytical corpus and record-level coding for the 260 studies included in the review, including bibliographic information, transport context, textual data sources and platforms, analytical methods, evidence roles, service-quality links, interpretive strength, validation approaches, and integration with survey and operational data; Supplementary Table S2: Search strategy and screening flow, including consolidated screening counts, database- and source-specific search strategies, search strings, search fields, filters, and records carried forward; Supplementary Table S3: Field definitions and coding rules, including operational definitions, allowed values, decision rules, indicators, and examples for the record-level fields and qualitative coding categories used in the review.

Author Contributions: Conceptualization, A.N. and P.P.; methodology, A.N. and P.P.; formal analysis, A.N.; investigation, A.N.; resources, A.N.; data curation, A.N.; writing—original draft preparation, A.N.; writing—review and editing, A.N., P.P., S.B. and I.P.; visualization, A.N.; supervision, P.P., S.B. and I.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data supporting this review are available in the Supplementary Materials. No new primary passenger, survey, or social media datasets were created or analyzed in this study.

Acknowledgments: During the preparation of this manuscript, the authors used Grammarly Pro (version 1.173.1.0; WebUI 2.16.13) for language editing and proofreading, and ChatGPT (GPT-5.6, OpenAI) for support in figure preparation. The authors reviewed, edited, and verified all outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABSA	Aspect-Based Sentiment Analysis
AIRQUAL	Airline Service Quality
APC	Automatic Passenger Count
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
CRM	Customer Relationship Management
EDT	Expectancy-Disconfirmation Theory
eWOM	Electronic Word-of-Mouth
GPS	Global Positioning System
GTFS	General Transit Feed Specification
IPA	Importance-Performance Analysis
IEEE	Institute of Electrical and Electronics Engineers
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
MaaS	Mobility as a Service

MUSA	Multicriteria Satisfaction Analysis
NLP	Natural Language Processing
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAILQUAL	Railway Service Quality
RoBERTa	Robustly Optimized BERT Pretraining Approach
SERVQUAL	Service Quality
SVM	Support Vector Machine
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
TRID	Transport Research International Documentation
VIKOR	VlseKriterijumska Optimizacija I Kompromisno Resenje
XAI	Explainable Artificial Intelligence

References

1. Barabino, B.; Di Francesco, M. Characterizing, Measuring, and Managing Transit Service Quality. *J. Adv. Transp.* **2016**, *50*, 818–840. [CrossRef]
2. De Oña, R.; Machado, J.L.; De Oña, J. Perceived Service Quality, Customer Satisfaction, and Behavioral Intentions: Structural Equation Model for the Metro of Seville, Spain. *Transp. Res. Rec.* **2015**, *2538*, 76–85. [CrossRef]
3. Friman, M.; Fellesson, M. Service Supply and Customer Satisfaction in Public Transportation: The Quality Paradox. *J. Public Transp.* **2009**, *12*, 57–69. [CrossRef]
4. Eboli, L.; Mazzulla, G. Service Quality Attributes Affecting Customer Satisfaction for Bus Transit. *J. Public Transp.* **2007**, *10*, 21–34. [CrossRef]
5. Eboli, L.; Mazzulla, G. Customer Satisfaction as a Measure of Service Quality in Public Transport Planning. In *International Encyclopedia of Transportation*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 220–224.
6. Luo, S.; He, S.Y. Understanding Gender Difference in Perceptions toward Transit Services across Space and Time: A Social Media Mining Approach. *Transp. Policy* **2021**, *111*, 63–73. [CrossRef]
7. Becker, J.U.; Albers, S. The Limits of Analyzing Service Quality Data in Public Transport. *Transportation* **2016**, *43*, 823–842. [CrossRef]
8. King, M.F.; Bruner, G.C. Social Desirability Bias: A Neglected Aspect of Validity Testing. *Psychol. Mark.* **2000**, *17*, 79–103. [CrossRef]
9. Eboli, L.; Mazzulla, G. Relationships between Rail Passengers' Satisfaction and Service Quality: A Framework for Identifying Key Service Factors. *Public Transp.* **2015**, *7*, 185–201. [CrossRef]
10. dell'Olio, L.; Ibeas, Á.; Cecín, P. The Quality of Service Desired by Public Transport Users. *Transp. Policy* **2011**, *18*, 217–227. [CrossRef]
11. Agrawal, A.W.; Granger-Bevan, S.; Newmark, G.L.; Nixon, H. Comparing Data Quality and Cost from Three Modes of On-Board Transit Surveys. *Transp. Policy* **2017**, *54*, 70–79. [CrossRef]
12. Axhausen, K.W.; Löchl, M.; Schlich, R.; Buhl, T.; Widmer, P. Fatigue in Long-Duration Travel Diaries. *Transportation* **2007**, *34*, 143–160. [CrossRef]
13. Carreira, R.; Patrício, L.; Natal Jorge, R.; Magee, C. Understanding the Travel Experience and Its Impact on Attitudes, Emotions and Loyalty towards the Transportation Provider—A Quantitative Study with Mid-Distance Bus Trips. *Transp. Policy* **2014**, *31*, 35–46. [CrossRef]
14. Sarker, R.I.; Kaplan, S.; Mailer, M.; Timmermans, H.J.P. Applying Affective Event Theory to Explain Transit Users' Reactions to Service Disruptions. *Transp. Res. Part A Policy Pract.* **2019**, *130*, 593–605. [CrossRef]
15. Gong, S.-H.; Teng, J.; Duan, C.-Y.; Liu, S.-J. Framework for Evaluating Online Public Opinions on Urban Rail Transit Services through Social Media Data Classification and Mining. *Res. Transp. Bus. Manag.* **2024**, *56*, 101197. [CrossRef]
16. Li, C.; Zegras, P.C.; Zhao, F.; Qin, Z.; Shahid, A.; Ben-Akiva, M.; Pereira, F.; Zhao, J. Enabling Bus Transit Service Quality Co-Monitoring through Smartphone-Based Platform. *Transp. Res. Rec.* **2017**, *2649*, 42–51. [CrossRef]
17. Dou, M.; Gu, Y.; Gong, J. How Do People Perceive the Quality of Urban Transport Service? New Insights from Online Reviews of Shanghai Metro System. *J. Urban Manag.* **2024**, *13*, 705–719. [CrossRef]
18. Collins, C.; Hasan, S.; Ukkusuri, S. A Novel Transit Rider Satisfaction Metric: Rider Sentiments Measured from Online Social Media Data. *J. Public Transp.* **2013**, *16*, 21–45. [CrossRef]
19. Osorio-Arjona, J.; Horak, J.; Svoboda, R.; García-Ruíz, Y. Social Media Semantic Perceptions on Madrid Metro System: Using Twitter Data to Link Complaints to Space. *Sustain. Cities Soc.* **2021**, *64*, 102530. [CrossRef]
20. Luo, S.; He, S.Y. Using Data Mining to Explore the Spatial and Temporal Dynamics of Perceptions of Metro Services in China: The Case of Shenzhen. *Environ. Plan. B Urban Anal. City Sci.* **2021**, *48*, 449–466. [CrossRef]

21. Casas, I.; Delmelle, E.C. Tweeting about Public Transit—Gleaning Public Perceptions from a Social Media Microblog. *Case Stud. Transp. Policy* **2017**, *5*, 634–642. [CrossRef]
22. Etu, E.-E.; Agrawal, A.W.; Larot, J.; Tatipigari, C.; Tenebe, I.T. Learning about the Transit Passenger Experience from Microblogging Posts. *Findings* **2025**, *2025*. [CrossRef]
23. Song, C.; Guo, J.; Zhuang, J. Analyzing Passengers' Emotions Following Flight Delays—A 2011–2019 Case Study on SKYTRAX Comments. *J. Air Transp. Manag.* **2020**, *89*, 101903. [CrossRef]
24. Martin-Domingo, L.; Martín, J.C.; Mandsberg, G. Social Media as a Resource for Sentiment Analysis of Airport Service Quality (ASQ). *J. Air Transp. Manag.* **2019**, *78*, 106–115. [CrossRef]
25. Wang, Y.; Yang, M.; Cheng, L.; Qin, B.; Zhang, R. Investigating Service Perception Dynamics of Transportation Hub Users through Social Media Data. *Transp. Res. Part D Transp. Environ.* **2025**, *142*, 104716. [CrossRef]
26. Leong, M.; Abdelhalim, A.; Ha, J.; Patterson, D.; Pincus, G.L.; Harris, A.B.; Eichler, M.; Zhao, J. MetRoBERTa: Leveraging Traditional Customer Relationship Management Data to Develop a Transit-Topic-Aware Language Model. *Transp. Res. Rec.* **2024**, *2678*, 215–229. [CrossRef]
27. Korfiatis, N.; Stamolampros, P.; Kourouthanassis, P.; Sagiadinos, V. Measuring Service Quality from Unstructured Data: A Topic Modeling Application on Airline Passengers' Online Reviews. *Expert Syst. Appl.* **2019**, *116*, 472–486. [CrossRef]
28. Kyriakidis, A.; Tsafarakis, S. Extracting Knowledge from Customer Reviews: An Integrated Framework for Digital Platform Analytics. *Int. Trans. Oper. Res.* **2025**, *32*, 2061–2086. [CrossRef]
29. Luo, S.; He, S.Y.; Song, L.; Jian, S.; Yao, Y. Exploring the Potential of Large Language Models (LLMs) in Analyzing Passengers' Perceptions of Transit Service Quality. *Environ. Plan. B Urban Anal. City Sci.* **2026**, *53*, 90–106. [CrossRef]
30. Wandelt, S.; Zheng, C.; Wang, S.; Liu, Y.; Sun, X. Large Language Models for Intelligent Transportation: A Review of the State of the Art and Challenges. *Appl. Sci.* **2024**, *14*, 7455. [CrossRef]
31. Xie, H.; Li, Y.; Pu, Y.; Zhang, C.; Huang, J. Evaluating Airline Service Quality through a Comprehensive Text-Mining and Multi-Criteria Decision-Making Analysis. *J. Air Transp. Manag.* **2024**, *120*, 102655. [CrossRef]
32. Raad, N.; Fick, J.; Shah, N.; Srinivas, S. From Voices to Value: Leveraging Online Reviews to Benchmark Airline Efficiency with GenAI-Assisted Text Analytics and DEA. *J. Oper. Res. Soc.* **2025**, *77*, 1843–1866. [CrossRef]
33. El-Diraby, T.; Shalaby, A.; Hosseini, M. Linking Social, Semantic and Sentiment Analyses to Support Modeling Transit Customers' Satisfaction: Towards Formal Study of Opinion Dynamics. *Sustain. Cities Soc.* **2019**, *49*, 101578. [CrossRef]
34. Luo, S.; He, S.Y.; Grant-Muller, S.; Song, L. Influential Factors in Customer Satisfaction of Transit Services: Using Crowdsourced Data to Capture the Heterogeneity across Individuals, Space and Time. *Transp. Policy* **2023**, *131*, 173–183. [CrossRef]
35. Tran, M.; Draeger, C.; Wang, X.; Nikbakht, A. Monitoring the Well-Being of Vulnerable Transit Riders Using Machine Learning Based Sentiment Analysis and Social Media: Lessons from COVID-19. *Environ. Plan. B Urban Anal. City Sci.* **2023**, *50*, 60–75. [CrossRef] [PubMed]
36. Parasuraman, A.; Zeithaml, V.A.; Berry, L.L. SERVQUAL A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. *J. Retail.* **1988**, *64*, 12–40.
37. Oliver, R.L. A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *J. Mark. Res.* **1980**, *17*, 460–469. [CrossRef]
38. Anderson, E.W. Customer Satisfaction and Word of Mouth. *J. Serv. Res.* **1998**, *1*, 5–17. [CrossRef]
39. Hosseini, M.; El-Diraby, T.; Shalaby, A. Supporting Sustainable System Adoption: Socio-Semantic Analysis of Transit Rider Debates on Social Media. *Sustain. Cities Soc.* **2018**, *38*, 123–136. [CrossRef]
40. Grant-Muller, S.M.; Gal-Tzur, A.; Minkov, E.; Kuflik, T.; Nocera, S.; Shoor, I. Transport Policy: Social Media and User-Generated Content in a Changing Information Paradigm. In *Social Media for Government Services*; Nepal, S., Paris, C., Georgakopoulos, D., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 325–366.
41. Lock, O.; Pettit, C. Social Media as Passive Geo-Participation in Transportation Planning—How Effective Are Topic Modeling & Sentiment Analysis in Comparison with Citizen Surveys? *Geo-Spat. Inf. Sci.* **2020**, *23*, 275–292. [CrossRef]
42. Haghighi, N.N.; Liu, X.C.; Wei, R.; Li, W.; Shao, H. Using Twitter Data for Transit Performance Assessment: A Framework for Evaluating Transit Riders' Opinions about Quality of Service. *Public Transp.* **2018**, *10*, 363–377. [CrossRef]
43. Das, S.; Zubaidi, H.A. City Transit Rider Tweets: Understanding Sentiments and Politeness. *J. Urban Technol.* **2023**, *30*, 111–126. [CrossRef]
44. Srinivas, S.; Ramachandiran, S. Passenger Intelligence as a Competitive Opportunity: Unsupervised Text Analytics for Discovering Airline-Specific Insights from Online Reviews. *Ann. Oper. Res.* **2024**, *333*, 1045–1075. [CrossRef]
45. Barberá, P.; Rivero, G. Understanding the Political Representativeness of Twitter Users. *Soc. Sci. Comput. Rev.* **2015**, *33*, 712–729. [CrossRef]
46. Mellon, J.; Prosser, C. Twitter and Facebook Are Not Representative of the General Population: Political Attitudes and Demographics of British Social Media Users. *Res. Politics* **2017**, *4*, 2053168017720008. [CrossRef]

47. Torres, E.C.M.; De Picado-Santos, L.G. Sentiment Analysis and Topic Modeling in Transportation: A Literature Review. *Appl. Sci.* **2025**, *15*, 6576. [CrossRef]
48. Vencovský, F. Service Quality Evaluation Using Text Mining: A Systematic Literature Review. In *Perspectives in Business Informatics Research*; Buchmann, R.A., Polini, A., Johansson, B., Karagiannis, D., Eds.; Lecture Notes in Business Information Processing; Springer International Publishing: Cham, Switzerland, 2020; Volume 398, pp. 159–173.
49. Nikolaidou, A.; Papaioannou, P. Utilizing Social Media in Transport Planning and Public Transit Quality: Survey of Literature. *J. Transp. Eng. Part A Syst.* **2018**, *144*, 04018007. [CrossRef]
50. Güner, S.; Taşkın, K.; Cebeci, H.İ.; Aydemir, E. Service Quality in Rail Systems: Listen to the Voice of Social Media. *Transp. Res. Rec.* **2024**, *2678*, 824–847. [CrossRef]
51. Al-Sahar, R.; Klumpenhouwer, W.; Shalaby, A.; El-Diraby, T. Using Twitter to Gauge Customer Satisfaction Response to a Major Transit Service Change in Calgary, Canada. *Transp. Res. Rec.* **2024**, *2678*, 190–206. [CrossRef] [PubMed]
52. Kuflik, T.; Minkov, E.; Nocera, S.; Grant-Muller, S.; Gal-Tzur, A.; Shoor, I. Automating a Framework to Extract and Analyse Transport Related Social Media Content: The Potential and the Challenges. *Transp. Res. Part C Emerg. Technol.* **2017**, *77*, 275–291. [CrossRef]
53. Gao, L.; Yu, Y.; Liang, W. Public Transit Customer Satisfaction Dimensions Discovery from Online Reviews. *Urban Rail Transit* **2016**, *2*, 146–152. [CrossRef]
54. Qi, B.; Costin, A.; Jia, M. A Framework with Efficient Extraction and Analysis of Twitter Data for Evaluating Public Opinions on Transportation Services. *Travel Behav. Soc.* **2020**, *21*, 10–23. [CrossRef]
55. Rashidi, T.H.; Abbasi, A.; Maghrebi, M.; Hasan, S.; Waller, T.S. Exploring the Capacity of Social Media Data for Modelling Travel Behaviour: Opportunities and Challenges. *Transp. Res. Part C Emerg. Technol.* **2017**, *75*, 197–211. [CrossRef]
56. Li, Z.; Yang, C.; Huang, C. A Comparative Sentiment Analysis of Airline Customer Reviews Using Bidirectional Encoder Representations from Transformers (BERT) and Its Variants. *Mathematics* **2024**, *12*, 53. [CrossRef]
57. Li, D.; Zhang, Y.; Li, C. Mining Public Opinion on Transportation Systems Based on Social Media Data. *Sustainability* **2019**, *11*, 4016. [CrossRef]
58. Hoang, T.; Cher, P.H.; Prasetyo, P.K.; Lim, E.-P. Crowdsensing and Analyzing Micro-Event Tweets for Public Transportation Insights. In *Proceedings of the 2016 IEEE International Conference on Big Data (Big Data)*; IEEE: Washington, DC, USA, 2016; pp. 2157–2166.
59. Kabbani, O.; Klumpenhouwer, W.; El-Diraby, T.; Shalaby, A. What Do Riders Say and Where? The Detection and Analysis of Eyewitness Transit Tweets. *J. Intell. Transp. Syst.* **2023**, *27*, 347–363. [CrossRef]
60. Guo, Y.; Sun, S.; Schuckert, M.; Law, R. Online Feedback and Attraction Management: An Exploration of the Critical Factors in Effective Operations. *Asia Pac. J. Tour. Res.* **2016**, *21*, 883–904. [CrossRef]
61. Garcia-Ledezma, H.; Zubieta, J. Twitter in Improving the Quality of a BRT. *TQM J.* **2025**, *38*, 153–174. [CrossRef]
62. Rahman, R.; Redwan Shabab, K.; Chandra Roy, K.; Zaki, M.H.; Hasan, S. Real-Time Twitter Data Mining Approach to Infer User Perception Toward Active Mobility. *Transp. Res. Rec.* **2021**, *2675*, 947–960. [CrossRef]
63. Monmousseau, P.; Marzuoli, A.; Feron, E.; Delahaye, D. Impact of Covid-19 on Passengers and Airlines from Passenger Measurements: Managing Customer Satisfaction While Putting the US Air Transportation System to Sleep. *Transp. Res. Interdiscip. Perspect.* **2020**, *7*, 100179. [CrossRef] [PubMed]
64. Schweitzer, L. Planning and Social Media: A Case Study of Public Transit and Stigma on Twitter. *J. Am. Plan. Assoc.* **2014**, *80*, 218–238. [CrossRef]
65. Brochado, A.; Rita, P.; Oliveira, C.; Oliveira, F. Airline Passengers' Perceptions of Service Quality: Themes in Online Reviews. *Int. J. Contemp. Hosp. Manag.* **2019**, *31*, 855–873. [CrossRef]
66. Chatterjee, S. Explaining Customer Ratings and Recommendations by Combining Qualitative and Quantitative User Generated Contents. *Decis. Support Syst.* **2019**, *119*, 14–22. [CrossRef]
67. Lucini, F.R.; Tonetto, L.M.; Fogliatto, F.S.; Anzanello, M.J. Text Mining Approach to Explore Dimensions of Airline Customer Satisfaction Using Online Customer Reviews. *J. Air Transp. Manag.* **2020**, *83*, 101760. [CrossRef]
68. Sezgen, E.; Mason, K.J.; Mayer, R. Voice of Airline Passenger: A Text Mining Approach to Understand Customer Satisfaction. *J. Air Transp. Manag.* **2019**, *77*, 65–74. [CrossRef]
69. Stamolampros, P.; Dousios, D.; Korfiatis, N. Evaluating Domestic Bias on Airline Passengers' Ratings: The Moderating Effect of Cultural Value Orientation. *Int. J. Hosp. Manag.* **2020**, *87*, 102466. [CrossRef]
70. Aman, J.J.C.; Smith-Colin, J. Application of Crowdsourced Data to Infer User Satisfaction with Mobility as a Service (MaaS). *Transp. Res. Interdiscip. Perspect.* **2022**, *15*, 100672. [CrossRef]
71. Aman, J.J.C.; Smith-Colin, J.; Zhang, W. Listen to E-Scooter Riders: Mining Rider Satisfaction Factors from App Store Reviews. *Transp. Res. Part D Transp. Environ.* **2021**, *95*, 102856. [CrossRef]
72. Bardutz, E.; Bigazzi, A. Communicating Perceptions of Pedestrian Comfort and Safety: Structural Topic Modeling of Open Response Survey Comments. *Transp. Res. Interdiscip. Perspect.* **2022**, *14*, 100600. [CrossRef]

73. Liu, X.; Ye, Q.; Li, Y.; Fan, J.; Tao, Y. Examining Public Concerns and Attitudes toward Unfair Events Involving Elderly Travelers during the COVID-19 Pandemic Using Weibo Data. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1756. [CrossRef] [PubMed]
74. Gal-Tzur, A.; Bar-Lev, S.; Shiftan, Y. Using Question & Answer Forums as a Platform for Improving Transport-Related Information for Tourists. *J. Travel Res.* **2020**, *59*, 1221–1237. [CrossRef]
75. Pineda-Jaramillo, J.; Pineda-Jaramillo, D. Analysing Travel Satisfaction of Tourists towards a Metro System from Unstructured Data. *Res. Transp. Bus. Manag.* **2022**, *43*, 100746. [CrossRef]
76. Vasquez-Henriquez, P.; Graells-Garrido, E.; Caro, D. Characterizing Transport Perception Using Social Media: Differences in Mode and Gender. In *Proceedings of the 10th ACM Conference on Web Science*; ACM: Boston, MA, USA, 2019; pp. 295–299.
77. Ali, F.; Kwak, D.; Khan, P.; El-Sappagh, S.; Ali, A.; Ullah, S.; Kim, K.H.; Kwak, K.-S. Transportation Sentiment Analysis Using Word Embedding and Ontology-Based Topic Modeling. *Knowl.-Based Syst.* **2019**, *174*, 27–42. [CrossRef]
78. Zhang, C.; Li, W.; Shou, J.; Ma, H.; Xiang, J.; Skitmore, M.; Liu, X.; Rong, D.; Yao, W. Exploring Bus Service Demand in Emerging Business Formats: A Text Mining Analysis. *Transp. Policy* **2026**, *179*, 104010. [CrossRef]
79. Mendez, J.T.; Lobel, H.; Parra, D.; Herrera, J.C. Using Twitter to Infer User Satisfaction with Public Transport: The Case of Santiago, Chile. *IEEE Access* **2019**, *7*, 60255–60263. [CrossRef]
80. Gupta, M.; Singh, A.; Jain, R.; Saxena, A.; Ahmed, S. Multi-Class Railway Complaints Categorization Using Neural Networks: RailNeural. *J. Rail Transp. Plan. Manag.* **2021**, *20*, 100265. [CrossRef]
81. Rani, M.S.; Sumathy, S. Opinion Mining on Social Media Transit Tweets Using Text Pre-Processing and Machine Learning Techniques. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *9*, 1015–1025. [CrossRef]
82. Gao, S.; Ran, Q.; Su, Z.; Wang, L.; Ma, W.; Hao, R. Evaluation System for Urban Traffic Intelligence Based on Travel Experiences: A Sentiment Analysis Approach. *Transp. Res. Part A Policy Pract.* **2024**, *187*, 104170. [CrossRef]
83. Rita, P.; Moro, S.; Cavalcanti, G. The Impact of COVID-19 on Tourism: Analysis of Online Reviews in the Airlines Sector. *J. Air Transp. Manag.* **2022**, *104*, 102277. [CrossRef]
84. Mogaji, E.; Erkan, I. Insight into Consumer Experience on UK Train Transportation Services. *Travel Behav. Soc.* **2019**, *14*, 21–33. [CrossRef]
85. Mishra, D.N.; Panda, R.K. Decoding Customer Experiences in Rail Transport Service: Application of Hybrid Sentiment Analysis. *Public Transp.* **2023**, *15*, 31–60. [CrossRef]
86. Rane, A.; Kumar, A. Sentiment Classification System of Twitter Data for US Airline Service Analysis. In *Proceedings of the 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*; IEEE: New York, NY, USA, 2018; pp. 769–773.
87. Şencan, Ö.A.; Atacak, İ. Social Media User Opinion Analysis Using Deep Learning and Machine Learning Methods: A Case Study on Airlines. *Turk. J. Math. Comput. Sci.* **2023**, *15*, 449–463. [CrossRef]
88. Chang, Y.-C.; Ku, C.-H.; Nguyen, D.-D.L. Predicting Aspect-Based Sentiment Using Deep Learning and Information Visualization: The Impact of COVID-19 on the Airline Industry. *Inf. Manag.* **2022**, *59*, 103587. [CrossRef]
89. Kiliç, S.; Çadirci, T.O. An Evaluation of Airport Service Experience: An Identification of Service Improvement Opportunities Based on Topic Modeling and Sentiment Analysis. *Res. Transp. Bus. Manag.* **2022**, *43*, 100744. [CrossRef]
90. Li, X.; He, G.; Guo, P.; Guo, Z.; Lin, S.; Du, S. Topic Modeling Help Enhancing Sustainable Mobility. *J. Clean. Prod.* **2025**, *534*, 147068. [CrossRef]
91. Mansouri, S.; Moradpour, S.; Prentice, C. An Alternative Approach to Understanding Airline Customers' Attitudes and Behaviors. *J. Air Transp. Manag.* **2026**, *133*, 102960. [CrossRef]
92. Agrawal, A.; Kuriakose, P.N. Implications of a Twitter Data-Centred Methodology for Assessing Commuters' Perceptions of the Delhi Metro in India. *Comput. Urban Sci.* **2022**, *2*, 38. [CrossRef] [PubMed]
93. Siering, M.; Deokar, A.V.; Janze, C. Disentangling Consumer Recommendations: Explaining and Predicting Airline Recommendations Based on Online Reviews. *Decis. Support Syst.* **2018**, *107*, 52–63. [CrossRef]
94. Serna, A.; Gerrikagoitia, J.K.; Bernabé, U.; Ruiz, T. Sustainability Analysis on Urban Mobility Based on Social Media Content. *Transp. Res. Procedia* **2017**, *24*, 1–8. [CrossRef]
95. Lee, K.; Yu, C. Assessment of Airport Service Quality: A Complementary Approach to Measure Perceived Service Quality Based on Google Reviews. *J. Air Transp. Manag.* **2018**, *71*, 28–44. [CrossRef]
96. Barakat, H.; Yeniterzi, R.; Martin-Domingo, L. Applying Deep Learning Models to Twitter Data to Detect Airport Service Quality. *J. Air Transp. Manag.* **2021**, *91*, 102003. [CrossRef]
97. Rasool, G.; Pathania, A. Reading between the Lines: Untwining Online User-Generated Content Using Sentiment Analysis. *J. Res. Interact. Mark.* **2021**, *15*, 401–418. [CrossRef]
98. Gitto, S.; Mancuso, P. Improving Airport Services Using Sentiment Analysis of the Websites. *Tour. Manag. Perspect.* **2017**, *22*, 132–136. [CrossRef]
99. Yu, L.J.; Saad, N.H.M.; Kun, Z.; El-Sayad, G. Evaluating Airline Passengers' Satisfaction During the COVID-19 Pandemic: A Case Study of AirAsia Services Through Sentiment Analysis and Topic Modelling. *CommIT J.* **2024**, *18*, 137–150. [CrossRef]

100. Aman, J.J.C.; Smith-Colin, J. Leveraging Social Media to Understand Public Perceptions toward Micromobility Policies: The Dallas Scooter Ban Case. *Findings* **2021**. [CrossRef]
101. Wang, H.; Chen, Y. Research on the Classification and Optimization Strategies of Civil Aviation Customer Service Based on BERTopic and the Kano Model. *J. Air Transp. Manag.* **2026**, *133*, 102959. [CrossRef]
102. Mahmud, T.; Uddin, B.; Akter, T.; Majumder, A.; Aziz, M.F.B.A.; Shahadat Hossain, M.; Andersson, K. Exploring Transformer Models for Sentiment Analysis in Airline Service Reviews. In *Proceedings of the 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*; IEEE: New York, NY, USA, 2024; pp. 1–7.
103. Benaddi, L.; Ouaddi, C.; Naimi, L.; Souha, A.; Jakimi, A.; Ouchao, B. A Decision Support System for Enhancing Transportation Services Using Aspect-Based Sentiment Analysis. In Proceedings of the 2024 International Conference on Decision Aid Sciences and Applications (DASA), Manama, Bahrain, 11–12 December 2024; pp. 1–5.
104. Çallı, L.; Çallı, F. Understanding Airline Passengers during Covid-19 Outbreak to Improve Service Quality: Topic Modeling Approach to Complaints with Latent Dirichlet Allocation Algorithm. *Transp. Res. Rec. J. Transp. Res. Board* **2023**, *2677*, 656–673. [CrossRef] [PubMed]
105. Alamsyah, A.; Rachmadiansyah, I. Mapping Online Transportation Service Quality and Multiclass Classification Problem Solving Priorities. *Proc. J. Phys. Conf. Ser.* **2018**, *971*, 012021. [CrossRef]
106. Liu, S.-J.; Teng, J.; Gong, Y. Extraction Method and Integration Framework for Perception Features of Public Opinion in Transportation. *Sustainability* **2021**, *13*, 254. [CrossRef]
107. Kamga, C.; Kish, R.; Mudigonda, S.; Tchamna, R. Utilizing Social Media Data for Estimating Transit Performance Metrics. *Sustainability* **2023**, *15*, 16183. [CrossRef]
108. Wang, X.; Zheng, J.; Tang, L.; Luo, Y. Recommend or Not? The Influence of Emotions on Passengers' Intention of Airline Recommendation during COVID-19. *Tour. Manag.* **2023**, *95*, 104675. [CrossRef] [PubMed]
109. Farzadnia, S.; Raeesi Vanani, I. Identification of Opinion Trends Using Sentiment Analysis of Airlines Passengers' Reviews. *J. Air Transp. Manag.* **2022**, *103*, 102232. [CrossRef]
110. Beck, D.; Teixeira, M.; Ferasso, M. From Commute to Contentment: Evaluating Service Quality in Lisbon's Rapid Transit System through Non-Solicited User Feedback. *Case Stud. Transp. Policy* **2025**, *20*, 101460. [CrossRef]
111. Lu, L.; Mitra, A.; Wang, Y.-Y.; Wang, Y.; Xu, P. Use of Electronic Word of Mouth as Quality Metrics: A Comparison of Airline Reviews on Twitter and Skytrax. In Proceedings of the Annual Hawaii International Conference on System Sciences, Maui, HI, USA, 4–7 January 2022; pp. 1349–1357.
112. Allen, J.; Muñoz, J.C.; Ortúzar, J.D.D. Modelling Service-Specific and Global Transit Satisfaction under Travel and User Heterogeneity. *Transp. Res. Part A Policy Pract.* **2018**, *113*, 509–528. [CrossRef]
113. Vasquez-Henriquez, P.; Graells-Garrido, E.; Caro, D. Tweets on the Go: Gender Differences in Transport Perception and Its Discussion on Social Media. *Sustainability* **2020**, *12*, 5405. [CrossRef]
114. Pratt, A.N.; Morris, E.A.; Zhou, Y.; Khan, S.; Chowdhury, M. What Do Riders Tweet about the People That They Meet? Analyzing Online Commentary about UberPool and Lyft Shared/Lyft Line. *Transp. Res. Part F Traffic Psychol. Behav.* **2019**, *62*, 459–472. [CrossRef]
115. Taylor, S.E. Asymmetrical Effects of Positive and Negative Events: The Mobilization-Minimization Hypothesis. *Psychol. Bull.* **1991**, *110*, 67–85. [CrossRef] [PubMed]
116. Dike, S.E.; Davis, Z.; Abrahams, A.; Anjomshoae, A.; Ractham, P. Evaluation of Passengers' Expectations and Satisfaction in the Airline Industry: An Empirical Performance Analysis of Online Reviews. *Benchmarking Int. J.* **2024**, *31*, 611–639. [CrossRef]
117. Ho, S.; Choi, K.; Yang, F. Harnessing Aspect-Based Sentiment Analysis: How Are Tweets Associated with Forecast Accuracy? *J. Assoc. Inf. Syst.* **2019**, *20*, 1174–1209. [CrossRef]
118. Kim, D.; Lim, C.; Ha, H.-K. Comparative Analysis of Changes in Passenger's Perception for Airline Companies' Service Quality before and during COVID-19 Using Topic Modeling. *J. Air Transp. Manag.* **2024**, *115*, 102542. [CrossRef]
119. Lim, J.; Lee, M.; Jang, J. Revisiting Passengers' Perceptions of Airline Service Quality: A Theory-Driven Machine Learning Approach Using Big Data. *J. Qual. Assur. Hosp. Tour.* **2024**, 1–22. [CrossRef]
120. Chen, X.; Wang, Z.; Di, X. Sentiment Analysis on Multimodal Transportation during the COVID-19 Using Social Media Data. *Information* **2023**, *14*, 113. [CrossRef]
121. Senthilnathan, V.; Rajendran, S. Understanding the Voice of Customers and Employees in Ride-Sharing Companies Using Text Analytics. *Transp. Res. Interdiscip. Perspect.* **2024**, *23*, 101002. [CrossRef]
122. Schnieder, M.; Dora, M. Trains or Planes? An Aspect-Based Sentiment Analysis of Travel Satisfaction. In Proceedings of the 2024 IEEE International Conference on Smart Mobility (SM), Niagara Falls, ON, Canada, 16–18 September 2024; pp. 242–247.
123. Myneni, M.B.; Dandamudi, R. Harvesting Railway Passenger Opinions on Multi Themes by Using Social Graph Clustering. *J. Rail Transp. Plan. Manag.* **2020**, *13*, 100151. [CrossRef]
124. Politis, I.; Georgiadis, G.; Kopsacheilis, A.; Nikolaidou, A.; Papaioannou, P. Capturing Twitter Negativity Pre- vs. Mid-COVID-19 Pandemic: An LDA Application on London Public Transport System. *Sustainability* **2021**, *13*, 13356. [CrossRef]

125. Gupta, M.; Kumar, R.; Walia, H.; Kaur, G. Airlines Based Twitter Sentiment Analysis Using Deep Learning. In Proceedings of the 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 22–23 October 2021; pp. 1–6.
126. Ram, Y.; Gal-Tzur, A.; Rechavi, A. Identifying Attributes of Public Transport Services for Urban Tourists: A Data-Mining Method. *J. Transp. Geogr.* **2021**, *93*, 103069. [CrossRef]
127. Basu, B.; Sebastian, M.P.; Kar, A.K. What Affects User Experience of Shared Mobility Services? Insights from Integrating Signaling Theory and Value Framework. *Inf. Syst. Front.* **2025**, *27*, 1789–1811. [CrossRef]
128. Sulu, D.; Arasli, H.; Saydam, M.B. Air-Travelers' Perceptions of Service Quality during the COVID-19 Pandemic: Evidence from Tripadvisor Sites. *Sustainability* **2021**, *14*, 435. [CrossRef]
129. Hsiao, Y.-H.; Li, B.-X. Exploring Passenger Perceptions of Services and Sustainability via Online Review Analytics for Airport Assessing and Diagnosing. *J. Retail. Consum. Serv.* **2026**, *90*, 104664. [CrossRef]
130. Shin, E.J. A Comparative Study of Bike-Sharing Systems from a User's Perspective: An Analysis of Online Reviews in Three U.S. Regions between 2010 and 2018. *Int. J. Sustain. Transp.* **2021**, *15*, 908–923. [CrossRef]
131. Zulkarnain, Z.; Surjandari, I.; Wayasti, R.A. Sentiment Analysis for Mining Customer Opinion on Twitter: A Case Study of Ride-Hailing Service Provider. In Proceedings of the 2018 5th International Conference on Information Science and Control Engineering (ICISCE), Zhengzhou, China, 20–22 July 2018; pp. 512–516.
132. Yadav, A.; Chaudhary, A.; Gupta, A. Analyzing Airline Sentiments: A Transformer Model Approach. In Proceedings of the 2024 International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit), Greater Noida, India, 4–6 September 2024; pp. 125–130.
133. Zeng, L.; Chen, Z. A Neural Network-Based Model for Classifying the Opinions of Airline Passengers Published on Social Media. *Intell. Decis. Technol.* **2025**, 18724981251387992. [CrossRef]
134. Xu, Y.; Krishnakumari, P.; Yorke-Smith, N.; Hoogendoorn, S. Augmenting Ridership Data with Social Media Data to Analyse the Long-Term Effect of COVID-19 on Public Transport. In Proceedings of the 2023 8th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), Nice, France, 14–16 June 2023; pp. 1–6.
135. Misopoulos, F.; Mitic, M.; Kapoulas, A.; Karapiperis, C. Uncovering Customer Service Experiences with Twitter: The Case of Airline Industry. *Manag. Decis.* **2014**, *52*, 705–723. [CrossRef]
136. Almuayqil, S.N.; Humayun, M.; Jhanjhi, N.Z.; Almufareh, M.F.; Javed, D. Framework for Improved Sentiment Analysis via Random Minority Oversampling for User Tweet Review Classification. *Electronics* **2022**, *11*, 3058. [CrossRef]
137. Mahfudiyah, N.; Alamsyah, A. Understanding User Perception of Ride-Hailing Services Sentiment Analysis and Topic Modelling Using IndoBERT and BERTopic. In Proceedings of the 2023 International Conference on Digital Business and Technology Management, ICONDBTM 2023, Lombok, Indonesia, 2–3 August 2023.
138. Surjandari, I.; Wayasti, R.A.; Laoh, E.; Zulkarnain, Z.; Rus, A.M.M.; Prawiradinata, I. Mining Public Opinion on Ride-Hailing Service Providers Using Aspect-Based Sentiment Analysis. *Int. J. Technol.* **2019**, *10*, 818–828. [CrossRef]
139. Giancristofaro, G.T.; Panangadan, A. Predicting Sentiment toward Transportation in Social Media Using Visual and Textual Features. In Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 1–4 November 2016; pp. 2113–2118.
140. Dashtipour, K.; Poria, S.; Hussain, A.; Cambria, E.; Hawalah, A.Y.A.; Gelbukh, A.; Zhou, Q. Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. *Cogn. Comput.* **2016**, *8*, 757–771. [CrossRef] [PubMed]
141. Rahimi, M.M.; Naghizade, E.; Stevenson, M.; Winter, S. SentiHawkes: A Sentiment-Aware Hawkes Point Process to Model Service Quality of Public Transport Using Twitter Data. *Public Transp.* **2023**, *15*, 343–376. [CrossRef]
142. Welch, T.F.; Widita, A. Big Data in Public Transportation: A Review of Sources and Methods. *Transp. Rev.* **2019**, *39*, 795–818. [CrossRef]
143. Kulkarni, G.; Abellera, L.; Panangadan, A. Unsupervised Classification of Online Community Input to Advance Transportation Services. In Proceedings of the 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 8–10 January 2018; pp. 261–267.
144. Candelieri, A.; Archetti, F. Analyzing Tweets to Enable Sustainable, Multi-Modal and Personalized Urban Mobility: Approaches and Results from the Italian Project TAM-TAM. In Proceedings of the Urban Tranport 2014, Algarve, Portugal, 28–30 May 2014; pp. 373–379.
145. Singh, N.; Upreti, M. HMRFLR: A Hybrid Model for Sentiment Analysis of Social Media Surveillance on Airlines. *Wirel. Pers. Commun.* **2023**, *132*, 97–112. [CrossRef]
146. Ahmed, A.Z.; Rodríguez-Díaz, M. Significant Labels in Sentiment Analysis of Online Customer Reviews of Airlines. *Sustainability* **2020**, *12*, 8683. [CrossRef]
147. Kwon, H.-J.; Ban, H.-J.; Jun, J.-K.; Kim, H.-S. Topic Modeling and Sentiment Analysis of Online Review for Airlines. *Information* **2021**, *12*, 78. [CrossRef]

148. Park, J. Combined Text-Mining/DEA Method for Measuring Level of Customer Satisfaction from Online Reviews. *Expert Syst. Appl.* **2023**, *232*, 120767. [CrossRef]
149. Lu, L.; Xu, P.; Wang, Y.-Y.; Wang, Y. Measuring Service Quality with Text Analytics: Considering Both Importance and Performance of Consumer Opinions on Social and Non-Social Online Platforms. *J. Bus. Res.* **2023**, *169*, 114298. [CrossRef]
150. Nisar, T.M.; Prabhakar, G. Trains and Twitter: Firm Generated Content, Consumer Relationship Management and Message Framing. *Transp. Res. Part A Policy Pract.* **2018**, *113*, 318–334. [CrossRef]
151. Alanazi, M.S.M.; Li, J.; Jenkins, K.W. Multiclass Sentiment Prediction of Airport Service Online Reviews Using Aspect-Based Sentimental Analysis and Machine Learning. *Mathematics* **2024**, *12*, 781. [CrossRef]
152. Falatouri, T.; Hrušecká, D.; Fischer, T. Harnessing the Power of LLMs for Service Quality Assessment From User-Generated Content. *IEEE Access* **2024**, *12*, 99755–99767. [CrossRef]
153. Serna, A.; Soroa, A.; Agerri, R. Applying Deep Learning Techniques for Sentiment Analysis to Assess Sustainable Transport. *Sustainability* **2021**, *13*, 2397. [CrossRef]
154. Ce, P.; Tie, B. An Analysis Method for Interpretability of CNN Text Classification Model. *Future Internet* **2020**, *12*, 228. [CrossRef]
155. Ruan, K.; Wang, X.; Di, X. From Twitter to Reasoner: Understand Mobility Travel Modes and Sentiment Using Large Language Models. In Proceedings of the 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), Edmonton, AB, Canada, 24–27 September 2024; pp. 454–459.