

Article

Quantifying Airline Reputation from Multilingual Online Content: Artificial Intelligence and a Practical Application in a Lightweight Reputation-Intelligence Framework

Luís F. F. M. Santos ^{1,2} 

¹ School of Management, Engineering and Aeronautics, ISEC Lisboa, Alameda das Linhas de Torres 179, 1750-142 Lisbon, Portugal; luis.santos@iseclisboa.pt

² AEROG-LAETA—Aeronautics and Astronautics Research Center, Universidade da Beira Interior, 6201-001 Covilhã, Portugal

Abstract

This paper presents “Quantifying Airline Reputation from Multilingual Online Content: Artificial Intelligence and a Practical Application in a Lightweight Reputation-Intelligence Framework.” Airline reputation is increasingly shaped by multilingual digital narratives that evolve faster than conventional survey cycles, creating a need for timely and interpretable monitoring tools. This study develops and evaluates a lightweight reputation-intelligence framework that integrates brand-safe retrieval, multilingual transformer-based sentiment inference, zero-shot natural-language-inference aspect categorization, TF-IDF/KMeans topic induction, and short-horizon forecasting. The framework formalizes document-level outputs into managerial indicators, including a Net Sentiment Index, polarity shares, aspect scores, topic summaries, and projected sentiment trajectories. On a 3990-document annotated sentiment subset, the multilingual transformer achieved 0.9015 accuracy, 0.9048 macro-F1, 0.9050 weighted-F1, a Cohen’s kappa of 0.8492, and a Net Sentiment Index of 48.5%, while errors were concentrated between adjacent polarity classes. Aspect evaluation showed that supervised in-domain learning substantially outperformed zero-shot inference, clarifying the trade-off between cold-start portability and benchmark accuracy. The results support the framework as a pilot decision-support architecture for airline reputation sensing rather than as a definitive large-scale benchmark. The approach offers scalable and CPU-friendly monitoring capability for airlines, airports, consultants, and public-sector users, while future work should expand multilingual annotation and domain adaptation.

Keywords: air transport management; airline reputation; multilingual sentiment analysis; transformers; zero-shot classification; artificial intelligence; transport economics; marketing



Academic Editor: Jose María Alvarez Rodríguez

Received: 2 June 2026

Revised: 23 June 2026

Accepted: 30 June 2026

Published: 2 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Airline reputation is increasingly produced in a continuous public information environment rather than in isolated survey waves. Customers, employees, journalists, regulators, and other stakeholders generate multilingual narratives at high frequency, and those narratives reinterpret operational events in near real time. For this reason, reputation in aviation is better understood as a dynamic informational state than as a static communication attribute [1–6].

The strategic significance of this state is as much economic as communicational. In air transport markets, perceived service quality and reliability influence customer acquisition,

willingness to pay, loyalty retention, and resilience after disruptions. Prior research on electronic word of mouth shows that public opinion signals can shape demand and valuation outcomes, particularly when customer uncertainty is high [7–9]. Airlines are especially exposed to these effects because purchase occurs before consumption and service failures are highly visible [10,11].

Conventional survey instruments remain valuable, but they are structurally limited for high-frequency sensing. They are periodic, costly, and often delayed relative to rapidly evolving operational narratives. Digital text streams provide much greater temporal and thematic granularity, yet they also introduce methodological risk because language, source structure, and event intensity vary substantially across contexts [12–14].

The central challenge is therefore systems integration. A robust airline monitoring framework must scan the internet for query-relevant information, enforce brand-safe filtering, infer multilingual sentiment and aspects, induce interpretable topics, and aggregate outputs into decision-ready indicators. Isolated model components are insufficient if their outputs cannot be linked through transparent definitions and consistent temporal logic.

This research addresses that challenge by proposing a mathematically explicit, implementation-grounded, and computationally lightweight pilot framework for airline reputation intelligence. The framework combines rule-constrained retrieval, transformer sentiment inference, zero-shot natural language inference (NLI) aspect assignment, topic induction, and short-horizon forecasting to construct a time-indexed reputation state. Rather than introducing a new Natural Language Processing (NLP) or forecasting algorithm, it adapts and integrates generic computational components into an applied airline-reputation setting. The goal is to connect methodological rigor with managerial interpretability while preserving deployability on ordinary CPU-based hardware without dependence on dedicated GPU infrastructure.

The scientific contribution is therefore architectural and measurement-oriented rather than algorithm-inventive. The paper contributes a formal systems design that specifies how heterogeneous online aviation narratives are acquired, filtered, inferred, aggregated, benchmarked, and translated into transport-management indicators under explicit computational constraints. This positioning is important because many airline-reputation studies evaluate isolated NLP components, whereas managerial decision support requires an auditable chain from source selection to actionable indicators.

The main contributions are:

- (i) Formalization of a complete end-to-end reputation intelligence model;
- (ii) Translation and adaptation of generic model components into a reproducible and lightweight computational architecture;
- (iii) Definition of a multi-dimensional aspect structure suitable for airline management and transport economics;
- (iv) Derivation of interpretable aggregate indicators from document-level outputs;
- (v) Integration of CPU-friendly analytical components that lower adoption barriers for organizations without GPU-dependent infrastructure; and
- (vi) Specification of a validation pathway for settings without an initial gold-standard corpus.

These contributions are intended to be cumulative: the framework is useful not because any single module is novel in isolation but because the modules are connected through explicit equations, retained metadata, model comparisons, and managerial interpretation rules. The resulting contribution is a reproducible decision-support methodology for airline reputation sensing that can be inspected, reproduced, and progressively strengthened as larger labeled corpora become available.

The remainder of the paper is organized as follows. Section 1 synthesizes the interdisciplinary foundations and identifies the unresolved research gap. Section 2 presents the

methodology and architecture that structure the reputation-intelligence pipeline. Section 3 reports the empirical results and the associated reporting artifacts. Section 4 discusses the managerial and methodological implications of the findings. Section 5 concludes the article, summarizes its main limitations, and outlines directions for future work.

Literature Review

Historically, airline service evaluation emerged from customer-satisfaction and perceived-service-quality traditions, where structured questionnaires and periodic instruments were used to model expectation–performance gaps and behavioral intentions [12,13,15]. These frameworks remain conceptually important because they define many constructs still used by airline managers, reliability, responsiveness, assurance, and value fairness, but they were designed for low-frequency measurement environments. As communication shifted from private complaint channels to continuously updated public text streams, the empirical basis of reputation analysis expanded beyond curated survey responses and became inherently more heterogeneous.

This shift mirrors the broader development of electronic word-of-mouth research. Early work on online feedback mechanisms demonstrated that digital evaluations alter information asymmetry and reduce search uncertainty for future consumers [7]. Subsequent studies showed measurable effects of online ratings and review valence on purchasing behavior, with stronger impacts in experience goods and high-risk service categories [8]. Importantly, market-oriented evidence also linked user-generated content to firm-level outcomes, including stock-market relevance, which supports the argument that online narratives are not merely communication artifacts but economically meaningful signals [9]. For aviation, this perspective is particularly salient because customer risk, uncertainty, and service heterogeneity are all structurally high.

Within airline studies, the empirical move from surveys to online text has been progressive rather than abrupt. A first wave used descriptive and frequency-based text mining to identify dominant themes in passenger feedback. A second wave introduced polarity classification, usually at document level, to summarize positive and negative reactions to carriers, routes, and service episodes. More recent work has evolved toward richer analytical goals: identifying latent service dimensions, linking sentiment to operational categories, and extracting managerial priorities from large corpora [2–6,16]. The trajectory of this literature suggests two robust findings: unsolicited text provides much higher topical granularity than traditional survey averages, and multilingual coverage is increasingly unavoidable for globally operating airlines.

Methodologically, this literature tracks the broader evolution of NLP. Lexicon-based approaches and shallow supervised models offered early interpretability and low computational cost, but they struggled with context sensitivity, domain shift, and multilingual transfer [17–19]. Deep learning then improved representation quality through distributed embeddings and sequence architectures, enabling better handling of compositional semantics and long-range dependencies [20–23]. Even so, pre-transformer models typically required substantial task-specific supervision and re-training, which limited portability across brands, languages, and aspect taxonomies.

Transformer architectures fundamentally altered this constraint set by enabling transfer learning at scale and stronger contextual representation [24–26]. In aviation, related safety-report studies show that NLP and machine-learning pipelines can classify human-factor categories and fatality-related outcomes using SVM, RF, ANN, KNN, Extra Trees, XGBoost, and BERT-adapted strategies [27–31]. Multilingual variants, especially XLM-RoBERTa (XLM-R), created a practical basis for cross-lingual sentiment inference in environments where labeled corpora are sparse for many languages [32]. Unified text-to-text

paradigms further normalized task transfer and flexible prompting [33]. Large-scale autoregressive and instruction-following models then expanded zero-shot and few-shot capabilities, making it possible to infer categories, summarize themes, and generate analytic explanations with minimal additional labeling [34–37]. In practical airline monitoring systems, however, this does not imply that one large model should replace all pipeline components. Instead, literature increasingly supports hybrid designs that combine strong semantic models with explicit aggregation and governance logic [38].

A central bridge between managerial usefulness and model capability is aspect-level interpretation. Airline decisions are rarely made on a single scalar sentiment score; they require decomposition by operational domain. Aspect-based sentiment analysis research provides the methodological foundation for this decomposition, from early unsupervised aspect discovery to benchmarked shared tasks and taxonomy-driven classification settings [39–41]. In aviation contexts, aspect definitions must be broader than hospitality-style categories because managers monitor not only service touchpoints but also labor narratives, safety communication, strategic announcements, and sustainability claims. This creates a recurring trade-off in the literature between taxonomy specificity (managerial relevance) and model portability (cross-brand and cross-lingual generalization).

Zero-shot NLI-based classification has emerged as an operational response to this trade-off. Instead of training a dedicated classifier for each new aspect ontology, the model estimates entailment between text and hypothesis templates, enabling rapid deployment under weak supervision [42]. The approach is attractive for airline intelligence programs because category sets evolve with regulatory agendas, macro shocks, and corporate strategy cycles. Yet the literature also warns that zero-shot outputs are sensitive to label phrasing, class imbalance, and domain mismatch. Accordingly, robust system design requires threshold calibration, human-in-the-loop auditing, and periodic back-testing against annotated subsets—an evaluation philosophy that aligns with reliability-centered transport analytics [43–46].

Topic extraction constitutes another crucial layer in the literature because it transforms thousands of granular comments into strategic themes. Probabilistic models such as Latent Dirichlet Allocation (LDA) established a canonical framework for latent thematic structure [47], while recent embedding-based methods improved semantic coherence in many settings [48]. Nevertheless, several applied studies report that term frequency–inverse document frequency (TF–IDF) plus centroid clustering remains competitively useful in production dashboards due to transparency, speed, and low maintenance requirements [49–51]. For airline management, where practitioners often need auditable and explainable outputs under tight reporting timelines, this pragmatic preference is understandable. The methodological implication is that “state-of-the-art” should be evaluated not only by benchmark accuracy but also by reproducibility, interpretability, and operational cost.

Another recurrent theme concerns data quality and platform heterogeneity. Public text streams differ in moderation rules, user demographics, language mix, bot prevalence, and temporal burstiness. These factors can distort sentiment distributions and topic prominence if not explicitly modeled [52,53]. In airline datasets, additional distortions arise from event-driven spikes (weather disruptions, strikes, safety incidents) and from source-specific discourse styles (journalistic articles versus short informal comments). Literature on social and review analytics therefore emphasizes preprocessing rigor, source metadata retention, and sensitivity analysis across time windows. This strengthens the case for pipeline architectures that separate ingestion, filtering, inference, and aggregation so that each stage can be audited and tuned independently.

From a transport-economics perspective, the central question is not whether online sentiment exists, but whether and how it maps to economically relevant outcomes. Airline

demand models already account for fares, schedule quality, network effects, and macro conditions [10,11]. Reputation enters this system as an informational mediator that can influence perceived quality, risk, and expected service consistency. In practical terms, negative reputation shocks may reduce conversion at unchanged prices, force discounting to restore load factors, or weaken premium-cabin willingness to pay. Positive reputation episodes may support price resilience, loyalty retention, and stronger recovery after operational disruptions. Although causal identification remains difficult in observational data, the accumulated evidence across digital markets and tourism supports treating online reputation as a leading indicator rather than a purely descriptive metric [8,9].

A further strand of recent literature emphasizes that operational usefulness depends not only on model choice but also on evaluation governance. In multilingual settings, benchmark scores can be misleading when label semantics, source composition, or language coverage differ substantially across studies. As a result, comparability requires explicit reporting of sampling frames, annotation protocols, error typologies, and confidence intervals, rather than accuracy alone [43–46]. The same concern appears in current discussions on foundation models: strong generative or zero-shot performance does not remove the need for calibration, domain adaptation, and auditability when outputs are used in organizational decision processes [32,37,38]. For airline intelligence programs, this implies that methodological rigor must extend beyond inference quality to include reproducible data lineage, stable metric definitions, and transparent exception handling. In other words, the literature increasingly treats reputation analytics as a socio-technical measurement system rather than a standalone text-classification task [54,55].

Related research in transport and service operations also highlights the importance of temporal structure. Reputation shocks are often event-driven and nonlinear: severe delays, cancellation waves, labor actions, safety incidents, or abrupt pricing controversies can generate asymmetric reactions that persist beyond the original trigger. Consequently, static average sentiment can obscure managerial risk if temporal concentration and recovery dynamics are ignored. Studies in airline economics and digital markets suggest that high-frequency narrative indicators are especially valuable when integrated with operational and commercial variables such as load factor, yield, route-level competition, and seasonality [7,9–11]. This perspective strengthens the argument for architectures that model time explicitly and support rolling-window analysis, cross-period comparability, and shock-sensitive diagnostics. It also motivates the design choice adopted in this paper: to combine multilingual semantic inference with interpretable aggregate indicators that can be aligned with managerial planning cycles and post-event review processes [1,6,16].

These interdisciplinary developments reveal a persistent gap in the current airline literature. Many studies still emphasize one language, one source type, or one task (often sentiment polarity), which limits guidance for integrated managerial monitoring. Other studies report strong model-level performance without explicitly linking outputs to decision-facing indicators. Conversely, practitioner reporting systems may provide useful operational summaries but without full methodological formalization. What remains underdeveloped is a fully integrated framework that connects internet-scoped multilingual acquisition, brand-safe filtering, sentiment and aspect inference, topic aggregation, and transport-economics interpretation within one coherent methodology.

The contributions are to address that gap by proposing a mathematically explicit and implementation-grounded framework for airline reputation intelligence. The literature suggests that no single methodological tradition is sufficient: service-quality theory provides managerial constructs, NLP provides scalable inference tools, and transport economics provides decision relevance. The contribution of this study is to integrate these strands into a reproducible architecture that supports both scientific evaluation and operational deployment.

Table 1 consolidates the main methodological families cited in this literature review and summarizes their relevance for the proposed airline reputation-intelligence framework. The table is intentionally organized by groups of related studies where the cited works share comparable objectives, methods, or limitations.

Table 1. Summary of cited literature, methodological advantages, and limitations.

Cited Literature	Main Advantages for This Study	Main Limitations or Gaps
Service-quality and satisfaction foundations [12,13,15]	Provide established constructs for perceived service quality, satisfaction, and behavioral consequences, giving managerial meaning to reputation indicators.	Designed mainly for structured and periodic measurement; they do not directly address high-frequency multilingual online text streams.
Digital word of mouth and market outcomes [7–9]	Demonstrate that online evaluations and user-generated content can be economically meaningful signals for demand, sales, and firm-level outcomes.	Evidence is not airline-specific in all cases and often requires careful treatment of platform bias, endogeneity, and causal interpretation.
Airline review and sentiment studies [1–6,16]	Confirm the relevance of online airline text for identifying satisfaction dimensions, sentiment patterns, emotions, review summaries, and customer attitudes.	Many studies remain bounded by a specific data source, language setting, review platform, or prediction task, leaving limited guidance for integrated monitoring pipelines.
Classical opinion mining and lexicon-based sentiment analysis [17–19,55]	Offer conceptual foundations and transparent baseline methods that are easy to audit and useful for comparison against transformer-based models.	Lexicon and shallow approaches are less robust to context, domain shift, irony, implicit sentiment, and multilingual transfer.
Pre-transformer neural representation learning [20–23]	Introduce distributed and sequence-aware representations that improved text classification beyond purely lexical models.	Typically require task-specific supervision and offer weaker cross-lingual portability and contextual modeling than modern transformer architectures.
Transformer, multilingual, and foundation-model research [24–26,32–38]	Support contextual representation, transfer learning, multilingual inference, and zero/few-shot task adaptation, which are central to scalable reputation intelligence.	Require calibration, governance, and computational control; strong generic performance does not remove the need for domain validation and transparent aggregation.
Aviation safety NLP and machine-learning studies [27–31]	Show that NLP and machine learning can classify aviation-specific textual or operational records with high practical relevance.	Focus mainly on safety reports, human factors, fatality prediction, or aviation data management rather than public reputation narratives and customer-facing discourse.
Aspect-based and zero-shot classification literature [39–42]	Provides the basis for decomposing aggregate sentiment into managerial dimensions and for deploying flexible aspect taxonomies under limited supervision.	Aspect labels can be sparse, overlapping, and sensitive to taxonomy design, threshold choice, hypothesis phrasing, and domain mismatch.
Evaluation, agreement, and reliability metrics [43–46]	Establish appropriate reporting of accuracy, F-measure, agreement, and annotation reliability, supporting transparent validation of the framework.	Metrics are diagnostic rather than substantive; their interpretation depends on class balance, annotation protocol, sample size, and error distribution.
Topic modeling, term weighting, and clustering [47–51]	Enable large text collections to be summarized into interpretable themes; TF-IDF and KMeans are especially transparent and lightweight for dashboards.	Topic quality depends on preprocessing, parameter choices, corpus size, and analyst interpretation; more semantic methods may improve coherence but can reduce simplicity.
Data-quality, platform heterogeneity, and recommender-system literature [52,53]	Highlights sampling, platform, user-demographic, and algorithmic biases that must be considered when interpreting public online discourse.	These works are broader than airline reputation monitoring, so their lessons must be adapted to aviation-specific data sources and managerial questions.
Transport economics and language-processing reference frameworks [10,11,54]	Connect reputation indicators to airline economics, demand, pricing, network decisions, and general NLP system design.	They provide essential theoretical context but do not by themselves specify an end-to-end multilingual reputation-intelligence pipeline.

2. Materials and Methods

This section formalizes the reputation-intelligence pipeline as a sequence of linked mappings from query specification to managerial indicators. The objective is to ensure that each computational stage has a clear mathematical definition and a transparent dependency on prior stages in order to have a clean and fully dependent algorithm.

2.1. Query-Scoped Corpus Construction

Let b denote the target airline brand, let $I = [t_1, t_2]$ denote the observation interval, and let Ω denote the admissible internet scope used by the study design. The candidate corpus obtained by scoping the internet for query-relevant information is defined by Equation (1):

$$\tilde{\mathcal{D}}(b, I, \Omega) = \{\tilde{d}_i : \text{retrieve}(b, I, \Omega)\} \quad (1)$$

Each raw candidate document is represented by Equation (2):

$$\tilde{d}_i = (u_i, \tau_i, x_i^{(0)}, m_i) \quad (2)$$

where u_i is the source uniform resource locator (URL), τ_i is publication timestamp, $x_i^{(0)}$ is raw text, and m_i stores source metadata. Preprocessing is represented by Equation (3):

The metadata field m_i is retained specifically to preserve source type, language-detection output, timestamp provenance, and traceability through the filtering pipeline. In the present pilot application these metadata are used for auditability and preprocessing control rather than for claims of statistical representativeness by language or platform.

$$\mathcal{D}(I) = \mathcal{P}(\tilde{\mathcal{D}}(b, I, \Omega)) \quad (3)$$

where $\mathcal{P}(\cdot)$ includes normalization, deduplication, and language detection.

To formalize the downstream document state, each processed record $d_i \in \mathcal{D}(I)$ is represented by Equation (4),

$$d_i = (u_i, \tau_i, \ell_i, x_i, s_i, \mathcal{A}_i, c_i) \quad (4)$$

where u_i is the source URL, τ_i is the publication timestamp, ℓ_i is the language label, x_i is the cleaned text string, s_i is the scalar sentiment score, $\mathcal{A}_i$ is the inferred set of aspect labels, and c_i is the assigned topic cluster. Once the document state in Equation (4) is established, the first analytical requirement is to enforce brand relevance before sentiment and aspect inference.

2.2. Brand Relevance and Disambiguation

Let $\mathcal{B}$ denote the alias set for the brand, $\mathcal{N}$ the set of negative exclusion phrases, and $\mathcal{C}$ the set of airline-specific context terms. Let $m(b, x)$ be a binary alias occurrence function equal to 1 if alias $b \in \mathcal{B}$ appears in text x and 0 otherwise, and let $n(n_j, x)$ equal 1 if exclusion phrase $n_j \in \mathcal{N}$ appears in x . The brand-match predicate is defined by Equation (5),

$$M(x) = \mathbb{I}\left(\sum_{b \in \mathcal{B}} m(b, x) \geq 1\right) \cdot \mathbb{I}\left(\sum_{n_j \in \mathcal{N}} n(n_j, x) = 0\right) \cdot \Psi(x) \quad (5)$$

where the disambiguation term $\Psi(x)$ is given by Equation (6):

$$\Psi(x) = \begin{cases} 1, & \text{if the brand name is unambiguous,} \\ \mathbb{I}(\sum_{c \in \mathcal{C}} q(c, x) \geq 1), & \text{otherwise.} \end{cases} \quad (6)$$

After filtering with Equations (5) and (6), each retained text x_i is tokenized and processed by the multilingual sentiment model.

2.3. Multilingual Sentiment Inference

Let T_θ denote a multilingual transformer encoder with parameters θ . Given tokenized input sequence $\mathbf{z}_i$, the encoder produces hidden states according to Equation (7),

$$\mathbf{H}_i = T_\theta(\mathbf{z}_i) \in \mathbb{R}^{L_i \times h} \quad (7)$$

where L_i is sequence length and h is hidden dimensionality. Let $\mathbf{h}_i^{[\text{cls}]}$ denote the pooled representation. The linear classification head then maps this representation to logits through Equation (8):

$$\mathbf{o}_i = \mathbf{W}_s \mathbf{h}_i^{[\text{cls}]} + \mathbf{b}_s \quad (8)$$

These logits are converted to class posteriors by the softmax transform in Equation (9):

$$\mathbf{p}_i = \text{softmax}(\mathbf{o}_i) \quad (9)$$

whose components are $p_{i,\text{neg}}$, $p_{i,\text{neu}}$, and $p_{i,\text{pos}}$. The predicted class is therefore obtained by Equation (10),

$$y_i = \arg \max_{k \in \{\text{neg}, \text{neu}, \text{pos}\}} p_{i,k} \quad (10)$$

and the signed sentiment score is computed by Equation (11):

$$s_i = \begin{cases} +p_{i,\text{pos}}, & \text{if } y_i = \text{positive}, \\ -p_{i,\text{neg}}, & \text{if } y_i = \text{negative}, \\ 0, & \text{if } y_i = \text{neutral}. \end{cases} \quad (11)$$

For temporal diagnostics, the average signed sentiment in interval I is defined by Equation (12):

$$\bar{s}(I) = \frac{1}{N_I} \sum_{i=1}^{N_I} s_i \quad (12)$$

2.4. Zero-Shot Aspect Inference

Given the sentiment stage, the next step is to infer semantic aspects without task-specific retraining. Let the aspect set be $\mathcal{K} = \{a_1, \dots, a_K\}$ and define the hypothesis template in Equation (13):

$$h_k = \text{"This text is about } a_k\text{"} \quad (13)$$

For each document–aspect pair, the NLI model returns the triplet in Equation (14):

$$\boldsymbol{\pi}_{ik} = (\pi_{ik}^{\text{contr}}, \pi_{ik}^{\text{neut}}, \pi_{ik}^{\text{ent}}) \quad (14)$$

The entailment-based aspect score is then defined by Equation (15),

$$\alpha_{ik} = \pi_{ik}^{\text{ent}} \quad (15)$$

and the final multi-label aspect assignment follows Equation (16):

$$\mathcal{A}_i = \{a_k \in \mathcal{K} : \alpha_{ik} \geq \gamma\} \quad (16)$$

Aspect prevalence in interval I is defined by Equation (17):

$$\rho_k(I) = \frac{1}{N_I} \sum_{i=1}^{N_I} \mathbb{I}(a_k \in \mathcal{A}_i) \quad (17)$$

where γ is the multi-label threshold.

Because the aspect module is zero-shot in the baseline architecture, the resulting labels are treated as exploratory diagnostic signals. They are useful for structuring analyst attention and testing whether the taxonomy captures relevant airline narratives, but they should not be interpreted as fully validated aspect-level measurements until calibrated against a larger domain-labeled corpus.

2.5. Topic Induction and Thematic Salience

After sentiment and aspect inference, topic structure is induced over the filtered corpus. Let interval I contain N_I non-empty texts. The TF-IDF representation is defined in Equation (18):

$$\mathbf{X}_I \in \mathbb{R}^{N_I \times V} \quad (18)$$

with entries computed by Equation (19):

$$X_{iv} = \text{tf}_{iv} \cdot \log\left(\frac{N_I}{1 + \text{df}_v}\right) \quad (19)$$

where tf_{iv} is term frequency and df_v is document frequency. Cluster assignment is obtained by minimizing the KMeans objective in Equation (20):

$$\min_{\{\mu_j\}_{j=1}^J} \sum_{i=1}^{N_I} \sum_{j=1}^J r_{ij} \|\mathbf{x}_i - \mu_j\|_2^2 \quad (20)$$

with binary assignments r_{ij} satisfying $\sum_j r_{ij} = 1$. The mean sentiment per topic cluster is then computed by Equation (21):

$$\bar{s}_j = \frac{1}{|C_j|} \sum_{i \in C_j} s_i \quad (21)$$

Topic prevalence is defined by Equation (22):

$$\omega_j(I) = \frac{|C_j|}{N_I} \quad (22)$$

2.6. Temporal Reputation State and Aggregate Indicators

Because reputation is time-dependent, the interval-specific state is defined by Equation (23), which combines the Net Sentiment Index (NSI), positive share (PS), negative share (NS), aspect-level scores, and topic-level mean sentiment:

$$R(I) = \left(\text{NSI}(I), \text{PS}(I), \text{NS}(I), \{\text{AScore}_k(I)\}_{k=1}^K, \{\bar{s}_j(I)\}_{j=1}^{J_I} \right) \quad (23)$$

where J_I is the effective number of topics.

To operationalize the temporal state in Equation (23), let N_I^+ , N_I^0 , and N_I^- denote the positive, neutral, and negative document counts. Their identity is given by Equation (24):

$$N_I = N_I^+ + N_I^0 + N_I^- \quad (24)$$

Using this total, the positive and negative shares are defined by Equations (25) and (26), respectively:

$$\text{PS}(I) = 100 \cdot \frac{N_I^+}{N_I} \quad (25)$$

$$NS(I) = 100 \cdot \frac{N_I^-}{N_I} \quad (26)$$

The NSI is then computed by Equation (27):

$$NSI(I) = 50 \left(1 + \frac{N_I^+ - N_I^-}{N_I} \right) \quad (27)$$

If $N_I = 0$, the implementation sets $NSI(I) = 50$. At aspect level, the score is defined by Equation (28):

$$AScore_k(I) = \begin{cases} 50 \left(1 + \frac{N_{I,k}^+ - N_{I,k}^-}{N_{I,k}} \right), & \text{if } N_{I,k} > 0, \\ \emptyset, & \text{otherwise,} \end{cases} \quad (28)$$

where $N_{I,k}$ is the number of documents assigned to aspect a_k within interval I .

To rank topics by both polarity and volume, positive-topic strength is defined by Equation (29) and negative-topic strength by Equation (30):

$$\eta_j^+ = \bar{s}_j |C_j| \quad \text{for } \bar{s}_j > 0 \quad (29)$$

$$\eta_j^- = |\bar{s}_j| |C_j| \quad \text{for } \bar{s}_j < 0 \quad (30)$$

The software applies strict thresholds first and then relaxed sign-based thresholds when the corpus is too small to yield strongly signed topics.

2.7. Zero-Shot Supervision Evaluation and Computational Profile

In the absence of a human-labeled gold-standard corpus, let $\hat{y}_i$ denote the machine label and $\tilde{y}_i$ a zero-shot external label obtained from controlled heuristics, lexicon rules, external ratings, or a small expert-labeled sample. Proxy accuracy is estimated by Equation (31):

$$\widehat{\text{Acc}} = \frac{1}{M} \sum_{i=1}^M \mathbb{I}(\hat{y}_i = \tilde{y}_i) \quad (31)$$

where M is the validation-subset size. Bootstrap confidence intervals are then obtained by repeated resampling.

With N documents, average token length L , K aspects, and vocabulary size V , sentiment inference scales approximately as $O(NLh)$ for hidden size h , aspect inference as $O(NKLh)$, TF-IDF construction as $O(NV)$ in sparse form, and KMeans as $O(NJVq)$ for J clusters and q iterations.

2.8. Forecast

Because weekly sentiment is a temporally ordered signal, the framework extends descriptive reputation measurement with a short-horizon forecasting layer. Let I_t denote the t -th weekly interval and let the raw weekly sentiment mean be defined by Equation (32):

$$\tilde{y}_t = \bar{s}(I_t), \quad t = 1, \dots, T \quad (32)$$

where $\bar{s}(I_t)$ is the interval-level average signed sentiment from Equation (12). Following the implementation, missing weekly values are first interpolated and then the resulting signal is smoothed by a rolling operator of width ω . The observed series used for forecasting is therefore given by Equation (33):

$$y_t = (\mathcal{R}_\omega \circ \mathcal{I})(\tilde{y}_t) \quad (33)$$

where $\mathcal{I}(\cdot)$ denotes interpolation and $\mathcal{R}_\omega(\cdot)$ denotes rolling smoothing. The sequence $\{y_t\}_{t=1}^T$ thus represents the preprocessed weekly average sentiment series used in model fitting and forecast generation.

The forecasting module evaluates Holt's linear damped trend exponential smoothing together with a compact set of benchmark models spanning persistence-based, deterministic-trend, exponential-smoothing, and stochastic time-series approaches. The benchmark family is summarized by Equation (34):

$$\begin{aligned} \hat{y}_{t+h|t}^{\text{naive}} &= y_t, \\ y_t^{\text{lin}} &= c_0 + c_1 t + \varepsilon_t, \\ \hat{y}_{t+h|t}^{\text{lin}} &= \hat{c}_0 + \hat{c}_1(t+h), \\ y_t^{\text{quad}} &= c_0 + c_1 t + c_2 t^2 + \varepsilon_t, \\ \hat{y}_{t+h|t}^{\text{quad}} &= \hat{c}_0 + \hat{c}_1(t+h) + \hat{c}_2(t+h)^2, \\ \ell_t^{\text{ses}} &= \alpha_s y_t + (1 - \alpha_s) \ell_{t-1}^{\text{ses}}, \\ \hat{y}_{t+h|t}^{\text{ses}} &= \ell_t^{\text{ses}}, \\ (1 - \varphi B)(1 - B)y_t &= c + (1 + \vartheta B)\varepsilon_t. \end{aligned} \quad (34)$$

Equation (34) includes the naive last-value forecast, linear trend regression, quadratic trend regression, simple/additive exponential smoothing, and an autoregressive integrated moving-average model of order (1,1,1). In what follows, this specification is denoted ARIMA(1,1,1), and B is the backshift operator. These comparators were selected because they represent different structural assumptions: no dynamics beyond persistence, deterministic trend extrapolation, local-level smoothing without explicit trend, and stochastic differenced dynamics. They therefore provide a compact but methodologically meaningful benchmark against which the selected damped-trend specification can be justified.

Model selection is conducted by holdout validation. Let the training segment and validation segment be defined by Equation (35):

$$\mathcal{T}_{\text{train}} = \{1, \dots, T_0\}, \quad \mathcal{T}_{\text{val}} = \{T_0 + 1, \dots, T\}, \quad H = T - T_0 \quad (35)$$

where H is the number of holdout observations. Each candidate model is estimated on $\mathcal{T}_{\text{train}}$ and then used to generate forecasts $\hat{y}_{T_0+h|T_0}^{(m)}$ for $h = 1, \dots, H$. Forecast errors are evaluated by root mean squared error (RMSE), mean absolute error (MAE), and symmetric mean absolute percentage error (sMAPE), defined by Equations (36)–(38):

$$\text{RMSE}(m) = \sqrt{\frac{1}{H} \sum_{h=1}^H \left(y_{T_0+h} - \hat{y}_{T_0+h|T_0}^{(m)} \right)^2} \quad (36)$$

$$\text{MAE}(m) = \frac{1}{H} \sum_{h=1}^H \left| y_{T_0+h} - \hat{y}_{T_0+h|T_0}^{(m)} \right| \quad (37)$$

$$\text{sMAPE}(m) = \frac{100}{H} \sum_{h=1}^H \frac{2 \left| y_{T_0+h} - \hat{y}_{T_0+h|T_0}^{(m)} \right|}{\left| y_{T_0+h} \right| + \left| \hat{y}_{T_0+h|T_0}^{(m)} \right|} + \epsilon \quad (38)$$

where $\epsilon > 0$ is a small stabilizing constant used only to avoid division by zero when the denominator is numerically negligible. The primary selection criterion is the lowest out-of-sample RMSE, while MAE and sMAPE are retained as complementary diagnostics

of absolute error magnitude and relative forecast accuracy. Let $\mathcal{M}$ denote the set containing Holt's damped-trend model together with the benchmark specifications in Equation (34). The selected model therefore follows Equation (39):

$$m^* = \arg \min_{m \in \mathcal{M}} \text{RMSE}(m) \quad (39)$$

Under this rule, the selected specification is Holt's linear damped trend exponential smoothing, implemented as holt linear damped.

Holt's linear damped trend model represents the sentiment series through a local level ℓ_t and a local trend b_t , with damping parameter ϕ attenuating trend extrapolation as the forecast horizon increases. The recursive updating equations are given by Equations (40) and (41):

$$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1}) \quad (40)$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1 - \beta)\phi b_{t-1} \quad (41)$$

where ℓ_t is the smoothed local level at time t , b_t is the smoothed local slope, $\alpha \in (0, 1)$ is the level smoothing parameter, $\beta \in (0, 1)$ is the trend smoothing parameter, and $\phi \in (0, 1]$ is the damping parameter. In substantive terms, ℓ_t captures the current baseline of weekly sentiment, whereas b_t captures the short-run direction of change.

Given the state updates in Equations (40) and (41), the one-step-ahead forecast is defined by Equation (42):

$$\hat{y}_{t+1|t} = \ell_t + \phi b_t \quad (42)$$

and the h -step-ahead forecast follows Equation (43):

$$\hat{y}_{t+h|t} = \ell_t + \left(\sum_{j=1}^h \phi^j \right) b_t = \ell_t + \phi \left(\frac{1 - \phi^h}{1 - \phi} \right) b_t \quad (43)$$

for $0 < \phi < 1$. When $\phi = 1$, Equation (43) reduces to the undamped Holt linear trend forecast $\hat{y}_{t+h|t} = \ell_t + hb_t$. The damping mechanism is practically important in sentiment applications because it prevents short-run upward or downward movements from being extrapolated indefinitely. An equivalent additive-error state-space interpretation is provided by Equations (44)–(46):

$$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t \quad (44)$$

$$\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t \quad (45)$$

$$b_t = \phi b_{t-1} + \beta \varepsilon_t \quad (46)$$

where ε_t is an innovation term with zero mean and finite variance. This interpretation makes explicit that the observed weekly sentiment is generated by a latent local level and a damped local trend, both of which are updated when new information arrives. For smoothed weekly sentiment series, the model is appropriate when the underlying process exhibits short-run persistence and directional movement and also partial mean reversion over longer horizons.

In practical estimation, the parameters $(\alpha, \beta, \phi, \ell_0, b_0)$ are typically obtained by numerical optimization on the training segment, either by minimizing the sum of squared one-step-ahead forecast errors or, under Gaussian innovation assumptions, by maximizing the corresponding likelihood. The standard optimization problem is stated in Equation (47):

$$(\hat{\alpha}, \hat{\beta}, \hat{\phi}, \hat{\ell}_0, \hat{b}_0) = \arg \min_{\alpha, \beta, \phi, \ell_0, b_0} \sum_{t \in \mathcal{T}_{\text{train}}} (y_t - \hat{y}_{t|t-1})^2 \quad (47)$$

subject to $0 < \alpha < 1$, $0 < \beta < 1$, and $0 < \phi \leq 1$. In the alternative models, regression coefficients are estimated by ordinary least squares, simple/additive exponential smoothing is fitted by analogous smoothing-parameter optimization, and ARIMA parameters are typically estimated by maximum likelihood.

The methodological rationale for preferring Holt's damped trend over the alternative models is therefore twofold. Relative to the naive and simple/additive exponential smoothing benchmarks, it can accommodate short-run directional movement in sentiment. Relative to linear and quadratic trend regression, it updates recursively and does not impose a globally fixed deterministic trend. Relative to ARIMA(1, 1, 1), it remains more transparent, easier to audit, and better aligned with the managerial interpretation of locally evolving sentiment dynamics. Because model choice is based primarily on the lowest out-of-sample RMSE, the final forecast used in the pipeline is the one generated by the specification that offers the best empirical predictive performance on the holdout segment while preserving interpretability for decision support.

The forecast is not used as an automatic prescription. Its decision-support role is to flag whether the latest reputation movement is likely to persist, reverse, or deteriorate over a short horizon, thereby helping managers decide whether monitoring is sufficient or whether communication, service-recovery, staffing, pricing, or operational interventions should be escalated.

2.9. Architecture

The developed architecture is presented in Figure 1; it is designed as a modular dependency where each module corresponds to an explicit mathematical framework, and each directed edge represents a required information dependency between stages. The workflow starts with query-scoped acquisition from the admissible online scope, then applies brand-profile filtering for disambiguation and relevance control. Filtered documents are processed by multilingual sentiment inference and zero-shot aspect inference, after which topic induction organizes discourse into interpretable thematic clusters. These streams are fused in the aggregation module to compute interval-level outputs such as NSI, positive and negative shares, aspect scores, and ranked topic signals.

The reporting stage exports structured outputs for analyst interpretation and period-over-period comparison. In this design, Figure 1 should be read as a logical system map, while Figure 2 should be read as an execution-order specification. Using both views improves reproducibility because architectural coherence and operational sequencing can be audited separately.

Figure 1 should be interpreted as a dependency graph rather than a simple flowchart. The left-to-right trajectory represents the primary transformation chain from acquisition to reporting. The branch from filtering to aspect inference makes explicit that aspect attribution is only meaningful after relevance control. The bidirectional linkage between sentiment and topic modules reflects analytical coupling: topic structure contextualizes polarity interpretation, while polarity distributions improve thematic ranking. The convergence into aggregation indicates that final KPIs are composite indicators, not outputs of a single model block.

To make this mapping explicit, Table 2 links each architecture module to the corresponding equations in Section 2. This table clarifies reproducibility because any reported metric can be traced to its formal definition and upstream computational dependencies.

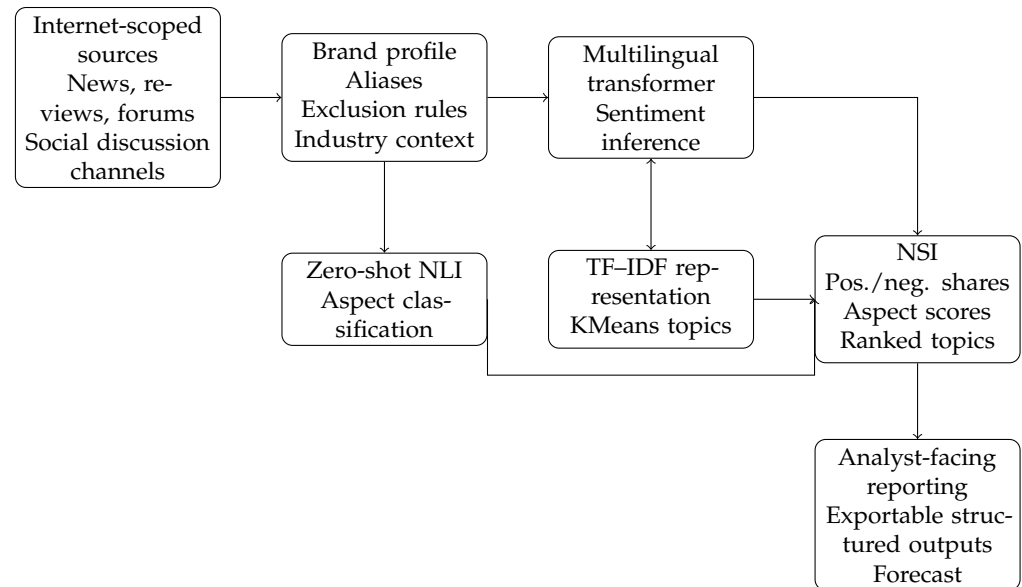


Figure 1. End-to-end architecture of the airline reputation intelligence framework.

Table 2. Operational mapping between system modules, formal definitions, and managerial interpretation.

Module	Formal Anchor	Managerial Interpretation
Internet-scoped acquisition	Equations (1)–(3)	Defines information boundaries, temporal coverage, and traceability of incoming records.
Brand relevance filtering	Equations (5) and (6)	Reduces false positives and increases interpretive validity of brand-level analytics.
Sentiment inference	Equations (7)–(12)	Quantifies directional narrative intensity for risk/opportunity diagnostics.
Aspect inference	Equations (13)–(17)	Decomposes sentiment into operationally interpretable dimensions.
Topic induction	Equations (18)–(22)	Identifies emergent themes and their relative prominence in discourse.
Aggregation and ranking	Equations (23)–(30)	Produces period-level KPIs suitable for benchmarking and decision review.
Forecasting	Equations (32)–(47)	Extends static reporting with short-horizon sentiment outlooks for early warning, intervention timing, and assessment of whether reputation shifts are temporary or persistent.

Although the aspect taxonomy is fully configurable, the presented analytical workflow applies only the predefined taxonomy selected for the study after query-scoped retrieval and preprocessing. This sequencing improves methodological consistency because documents are first filtered for relevance and only then mapped to a stable set of managerial categories. In the present configuration, the evaluated taxonomy includes punctuality and delays, ground and baggage experience, on-board comfort, on-board service, pricing fairness, safety and reliability, network connectivity, labor relations, corporate image, sustainability, innovation, digital experience, loyalty programs, and financial-strategic narratives. This breadth is deliberate: airline reputation risk is intrinsically multi-dimensional and spans operational execution, passenger experience, commercial policy, organizational stability, and long-horizon brand positioning.

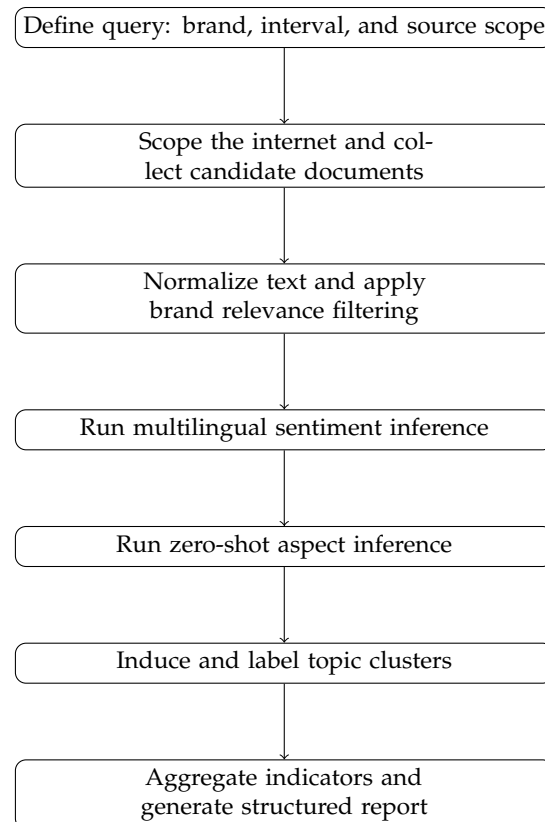


Figure 2. Sequential implementation workflow of the proposed system.

Figure 2 complements Figure 1 by emphasizing execution order instead of structural dependency. The workflow begins with query definition, since the brand aliases, observation interval, and admissible source scope determine the empirical boundaries of the corpus. Candidate documents are then collected, normalized, and filtered for brand relevance before any semantic inference is performed, reducing the risk that unrelated online content contaminates the sentiment, aspect, or topic outputs. Once the retained texts have passed this relevance-control stage, multilingual sentiment inference estimates polarity, zero-shot aspect inference assigns managerial dimensions, and topic induction groups recurring narratives into interpretable clusters. The final aggregation stage converts these document-level outputs into structured indicators such as the NSI, polarity shares, aspect scores, ranked topics, and reportable summaries. In validation settings, this sequential reading is useful because it makes each processing dependency auditable: errors can be traced to corpus scoping, filtering, inference, topic modeling, or aggregation rather than being treated as undifferentiated system-level failures.

Table 3 summarizes, for the illustrative airline application, how detailed aspects are grouped into broader managerial families. Retaining this hierarchical structure is analytically important because it reduces interpretive ambiguity in the reporting of results by allowing polarity shifts to be examined at multiple levels of abstraction: fine-grained labels for diagnostic analysis and family-level groupings for strategic decision-making.

Table 3. Aspect families used for managerial interpretation of reputation outputs.

Aspect Family	Representative Aspects	Strategic Interpretation
Operational reliability	Punctuality, delays, ground handling, baggage	Reflects execution consistency and disruption exposure.

Table 3. *Cont.*

Aspect Family	Representative Aspects	Strategic Interpretation
Passenger experience	Cabin comfort, in-flight service, digital experience, loyalty	Captures perceived quality and retention potential.
Commercial value	Price fairness and fare-value narratives	Signals pricing power and discount pressure.
Safety and trust	Safety, reliability, crisis-related discourse	Measures perceived risk and assurance.
Network and organization	Connectivity, labor relations, financial strategy	Links reputation to structural competitiveness and organizational stability.
Corporate positioning	Brand image, sustainability, innovation	Reflects long-horizon legitimacy and differentiation.

3. Results

The proposed framework is evaluated empirically as a pilot implementation rather than being presented solely as a hypothetical reporting protocol. To this end, a pilot reference subset was constructed through double review and used to benchmark the core sentiment and aspect modules under controlled conditions. For sentiment classification, the annotated sample comprises 296 documents. Although still modest relative to large benchmark corpora, this size provides a stronger basis for assessing agreement patterns and diagnosing class-specific errors through the confusion structure reported in Figure 3. For aspect assignment, the same annotation effort was recast as a multi-label benchmark to compare the zero-shot NLI configuration with rule-based and supervised alternatives. Accordingly, the current validation should be interpreted as preliminary and based on modestly sized annotated subsets rather than as a final large-scale external benchmark. It establishes relative model performance, identifies dominant error modes, and provides an empirically grounded basis for subsequent dataset expansion. The operational outputs reported later in this section should likewise be read as illustrations of end-to-end pipeline functionality on the filtered corpus rather than as definitive evidence of full operational robustness.

For transparency, the validation subset is defined at the document level. Each document was assigned one mutually exclusive sentiment class (negative, neutral, or positive), while the aspect evaluation used the same material in a sparse multi-label setting because one text can refer simultaneously to several airline dimensions. The source URL, timestamp, detected language, and source metadata are retained in the document representation defined in Equation (4), but the present pilot subset was not constructed as a stratified sample by language or platform. Consequently, the reported performance values should be read as controlled module-validation evidence for the annotated subset, not as population estimates for all airline-related online discourse.

The comparative strategy follows a structured benchmarking logic designed to evaluate the proposed framework against alternatives that represent distinct methodological families rather than an arbitrary set of models. For sentiment classification, the multilingual transformer is contrasted with a lexicon-based baseline because the latter provides a transparent and low-complexity reference point against which the incremental value of contextual multilingual representation learning can be assessed. For aspect detection, the comparison spans zero-shot NLI, keyword-based rules, and a supervised TF-IDF plus logistic-regression model in order to distinguish portability without task-specific training, interpretable heuristic matching, and performance under in-domain supervision. For topic

induction, the benchmark contrasts LDA, TF-IDF plus KMeans, and embedding-based clustering so that model choice can be justified not only by aggregate score but also by coherence, stability, interpretability, and implementation practicality.

Within this benchmarking design, Table 4 shows that the multilingual transformer substantially outperforms the lexicon baseline on the annotated sentiment subset, achieving an accuracy of 0.9045, a macro-F1 of 0.9048, and a weighted-F1 of 0.9050. The estimated Cohen's κ of 0.8492 further indicates near-perfect agreement beyond chance in a three-class sentiment task.

Table 4. Preliminary sentiment-classification validation results on the annotated sentiment subset.

Model	Accuracy	Macro-F1	Weighted-F1	Cohen's κ
Multilingual transformer	0.9045	0.9048	0.9050	0.8492
Lexicon baseline	0.5085	0.2443	0.3514	0.0183

To address the language-distribution and class-balance diagnostics, Table 5 reports the detected language composition and sentiment balance of an example for TAP Portugal, while Table 6 separates the operational corpus balance from the manually adjudicated validation-subset balance. The analyzed corpus is dominated by English-language material, but it also includes Portuguese, German, French, and additional lower-frequency languages. The operational corpus is strongly neutral, whereas the validation subset is more evenly distributed across sentiment classes, which is useful for estimating class-specific classifier behavior. TAP Portugal was selected as the pilot case because it offers a suitable balance between international exposure and analytical tractability: it operates across multiple markets and therefore generates multilingual public discourse, but its scale remains sufficiently bounded for transparent corpus inspection within a lightweight framework. This makes it large enough to produce heterogeneous reputation signals, yet not so large that the pilot evaluation is dominated by the exceptional media volume associated with the largest global carriers.

Table 5. Language distribution and sentiment balance in the TAP Portugal analyzed corpus.

Language	Docs	Share (%)	Positive	Neutral	Negative	NSI
English	6902	87.2	240	6082	580	47.5
Portuguese	462	5.8	170	290	2	68.2
Deutsch	208	2.6	3	196	9	48.6
French	129	1.6	32	20	77	32.6
Dutch	48	0.6	0	48	0	50.0
Spanish	41	0.5	6	31	4	52.4
Indonesian	34	0.4	0	34	0	50.0
Polish	18	0.2	0	18	0	50.0
Italian	17	0.2	0	15	2	44.1
Catalan	12	0.2	0	12	0	50.0
Japanese	12	0.2	2	8	2	50.0
Swedish	12	0.2	0	12	0	50.0
Finnish	8	0.1	0	6	2	37.5
Norwegian	4	0.1	0	0	4	0.0
Korean	3	0.0	0	3	0	50.0
Hungarian	2	0.0	0	2	0	50.0
Turkish	2	0.0	0	0	2	0.0

Table 6. Sentiment class balance in the operational corpus and validation subset.

Sample	Class	Documents	Share (%)
Analyzed corpus	Positive	453	5.7
Analyzed corpus	Neutral	6777	85.6
Analyzed corpus	Negative	684	8.6
Validation subset	Positive	1062	26.6
Validation subset	Neutral	2002	50.2
Validation subset	Negative	926	23.2

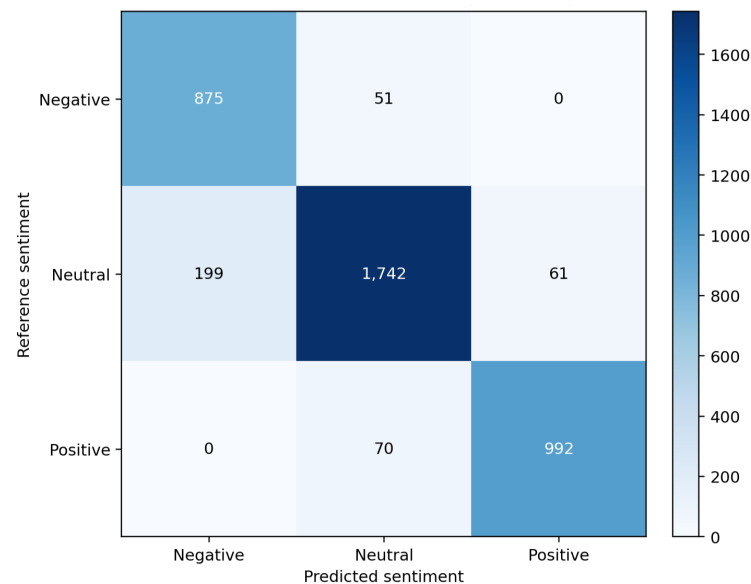
**Figure 3.** Confusion matrix for the multilingual transformer on the annotated sentiment subset ($n = 3990$).

Figure 3 shows that the model correctly classifies 282 of 296 documents, leaving only 14 off-diagonal errors. Most mistakes occur between adjacent classes rather than through extreme polarity reversals. Negative observations are never mistaken for positive ones, while positive documents are more often absorbed into the neutral class than into the negative class. This is a plausible pattern in airline discourse, where many positive comments are expressed in moderate terms and therefore sit close to neutrality at the class boundary.

Table 7 shows that aspect assignment is materially more challenging than document-level sentiment classification. The zero-shot NLI configuration remains attractive because of its taxonomy flexibility. The keyword baseline improves the micro-F1 only marginally, while the best micro-F1 result is achieved by the supervised TF-IDF plus logistic-regression model, although macro-F1 remains low across all configurations. This comparison does not invalidate the zero-shot design; instead, it clarifies the trade-off between portability and benchmark accuracy. The low zero-shot values are also methodologically understandable in this setting. Unlike the sentiment task, aspect detection is formulated here as a sparse multi-label problem in which a single document may activate several partially overlapping airline dimensions, while others remain rare in the annotated subset. Under such conditions, zero-shot entailment scores are penalized by label imbalance, hypothesis sensitivity, and semantic overlap between categories such as service, comfort, punctuality, digital experience, and price fairness. In addition, subset accuracy is an especially severe metric in multi-label classification because it requires an exact match of the full label set for each document; with a relatively fine-grained taxonomy and a small pilot sample, even partially correct predictions can still yield a value close to zero. The stronger supervised micro-F1

is important precisely because it shows that frequent aspect labels can be recovered more reliably with domain-specific supervision, while the low macro-F1 indicates that rare aspect categories remain difficult to model. Read in that way, the low zero-shot metrics are informative because they empirically justify the article’s claim that portability is gained at the cost of benchmark performance in cold-start aspect mapping.

To complement classifier validation, Table 8 compares alternative topic-induction strategies on the evaluated corpus using normalized pointwise mutual information (NPMI), topic diversity, stability, interpretability, coverage, and a composite score. In this comparison, diversity captures how distinct the retained topics are from one another, stability captures how consistently the topic solution is reproduced across repeated runs or perturbations, interpretability captures how clearly the resulting topics can be understood and labeled by an analyst, and coverage captures how much of the evaluated corpus is meaningfully represented by the retained topic set. The Score column in Table 8 is the descriptive ranking index exported by the application, computed as $S = \frac{1}{6}((\text{NPMI} + 1)/2 + \text{Div.} + \text{Stab.} + \text{Interp.} + \text{Cov.} + \text{Rel.})$. It is included to summarize several diagnostics in one row-level indicator. The TF-IDF plus KMeans configuration achieves the highest composite score, while the embedding-based KMeans alternative remains useful as a modern semantic comparator but does not outperform the simpler configuration in this run. LDA performs visibly worse on coherence, stability, and final score, which supports the decision to prioritize vector-space clustering approaches in the present framework.

Table 7. Preliminary multi-label aspect evaluation results on the annotated subset.

Model	Micro-F1	Macro-F1
Zero-shot NLI	0.185	0.024
Keyword baseline	0.198	0.035
Supervised TF-IDF + logistic regression	0.703	0.036

Table 8. Topic-model comparison on the evaluated corpus.

Model	NPMI	Div.	Stab.	Interp.	Cov.	Score
TF-IDF + KMeans	0.165	0.783	0.368	0.700	1.000	0.572
Embedding + KMeans	0.119	0.617	0.463	0.433	1.000	0.512
LDA	0.050	0.450	0.318	0.283	1.000	0.429

Table 9 adds a BERTopic-style embedding benchmark to the topic-model comparison. The BERTopic-style benchmark uses multilingual MiniLM sentence embeddings, PCA dimensionality reduction, HDBSCAN clustering, and class-based TF-IDF topic representation. It obtains a higher diagnostic score, but HDBSCAN assigns a large share of documents to noise, so its stronger semantic separation comes with lower corpus coverage. This comparison therefore clarifies the trade-off between semantic topic separation and full-corpus operational coverage.

Table 9. Extended topic-model comparison including a BERTopic-style embedding benchmark.

Model	Topics	NPMI	Div.	Stab.	Interp.	Cov.	Score
BERTopic-style c-TF-IDF	13	0.630	0.946	0.729	0.900	0.500	0.687
TF-IDF + KMeans	6	0.165	0.783	0.368	0.700	1.000	0.572
Embedding + KMeans	6	0.119	0.617	0.463	0.433	1.000	0.512
LDA	6	0.050	0.450	0.318	0.283	1.000	0.429

Table 10 summarizes the operational output space generated by the current system for the executed test and shows that the framework produces a structured set of indicators

rather than a single aggregate sentiment score. At the aggregate level, the reported metrics capture overall reputation direction through the NSI, positive share, neutral share, and negative share; at the diagnostic level, they provide pillar-specific score ranges and the most salient positive and negative thematic dimensions. The inclusion of the number of retrieved and analyzed documents is also methodologically important because it locates these outputs within their empirical evidence base and therefore supports interpretation of the reported values in relation to corpus size and coverage.

Table 11 indicates for the test conducted a near-neutral but mildly negative overall reputation state in the illustrative period. The NSI remains slightly below the neutral midpoint, and the negative share exceeds the positive share, although the corpus is dominated by neutral documents. In practice, this type of profile is often more informative for managers than a uniformly positive score because it reveals a mixed reputation environment rather than a one-directional signal.

Table 10. Operational output metrics produced with the test.

Metric	Description	Illustrative Value
Net Sentiment Index	0–100 scaled balance of positive and negative documents	48.5
Positive share	Percentage of documents classified as positive	5.7%
Neutral share	Percentage of documents classified as neutral	85.6%
Negative share	Percentage of documents classified as negative	8.6%
Pillar scores	0–100 pillar-specific polarity scores	8.3–73.6
Top positive pillars	Most positive thematic dimensions in the illustrative run	Customer care; safety; corporate Labor relations; financial; regulatory and public affairs
Top negative pillars	Most negative thematic dimensions in the illustrative run	
Documents retrieved	Size of the retrieved corpus before analysis filtering	8086
Documents analyzed	Size of the filtered corpus in the illustrative run	7914

Table 11. Illustrative runtime output for a selected airline brand and observation interval.

Indicator	Runtime Value
Retrieved documents	8086
Documents analyzed	7914
Net Sentiment Index	48.5
Positive share (%)	5.7
Neutral share (%)	85.6
Negative share (%)	8.6
Number of active reputation pillars	17
Number of active detailed aspects	59
Number of topics retained	6

To move beyond the aggregate NSI, Table 12 disaggregates the illustrative reputation profile across the principal evaluative dimensions. These counts refer to pillar-level assignments within the 7914 analyzed documents reported in Table 11; because the pillar module is multi-label, a single document may contribute to multiple pillar rows, so the totals are not expected to match the total number of retrieved labels. The strongest dimensions are customer care, safety, and corporate reputation, whereas the weakest are labor relations, financial performance, and regulatory and public affairs. This configuration is analytically plausible for an airline brand that performs comparatively well in some customer-facing and safety-related narratives while remaining exposed to criticism associated with labor, financial, and regulatory pressures.

Table 12. Illustrative pillar-level reputation profile for the selected airline and period (multi-label pillar counts).

Pillar	Positive	Neutral	Negative	Score
Onboard	91	2211	121	49.38
Commercial	152	1498	41	53.28
Connectivity	87	1226	20	52.51
Loyalty	68	755	4	53.87
Regulatory and public affairs	4	709	81	45.15
Corporate	146	505	51	56.77
Financial	0	399	107	39.43
Cargo and adjacencies	5	400	14	48.93
Ground experience	7	391	14	49.15
Safety	157	227	9	68.83
Competitive position	4	355	5	49.86
Digital experience	2	338	10	48.86
Operations	4	282	19	47.54
Customer care	108	74	15	73.60
Sustainability	1	85	4	48.33
Labor relations	0	9	45	8.33
Marketing and demand	4	7	0	68.18

The topic-model comparison further reinforces the practical orientation of the developed system. Although the TF-IDF plus KMeans configuration achieves the highest composite score in Table 8, the embedding-based KMeans alternative remains a useful comparator while not outperforming the simpler configuration in this run. This is not a trivial advantage. In organizational settings, especially in aviation, analytic systems are used not only to classify text but also to justify action. A topic model that produces slightly better abstract performance but cannot be interpreted by route managers, communications teams, or service executives may be less valuable than a transparent model whose outputs can be discussed and challenged. The reported results therefore support the broader design principle of the paper: reputation intelligence should be optimized not only for benchmark quality but also for interpretability, reproducibility, and operational adoption.

Using the forecasting framework defined in Section 2.8, Figure 4 adds a temporal perspective to the static tables. The observed series is volatile over 2025, with early negative values, a mid-2025 trough, and subsequent movement around neutrality through late 2025 and early 2026 before weakening again at the end of the observed window. The short-horizon forecast remains slightly negative but trends gently toward neutrality, moving from approximately -0.023 to -0.014 by the end of the projection range. In managerial terms, this pattern would be interpreted as a signal of mild residual reputational pressure rather than a sharp projected deterioration, suggesting continued monitoring of the weaker pillar dimensions.

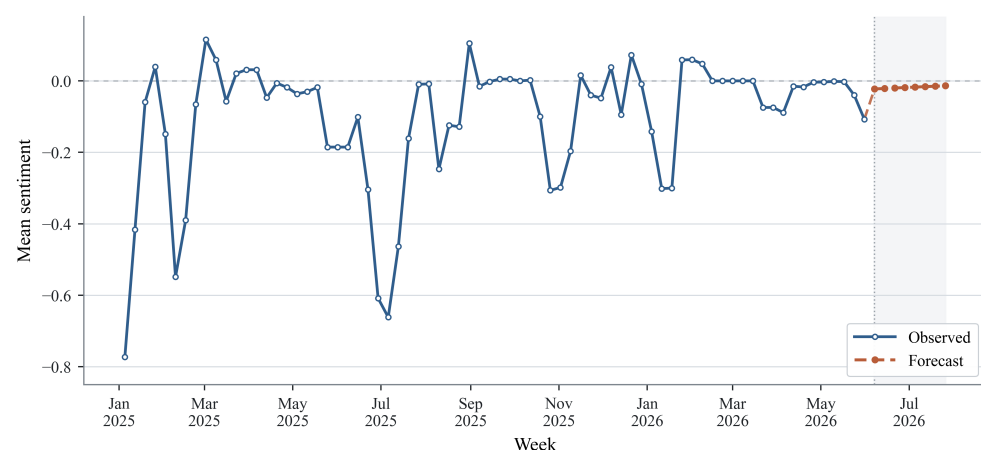
**Figure 4.** Observed and short-horizon forecast of mean sentiment.

Table 13 extends the validation of the predictive component by comparing the selected forecast against simple and stochastic alternatives over an eight-week holdout window. The selected quadratic-trend specification obtains the lowest RMSE, while the linear trend obtains a very similar error profile. The forecast should therefore be interpreted as a short-horizon decision-support signal rather than as a claim of long-range predictive superiority.

Table 13. Holdout validation of alternative short-horizon sentiment forecasting models.

Model	RMSE	MAE	sMAPE	Forecast SD	Selected
quadratic_trend	0.037	0.032	1.238	0.003	Yes
linear_trend	0.038	0.025	1.202	0.006	No
exponential_smoothing	0.055	0.050	1.543	0.021	No
holt_linear_damped	0.073	0.070	1.409	0.000	No
naive_last_value	0.073	0.070	1.409	0.000	No
arima_1_1_1	0.087	0.081	1.429	0.004	No

4. Discussion

Online reputation affects customer acquisition, retention, and pricing power. Let Q_t be booking demand at time t , P_t the fare level, and R_t the reputation state. A generic demand relation can be written as Equation (48):

$$Q_t = f(P_t, R_t, Z_t) \quad (48)$$

where Z_t is a vector of control variables such as seasonality, macroeconomic conditions, competition, and route supply. If R_t deteriorates sharply because of operational failures or labor disputes, the airline may experience downward pressure on conversion, yield, or repeat purchase.

The results provide encouraging pilot evidence that the developed framework can operate as a managerial sensing instrument, even though some modules remain at a preliminary validation stage. The multilingual transformer achieves markedly stronger sentiment performance than the lexicon baseline in Table 4, with an accuracy of 0.905, a macro-F1 of 0.905, and a Cohen's κ of 0.849. This performance gap is important in substantive terms because airline discourse is often context-dependent, multilingual, and lexically ambiguous: complaints may be expressed indirectly, and positive evaluations are often moderate rather than emphatic. Figure 3 is therefore encouraging because only 379 of 3990 cases fall outside the diagonal, and the residual errors are concentrated between adjacent classes, especially between positive and neutral, rather than between opposite polarities. For management, this suggests that the sentiment component is promising for identifying directional reputation shifts and for detecting the onset of negative narrative shocks without being dominated by extreme classification reversals.

Table 14 positions the present sentiment module against recent aviation-oriented benchmark studies, while deliberately treating the comparison as interpretive rather than as a strict leaderboard because the studies differ in data source, language coverage, label structure, class balance, and validation protocol. The safety-report and accident-classification literature provides a useful performance envelope: Madeira et al. [27] report micro-F1 values of 0.900, 0.779, and 0.875 across HFACS levels; Nogueira et al. [28] report 90% accuracy, 87% macro-F1, and 86% recall for fatal-accident prediction; Lázaro et al. [29] report best accuracy of 0.90 and macro-F1 up to 0.86; and Lázaro et al. [30] report KNN accuracies of 0.97, 0.82, and 0.94 with corresponding F1 values of 0.96, 0.80, and 0.92 across three HFACS levels. Against airline-review sentiment benchmarks, the present multilingual transformer, with 90.5% accuracy, 90.5% macro-F1, 90.5% weighted-F1, and Cohen's $\kappa = 0.849$, is above

the three-class BERT-family result reported by Li et al. [2] and the random-split accuracy reported by Syed et al. [6], below Syed et al.'s cross-validation result, and within the reported range of stronger aviation-adapted BERT results reported by Lázaro et al. [31]. The main implication is therefore not that the proposed model universally dominates prior work, but that it achieves competitive aviation-text performance while adding multilingual reputation monitoring, interpretable managerial indicators, topic diagnostics, and forecasting within a lightweight pipeline.

The language and class-balance diagnostics also qualify how the validation evidence should be interpreted. The operational corpus is heavily English and neutral, which is typical of broad news- and review-based airline monitoring but means that aggregate NSI values are influenced by a large neutral class. By contrast, the adjudicated validation subset is more balanced across negative, neutral, and positive cases, allowing the confusion matrix to test whether the classifier handles minority polarities rather than merely reproducing the majority operational class.

The aspect results are more demanding, but they are no less informative. Table 7 shows that zero-shot NLI aspect assignment performs far below the supervised benchmark, which is unsurprising given the absence of domain-specific training and the complexity of multi-label airline narratives. Yet this finding should not be read simply as a weakness. Instead, it identifies the operational role of the zero-shot layer within the present architecture: it is most valuable in cold-start settings where a firm needs immediate thematic visibility before a sufficiently large annotated corpus exists. In that role, zero-shot inference serves as a flexible structuring device that allows analysts to test taxonomies, discover recurrent issue areas, and prioritize where future manual annotation or model fine-tuning will have the highest return. The supervised result then indicates the performance gains that become attainable for frequent aspect labels once the organization invests in domain-labeled training data, though the low macro-F1 shows that rare aspect categories remain difficult. Accordingly, the aspect tables are used in this article to illustrate diagnostic decomposition and prioritization logic, not to claim definitive aspect prevalence in the underlying population of online airline discourse. In the present paper, aspect-level outputs should therefore be interpreted primarily as exploratory and illustrative managerial diagnostics rather than as fully validated bases for strong aspect-specific inference.

Table 14. Comparison with related aviation text-classification models.

Reference	Task and Data Context	Best/Selected Model	Reported Metrics
Madeira et al. [27]	Human-factor prediction from aviation incident reports	NLP with SVM classifiers	Micro-F1 = 0.900, 0.779, and 0.875 across the three reported HFACS levels.
Nogueira et al. [28]	Fatal-accident prediction using aviation human-factor data	Random Forest	Accuracy = 90%; macro-F1 = 87%; recall = 86%.
Lázaro et al. [29]	Fatality prediction from aviation accident and incident reports	Neural networks and RF	Best accuracy = 0.90; macro-F1 up to 0.86 for neural networks and 0.85 for RF.
Lázaro et al. [30]	Human-factor classification in aviation accidents	KNN, RF, Extra Trees, and XGBoost	Best KNN accuracies = 0.97, 0.82, and 0.94 across three HFACS levels; corresponding F1 = 0.96, 0.80, and 0.92.
Li et al. [2]	Sentiment analysis of airline customer reviews	RoBERTa among BERT-family variants	Accuracy = 96.97% for binary classification and 86.89% for three-class classification.
Syed et al. [6]	Rating-based sentiment classification and summarization of airline reviews	Deep transfer-learning configuration	Accuracy = 89% under random train–test splitting and 95% under cross-validation. Class-level precision = 0.65–0.97, F1 = 0.65–0.95, and AUC = 0.92–1.00 across reported aviation classes.
Lázaro et al. [31]	Aviation data classification and management	Aviation-adapted BERT strategy	
Present study	Multilingual airline-reputation sentiment classification on a 3990-document annotated subset	Multilingual transformer	Accuracy = 90.5%; macro-F1 = 90.5%; weighted-F1 = 90.5%; Cohen's κ = 0.849.

The topic-model comparison further reinforces the practical orientation of the developed system. Although the TF-IDF plus KMeans configuration achieves the highest

composite score in Table 8, the embedding-based KMeans alternative remains a useful modern semantic comparator but does not outperform the simpler configuration in this run. This is not a trivial advantage. In organizational settings, especially in aviation, analytic systems are used not only to classify text but also to justify action. A topic model that produces slightly better abstract performance but cannot be interpreted by route managers, communications teams, or service executives may be less valuable than a transparent model whose outputs can be discussed and challenged. The reported results therefore support the broader design principle of the paper: reputation intelligence should be optimized not only for benchmark quality but also for interpretability, reproducibility, and operational adoption.

The extended topic benchmark strengthens this interpretation. The BERTopic-style comparator introduces the main semantic ingredients of modern neural topic modeling, namely, multilingual sentence embeddings, density-based clustering, and class-based TF-IDF topic representation. It achieves the highest diagnostic score, but it only covers about half of the corpus because the density-based clustering step treats many documents as noise. The preference for TF-IDF + KMeans is therefore not based on the absence of an embedding alternative; it reflects a different operational trade-off, prioritizing complete corpus coverage, lower computational cost, easier reproducibility, and clearer auditability for managerial use.

This point becomes clearer when the preliminary validation results are connected to the illustrative runtime outputs. Table 11 shows a near-neutral but mildly negative aggregate state, with an NSI of 48.5 and a negative share that exceeds the positive share. These operational outputs should be read as an illustration of end-to-end pipeline functionality rather than as a standalone validation exercise. Even so, Table 12 shows why aggregate neutrality should not be overinterpreted. Weak scores in labor relations, financial performance, and regulatory and public affairs indicate that the brand remains exposed to dimensions that can trigger rapid reputational deterioration when operational disturbances, financial pressure, or regulatory controversies occur. In practical diagnostic terms, the framework shows how broad sentiment direction can be separated from possible drivers of vulnerability, which is useful for prioritizing corrective attention even though the aspect layer itself remains exploratory.

The forecasting layer adds a further level of strategic relevance, but in the present article it should be interpreted as an applied and functional demonstration rather than as conclusive evidence of predictive superiority. Figure 4 shows that a weakly negative observed endpoint can be followed by a short-horizon projection that remains slightly negative but moves gently toward neutrality. This is precisely the kind of information that static reporting cannot provide. If reputation analytics are treated only as retrospective scorekeeping, managers learn what has already happened; if they are treated as an early-warning system, managers gain a basis for deciding whether a temporary recovery is likely to persist or whether additional intervention is needed. In airline settings, where disruptions cascade quickly across customer experience, media coverage, labor discourse, and investor perception, this forward-looking capability is particularly important. In the present study, however, its role is to demonstrate how temporal projection can be integrated into the framework in an operationally interpretable way.

The new holdout comparison clarifies the evidential status of the predictive component. The selected model improves on naive, exponential-smoothing, Holt damped-trend, and ARIMA benchmarks in RMSE terms, but the margin over the linear trend is small. The forecast is therefore best read as a lightweight early-warning extension to descriptive monitoring, not as a stand-alone forecasting contribution.

Taken together, these findings support the view that the developed framework is best understood as an operational decision-support architecture. Its importance lies in com-

binning several kinds of visibility that are rarely available at the same time in conventional monitoring arrangements: broad directional sentiment measurement, exploratory aspect-level diagnosis, thematic clustering, and short-horizon forecasting. Surveys remain useful for controlled sampling and causal interpretation, but they are structurally too infrequent to serve as the sole instrument in a high-velocity information environment. Airlines increasingly operate under conditions in which customer narratives, media framing, employee discourse, and platform amplification can interact within hours. Under those conditions, the industry benefits from always-on tools that reduce reaction lag and make reputational change legible while it is still actionable.

A concrete illustration is the United Airlines Flight 3411 crisis in April 2017. After video of a passenger being forcibly removed from an overbooked flight spread rapidly across social media, the company's initial response was widely criticized as insufficient and misaligned with public expectations. Subsequent research on the crisis showed how online stakeholder reactions amplified reputational damage, while contemporary reporting indicated that the company lost hundreds of millions of dollars in market value and at one point nearly a billion dollars intraday before later announcing changes to customer-handling procedures [56–59]. This case illustrates that, once a negative narrative becomes platform-native and emotionally salient, the financial consequences of delayed recognition can materialize very quickly.

A useful theoretical exercise is to imagine the developed system operating in that situation. Query-scoped acquisition would have captured the rapidly growing volume of posts, articles, and comments within the first hours of the incident. The sentiment layer would have registered an abrupt negative shift; the aspect layer would likely have concentrated criticism in customer treatment, overbooking policy, staff behavior, and corporate response; the topic module would have surfaced themes around passenger removal, apology adequacy, compensation, and procedural legitimacy; and the forecasting layer would have indicated whether the negative trajectory was likely to persist beyond the first media cycle. The built model fed on those signals could have triggered escalation thresholds for senior management much earlier than manual monitoring alone, enabling a faster apology, clearer corrective messaging, immediate policy suspension, and better coordination between operations and communications. The claim is not that the tool would have eliminated the crisis, but that it could have reduced diagnostic delay and improved the precision of the response. In settings where a few hours of hesitation can convert online outrage into enduring demand, trust, and valuation losses, that reduction in response latency is economically and strategically meaningful.

A further implication of the developed architecture is computational accessibility. The framework was intentionally designed around a balance between semantic capability and modest infrastructure requirements: compact retrieval rules, multilingual sentiment inference that remains usable without GPU clusters, zero-shot aspect assignment for cold-start taxonomy deployment, transparent TF-IDF plus KMeans topic induction, and lightweight time-series forecasting. For moderate monitoring workloads, this combination can be operated on standard office machines and, in many practical settings, even on low-end CPU-only hardware. That matters strategically because many airlines, airport operators, consultants, and regulators do not lack interest in reputation intelligence; they lack the resources to maintain large AI stacks, expensive cloud inference, or dedicated data-engineering teams. In that sense, lightweight design is not a secondary engineering preference but part of the managerial contribution of the framework since it lowers the threshold for adoption and enables more continuous monitoring in organizations with limited technical capacity.

The use of public online content for reputation intelligence raises ethical and legal constraints that must be treated as part of the methodology rather than as an afterthought.

The framework should be deployed only within an admissible source scope, respecting platform terms of use, robots and access restrictions, applicable data-protection rules, and institutional review requirements where relevant. It should not attempt to deanonymize users, infer sensitive personal attributes, or publish raw user-level text when aggregate indicators are sufficient for managerial analysis. Operational reports should therefore emphasize aggregated sentiment, aspect, topic, and temporal patterns, with source-level records retained only for audit, deduplication, and error analysis under controlled access.

There are also fairness and validity risks. Public online discourse is not a representative sample of all passengers or stakeholders: it may overrepresent highly dissatisfied users, highly engaged loyal customers, media-amplified crises, particular languages, or platform-specific demographics. For that reason, the system should be interpreted as a high-frequency sensing complement to surveys and operational data not as a replacement for controlled customer research. Human review remains necessary when outputs may inform consequential decisions such as service recovery priorities, crisis communication, labor-response strategy, or public-sector assessment.

5. Conclusions

In this paper, a multilingual, computationally lightweight framework for airline reputation intelligence has been formalized. The framework is built around transformer-based sentiment analysis, zero-shot aspect categorization, topic clustering, and forecasting. Both a methodological and a practical contribution are provided: noisy public text is translated by the proposed system into interpretable transport-management indicators such as the NSI, positive and negative shares, aspect scores, ranked thematic clusters, and short-horizon sentiment outlooks. A brand-generic, time-aware, extensible architecture is therefore specified for deployment without a large dedicated infrastructure.

The view that lightweight AI-based reputation intelligence can serve as a promising complement to traditional survey-only monitoring is supported by the analysis. Immediacy, scale, thematic granularity, and deployability on ordinary CPU-based hardware without dedicated GPU infrastructure are identified as its strongest advantages. Through this portability, the practical use of reputation analytics is broadened beyond large carriers with specialized technical teams. At the same time, because the current validation is based on modestly sized annotated subsets, the present evidence should be interpreted as supporting proof-of-concept functionality, relative benchmarking, and applied demonstration rather than strong claims of operational superiority or substitution.

A viable architecture for supporting online market and reputation research in aviation is demonstrated by the developed system. Its main contribution is located in methodological architecture, reproducible formalization, lightweight integration, deployability, and the adaptation of generic computational components to an applied airline-reputation context. Precisely because the framework is lightweight, it can be developed not only as an internal strategic tool for large airlines but also as a deployable monitoring solution for smaller carriers, airports, consultancies, and public institutions that need timely intelligence without large-scale computing resources.

Several limitations are associated with the current framework. The data layer remains constrained by public scraping routes. Comment extraction is supported only for HyperText Markup Language (HTML), not for JavaScript-based systems. The aspect classifier is implemented as a zero-shot model rather than as a domain fine-tuned model. Topic induction via TF-IDF and KMeans is simpler than embedding-based methods such as BERTopic, although this simplicity is partly adopted as a deliberate design objective so that the lightweight and deployable character of the system can be preserved. Most importantly, the present validation evidence is still based on a relatively small annotated subset rather

than on a large, public, multilingual benchmark corpus. A further limitation is introduced by the legal and ethical sensitivity of public online content: platform rules, data-protection duties, anonymity expectations, and sampling bias constrain how raw text can be collected, stored, shared, and interpreted.

Future work should therefore be organized along three connected directions. First, a dedicated multilingual airline benchmark for sentiment and aspect annotation should be constructed so that model performance can be evaluated in future studies on a shared empirical basis. Second, the data layer should be expanded through the integration of additional platform connectors so that broader and more stable coverage of relevant public discourse can be obtained. Third, the current zero-shot aspect module should be progressively replaced or complemented by domain-adapted models that are trained on aviation-specific language and are therefore better aligned with the semantic structure of airline reputation narratives.

Funding: The present work was performed within the scope of activities at the Aeronautics and Astronautics Research Center (AEROG) of the Laboratório Associado em Energia, Transportes e Aeroespacial (LAETA), and was supported by the Fundação para a Ciência e a Tecnologia (Project doi: 10.54499/UIDB/50022/2020, 10.54499/UIDP/50022/2020 and 10.54499/LA/P/0079/2020).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The article reports a pilot workflow and illustrative outputs. No raw user-level online text or personally identifying material is released with the article.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Antunes, F.; Freire, M.; Melo, P.; Costa, J.P. From emotional data to decisions: A systematic review on how airlines use sentiments and emotions to stay ahead. *J. Air Transp. Manag.* **2026**, *131*, 102911. [\[CrossRef\]](#)
2. 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. [\[CrossRef\]](#)
3. Liu, Y.; Huang, J.; Huang, F.; Zhu, Z.; Ma, L. Deep learning-driven sentiment analysis for customer satisfaction optimization in airline reviews. *Ind. Manag. Data Syst.* **2025**, *126*, 833–864. [\[CrossRef\]](#)
4. Farzadnia, S.; Raeesi Vanani, I. Identification of opinion trends using sentiment analysis of airlines passengers' reviews. *J. Air Transp. Manag.* **2022**, *103*, 102232. [\[CrossRef\]](#)
5. 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. [\[CrossRef\]](#)
6. Syed, A.A.; Gaol, F.L.; Boediman, A.; Budiharto, W. Airline reviews processing: Abstractive summarization and rating-based sentiment classification using deep transfer learning. *Int. J. Inf. Manag. Data Insights* **2024**, *4*, 100238. [\[CrossRef\]](#)
7. Dellarocas, C. The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms. *Manag. Sci.* **2003**, *49*, 1407–1424. [\[CrossRef\]](#)
8. Chevalier, J.A.; Mayzlin, D. The effect of word of mouth on sales: Online book reviews. *J. Mark. Res.* **2006**, *43*, 345–354. [\[CrossRef\]](#)
9. Tirunillai, S.; Tellis, G.J. Does Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance. *Mark. Sci.* **2012**, *31*, 198–215. [\[CrossRef\]](#)
10. Doganis, R. *Flying Off Course: Airline Economics and Marketing*, 5th ed.; Routledge: Abingdon, UK, 2019.
11. Belobaba, P.; Odoni, A.; Barnhart, C. *The Global Airline Industry*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2024.
12. Fornell, C. A national customer satisfaction barometer: The Swedish experience. *J. Mark.* **1992**, *56*, 6–21. [\[CrossRef\]](#)
13. 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.
14. Verhoef, P.C.; Lemon, K.N.; Parasuraman, A.; Roggeveen, A.; Tsiros, M.; Schlesinger, L.A. Customer experience creation: Determinants, dynamics and management strategies. *J. Retail.* **2009**, *85*, 31–41. [\[CrossRef\]](#)
15. Zeithaml, V.A.; Berry, L.L.; Parasuraman, A. The behavioral consequences of service quality. *J. Mark.* **1996**, *60*, 31–46. [\[CrossRef\]](#)

16. Mansouri, S.; Moradpour, S.; Prentice, C. An alternative approach to understanding airline customers' attitudes and behaviors. *J. Air Transp. Manag.* **2026**, *133*, 102960. [\[CrossRef\]](#)
17. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135. [\[CrossRef\]](#)
18. Cambria, E.; Schuller, B.; Xia, Y.; Havasi, C. New avenues in opinion mining and sentiment analysis. *IEEE Intell. Syst.* **2013**, *28*, 15–21. [\[CrossRef\]](#)
19. Hutto, C.J.; Gilbert, E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; pp. 216–225.
20. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the EMNLP, Doha, Qatar, 25–29 October 2014; pp. 1746–1751. [\[CrossRef\]](#)
21. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the EMNLP, Doha, Qatar, 25–29 October 2014; pp. 1532–1543. [\[CrossRef\]](#)
23. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781. [\[CrossRef\]](#)
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
25. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [\[CrossRef\]](#)
26. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692. [\[CrossRef\]](#)
27. Madeira, T.; Melício, R.; Valério, D.; Santos, L.F.F.M. Machine Learning and Natural Language Processing for Prediction of Human Factors in Aviation Incident Reports. *Aerospace* **2021**, *8*, 47. [\[CrossRef\]](#)
28. Nogueira, R.P.R.; Melicio, R.; Valério, D.; Santos, L.F.F.M. Learning Methods and Predictive Modeling to Identify Failure by Human Factors in the Aviation Industry. *Appl. Sci.* **2023**, *13*, 4069. [\[CrossRef\]](#)
29. Lázaro, F.L.; Nogueira, R.P.R.; Melicio, R.; Valério, D.; Santos, L.F.F.M. Human Factors as Predictor of Fatalities in Aviation Accidents: A Neural Network Analysis. *Appl. Sci.* **2024**, *14*, 640. [\[CrossRef\]](#)
30. Lázaro, F.L.; Madeira, T.; Melicio, R.; Valério, D.; Santos, L.F.F.M. Identifying Human Factors in Aviation Accidents with Natural Language Processing and Machine Learning Models. *Aerospace* **2025**, *12*, 106. [\[CrossRef\]](#)
31. Lázaro, F.L.; Santos, L.F.F.M.; Valério, D.; Melicio, R. Artificial Intelligence and Aviation: A Deep Learning Strategy for Improved Data Classification and Management. *Appl. Sci.* **2025**, *15*, 9403. [\[CrossRef\]](#)
32. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the ACL, Online, 5–10 July 2020; pp. 8440–8451. [\[CrossRef\]](#)
33. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
34. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
35. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. *arXiv* **2022**, arXiv:2203.02155. [\[CrossRef\]](#)
36. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. [\[CrossRef\]](#)
37. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2023**, arXiv:2303.18223. [\[CrossRef\]](#)
38. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, arXiv:2108.07258. [\[CrossRef\]](#)
39. Brody, S.; Elhadad, N. An unsupervised aspect-sentiment model for online reviews. In Proceedings of the Human Language Technologies: The 2010 Annual Conference of the NAACL, Los Angeles, CA, USA, 2–4 June 2010; pp. 804–812.
40. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In Proceedings of the SemEval, San Diego, CA, USA, 16–17 June 2016; pp. 19–30. [\[CrossRef\]](#)
41. Schouten, K.; Frasinca, F. A survey on aspect-level sentiment analysis. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 813–830. [\[CrossRef\]](#)
42. Yin, W.; Hay, J.; Roth, D.; Schütze, H. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In Proceedings of the EMNLP-IJCNLP, Hong Kong, China, 3–7 November 2019; pp. 3914–3923. [\[CrossRef\]](#)
43. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Int. J. Mach. Learn. Technol.* **2011**, *2*, 37–63. [\[CrossRef\]](#)
44. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [\[CrossRef\]](#)

45. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [[CrossRef](#)]
46. Krippendorff, K. Reliability in content analysis: Some common misconceptions and recommendations. *Hum. Commun. Res.* **2004**, *30*, 411–433. [[CrossRef](#)]
47. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
48. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794. [[CrossRef](#)]
49. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [[CrossRef](#)]
50. MacQueen, J.B. Some methods for classification and analysis of multivariate observations. *Berkeley Symp. Math. Stat. Probab.* **1967**, *5.1*, 281–297.
51. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [[CrossRef](#)]
52. Mirzaalian, F.; Halpenny, E. Social media analytics in hospitality and tourism: A systematic literature review and future trends. *J. Hosp. Tour. Technol.* **2019**, *10*, 764–790. [[CrossRef](#)]
53. Zhang, S.; Yao, L.; Sun, A.; Tay, Y.L. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.* **2019**, *52*, 1–38. [[CrossRef](#)]
54. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*, 3rd ed.; Prentice Hall: Hoboken, NJ, USA, 2023.
55. Liu, B. *Sentiment Analysis and Opinion Mining*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2012. [[CrossRef](#)]
56. Ma, J.; Tse, Y.K.; Wang, X.; Zhang, M. Examining customer perception and behaviour through social media research—An empirical study of the United Airlines overbooking crisis. *Transp. Res. Part E Logist. Transp. Rev.* **2019**, *127*, 192–205. [[CrossRef](#)]
57. Cheng, Y.; Shen, H. United Airlines crisis from the stakeholder perspective: Exploring customers' ethical judgment, trust and distrust, and behavioral intentions. *Public Relat. Rev.* **2020**, *46*, 101908. [[CrossRef](#)]
58. CNNMoney. *United Loses \$250 Million of Its Market Value*; CNNMoney: New York, NY, USA, 2017.
59. United Airlines. *United Airlines Announces Changes to Improve Customer Experience*; United Airlines: Chicago, IL, USA, 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.