Identifying User Pain Points in Airline Mobile Apps: A Term Co-Occurrence Network Analysis of App-Store Reviews

Yu Chen*

Department of Information Management, College of Business, Chinese Culture University, Taipei, Taiwan

***Correspondions Author**

DOI: https://doi.org/10.51244/IJRSI.2026.1311000020

Received: 02 November 2026; Accepted: 07 November 2026; Published: 19 November 2026

ABSTRACT

Mobile apps are now a main touchpoint between airlines and passengers, and user reviews are a rich, unsolicited record of what goes wrong and what works. This study identifies the issues that airline-app users raise most often, using term co-occurrence network analysis of public app-store reviews. On 1 June 2026 I collected the most recent reviews of the apps of nine airlines (three Taiwan-based and six international carriers) from Google Play (n = 2,563) and the Apple App Store (n = 3,147) in Taiwan and the United States, covering January 2022 to May 2026. After removing 437 duplicate texts and 21 emoji-only reviews, 5,258 reviews remained, of which 62.3% were not written in Chinese and were machine-translated into Traditional Chinese. The 5,164 reviews with at least one retained keyword formed the analytic corpus. Eighty TF-IDF keywords were linked by co-occurrence (at least 25 reviews), visualised in VOSviewer, and grouped into seven interpretive themes, whose prevalence is the share of reviews containing their core terms: technical failures (32.8%), overall experience (26.6%), login and account management (23.2%), flight service (23.2%), booking and transactions (18.0%), membership and inquiry (10.8%), and check-in (7.4%). Failure terms were the most prominent, were linked to every other theme, were concentrated in low-rated reviews (44.4% of one- and two-star versus 8.6% of four- and five-star reviews), and were similar across platforms, carrier types, and languages (30.6% to 35.5%). The other themes differed strongly by language and storefront, though not by platform, and are provisional. Only 0.2% of reviews mentioned artificial intelligence or chatbots. The findings suggest that stability and access deserve priority and yield propositions, based on the DeLone and McLean model, for later survey research. Limitations include the capped, recency-weighted sample, machine translation, and the descriptive design.

Keywords: Airline Mobile App; User-Generated Content; Text Mining; Co-Occurrence Network; User Pain Points

INTRODUCTION

Mobile applications have become a central touchpoint in the airline customer journey. Passengers use them to search and book flights, check in, select seats, obtain electronic boarding passes, follow flight status, and manage loyalty accounts. For airlines, an app is both an operational lever and a competitive differentiator: it shifts routine tasks to self-service technology (SST), which allows customers to produce a service outcome without direct staff involvement (Meuter et al., 2000), and it provides a direct channel for member relationships. Recent studies describe the digitalisation of aviation processes as a response to complexity and efficiency pressure (Karafakıoğlu, 2026) and treat the airline app as an operant resource whose smooth technical operation decides whether value is created or destroyed (Onat & Devran, 2026).

A high number of downloads does not by itself show that an app works for its users. Whether an app is used again depends on whether it works reliably and meets users' needs at each stage of the journey. Traditional ways of collecting such feedback, such as interviews and questionnaires, are costly and time-consuming, and are subject to small samples, group bias, and observer effects. User-generated content (UGC), such as app-store

reviews, offers a complementary source. Reviews are written spontaneously, at scale, and close to the moment of use, and they have been used to identify the features that lead users to stay or to abandon an app (Wang et al., 2024).

Text-mining studies of airline apps exist, but they have mainly examined individual carriers or other regions. Balcıoğlu (2025) analysed reviews of Turkish Airlines apps, and Onat and Devran (2026) combined topic modelling and sentiment analysis for airline apps.

This study addresses three research questions:

RQ1. Which issues do users of airline mobile apps most frequently raise in their reviews?

RQ2. How are these issues related, that is, which terms and themes are central and which bridge different parts of the journey?

RQ3. What do the patterns imply for the improvement priorities of airline app managers?

I answer these questions with a term co-occurrence network analysis of 5,164 reviews from Google Play and the Apple App Store. The study contributes a transparent, reproducible description of user concerns, and a map that links the concerns to established quality dimensions of information systems and can guide later survey work. The remainder of the paper reviews related work (Section 2), describes the method (Section 3), reports the results (Section 4), and discusses implications and limitations (Section 5).

BACKGROUND

2.1 Airline apps as self-service technology

SST is a technological interface that lets customers produce a service independently of direct service-employee involvement (Meuter et al., 2000). In airline services the mobile app is among the most representative SST interfaces. It allows travellers to book, select seats, check in, obtain electronic boarding passes, and manage trips anywhere and at any time, and it reduces the workload at counters, which matters where labour is constrained. Text-mining of airline-app reviews shows that users are particularly concerned with functionality, efficiency, and smoothness, the dimensions that drive customer satisfaction (Balcıoğlu, 2025). The smoothness of digital services such as online check-in has also been identified as an important influence on journey satisfaction (Mirthipati, 2024).

Service-quality research gives a vocabulary for interpreting such concerns. In the information systems success model, system quality, information quality, and service quality are the core quality dimensions that influence user satisfaction and use (DeLone & McLean, 2003). In SERVQUAL, reliability and responsiveness are among the dimensions that shape perceived service quality (Parasuraman et al., 1988). An analysis of airline reviews with topic modelling found that reliability was the most salient dimension before COVID-19 and responsiveness during it (Kim, Lim, & Ha, 2024). These frameworks guide our interpretation of the clusters but are not imposed on the data.

2.2 User-generated content and text mining

User-generated content. UGC refers to text, images, audio, video, or other content created by users and published online. For organisations with a user-centred strategy, UGC is a cheaper and more spontaneous source of experience feedback than interviews or surveys, and it has become a key data source for analysing why users continue or stop using mobile apps (Wang et al., 2024). In the airline industry, reviews are also treated as a digital footprint that helps airlines improve service processes and track trends (Onat & Devran, 2026).

Topic modelling, sentiment analysis, and co-occurrence networks. Topic modelling is an unsupervised method that groups latent semantic structure in a text collection; the topics reflect different aspects of the user experience and can be linked to satisfaction and dissatisfaction (Wang et al., 2024). Previous airline-app work found that

experiences revolve around app performance, system updates, the booking and boarding process, and login problems, and that text mining can reveal hidden dimensions, such as crashes or features that fail after an update, that questionnaires rarely capture (Balcıoğlu, 2025). Sentiment analysis, also called opinion mining, identifies the feelings, opinions, or attitudes expressed in text (Onat & Devran, 2026). A term co-occurrence network represents keywords as nodes and links two keywords when they appear together in the same documents, so that central terms and clusters of related concerns can be seen at a glance. I use this approach because it is transparent and easy to inspect, and because it does not require a predefined number of topics.

METHODOLOGY

3.1 Research design

I used an exploratory, descriptive text-mining design. User reviews were collected from Google Play and the Apple App Store, cleaned, translated where necessary, segmented, and converted into a keyword co-occurrence network. The analysis identifies the issues users raise and how they are related; it does not test hypotheses or estimate causal effects.

3.2 Data collection and corpus

The sample covers the official apps of nine airlines (Table 1). These carriers were purposely selected to encompass both the primary home-market providers and the major international airlines most frequently utilized by travelers in Taiwan: the three Taiwan-based full-service carriers (EVA Air, China Airlines, and STARLUX) representing domestic aviation leadership and new-entrant operations, alongside six prominent international carriers spanning key regional and long-haul routes, namely Cathay Pacific (Hong Kong), Japan Airlines and ANA (Northeast Asia), Singapore Airlines (Southeast Asia), and Turkish Airlines and Emirates (long-haul). This selection provides a balanced representation of domestic market conditions and international travel experiences.

The raw corpus contains 5,710 reviews posted between 2 January 2022 and 31 May 2026: 2,563 from Google Play (914 Taiwan locale, 1,649 US locale) and 3,147 from the App Store (1,334 Taiwan storefront, 1,813 US storefront). Because only the newest reviews up to the caps were retrieved, the corpus is not a complete record of each app\u2019s reviews. It over-represents recent years (66% of cleaned reviews were posted in 2025 or 2026), and the number of reviews per app reflects the caps and the review volume of each store as well as the popularity of the app. The ratings are polarised: in the cleaned corpus 51.2% of reviews give one star and 26.4% give five stars (mean rating 2.46).

Table 1. Airline apps and raw review counts by platform

Airline	Google Play package name	Carrier type	Google Play	App Store	Total
China Airlines	com.chinaairlines.mobile30	Taiwan	397	619	1,016
EVA Air	com.evaair.android	Taiwan	400	405	805
STARLUX	com.starlux.cale	Taiwan	192	204	396
Cathay Pacific	com.xs2theworld.cxmobile	Hong Kong	400	702	1,102
Japan Airlines	jp.co.jal.dom	Northeast Asia	225	240	465
ANA	jp.co.ana.android.tabidachi	Northeast Asia	229	116	345
Singapore Airlines	com.amadeus.merci.client.ui	Southeast Asia	224	125	349
Turkish Airlines	com.turkishairlines.mobile	Long-haul	231	523	754
Emirates	com.emirates.ek.android	Long-haul	265	213	478
Total			2,563	3,147	5,710

Raw corpus (n = 5,710). The three Taiwan-based carriers contribute 2,217 reviews (38.8%) and the six international carriers 3,493 (61.2%).

Three corpora are distinguished throughout the paper. The raw corpus contains the 5,710 collected reviews, none of which shares a review identifier with another. The cleaned corpus (n = 5,258) is obtained by removing 437 reviews whose translated text is identical to that of an earlier review and 21 reviews consisting only of emoji or symbols; six reviews were both, so 452 reviews were removed in total. Reviews with identical text come from different reviewers and are typically short generic comments; they were removed so that they do not inflate keyword frequencies. Retaining them changes the share of each group in Table 3 by at most 1.5 percentage points. The analytic corpus (n = 5,164) contains the cleaned reviews that retain at least one keyword after segmentation and stop-word filtering (94 reviews retained none). All keyword frequencies and percentages in this paper (Tables 2\u20135) are computed on the analytic corpus. Descriptive counts that do not depend on keyword filtering, namely language composition, mentions of AI features, and the rating distribution, are computed on the cleaned corpus. Each table states its denominator.

3.3 Language and machine translation

Only 1,980 of the 5,258 cleaned reviews (37.7%) contained Chinese characters in their original text. The other 3,278 (62.3%) were written in English or other languages (3,263 contained Latin letters) and were machine-translated into Traditional Chinese with GPT-3.5 on 8 June 2026, with a prompt that asked the model to translate the original review. Language follows the storefront: 1,870 of the 2,094 cleaned reviews from the Taiwan sources (89.3%) are in Chinese, and 3,054 of the 3,164 from the US sources (96.5%) are not. Reviews that already contained Chinese were passed through the same step, which changed the wording of 756 of the 2,055 Chinese-language reviews in the raw file (for example, character normalisation and light editing).

Machine translation can introduce bias into keyword extraction, because one source word may be rendered by several Chinese words (for example, “login” as 登入 or 登錄) and because translated and native text differ in word choice. I therefore (1) report group frequencies separately for reviews originally written in Chinese and for translated reviews, and by carrier type and platform (Section 4.4), (2) report the effect of adding common translation variants to the term lists, and (3) validated a random sample of 100 translated reviews (random seed 2026) by comparing each translation with the original Of these, 98 translations were judged by the research team to accurately preserve the original meaning and key problem terms.

3.4 Pre-processing and stop words

The Traditional-Chinese text was segmented with the jieba package. Tokens of one character, tokens consisting only of digits, punctuation, or Latin letters, and URLs were removed. The stop-word list covered (1) function words and prepositions, (2) personal pronouns, (3) generic intensifiers such as “very” and “really”, (4) generic app terms such as “app”, “application”, and “program”, and (5) the full forms of generic company terms (such as “airline company”) and place names, in order to limit confounding by individual brands. The stop-word list removed the full form “airline company” (航空公司) but not the fragment “airline/aviation” (航空), which the segmenter also produces from carrier names and phrases such as “Turkish Airlines”. This fragment and the verb “use” (使用) are therefore generic terms that remain in the network (Figure 1). Because they carry no specific meaning, they are excluded from the interpretive group definitions (Table 3) and from the ranking of frequent terms (Table 2); their frequencies are reported separately. The stop-word list and code are kept by the authors and are not publicly shared.

3.5 TF-IDF keyword extraction and network construction

Keyword extraction used TF-IDF weighting (Salton & Buckley, 1988), computed per review with the scikit-learn implementation (raw term frequency, smoothed inverse document frequency, L2 normalisation) on the segmented tokens. Terms had to appear in at least 12 reviews and in no more than 65% of reviews. The 80 terms with the highest summed TF-IDF weight across reviews were retained as nodes . Two keywords co-occur when both appear in the same review; co-occurrence is counted once per review (binary), and the link strength of two

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue XI November 2026

keywords is the number of reviews in which both appear. Two keywords were linked when their link strength was at least 25. Network density is the number of links divided by the number of possible links.

The network was visualised and clustered in VOSviewer 1.6.x (van Eck & Waltman, 2010) with association strength normalization, a resolution parameter of 1.0, minimum cluster size of 2, 500 random starts, and random seed 2026.. VOSviewer assigned the nodes to nine clusters. Based on the semantic orientation of the core terms in each cluster and on their position in the network, we then grouped the keywords post hoc into seven interpretive groups (Table 3). The grouping is a researcher judgement. It was checked by reading original reviews that contained each group's core terms; no independent second coder was used reading a random sample of 20 reviews per group. The network of the 80 keywords has a density of 0.102 .

Centrality. We use two node measures: the number of links of a node (degree) and its total link strength, the sum of the link strengths of all its links. Both are reported by VOSviewer. These measures describe how widely a term co-occurs with other terms and do not imply causal or temporal order.

Group prevalence. For each interpretive group, we report the number and share of analytic reviews that contain at least one of its core terms (substring matching on the translated text). Because the term "mileage" belongs to two groups, reviews often contain terms of several groups, the shares overlap and do not sum to 100%. I did not run sentiment classification or topic-model inference, and we make no claim that co-occurrence shows causal relations.

RESULTS

4.1 Overall structure of the network

Figure 1 shows the co-occurrence network of the 80 keywords (node labels are the original Traditional-Chinese terms; English glosses are given in Table 3). The terms "unable" (無法) and "use" (使用) appear as the largest nodes with the most links, which shows that they co-occur with a wide range of other terms. They are tightly connected with "problem", "display", "error", and "retry". Table 2 gives the document frequencies of the ten most frequent content terms. The generic terms "use" (845 reviews, 16.4%) and "airline" (850 reviews, 16.5%) are excluded from the ranking.

Figure 1. Term co-occurrence network of airline app reviews (VOSviewer; node labels in Traditional Chinese).

Table 2. Most frequent content terms (document frequency in the analytic corpus, n = 5,164)

Rank	Term (Chinese)	English gloss	Reviews containing term	% of analytic reviews
1	無法	unable	1,011	19.6
2	航班	flight	655	12.7
3	預訂	booking	545	10.6
4	服務	service	405	7.8
5	更新	update	404	7.8
6	登入	login	402	7.8
7	糟糕	terrible	389	7.5
8	問題	problem	376	7.3
9	登機	boarding / check-in	346	6.7
10	輸入	input / enter	331	6.4

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue XI November 2026

Counts are substring matches on the translated text, computed on the analytic corpus. Generic terms ("use", "airline") are excluded.

4.2 Interpretive groups

Table 3 summarises the seven interpretive groups, with the share of analytic reviews that contain their core terms. The paragraphs below describe each group, using the cluster colours of Figure 1 where helpful.

Table 3. Interpretive groups of the co-occurrence network (analytic corpus, n = 5,164)

Group	Core terms (English gloss)	Focus of reviews	Related quality dimension (interpretive)	Reviews with ≥1 core term, n (%)
1. Technical failures	unable, problem, display, error, retry	Cannot use the app, errors, display faults	System quality	1,693 (32.8)
2. Login and account management	login, update, password, input, account, mileage	Login, password entry, app updates	System quality	1,197 (23.2)
3. Booking and transactions	booking, ticket, itinerary, cancel, change	Booking, changes, cancellation	System quality; information quality	928 (18.0)
4. Flight service and travel practice	flight, service, seat, boarding, airport, baggage	Flight information, seats, baggage, airport service	Information quality	1,199 (23.2)
5. Check-in	process/handle, boarding	Check-in and electronic boarding pass	System quality; information quality	384 (7.4)
6. Membership and inquiry	mileage, member, query, itinerary	Mileage, membership functions, queries	Information quality	559 (10.8)
7. Overall experience and ease of use	experience, terrible, try, website, convenient, easy, electronic, receive	Mixed positive and negative overall evaluations	System quality; overall satisfaction	1,374 (26.6)

The generic terms "use" and "airline" are not counted as core terms. The last-but-one column is the authors' interpretive link to the quality dimensions of DeLone and McLean (2003) and is not derived statistically. Shares overlap across groups. Adding common translation variants (登錄 for login, 訂票 for booking, 報到 for check-in) raises the shares of groups 2, 3, and 5 to 24.2%, 19.4%, and 9.4%, and leaves the other groups unchanged.

Group 1: Technical failures. The hub terms "unable" (blue node) and "use" (purple node; a generic term that is not counted as a failure term) co-occur with the cyan nodes "problem", "display", and "error", and with "retry". The group shows a high concentration of operational vocabulary and indicates general concern about whether the app can be used at all. Failure terms appear in about one in three analytic reviews (32.8%).

Group 2: Login and account management. This group, mostly yellow-green nodes, is centred on "login", "update", "input", "password", and "account". It is closely linked to Group 1 and to the cyan node "update", so login-related words often appear together with operational problem words. Account verification therefore appears as a separate area of concern.

Group 3: Booking and transactions. Centred on "booking", "change", and "cancel" (green nodes), this group sits in the middle of the network and is linked to both the technical-failure group and the flight-service group.

Because of this bridging position, booking terms often occur together with “ticket” (pink-purple node) and with the blue node for ticket ordering.

Group 4: Flight service and travel practice. Red nodes such as “service” connect with the orange node “seat” and the large green node “flight”. The group covers the core airline business, including flight information, baggage, and airport service. The generic fragment “airline” (red node) is shown in the figure but not counted in the group share.

Group 5: Check-in. The orange nodes “handle” and “boarding” lie next to Group 4 and extend it towards the check-in process and electronic boarding passes.

Group 6: Membership and inquiry. “mileage”, “member”, “query”, and “itinerary” form a loosely structured group at the periphery of the network, reflecting membership functions and related notifications.

Group 7: Overall experience and ease of use. The brown nodes “try”, “website”, “terrible”, and “experience” and the pink-purple nodes “electronic”, “convenient”, “easy”, and “receive” lie in the middle of the network and connect to the technical-failure, booking, and flight-service groups. The vocabulary mixes negative (“terrible”) and positive (“convenient”, “easy”) evaluations and also includes words on trying the app and receiving notifications. We did not classify the sentiment in this group and describe the co-occurring words only.

4.3 Customer support and artificial intelligence features

Customer-support vocabulary (客服, 客戶服務, and similar terms) appears in 226 analytic reviews (4.4%). Reviews rarely discuss AI features. A keyword search of the cleaned corpus (n = 5,258) for AI, chatbot, intelligent customer service, and similar terms matched 12 reviews (0.2%; keyword lists in Appendix A), six from Taiwan-based and six from international carriers. Eleven of the 12 gave one or two stars; the one five-star review praised an “AI function” for trip planning. On reading, seven of the 12 referred to an AI or chatbot feature of the app, and six of these were critical, for example describing the chatbot or AI assistant as unhelpful or unable to resolve a problem. The other five used the term in a different sense (for example, a reviewer who asked ChatGPT for help, or a comparison with low-quality AI output). A further 17 reviews mention robot-verification (CAPTCHA) steps, which concern login and payment flows and not AI services. In the reviews I collected, operational reliability is therefore a far more common concern than AI features.

4.4 Robustness: platform, carrier type, and language

Because the corpus pools two platforms, two countries, nine apps, and two groups of languages, and because 62.3% of the reviews were machine-translated, I repeated the group counts for subsets of the analytic corpus (Table 4). The prominence of technical failures is similar in every subset (30.6% to 35.5%). The other groups differ in some splits but not in others. The profile of the 34 core terms is almost identical on Google Play and the App Store (Spearman $\rho = 0.89$, $p < .001$). It is moderately similar for the Taiwan-based and the international carriers ($\rho = 0.55$, $p < .001$). It is uncorrelated between reviews originally written in Chinese and translated reviews ($\rho = -0.03$, $p = .88$) and between Taiwan and US sources ($\rho = 0.08$, $p = .66$). Language and source country therefore matter far more than platform. Language also matters within the same carriers: for the three Taiwan-based carriers, Chinese-language and translated reviews are uncorrelated ($\rho = 0.08$, $p = .67$), with login and membership terms more frequent in Chinese reviews and flight-service and overall-experience terms more frequent in translated reviews.

Table 4. Share of analytic reviews containing the core terms of each group, by subset (%)

Group	All	Taiwan carriers	International carriers	Original Chinese	Translated	Google Play	App Store
n of reviews	5,164	2,124	3,040	1,947	3,217	2,209	2,955
1. Technical failures	32.8	35.5	30.9	30.6	34.1	30.6	34.4

Group	All	Taiwan carriers	International carriers	Original Chinese	Translated	Google Play	App Store
2. Login and account	23.2	30.0	18.4	29.6	19.3	20.9	24.9
3. Booking and transactions	18.0	14.7	20.2	10.4	22.5	14.9	20.3
4. Flight service	23.2	16.3	28.1	10.0	31.2	18.7	26.6
5. Check-in	7.4	5.4	8.8	1.7	10.9	7.0	7.7
6. Membership and inquiry	10.8	16.6	6.8	17.5	6.8	7.7	13.2
7. Overall experience	26.6	20.8	30.7	15.0	33.6	24.7	28.1

Taiwan carriers = EVA Air, China Airlines, STARLUX. The language subsets and the platform subsets each cover all 5,164 reviews. Because the Taiwan sources are almost entirely in Chinese and the US sources almost entirely in English, language cannot be separated from storefront; and because the US sources describe the same apps for different readers, translation effects cannot be separated from population effects.

I conclude that the pooled prevalence of failure terms is robust, whereas the prevalence of the other six groups depends on the language and storefront of the reviewers and is provisional. It would be reasonable to analyse Chinese-language and English-language reviews separately in later work.

4.5 Themes by star rating

Star ratings are available for all reviews. In the cleaned corpus (n = 5,258) 2,692 reviews (51.2%) give one star, 504 (9.6%) two stars, 377 (7.2%) three stars, 299 (5.7%) four stars, and 1,386 (26.4%) five stars. Table 5 shows how the groups are distributed over rating bands. Failure terms occur in 44.4% of one- and two-star reviews and in only 8.6% of four- and five-star reviews, and login and account terms show the same pattern (31.4% versus 6.6%). In contrast, flight-service terms and overall-experience terms are about as frequent in four- and five-star reviews (22.1% and 22.3%) as in low-rated reviews, which suggests that these groups also capture positive comments. Customer-support vocabulary appears mainly in low-rated reviews: 179 of the 226 reviews that mention it (79.2%) give one or two stars. These are descriptive associations in self-selected reviews and not estimates of the effect of failures on ratings.

Table 5. Share of analytic reviews containing the core terms of each group, by star rating (%)

Group	1–2 stars (n = 3,173)	3 stars (n = 375)	4–5 stars (n = 1,616)
1. Technical failures	44.4	38.4	8.6
2. Login and account management	31.4	25.1	6.6
3. Booking and transactions	21.6	22.7	9.7
4. Flight service and travel practice	23.5	25.6	22.1
5. Check-in	9.0	8.0	4.1
6. Membership and inquiry	13.9	12.0	4.5
7. Overall experience and ease of use	29.0	24.8	22.3

4.6 Summary

Reviews cluster around operational usability, account access, the booking process, flight service, and

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue XI November 2026

functions. Failure vocabulary is the most prominent part of the network, is linked to nearly every other group, is concentrated in low-rated reviews, and is the only theme whose prevalence is similar across platforms, carrier types, and languages. This pattern suggests that technical problems are not confined to one function but are raised throughout the journey, from login to booking and check-in. As the analysis is descriptive, these observations identify topics for further study rather than confirm relationships.

DISCUSSION

5.1 Interpretation in light of the IS success model and prior studies

The findings can be read through the quality dimensions of the IS success model (DeLone & McLean, 2003). Technical failures, login, and update problems correspond to system quality. They are the most prominent theme, the most stable across subsets, and the most concentrated in low-rated reviews, and they co-occur with booking and check-in terms, which suggests that a failure at an entry point blocks the functions behind it. Flight information, membership, and booking content correspond to information quality. Service quality, understood as the support available when something goes wrong, appears mainly as complaints about customer-service channels (4.4% of reviews), including chatbots that did not resolve problems. Overall evaluations in Group 7 correspond to user satisfaction. Table 6 summarises these links and compares them with earlier empirical studies.

Table 6. Themes, IS success model dimensions, and related prior findings

Theme	Dimension (DeLone & McLean, 2003)	Evidence in this study	Related prior findings
Technical failures; login and update	System quality	Failure terms in 32.8% of reviews, similar across platforms, carrier types, and languages; 44.4% of 1–2-star versus 8.6% of 4–5-star reviews; login and update terms in 23.2%	Performance, updates, and login are recurring themes (Balcıoğlu, 2025); technical problems can destroy value (Onat & Devran, 2026); reliability is a salient service dimension (Kim, Lim, & Ha, 2024)
Booking and check-in	System quality; information quality	Booking terms in 18.0% and check-in terms in 7.4% of reviews; both co-occur with failure terms	Smooth online check-in influences journey satisfaction (Mirthipati, 2024); booking and boarding processes recur in reviews (Balcıoğlu, 2025)
Flight service; membership and inquiry	Information quality	Flight-service terms in 23.2% and membership terms in 10.8%; both differ strongly by language and storefront	Users value functionality and efficiency of self-service apps (Balcıoğlu, 2025; Meuter et al., 2000)
Customer-support complaints	Service quality	Support terms in 4.4% of reviews, 79.2% of them rated 1–2 stars; six of seven comments on AI or chatbot features were critical	Reliability and responsiveness shape perceived quality (Parasuraman et al., 1988); responsiveness gained salience during COVID-19 (Kim, Lim, & Ha, 2024)
Overall experience	User satisfaction	Mixed positive and negative evaluations in 26.6% of reviews	Satisfaction is the central outcome of the model (DeLone & McLean, 2003)

Compared with earlier work, the results agree on the prominence of performance, update, login, and booking problems (Balcıoğlu, 2025; Onat & Devran, 2026), and they add three points. First, they quantify how much of the review corpus failure vocabulary occupies and show that this is the one theme that does not depend on platform, carrier type, or language, and that it is concentrated in one- and two-star reviews. Second, they show that failure terms co-occur with terms of every other theme, so that system quality appears as a gateway to the other dimensions. Third, they show that the perceived quality of customer support, including chatbots, is raised mainly when a failure has already occurred.

5.2 Theoretical contribution

The study contributes to the IS success literature in two ways. It applies the quality dimensions to unsolicited, journey-wide user reviews instead of survey items, which shows where the dimensions become visible to users in a self-service airline context. It also generates propositions for later survey and structural-equation research, which this study does not test: (P1) system quality, especially at login and update steps, acts as a gateway dimension, so that its failure weakens the effect of information quality and service quality on satisfaction; (P2) the perceived quality of customer support is contingent on failure, because it becomes salient mainly after a problem has occurred; and (P3) the weight of information-related themes differs between user groups, so that quality measures should be tested for invariance across language or market groups before they are pooled. These propositions follow from the descriptive patterns above and need to be tested with validated scales.

5.3 Implications for practice

First, for airline app managers, the network suggests that stability and access to the app should take priority over new features, because failure terms are the most prominent and connect with all journey stages. Second, login and account-verification flows deserve separate attention, since about one in four reviews (23.2%) contains login or account terms and this group is closely tied to update-related problems. Third, booking, change, and cancellation flows (18.0% of reviews) sit between technical and service concerns and are natural places for user-experience improvement, for example through clearer error messages and a way to retry after a failed step. Fourth, because customer-support complaints, including those about chatbots, arise mainly after failures, a quick route to a human agent when the app fails may matter more than adding automated features. Check-in terms appear less often (7.4%), but this share cannot show whether check-in works well or is simply used less often.

5.4 Limitations

Several limitations apply. First, the corpus is a capped, recency-weighted sample: only the newest reviews were retrieved (up to 200 per app and Google Play locale, or 10 pages per app and App Store storefront), 66% of the cleaned reviews date from 2025 or 2026, the number of reviews per app reflects these caps, and reviewers are self-selected (51.2% of reviews give one star). The results therefore describe the concerns expressed in recent public reviews and not those of all users. Second, the corpus pools nine apps from two platforms and two countries. I report robustness by platform and carrier type (Section 4.4), but I did not analyse the apps separately, and the numbers of reviews per app (345 to 1,102 in the raw corpus) are small for such analyses. Third, 62.3% of the cleaned reviews were machine-translated. Section 4.4 shows that only the prominence of failure terms is similar in original-Chinese and translated reviews. Because language is almost entirely determined by storefront, translation effects, reviewer population, and the apps described cannot be separated, and the pooled prevalence of the other themes is provisional. The translation was checked on a sample of 100 reviews only. Fourth, the analysis is descriptive: co-occurrence does not show causality or strength of association, the associations with star ratings are not effects, the grouping of keywords into interpretive themes involves researcher judgement and was checked by reading original reviews without an independent second coder, and the seventh group was not sentiment-classified. Fifth, keyword shares are based on substring matching and overlap across groups; the thresholds (80 keywords, a co-occurrence threshold of 25) are analyst choices. In an independent re-run of the network in Python on the analytic corpus, varying the threshold (15, 25, 35) and the number of keywords (60, 80, 100) changed the network density (0.067 to 0.255) but the same five terms (unable, use, flight, booking, airline) remained the most connected in every setting. Future work could analyse the apps and the Chinese-language and English-language reviews separately, collect the complete review history of each app, validate the

machine translation more extensively, apply topic modelling and sentiment classification with validation against manually coded samples, and test the propositions above with survey data.

CONCLUSION

This study analysed 5,164 Google Play and App Store reviews of nine airline apps with a term co-occurrence network and identified seven themes: technical failures, login and account management, booking and transactions, flight service, check-in, membership and inquiry, and overall experience and ease of use. Failure-related vocabulary was the most prominent (32.8% of reviews), was connected with all other themes, was concentrated in one- and two-star reviews, and was the only theme whose prevalence was similar across platforms, carrier types, and languages. Operational reliability and access to the app therefore appear to be the most visible concern of users, and the other themes need confirmation in language-specific analyses. The results give airline app managers a data-driven view of where users meet problems and give researchers propositions for studying how system quality, information quality, and service quality shape satisfaction with airline apps.

Declarations

Ethics approval. This study analysed publicly available, anonymized app-store reviews. Institutional review board approval was not required as the data were non-interventional and publicly accessible.

Conflict of interest. The author declares no conflict of interest.

Data availability. The review texts were collected from a public platform and are not redistributed because of the platform's terms of service. The stop-word list and analysis code are not publicly shared. Summary results supporting the findings are reported in the article.

Funding. This research received no external funding.

Author contributions. Yu Chen is the sole author of this article and is responsible for all aspects of the research, including conceptualisation, data curation, formal analysis, visualisation, and writing.

REFERENCES

1. Balcıoğlu, Y. S. (2025). Decoding customer sentiments in Turkish Airlines mobile apps: A comprehensive text mining approach. Journal of Aviation, 9(2), 321–330. https://doi.org/10.30518/jav.1553809
2. DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. Journal of Management Information Systems, 19(4), 9–30.
3. Karafakıoğlu, E. (2026). The future of digitalization in the aviation industry: Trends and academic approaches. Journal of Aviation. Advance online publication. https://doi.org/10.30518/jav.1781183
4. Kim, D., Lim, C., & Ha, H.-K. (2024). Comparative analysis of changes in passenger's perception for airline companies' service quality before and during COVID-19 using topic modeling. Journal of Air Transport Management, 115, Article 102542. https://doi.org/10.1016/j.jairtraman.2024.102542
5. Meuter, M. L., Ostrom, A. L., Roundtree, R. I., & Bitner, M. J. (2000). Self-service technologies: Understanding customer satisfaction with technology-based service encounters. Journal of Marketing, 64(3), 50–64. https://doi.org/10.1509/jmkg.64.3.50.18024
6. Mirthipati, T. (2024). Enhancing airline customer satisfaction: A machine learning and causal analysis approach (arXiv:2405.09076). arXiv. https://doi.org/10.48550/arXiv.2405.09076
7. Onat, S., & Devran, B. B. (2026). Examining consumer reviews of airline mobile applications using topic modelling and sentiment analysis. Journal of Aviation. Advance online publication. https://doi.org/10.30518/jav.1866610
8. Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. Journal of Retailing, 64(1), 12–40.

9. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
10. van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3
11. Wang, T., Wang, W., Feng, J., Fan, X., Guo, J., & Lei, J. (2024). A novel user-generated content-driven and Kano model focused framework to explore the impact mechanism of continuance intention to use mobile APPs. Computers in Human Behavior, 157, Article 108252. https://doi.org/10.1016/j.chb.2024.108252

Appendix A. Keyword Lists For Ai-Related And Robot-Verification Mentions

Reviews of the cleaned corpus (n = 5,258) were searched in both the translated text and the original text, case-insensitively. AI and chatbot mentions (12 reviews): AI; 人工智慧; 智慧客服; 智能客服; 聊天機器人; 客服機器人; 機器人客服; chatbot; chat bot; ChatGPT; virtual assistant; 智能助理; 虛擬助理; 智慧助理. Of the 12 matched reviews, seven referred to an AI or chatbot feature of the app. Robot-verification mentions (17 reviews): 不是機器人; 非機器人; 人機驗證; 驗證 followed within six characters by 機器人; reCAPTCHA; captcha; robot (in the original text).

Appendix B. App Identifiers

Airline	App name	Google Play package name	App Store identifier
China Airlines	China Airlines App	com.chinaairlines.mobile30	1093021450
EVA Air	EVA AIR	com.evaair.android	443099960
STARLUX	STARLUX	com.starlux.cale	1450459916
Cathay Pacific	Cathay Pacific	com.xs2theworld.cxmobile	305038764
Japan Airlines	Japan Airlines	jp.co.jal.dom	351785536
ANA	ANA	jp.co.ana.android.tabidachi	382028327
Singapore Airlines	Singapore Airlines	com.amadeus.merci.client.ui	515287841
Turkish Airlines	Turkish Airlines Flight Ticket	com.turkishairlines.mobile	1283414961
Emirates	Emirates	com.emirates.ek.android	935307149