Information available at the Journal of Asia Business and Service Innovation (http://www.absi.kr/)

Asia Business and Service Innovation

ISSN: 3092-0558 (Print) | eISSN: 3092-1112 (Online)

Review Topics Driving Star Ratings in Food Delivery Apps: A Cross-Country Comparison of Baemin (Korea) and Grab (Indonesia)

Hak-Seon Kim[1,2], Hyunwoo (David) Joung[2*]

1 School of Hospitality & Tourism Management, Kyungsung University, Busan, Republic of Korea

2 Department of Nutrition and Hospitality Management, The University of Mississippi, Oxford 38677, USA

* Corresponding author: hjoung@olemiss.edu

ARTICLE INFO

Keywords:
food delivery applications;
online reviews;
topic modeling;
Latent Dirichlet Allocation;
cross-country comparison;
customer satisfaction;
super-app;
Baemin;
Grab;
Gojek;
Maxim

ABSTRACT

This study compares the latent themes of online consumer reviews and their effects on customer satisfaction, operationalized as star ratings, between leading mobile food-delivery applications (FDAs) in two Asian markets: Baemin in the Republic of Korea and three Indonesian super-apps that all bundle ride-hailing with food delivery — Grab, Gojek, and Maxim. Because the Indonesian super-apps pool reviews across functionally distinct services, we apply a lexical-filter procedure that retains only food-delivery-related reviews (containing tokens such as makanan, pesan, grabfood, gofood; excluding ride-hailing tokens such as taksi, grabcar, gocar). Five hundred reviews per app were collected from the Apple App Store on April 29, 2026; the food-delivery share of the review channel was 38.0% for Grab, 13.0% for Gojek, and 7.0% for Maxim, indicating that the ride-hailing-anchored super-apps (Gojek, Maxim) carry far less food-delivery signal than the food-delivery-leading Grab. After filtering, the analytic samples were N = 476 for Baemin and N = 290 for the combined Indonesian food-delivery corpus. Latent Dirichlet Allocation (LDA) extracted five topics per market; star ratings were regressed on the document-topic proportions with HC3 robust standard errors, and the Indonesian model included app fixed effects to absorb baseline differences across the three platforms. Both models were statistically significant (Baemin: $F(4,471) = 7.53$, $p < .001$, adj. $R^2 = .058$; Indonesia: $F(6, 283) = 6.09$, $p < .001$, adj. $R^2 = .094$). For Baemin, delivery-time delay ($B = -1.01$, $p = .002$) and customer-service handling ($B = -0.94$, $p = .005$) emerged as significant negative drivers of star ratings. For combined Indonesia, the long wait / order cancellation topic drove ratings down significantly ($B = -0.68$, $p = .025$) once app fixed effects were controlled. Topic prevalences also differed across the three Indonesian apps in theoretically interesting ways: Maxim users voiced disproportionate concern with delivery fees and pricing (42.5% topic mass), while Grab and Gojek users emphasized waiting and cancellation. Theoretical, methodological, and managerial implications are discussed.

1. Introduction

The global mobile food-delivery market has expanded into one of the largest segments of digital consumer commerce. Industry estimates place the worldwide online food-delivery market at approximately USD 288.84 billion in 2024, with projections of USD 505.50 billion by 2030 at a compound annual growth rate of 9.4%, and the Asia-Pacific region accounted for over 41% of global revenue in 2024 (Grand View Research, 2025). Within Asia, the Republic of Korea has become one of the most digitally penetrated single national markets: the Korean online food-delivery sector generated approximately KRW 27 trillion (about USD 16.6 billion) in 2024, and Baedal Minjok (Baemin) alone retained roughly 60% domestic market share with about 21.7 million users as of mid-2024 (Korea Economic Daily, 2024; Statista, 2024). Indonesia has simultaneously emerged as the largest food-delivery market in Southeast Asia, where three super-apps dominate the food-delivery channel: GrabFood led with a 47% national share (USD 2.54 billion gross merchandise value in 2024), followed by GoFood within the Gojek super-app at 35%, and ShopeeFood at 18% (Momentum Works, 2025); a smaller fourth player, Maxim,

operates a low-cost ride-hailing-and-delivery super-app that is growing rapidly across emerging-market cities. Regionally, Southeast Asia's food-delivery gross merchandise value reached USD 22.7 billion in 2025, growing 18% year-on-year, with Grab consolidating its leadership at a 55% six-country market share (Momentum Works, 2026). Mobile food-delivery applications (FDAs) have thus become a structural component of contemporary urban consumption rather than a peripheral convenience.

Within this market, FDAs occupy a distinctive position as multi-sided platforms that coordinate consumers, restaurant partners, and last-mile riders through a single mobile interface (Pillai et al., 2022). Two structurally different FDA designs dominate the Asia-Pacific landscape. The first is the single-purpose model exemplified by Baemin (Baedal Minjok) in Korea, which concentrates exclusively on food-delivery transactions. The second is the super-app model exemplified by Grab, Gojek, and Maxim in Indonesia and across Southeast Asia, which bundles ride-hailing, food delivery, parcel logistics, and digital payments under a single brand and login (Tan et al., 2018). The super-app design is conceptually significant because the user-review channel of a multi-service application mixes complaints from functionally distinct services and so cannot be assumed to map cleanly onto food-delivery satisfaction without further preprocessing. Importantly, super-apps are not equivalent to one another in this respect: each has a distinct service identity (food-led for Grab, ride-led for Gojek and Maxim) that shapes which service users tend to talk about in reviews. The Indonesian re-view setting in particular has been shown to yield rich, satisfaction-relevant signal when analyzed with text-mining tools, as in studies of Indonesian hospitality reviews during and after the COVID-19 period (Handani et al., 2022a). A second feature common to both FDA types is the centrality of user-generated reviews. App-store reviews now serve as a primary public archive of consumer experience and as a low-cost, real-time substitute for survey data in hospitality research (Berezina et al., 2016; Tao & Kim, 2022).

A growing literature has used probabilistic topic models, particularly Latent Dirichlet Allocation (LDA), to extract latent themes from such reviews and link those themes to satisfaction outcomes (Blei et al., 2003; Guo et al., 2017; Hu et al., 2019; Xu & Li, 2016). In the FDA domain specifically, Pillai et al. (2022) integrated the Theory of Planned Behavior, Perceived Risk Theory, and the Elaboration Likelihood Model to explain purchase intention, and Ray et al. (2019) applied a uses-and-gratifications perspective to FDA continuance. Closer to the present design, Kwon et al. (2021) combined LDA with sentiment scoring to predict airline satisfaction, and Shrivastav et al. (2025) com-pared review-based purchase intention between Baemin (Korea) and Foodmandu (Nepal). Despite this momentum, three gaps remain. First, no published study has combined unsupervised topic discovery with regression-based satisfaction modeling in a Korean-versus-Indonesian FDA comparison. Second, prior cross-country FDA work has compared single-purpose to single-purpose platforms and has not engaged the super-app review channel as an analytic challenge in its own right. Third, hospitality research has rarely explicitly filtered super-app review streams to isolate the food-delivery sub-corpus before modeling, leaving open the question of whether the topic-rating links re-ported in mixed-service reviews would survive when applied to a clean food-delivery sub-sample.

The purpose of this study is therefore to identify the latent review topics that drive star ratings in two leading Asian FDA markets, Baemin in Korea and a combined corpus of three Indonesian super-apps (Grab, Gojek, Maxim), under a methodological design that explicitly addresses the super-app channel issue by filtering each Indonesian app to its food-delivery sub-stream. Three research questions follow: (RQ1) which latent topics structure user reviews of Baemin and the food-delivery sub-corpus of the three Indonesian super-apps; (RQ2) to what extent those topics predict the star rating consumers assign once differences across the Indonesian platforms are controlled with app fixed effects; and (RQ3) whether the topic-rating relation-ships differ between the two markets, and whether the three Indonesian super-apps differ from each other in topic prevalence. To address these questions, we collected 500 reviews per app from the Apple App Store, applied a lexical-filter procedure to retain only the food-delivery-relevant Indonesian reviews, fitted LDA models with K = 5 to each market corpus, and regressed the five-point star rating on the document-topic proportions using ordinary least squares with HC3 heteroskedasticity-robust standard errors. The study contributes methodologically by linking unsupervised topic modeling to compositional regression in a single reproducible pipeline that explicitly handles super-app channel heterogeneity; substantively by documenting both the divergence in food-delivery review composition across three Indonesian super-apps and the differential topic-rating drivers between Korea and Indonesia; and managerially by translating the regression coefficients into platform-specific operational priorities

2. Literature Review

2.1. Online Reviews as Antecedents of Customer Satisfaction

A consistent finding in hospitality big-data research is that aggregated online reviews carry measurable signals about service quality and customer satisfaction. Berezina et al. (2016) showed that text-mined attributes from hotel reviews systematically distinguished satisfied from dissatisfied guests, and subsequent work has confirmed the pattern across multiple sub-sectors. Xu and Li (2016) used probabilistic topic modeling to identify hotel attributes whose

presence in reviews predicted overall ratings, and Guo et al. (2017) demonstrated that LDA topics extracted from TripAdvisor reviews captured the dimensionality of guest experience as well as theory-driven scales. Within Asian hospitality contexts, Tao and Kim (2022) showed that the textual content of café reviews predicts overall satisfaction beyond what the star rating alone conveys, and Hu et al. (2019) confirmed similar dynamics for Chinese restaurant platforms. Common to these studies is the observation that the content dimension of reviews encodes specific service attributes, and that those attributes can be recovered with text-mining tools and then carried into inferential models. Conceptually, these dynamics situate the present study within two overlapping frameworks.

First, app-store reviews are a canonical form of electronic word-of-mouth (eWOM): publicly posted, experience-based consumer evaluations that influence the decisions of prospective users and that platforms cannot fully control (Berezina et al., 2016; Tao & Kim, 2022). The latent topics recovered by LDA can therefore be read as the salient content dimensions of the eWOM signal that a given platform generates. Second, the recovered topics map onto established service-quality dimensions in hospitality, most directly the reliability and responsiveness dimensions of SERVQUAL-type frameworks, because the dominant themes in both corpora concern temporal reliability (delivery delay, long wait, cancellation) and responsiveness (customer-service handling, courier behavior). Framing the extracted topics as content dimensions of eWOM that index underlying service-quality attributes clarifies why topic proportions should predict the star rating and links the unsupervised topics to the broader satisfaction literature rather than treating them as purely data-driven artifacts. This logic is consistent with prior text-mining studies that recover satisfaction-relevant service attributes from the textual content of hospitality reviews (Handani et al., 2022b).

2.2. Topic Modeling in Hospitality and FDA Research

Latent Dirichlet Allocation, introduced by Blei et al. (2003), assumes that each document is a mixture of latent topics and each topic is a distribution over words, allowing researchers to recover thematic structure from large unlabeled corpora. The method has become standard in hospitality review analysis (Guo et al., 2017; Hu et al., 2019; Xu & Li, 2016) and was introduced to the airline-review setting by Kwon et al. (2021), who combined LDA topics with sentiment scoring to predict satisfaction. The advantage of LDA over pure semantic-network or word-frequency approaches is that document-topic proportions can be carried directly into inferential models, allowing researchers to quantify how each theme contributes to outcome variables such as satisfaction or behavioral intention. For FDA contexts specifically, Ray et al. (2019) used a mixed-method design to identify drivers of FDA continuance, and Pillai et al. (2022) demonstrated that integrating perceived risk and elaboration likelihood explains a substantial share of FDA purchase-intention variance. The LDA-plus-regression pipeline used here builds directly on this tradition.

2.3. Super-Apps and Cross-Country FDA Comparison

Super-apps—applications that bundle multiple consumer services under a single brand and login—have become a defining feature of Southeast Asian digital ecosystems, with Grab, Gojek, and Shopee as leading regional examples (Tan et al., 2018). The super-app design has direct implications for review analysis. Because users access several functionally distinct services through the same interface, app-store reviews of a super-app may mix complaints originating in unrelated service lines, and individual reviews seldom self-identify which service triggered the post (Tsai et al., 2020). This is in sharp contrast to a single-purpose application such as Baemin, where the entire review stream concerns one service. The methodological consequence is that any cross-country comparison between a super-app and a single-purpose app must explicitly isolate the service of interest before modeling. The most directly relevant prior study is Shrivastav et al. (2025), who compared Baemin (Korea) and Foodmandu (Nepal) and showed that the antecedents of purchase intention vary across markets even when both platforms are single-purpose. Indonesia is a theoretically interesting comparator because Grab operates as a super-app, so consumer expectations and failure modes are likely to differ from those of a single-purpose delivery brand, and because no published study to our knowledge has compared topic-level satisfaction drivers between a Korean and an Indonesian FDA using a unified LDA-plus-regression pipeline that explicitly filters for food-delivery content.

3. Methodology

3.1. Data Collection

Reviews were scraped from the Apple App Store on April 29, 2026 for four leading FDAs: Baemin (app id 378084485) for the Korean market, and three Indonesian super-apps that all bundle food delivery with ride-hailing — Grab: Taxi Ride, Food Delivery (app id 647268330), Gojek (app id 944875099), and Maxim: order a taxi & delivery (app id 579985456). Five hundred reviews per app were retained, all posted between February 1, 2026 and April 28, 2026 (Baemin) and between March 12, 2026 and April 28, 2026 (Indonesian apps). Each record contained the review identifier, country, author, star rating (1-5), title, content, language, and timestamp. Korean (ko) was the dominant language for Baemin (99.8%), and Indonesian-English code-mixed text (recorded as en) dominated the three Indonesian samples (99% or more in each). This review-mining design follows a growing stream of tourism and hospitality research that treats large-scale user-generated content as a

primary data source for understanding consumer experience and destination performance (Kim et al., 2025).

3.2. Lexical-Filter Procedure for the Indonesian Super-Apps

Because all three Indonesian apps are super-apps whose review channels pool complaints across ride-hailing, food delivery, parcel logistics, and digital payments, we applied a two-stage lexical filter to each Indonesian corpus before modeling. The inclusion list consisted of food-related tokens drawn from Indonesian and Indonesian-English code-mixed usage: makan, makanan, pesan, pesen, pesanan, orderan, grabfood, gofood, shopeefood, restoran, resto, restaurant, menu, beli, cafe, kafe, minuman, drink, food, eat, ayam, nasi, mie, kopi, coffee, snack, kue, roti, bakso, soto, and several others. The exclusion list consisted of ride-hailing tokens: taksi, taxi, grabcar, grabbike, grabtaxi, gocar, goride, motor, mobil, ojek, rute, route, penumpang, passenger, jemput, naik, nganter, bike. A review was retained only if it contained at least one inclusion-list token and no exclusion-list token. The inclusion and exclusion lists were finalized in three steps rather than chosen a priori. We first seeded each list with the official service names (grabfood, gofood, shopeefood; grabcar, gocar, goride) and with high-frequency domain terms from the Indonesian and Indonesian-English food and ride-hailing lexicons. Second, we generated a ranked unigram frequency table from the pooled unfiltered Indonesian corpus and manually inspected the top several hundred tokens, assigning each clearly food-related or clearly ride-related term to the corresponding list and discarding ambiguous tokens (for example, generic service words that occur in both channels) so that they would not by themselves trigger retention or exclusion. Third, two coders independently reviewed a random sample of 100 retained and 100 excluded reviews to estimate face validity; disagreements were used to add a small number of additional tokens before the final pass. We acknowledge, consistent with the reviewers' concern, that a keyword-dependent filter of this kind is necessarily imperfect in two directions. It can exclude genuine food-delivery reviews that happen to use none of the inclusion tokens, and it can admit reviews whose food-related token is incidental rather than central. The exclusion risk is the more serious of the two here, because service-generic complaints that mention neither food nor ride tokens account for the majority of each Indonesian channel and are dropped entirely; the retained food-delivery sub-corpus should therefore be read as a high-precision rather than high-recall sample. We treat this limitation explicitly in Section 6 and frame supervised classification as the preferred replacement for keyword filtering in future work.

The audit of the three unfiltered Indonesian corpora (N = 500 each) yielded a striking pattern of channel composition: food-related reviews accounted for 38.0% of Grab's review stream but only 13.0% of Gojek's and 7.0% of Maxim's, with the remainder split between ride-related, mixed, and service-generic complaints (Figure 4). The ordering aligns with the publicly stated brand identity of each platform: Grab, despite operating ride-hailing, holds the largest food-delivery market share in Indonesia (47% as of 2024) and consequently attracts proportionally more food-delivery voice in its review channel; Gojek originated as a ride-hailing app and continues to lead in motorbike-taxi mind-share; and Maxim is a low-cost ride-hailing entrant whose food-delivery service is comparatively peripheral. This finding is itself a methodological contribution of the present study: the same data-collection design applied to three nominally similar super-apps produced markedly different food-delivery yields, implying that any cross-app or cross-country review study that does not explicitly inspect channel composition risks pooling structurally different signals.

After applying the filter, the three Indonesian samples were combined into a single Indonesian food-delivery corpus to ensure comparability with the single-purpose Baemin sample. The combined corpus retained 290 reviews after token-level preprocessing (Grab N = 190, Gojek N = 65, Maxim N = 35). The Korean Baemin sample, which does not require filtering because the platform is single-purpose, retained 476 reviews. Mean star ratings were 1.64 (SD = 1.28) for Baemin and 1.75 (SD = 1.36) for the combined Indonesian corpus; the difference was not statistically significant (Welch t = -1.05, p = .295), indicating broadly comparable evaluative environments. App-level mean ratings within the Indonesian sample varied (Grab M = 1.52, Gojek M = 1.97, Maxim M = 2.57), motivating the inclusion of app fixed effects in the Indonesian regression model (see Section 3.4).

3.3. Text Preprocessing

Korean Baemin reviews were normalized by stripping non-Hangul characters, segmenting into morphemes, and extracting nouns using the Okt tokenizer of the KoNLPy library. A custom stop-word list removed generic platform terms (e.g., the words for "delivery," "app," "review") and Korean function morphemes. Indonesian reviews from the three super-app corpora, predominantly written in colloquial Indonesian and Indonesian-English code-mixing, were lower-cased, restricted to alphabetic strings of three or more characters, and filtered against a combined Indonesian-English stop list including platform-generic tokens (e.g., app, aplikasi, grab, gojek, maxim, use). Document-term matrices were built with a minimum-document frequency of three and a maximum-document frequency of 0.85, capping the vocabulary at 500 terms for each market corpus.

3.4. Latent Dirichlet Allocation and Topic-Level Regression

The analytic core of the study combines unsupervised topic modeling with a topic-level regression of customer satisfaction. Following standard practice (Blei et al., 2003; Kwon et al., 2021), we first estimated Latent Dirichlet Allocation (LDA) models with K {3, 4, 5, 6, 7} on each market corpus using the variational batch algorithm with doc-topic prior of 0.10, topic-word prior of 0.01, and a maximum of 50 iterations. Perplexity rose monotonically with K from K = 4 onward in both corpora, while interpretability of the topic word lists was substantially better at K = 5 than at lower values. Because comparable cross-country interpretability requires the same K, we adopted K = 5 for both market corpora. Topics were labeled by examining the top 12 weighted words and validated by independent inspection of high-probability documents per topic. The choice of K was therefore guided by three criteria applied jointly rather than by a single fit statistic: (a) held-out perplexity, which favored smaller K but did not by itself identify an interpretable solution; (b) semantic coherence and distinctiveness of the top-weighted word lists, assessed by two coders; and (c) the requirement of an identical K across the two market corpora to keep the cross-country comparison on common footing. Solutions at K = 3 and K = 4 merged conceptually separate failure modes (for example, collapsing delivery delay and cancellation into a single undifferentiated topic), whereas K = 6 and K = 7 split topics into near-duplicate fragments with overlapping top words and very low mean proportions, offering no interpretive gain. K = 5 was the smallest solution that produced topics that were both internally coherent and mutually distinct in both corpora, which is why it was retained. To reduce the subjectivity of topic labeling, each topic was independently labeled by two coders from the top 12 weighted words and a sample of high-probability documents; disagreements were resolved by discussion, and the two reviewer-flagged Indonesian topics whose labels could appear to overlap (I3 long wait/order cancellation and I5 cancellation by app/courier) were retained as separate topics only after confirming that their high-probability documents described distinct events: I3 captures consumer-perceived waiting that culminates in the consumer abandoning or cancelling the order, whereas I5 captures one-sided cancellation initiated by the app or courier without the consumer's consent. We note this distinction is interpretive and report it transparently rather than claiming the two topics are fully orthogonal.

The document-topic matrix produced by LDA was then used to construct predictor variables for an ordinary least squares (OLS) regression on the five-point star rating. Because document-topic proportions sum to one within a document and therefore form a simplex, the topic with the smallest mean proportion was excluded as the reference category to avoid perfect collinearity. Star rating was regressed on the remaining four topic proportions using HC3 heteroskedasticity-robust standard errors, and the model was estimated separately for each market. We model the five-point star rating with OLS rather than ordinal logistic regression for three reasons. First, the analytic interest is in the marginal effect of a unit change in a topic's document proportion on the expected rating, which OLS estimates directly and interpretably, whereas ordinal-logit coefficients are expressed on a latent log-odds scale that is harder to translate into the managerial priorities the study reports. Second, the rating distribution in both corpora is strongly concentrated at one star (about 75 to 79 percent of reviews), which leaves several categories sparsely populated and can make the proportional-odds assumption of ordinal logistic regression unstable; treating the heavily skewed rating as a quasi-continuous outcome with heteroskedasticity-robust (HC3) standard errors is a more conservative choice under these conditions. Third, OLS with robust standard errors on review star ratings is an established convention in hospitality and tourism review research that links text-derived predictors to ratings (Guo et al., 2017; Tao & Kim, 2022; Xu & Li, 2016), which keeps the present results comparable with that body of work. We treat the resulting coefficients as descriptive effect sizes rather than as estimates from a fully specified ordinal data-generating process. For the combined Indonesian model, two app-dummy variables (Gojek and Maxim, with Grab as the reference category) were added as fixed effects to absorb baseline differences in star ratings across the three super-apps that are not captured by topic content. Linking LDA topics to a regression in this way preserves the interpretability of unsupervised topic discovery while yielding effect-size estimates and significance tests for each topic, an approach increasingly used in hospitality review research (Guo et al., 2017; Hu et al., 2019; Kwon et al., 2021).

4. Results

4.1. Descriptive Findings

Both market samples exhibit a strongly bottom-skewed rating distribution typical of complaint-dominated app review samples (Tao & Kim, 2022). In the Baemin sample, 79.2% of reviews carry a one-star rating; in the combined Indonesian food-delivery sample the corresponding share is 75.5%. Mean ratings differ only modestly across the three Indonesian apps within the filtered subsample (Grab M = 1.52, Gojek M = 1.97, Maxim M = 2.57), motivating the inclusion of app fixed effects in the Indonesian regression model. Figure 1 displays the rating distributions for the two market corpora side by side.

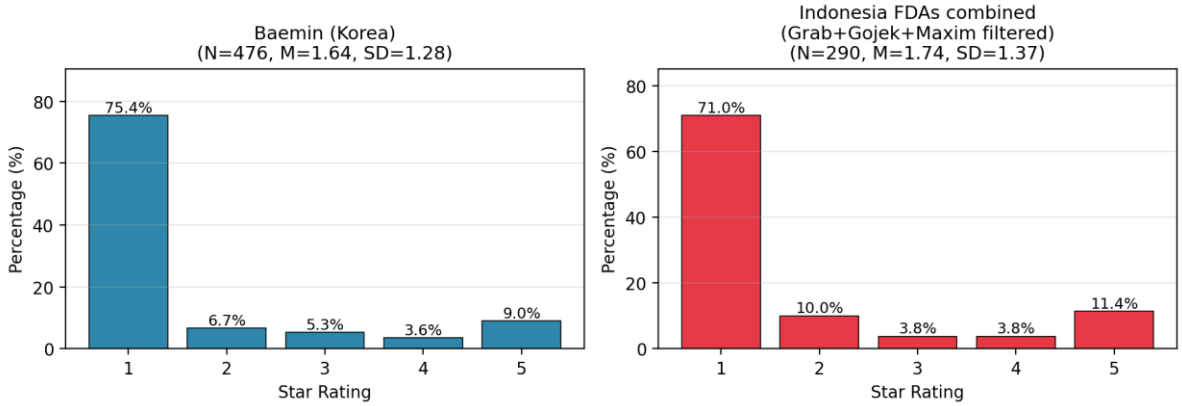

Figure 1. Star-rating distributions for Baemin (Korea) and the combined Indonesian food-delivery corpus.

4.2. Channel Composition Across Three Indonesian Super-Apps

Before turning to the LDA results, we report a methodologically important descriptive finding on the service-channel composition of the three Indonesian super-app review streams. As visualized in Figure 4, food-related reviews accounted for 38.0% of Grab's review channel, 13.0% of Gojek's, and 7.0% of Maxim's. Service-generic complaints — reviews containing neither food nor ride-hailing tokens — accounted for 53.4%, 80.4%, and 85.8% of the three corpora, respectively. The asymmetry tracks the publicly stated brand identity of each platform: Grab leads Indonesian food delivery and consequently attracts proportionally more food-delivery voice in its review channel, whereas Gojek and Maxim originate as ride-hailing platforms and their review channels remain dominated by ride-related and generic content. This finding is methodologically consequential: nominally similar super-app review datasets can carry very different food-delivery signal-to-noise ratios, and any cross-app or cross-country comparison must inspect channel composition before pooling.

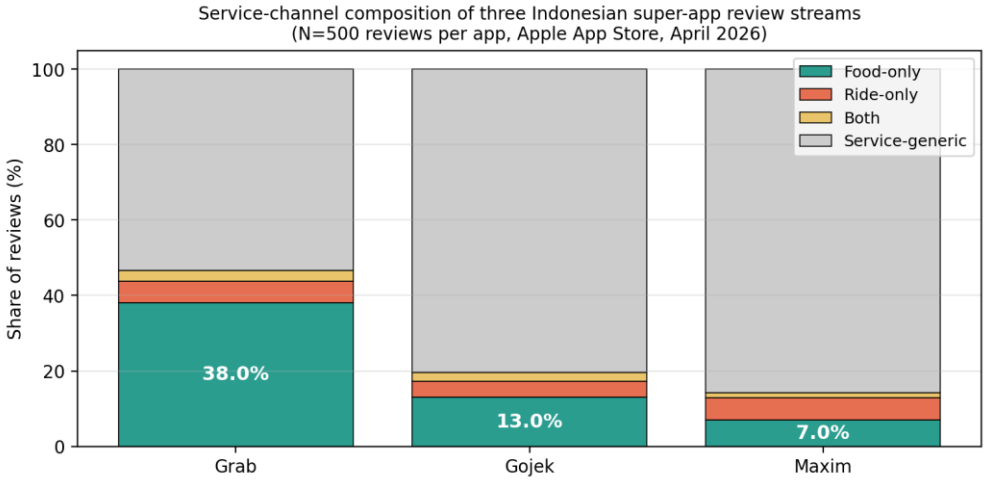

Figure 4. Service-channel composition of three Indonesian super-app review streams.

4.3. LDA Topics

Five interpretable topics emerged for each market corpus. Table 1 lists the topic labels assigned, the top six high-weight terms (translated where applicable), and the mean document-topic proportion across the corpus. A note on terminology is in order: in Indonesian-language reviews of food-delivery services, the word "driver" is used to refer to the food courier who delivers the order on a motorbike, not to a ride-hailing driver. We therefore translate "driver" as "courier" in the topic labels and discussion below to avoid misreading; the underlying token in the LDA word lists is the original Indonesian "driver."

For Baemin the dominant topic is delivery time delay (T4, mean = 0.295), characterized by terms such as "time," "rider assignment," "delay," "arrival," and "Coupang," reflecting frequent comparisons that Korean consumers draw

between Baemin and the rapidly growing competitor Coupang Eats. For the combined Indonesian food-delivery corpus the dominant topic is long wait and order cancellation (I3, mean = 0.274), with key terms "courier," "nunggu" (wait), "jam" (hour), "cancel," and "lama" (long). The remaining four Indonesian topics each capture a distinct food-delivery failure mode: restaurant–courier service conflicts (I1, covering reviews about how the restaurant and courier interact and how customer service mediates), delivery-fee and pricing concerns (I2), food-courier pickup failure (I4, capturing reviews where the courier accepts but does not pick up the order), and one-sided cancellation of the order by the app or courier (I5). All five Indonesian topics are saturated with food-delivery-specific vocabulary (grabfood, gofood, makanan, pesan, resto, pesanan, ongkir, menunggu), confirming that the lexical filter successfully isolated food-delivery content. Topic prevalence also varied across the three Indonesian apps within the combined corpus: Maxim users showed disproportionate concern with the delivery-fee/pricing topic (I2 prevalence = 0.425, vs. 0.127 for Grab and 0.210 for Gojek), reflecting Maxim's positioning as a low-cost platform; Grab and Gojek users emphasized waiting and cancellation themes more than Maxim users.

Table 1. LDA topic labels, top words, and topic prevalence

Market	Topic	Label	Top words	Mean prop.
aemin	T1	App / Tech & Update Issues	update, payment, version, screen, iPhone, language	0.178
aemin	T2	Riders & Coupang Competition	store, rider, function, Coupang, assignment, club	0.203
aemin	T3	Restaurant Service	restaurant, event, modify, block, improve, period	0.119
aemin	T4	Delivery Time Delay	time, dispatch, delay, rider, cooking, arrival	0.295
aemin	T5	Customer Service / Complaint	customer, coupon, center, consultation, cancel, problem	0.204
ndonesia	I1	Restaurant–Courier Service Conflicts	grabfood, sesuai (matching), makanan (food), marah (angry), resto, service	0.142
ndonesia	I2	Delivery Fee & Pricing	courier (driver), pesan, bayar (pay), ongkir (delivery fee), gofood, jauh (far)	0.181
ndonesia	I3	Long Wait & Order Cancellation	courier, nunggu (wait), jam (hour), cancel, pesen, lama (long)	0.274
ndonesia	I4	Food Courier Pickup Failure	courier, jam, makanan, susah (hard), tiba (arrive), dapet (get)	0.252
ndonesia	I5	Cancellation by App / Courier	courier, jam, lama, cancel, menunggu (waiting), dibatalkan (cancelled)	0.151

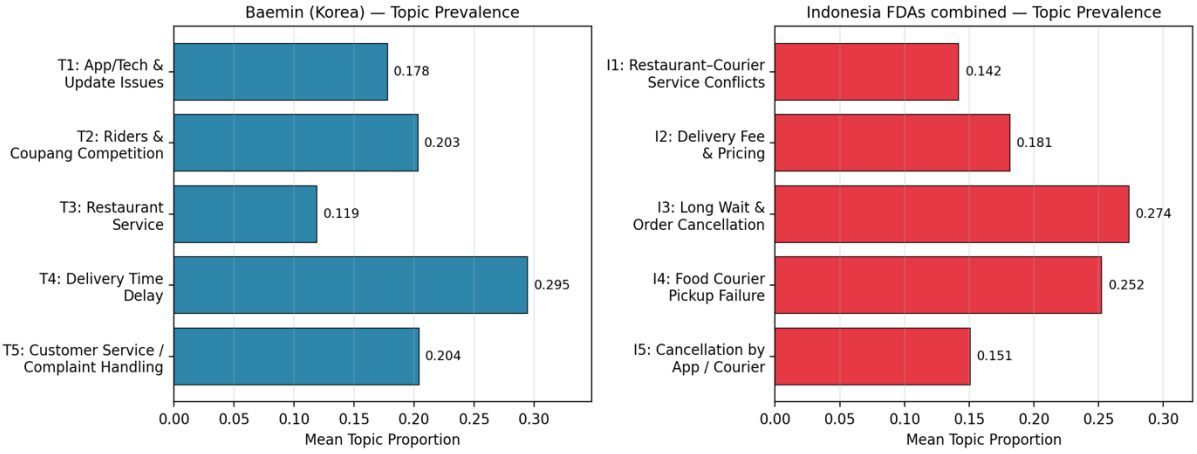

Figure 2. Mean topic prevalence by market corpus.

4.4. Regression of Star Rating on Topic Proportions

Table 2 reports the OLS regressions with HC3 robust standard errors. The Baemin model is statistically significant overall ($F(4, 471) = 7.53$, $p < .001$, adj. $R^2 = .058$). Two topics carry significant negative coefficients: delivery-time delay (B = -1.01, $p = .002$) and customer-service handling (B = -0.94, $p = .005$). The app-tech (T1) and rider/Coupang (T2) topics are not significant once the other topics are held constant, indicating that the headline driver of low ratings on Baemin is temporal failure on the order itself rather than the app or competitive context.

The combined Indonesian model with app fixed effects is also statistically significant overall ($F(6,283) = 6.09$, $p < .001$, adj. $R^2 = .094$) and explains a slightly larger share of rating variance than the Baemin model. The most consequential topic-level finding is for the long wait / order cancellation topic (I3, $B = -0.68$, $p = .025$), which significantly depresses star ratings across the three Indonesian super-apps once baseline app differences are controlled. The other three estimable topics (I2 delivery-fee/pricing, I4 food-courier pickup failure, I5 cancellation by app/courier) are not individually significant. The model also includes app fixed effects to absorb baseline rating differences across the three super-apps that are not attributable to topic content; these controls are not of substantive interest and are omitted from Table 2 for clarity.

Table 2. Topic-level regression of star rating

Predictor	B	SE	*t*	*p*	Sig.
Panel A. Baemin (Korea), N = 476					
Constant	2.20	0.29	7.57	< .001	***
T1 App / Tech	-0.38	0.36	-1.07	.285	
T2 Rider/Coupang	-0.01	0.37	-0.03	.977	
T4 Delivery Delay	-1.01	0.32	-3.14	.002	**
T5 Customer Service	-0.94	0.33	-2.83	.005	**
$R^2 = .066$, adj. $R^2 = .058$, $F(4,471) = 7.53$, $p < .001$					
Panel B. Indonesia combined with app fixed effects, N = 290					
Constant	1.68	0.27	6.13	< .001	***
I2 Delivery Fee / Pricing	0.11	0.38	0.30	.767	
I3 Long Wait / Cancellation	-0.68	0.30	-2.24	.025	*
I4 Food Courier Pickup Failure	0.14	0.34	0.41	.684	
I5 Cancellation by App / Courier	-0.12	0.37	-0.33	.740	
$R^2 = .113$, adj. $R^2 = .094$, $F(6, 283) = 6.09$, $p < .001$					

*Note. Reference (excluded) topics are T3 Restaurant Service for Baemin and I1 Order Accuracy for the Indonesian model, each chosen as the topic with the smallest mean proportion. The Indonesian model also includes app fixed effects (Grab as the reference category) to absorb baseline rating differences across the three super-apps; the fixed-effect coefficients are omitted from the table for clarity. Standard errors are HC3 robust. † p < .10, * p < .05, ** p < .01, *** p < .001.*

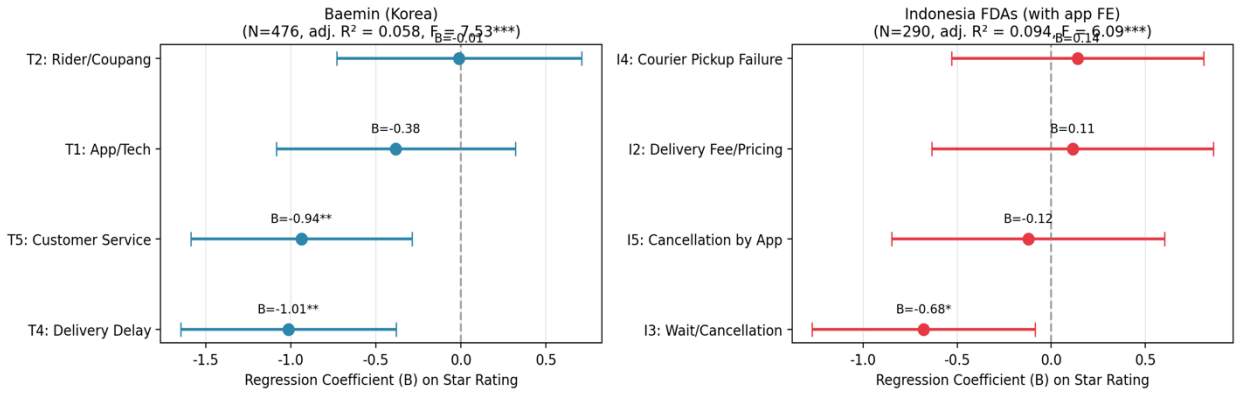

Figure 3. Regression coefficients for topic predictors with 95% CIs.

5. Discussion

5.1. What the Combined Indonesian Comparison Reveals

The first contribution is to characterize what is shared and what differs between the two markets once the three Indonesian super-app corpora have been filtered to their food-delivery sub-streams and combined into a single analytic sample. For Baemin, delivery-time delay (T4) and customer-service handling (T5) emerge as significant negative drivers of star ratings, with substantial effect sizes (B around -1.0 and -0.94, respectively). This pattern is consistent with the broader hospitality finding that time-related service failure has the largest single negative effect on review valence (Berezina et al., 2016; Hu et al., 2019), and is

also consistent with Pillai et al. (2022), who found that Korean FDA users place a heavy cognitive weight on perceived risk and on how the platform responds when something goes wrong.

For the combined Indonesian food-delivery corpus the long wait / order cancellation topic (I3) is the only topic with a statistically significant negative coefficient (B = -0.68, p = .025) once baseline differences across the three super-apps are controlled. The point estimate is meaningfully smaller than the comparable Baemin temporal-failure coefficient, suggesting that although temporal failure depresses ratings in both markets, its absolute weight on Indonesian food-delivery satisfaction is attenuated relative to the Korean case. The descriptive prevalence of the delivery-fee/pricing topic (I2) also varied dramatically across the three apps — 42.5% of topic mass for Maxim users, 21.0% for Gojek, and only 12.7% for Grab — reflecting Maxim's positioning as a low-cost super-app whose user base is more price-sensitive than the user base of the food-delivery-leading Grab. Although the I2 coefficient itself is not significant in the pooled regression, its differential prevalence across apps is itself a substantively meaningful finding for managers.

5.2. Theoretical Contributions

The study makes three theoretical contributions. First, it demonstrates that LDA topics, when interpreted carefully and entered as compositional regressors, are valid antecedents of digital satisfaction in both single-purpose and super-app FDAs (Guo et al., 2017; Kwon et al., 2021; Xu & Li, 2016). Second, by documenting the divergent service-channel composition of three Indonesian super-app review streams (38.0% food-related for Grab vs. 13.0% for Gojek and 7.0% for Maxim), it shows that platform identity rather than platform architecture alone determines the food-delivery signal-to-noise ratio of a super-app review channel. This refines the contingency view of digital service quality (Tan et al., 2018; Tsai et al., 2020) by suggesting that "super-app" is not a homogeneous category for review-based research. Third, by replicating the broader finding that time-failure dominates customer reviews in the Baemin sample and recovering a temporally framed driver (I3 long wait / order cancellation) in the combined Indonesian sample, the study strengthens the position that temporal reliability is a near-universal hospitality phenomenon while its absolute weight on satisfaction varies across markets.

5.3. Managerial Implications

For Baemin, the operational priority indicated by the regression is to compress the variance — not merely the mean — of delivery times, and to invest in complaint-handling channels that can resolve individual order failures before consumers post a low rating. The non-significance of the Coupang-comparison topic (T2) once delivery time is controlled for suggests that competitor salience is largely a downstream symptom of operational failure rather than an independent driver. For the three Indonesian super-apps, the significant negative coefficient on the long wait / order cancellation topic implies that dispatch reliability and timely courier assignment are the highest-priority operational levers across the food-delivery sub-stream. Differential topic prevalences across the three apps further suggest app-specific priorities: Grab and Gojek managers should focus on waiting and cancellation episodes, whereas Maxim managers should prioritize transparent communication of delivery fees and distance pricing (the dominant complaint cluster among Maxim food-delivery reviewers).

6. Limitations and Future Research

Several limitations should be noted. The most consequential arises from the lexical-filter procedure itself. Although filtering successfully isolated a food-delivery-specific Indonesian sub-corpus, it reduced the analytic Indonesian sample from 1,500 (combined unfiltered) to 290, with disproportionate loss for Gojek (N = 65) and Maxim (N = 35) given their predominantly ride-hailing review streams. Pooled estimation with app fixed effects partially mitigates this concern, but the small per-app sub-samples for Gojek and Maxim mean that app-specific topic-rating patterns cannot be estimated reliably. A keyword-based filter is also imperfect: it cannot recover food-delivery reviews that fail to mention any inclusion-list token (the service-generic complaints that account for 53–86% of each Indonesian corpus), and it may admit reviews whose food-related token is incidental rather than central. A more rigorous extension would replace lexical filtering with supervised classification: hand-label a balanced training subset of several hundred reviews per app as food-delivery, ride-hailing, financial-service, or general, and then train a multilingual classifier (for instance, a fine-tuned transformer in Indonesian and English) to label the full corpus. A further alternative, where data partnerships permit, is to obtain reviews scoped to GrabFood, GoFood, or comparable transactions only via the platforms' internal review channels rather than the cross-service App Store channel; such transaction-scoped reviews would eliminate the channel ambiguity entirely and would also enable larger-N replication of the regression model.

Second, all four samples are drawn exclusively from the Apple App Store. This is a non-trivial restriction, particularly for the Indonesian samples: Android holds the dominant share of mobile operating systems in Indonesia and in most Southeast Asian markets, so the iOS-only corpora may overrepresent reviewers in higher-income urban segments and underrepresent the modal Indonesian super-app user. The Korean Baemin sample is less affected by this bias because iOS market share in Korea is substantially higher, but the asymmetry in cross-platform coverage between the

two market samples is itself a comparability concern. A direct extension would replicate the analysis on Google Play reviews for all four apps, and on review streams from third-party app aggregators where available, and test whether the topic structure and topic-rating coefficients reported here generalize across review channels. A pooled cross-platform design (Apple App Store + Google Play, with the platform of origin entered as a covariate) would allow direct estimation of the channel effect and would also help recover the statistical power lost to lexical filtering by enlarging the Indonesian analytic sample, particularly for the Gojek and Maxim sub-corpora that are currently small.

Third, the Baemin model, although significant, explains a modest share of rating variance (adj. R^2 near .06), and the Indonesian model with app fixed effects explains a slightly larger but still modest share (adj. R^2 near .09). Both figures are consistent with the high latent-variance nature of star ratings reported in prior review studies (Tao & Kim, 2022) but suggest that incorporating reviewer demographics, app version, or order-level covariates would improve predictive power. Fourth, the cross-sectional snapshot covers a roughly three-month window in early 2026; longitudinal designs that track topic salience around platform updates or competitor entries would clarify causality. Future work could also extend the comparison to other Asian markets (e.g., Foodpanda in Thailand, Zomato/Swiggy in India) and to additional super-app versus single-purpose contrasts.

7. Conclusion

Using LDA-extracted topics as predictors in robust OLS regressions, this study examined the structural drivers of low star ratings in two leading Asian FDA markets — single-purpose Baemin in Korea and a combined corpus of three Indonesian super-apps (Grab, Gojek, Maxim) — under a methodological design that explicitly filtered the Indonesian super-app reviews to their food-delivery substreams and pooled them with app fixed effects. For Baemin, consumer dissatisfaction is anchored in temporal failure of the order and in customer-service handling, both of which significantly depress star ratings. For the combined Indonesian sample, the long wait / order cancellation topic significantly depresses ratings once baseline app differences are controlled, and topic prevalences differ in theoretically interpretable ways across the three super-apps. A separate descriptive finding — that food-related vocabulary appears in 38% of Grab reviews but only 13% of Gojek reviews and 7% of Maxim reviews — illustrates that the "super-app" label conceals substantial heterogeneity in review-channel composition and argues against pooling super-app review datasets without explicit channel inspection. Methodologically, the paper demonstrates that combining topic modeling with compositional regression and app fixed effects provides a transparent, reproducible pipeline for cross-country review research in hospitality and tourism, while also pointing toward supervised classification, multi-platform sampling, and transaction-scoped review data as the most promising next steps.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the use of publicly available, anonymized online data.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

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

References

Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: Text mining of online hotel reviews. *Journal of Hospitality Marketing & Management*, *25*(1), 1-24.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Grand View Research. (2025). *Online food delivery market size, share & trends analysis report, 2025-2030. Grand View Research.* https://www.grandviewresearch.com/industry-analysis/online-food-delivery-market-report

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using Latent Dirichlet Allocation. *Tourism Management*, *59*, 467-483.

Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management*, *72*, 417-426.

Korea Economic Daily. (2024, November 14). *South Korea's food delivery apps relish fresh boom in demand*. KED Global. https://www.kedglobal.com/retail/newsView/ked202411140002

Momentum Works. (2025). *Food delivery platforms in Southeast Asia 5.0*. Momentum Works.

Momentum Works. (2026). *Food delivery platforms in Southeast Asia 6.0*. Momentum Works.

Ray, A., Dhir, A., Bala, P. K., & Kaur, P. (2019). Why do people use food delivery apps (FDA)? A uses and gratification theory perspective. *Journal of Retailing and Consumer Services*, *51*, 221-230.

Statista. (2024*). Number of users for the most downloaded food delivery applications in South Korea as of May 2024*. https://www.statista.com/statistics/1103232/south-korea-leading-food-delivery-app-user-numbers/

Tan, F. T. C., Salo, J., Juntunen, J., & Kumar, A. (2018). A catalytic strategic approach for the development of a digital platform business: The case of Grab. *Journal of Strategic Information Systems*, *27*(4), 366-379.

Tsai, T. H., Chen, W. C., Hu, S. Y., & Wang, J. C. (2020). Cross-cultural differences in user perceptions of mobile applications: A comparison of Asian markets. *International Journal of Information Management*, *51*, 102019.

Xu, X., & Li, Y. (2016). The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. *International Journal of Hospitality Management*, *55*, 57-69.

Kwon, H. J., Ban, H. -J., Jun, J. K., & Kim, H. -S. (2021). Topic modeling and sentiment analysis of online review for airlines. *Information*, *12*(2), 78.

Pillai, S. G., Kim, W. G., Haldorai, K., & Kim, H. -S. (2022). Online food delivery services and consumers' purchase intention: Integration of theory of planned behavior, theory of perceived risk, and the elaboration likelihood model. *International Journal of Hospitality Management*, *105*, 103275.

Riswanto, A. L., Kim, S., & Kim, H. -S. (2023). Analyzing online reviews to uncover customer satisfaction factors in Indian cultural tourism destinations. *Behavioral Sciences*, *13*(11), 923.

Shrivastav, S., Riswanto, A. L., Kim, H. -S., Choi, J. M., & Wang, J. (2025). Understanding the influence of online reviews on customer purchase intention: A cross-country comparison of food delivery apps in Korea (Baemin) and Nepal (Foodmandu). *Culinary Science & Hospitality Research*, *31*(9), 153-166.

Tao, S., & Kim, H. -S. (2022). Online customer reviews: Insights from the coffee shops industry and the moderating effect of business types. *Tourism Review*, *77*(5), 1349-1364.

Williady, A., Kim, W. G., Haldorai, K., & Kim, H. -S. (2025). Towards a greener bite: Unraveling consumer intentions to embrace sustainable online food delivery services. *Journal of Foodservice Business Research*, 1-28.

Received 2026.02.28 - Revised 2026.04.24 - Accepted 2026.05.07