IRJEMS International Research Journal of Economics and Management Studies
ISSN: 2583 – 5238 / Volume 5 Issue 6 June 2026 / Pg. No: 120-131
Paper Id: IRJEMS-V5I6P112, Doi: 10.56472/25835238/IRJEMS-V5I6P112

Original Article

Crashes, Fees, and Customer Service: Theme-Level Correlates of E-Wallet App Ratings in Indonesia

[1]Nana Varian Januardi, [2]Afina Hasya
[1, 2]Management Department, Faculty of Economics and Business, Universitas Diponegoro, Indonesia.

Received Date: 16 May 2026 Revised Date: 30 May 2026 Accepted Date: 04 June 2026 Published Date: 07 June 2026

Abstract: *A star rating signals that users are unhappy without saying why. We open that black box for Indonesian e-wallets by mining 69,006 Indonesian-language Google Play reviews of six leading apps and asking which review themes move with the star a user leaves. We recover 10 latent themes using topic modeling, map them to e-service-quality (e-S-QUAL) dimensions, and regress the rating on theme prevalence, controlling for app and month fixed effects, within a balanced seven-month window (November 2025 to May 2026; standalone wallets as the primary, the Gojek super-app as a comparison). Relative to a general satisfaction theme, reviews dominated by e-S-QUAL failure dimensions fulfillment (failed top-ups, transfers, fees), system availability (crashes, lost access), and responssiveness (customer service) carry lower ratings; a realistic interquartile shift maps to roughly a tenth of a star. The association is not a within-review artifact: theme prevalence computed from other reviews in the same app-month still predicts a review's rating for 8 of 9 themes, and the pattern survives after removing all sentiment words from the topic model. Eight of nine associations remain significant under a Webb six-point wild-cluster bootstrap across the six apps, and the result is stable to dropping any app, equal-weighting apps, dropping the crash-spike months, and including single-token reviews. Two qualifications are explicit: the broad ordering (failure dimensions below satisfaction) is robust, but the fine ranking among failure dimensions is sensitive to the topic-model seed; and promotional content, often assumed central, appears in only ~4% of reviews and skews toward complaints. We read the results as a reproducible, associative map of where rating risk concentrates across e-S-QUAL dimensions, not as causal effects.*

Keywords: *E-Wallet, Digital Payment, Online Reviews, Topic Modeling, E-Service Quality, Indonesia JEL, G23, G41, M31, O33, L86.*

I. INTRODUCTION

Indonesia is one of the most active mobile payment markets in the world. The national QR standard (QRIS) and a cohort of well-funded e-wallets moved digital payments from a niche convenience to an everyday default in under a decade, and adoption studies have tracked that shift closely (Rachman et al., 2024; Sihaloho et al., 2020; Widodo et al., 2019). For the firms behind these apps, the app store star rating is both a public scoreboard and a discovery lever: it shapes store rankings, install rates, and the first impression a prospective user forms. A single 1-to-5 number, though, compresses a great deal of experience into one digit. A 2-star rating could reflect a failed top-up, a surprise fee, a login lockout, or a crash, and the product's response to each case is different.

The stakes are not small. Indonesia hosts dozens of licensed e-money issuers, QRIS is accepted at tens of millions of merchants, and a handful of wallets account for most retail digital-payment volume. For years, these wallets competed on subsidy, pouring cashback and discounts into acquiring users. As that competition has matured, attention has shifted from acquisition to retention, and retention turns on whether the everyday experience holds up. The star rating is the most visible public proxy for that experience.

What the rating hides, the written review often shows. App stores have been recognized as research repositories in their own right (Harman et al., 2012), and within that data, reviews carry information about features, defects, and service quality (Genc-Nayebi & Abran, 2016; Jha & Mahmoud, 2019; Leem & Eum, 2021). Two limitations recur as this work touches payments. Most studies examine a single app, which makes it hard to separate what is specific to one product from what is general to the category. And many stop at description: they surface topics or Sentiment but do not connect them, at scale and with controls, to the rating users actually leave. For Indonesian e-wallets, the question of which service-quality dimensions are most strongly associated with the star net of which app and which month a review belongs to is largely open.

This paper takes up that question. We assemble 69,006 Indonesian-language Google Play reviews for six widely used apps, recover ten latent themes with topic modeling, map them to the e-S-QUAL service-quality dimensions of Parasuraman et al. (2005), and estimate how theme prevalence relates to the rating while holding app and calendar month fixed. The framing is associative and predictive, not causal. The central hazard is that theme content and the rating come from the same review written

by the same user, so a naive regression risks rediscovering that complaint words are associated with low stars. We confront that hazard with two tests that most review-mining work omits: refitting the topic model on text with every sentiment word removed, and more decisively a cross-text design in which a review's rating is predicted from the theme content of *other* reviews in the same app-month.

The contribution is a knowledge claim, not a list of techniques. We show that, among users who write substantive Indonesian e-wallet reviews, low ratings attach to e-S-QUAL *failure* dimensions fulfillment, system availability, and responsiveness rather than to promotions, that this attachment is not a within-text artifact, and that it is a property of the category that holds across six structurally different apps. We are deliberately careful about what is *not* claimed: the broad ordering (failure dimensions sit below satisfaction) is robust, but the precise ranking among failure dimensions is sensitive to the topic-model solution, and the design is associative. Methodologically, cross-text validation, seed-stability accounting for a topic-model-generated regressor, and few-cluster inference appropriate to six apps are transferable components.

Section 2 situates the study and states the research questions. Section 3 describes the data and estimation, including the same-text hazard. Section 4 reports the themes, the main associations, and the robustness battery. Section 5 discusses what the pattern means and what it does not. Section 6 concludes.

II. BACKGROUND AND RESEARCH QUESTIONS

A) Satisfaction, service quality, and ratings

Research on mobile payments has concentrated on adoption and continuance. Continuance models grounded in expectation-confirmation and trust find that confirmation, perceived usefulness, satisfaction, and security predict whether users continue to pay with an app (Cao et al., 2018; Chen & Li, 2016; Humbani & Wiese, 2019; Kumar et al., 2018; Zhou, 2012). In Indonesia, extensions of UTAUT and the technology acceptance tradition reach similar conclusions for digital wallets and QRIS (Furinto et al., 2023; Paramita & Cahyadi, 2024; Soelasih & Sumani, 2022; Widodo et al., 2019; Widyanto et al., 2021).

We organize the empirical themes with the e-service-quality framework. Parasuraman et al. (2005) decompose electronic service quality into efficiency, fulfillment, system availability, and privacy, with responsiveness, compensation, and contact governing service recovery. This framework is the natural lens for app reviews: a failed top-up is a *fulfillment* failure, a crash is a *system availability* failure, a slow help channel is a *responsiveness* failure, and a clean, fast flow is *efficiency*. Conceptual work on digital-wallet satisfaction makes the same service-quality point (Fainusa et al., 2019). Two implications follow. If satisfaction is built from discrete service-quality experiences, the rating should be decomposable into those dimensions, and the useful question is which carries the most weight. And because reliability, access, and responsiveness are repeatedly identified as satisfaction's load-bearing components (Chen & Li, 2016; Kumar et al., 2018), an empirical ranking that places service-quality failures ahead of promotions would corroborate the survey literature with behavioral data.

The e-S-QUAL lens also fixes what a "theme" is in this study. Rather than leaving the ten data-driven topics as ad hoc labels, we read each as an indicator of a service-quality dimension: efficiency (the app is fast and easy), fulfillment (transactions complete and charges are correct), system availability (the app works and access is reliable), and the recovery dimensions of responsiveness and contact (help arrives when something fails). That move turns a descriptive topic list into a test of which service-quality dimensions written dissatisfaction loads onto, and it makes the result comparable to the broader service-quality literature rather than specific to one coding of one corpus.

B) Mining meaning from reviews

Methodologically, we build on the practice of extracting structure from review text. Opinion mining and aspect-level analysis turn free text into ratings of features and sentiments (Guzmán & Maalej, 2014; Hu & Liu, 2004; Luiz et al., 2018; Pang & Lee, 2008). Latent Dirichlet allocation is the workhorse for discovering latent themes in large review corpora (Blei et al., 2003; Blei, 2012; Jelodar et al., 2018), with coherence measures standard for selecting topic solutions (Mimno et al., 2011; Röder et al., 2015). Review mining has been applied across domains, including tourism, mobile banking, health apps, and government services (Guo et al., 2016; Leem & Eum, 2021; Prasetyo & Irawan, 2025; Uncovska et al., 2023), and a stream of recent Indonesian work mines app reviews for Sentiment and topics, often pairing a lexicon with a transformer model (Aryanti et al., 2025; Asri et al., 2025; Pranatawijaya et al., 2024; Sulistiyani et al., 2024). For Indonesian text, the InSet lexicon and the IndoBERT family are the standard building blocks (Koto et al., 2020; Koto & Rahmaningtyas, 2017). Within payments and banking specifically, review mining has extracted service-quality dimensions and ranked feature requests, usually for one app and usually without tying the structure back to the rating in a controlled way (Çallı, 2022; Leem & Eum, 2021; Oh & Kim, 2022; Sulistiyani et al., 2024); we connect themes to the rating in a regression with time controls across several apps, rather than reporting frequencies or Sentiment alone (Jha & Mahmoud, 2019; Kumar et al., 2023; Nayebi et al., 2018).

This connection inherits a caution from the broader text-as-data literature, which warns that quantities read off an unsupervised model are measurements with error rather than ground truth, and that the analyst's choices preprocessing, the number of topics, the random seed propagate into whatever regression follows (Gentzkow et al., 2019; Grimmer & Stewart, 2013). The same literature specifically warns against using a text-derived variable to explain an outcome in the same text without a validation step (Egami et al., 2022). We take both warnings seriously: we report how the estimates change when the topic model is refit with different seeds rather than trusting a single solution, and we validate the theme-rating link using text disjoint from the text used to score the rating. The point is not that LDA is unreliable, but that its outputs deserve the same scrutiny as those of any other constructed regressor.

C) The same-text hazard

One feature of this design must be named at the outset. The regressor (theme prevalence) and the outcome (the star) are both derived from the same review, written by the same person at the same moment. A user who types "saldo kepotong, gagal terus" ("balance deducted, keeps failing") both loads the fee and fulfillment themes and selects a low star. The text-as-data literature is explicit that using the same documents to construct a measure and an outcome invites circularity and recommends out-of-sample or split designs as remedies (Egami et al., 2022; Gentzkow et al., 2019). We adopt both a within-text check (a topic model refit with all sentiment words removed) and, decisively, a cross-text design that builds theme prevalence from reviews *disjoint* from the one whose rating is modeled (Section 3.5, Section 4.4).

D) Research questions

Because the themes are discovered, we frame the work around research questions rather than directional hypotheses. (RQ1) What latent themes, and which e-S-QUAL dimensions, dominate Indonesian e-wallet reviews? (RQ2) Which dimensions are most strongly associated with the star once app and month are held fixed, and is that association more than a within-text artifact? (RQ3) Does it differ across apps and periods? Our prior is that failure dimensions travel with lower ratings and a satisfaction theme anchors the top; recovering which dimensions, and how robustly, is the contribution.

III. DATA AND METHODOLOGY

A) Sample and collection

We collected public Google Play reviews using the open-source Google Play Scraper (no API key) in the Indonesian locale. The roster spans the market a user actually chooses from: bank-neutral wallets (DANA, OVO), a telco wallet (LinkAja), the standalone wallet of the largest ride-hailing platform (GoPay), a bank-issued wallet (Sakuku), and the Gojek super-app, which embeds GoPay. Two candidates were excluded after collection: a mobile-banking app whose package no longer resolved, and an e-commerce app whose reviews overwhelmingly concern shopping. Because a Gojek star rates a whole-app experience well beyond payment, the five standalone wallets are our primary Sample, and Gojek enters as an explicit comparison (Section 4.5).

The six apps differ in backing and business model bank-neutral, telco, ride-hailing- affiliated, bank-issued, and a super-app which is useful: a pattern that holds across this spread is unlikely to be an artifact of one product's defect profile. The analysis dataset records, for each review, the app, the star rating, the date, the anonymized text, and a token count; no demographic or device fields are used, and nothing that could identify a reviewer is retained. Because the densest apps could be scraped only a limited distance back under platform rate limiting, the balanced window is the price of comparability: a shallower span that every app covers in full is preferable to a deeper but ragged panel in which app and time coverage would be confounded.

We drew reviews newest-first and capped storage at 1,500 per app-month with seeded reservoir sampling (uniform within month, bounded memory), and froze the snapshot. We verified authenticity by re-fetching fresh reviews and matching opaque review identifiers against the stored Sample. The match rate varies with app volume, and we report the full range, not the best cases: 150 of 150 (100%) for Sakuku, 98% for OVO, LinkAja, and Gojek, but 66% for GoPay and 48% for DANA. The lower rates for the two highest-volume apps are expected the reservoir keeps a 1,500/month sample, and new reviews arrive between scrapes and re-fetches and any exact identifier match (which cannot be fabricated) confirms genuine platform provenance; the authenticity check passes, but the framing is the full range.

B) Cleaning, window, and the articulability filter

Cleaning was deterministic (seed 20260605). We removed duplicate identifiers, parsed timestamps, dropped empty text, and discarded reviewer usernames, so that only an anonymized, truncated SHA-1 key, the app, the star, the date, and the text survive. Indonesian-language filtering used a marker-word rule with statistical detection on the ambiguous minority. Text was normalized into a lightly cleaned form (negators kept) for lexicon sentiment and a stopword-removed, Sastrawi-stemmed form for topic modeling. The funnel runs from 106,593 raw records to 69,006 cleaned Indonesian reviews. The main window is the balanced span every app covers November 2025 to May 2026, seven complete months excluding the partial scrape month. We restrict the theme regression to reviews with at least 2 content tokens (32,140 reviews); single-word reviews, such as "mantap,"

carry a rating but no thematic content. This filter removes 38% of five-star but only 5% of one-star reviews, so we treat it as a selection on articulability and report the unfiltered Sample (42,829 reviews) as a check (Section 4.4).

C) Sentiment as a tuned sanity check

We scored Sentiment with the InSet lexicon (Koto & Rahmaningtyas, 2017), keeping only unambiguously signed words and dropping intensifiers and app-name tokens. We explicitly state that this construction was tuned to the observed rating we chose the scheme that best agreed with the star rating so the resulting Spearman of 0.45 is an in-sample upper bound, not an out-of-sample validation. Sentiment is therefore used only as a sanity check that the text carries valence; it is not the outcome and not a control in the primary model.

D) Themes

We recovered themes with LDA on stemmed text (count vectorizer, minimum document frequency 30), comparing LDA and non-negative matrix factorization (Lee & Seung, 1999) across topic counts from six to fourteen and selecting by C_v coherence (Röder et al., 2015). LDA dominated NMF at every count, and coherence peaked at 10 topics (C_v = 0.61) under the deployment iteration setting, so we fit a 10-topic LDA model and labeled themes from the top terms, mapping each to an e-S-QUAL dimension (Table 2). The topic solution is only moderately stable across random seeds (mean best-match top-term Jaccard of 0.31), which we treat seriously: Section 4.4 reports a seed-bootstrap of the entire pipeline, and we frame the *fine* ranking among themes as indicative rather than exact.

E) Estimation

At the review level, we estimate

$$R_{iat} = \alpha_a + \tau_t + \sum_{k \neq r} \beta_k \, \theta_{ik} + \varepsilon_{iat},$$

With R_{iat} the star, α_a app fixed effects, τ_t month fixed effects, and θ_{ik} theme prevalence. Theme proportions sum to one, so we omit the most prevalent (satisfaction) theme as reference; we confirm this drives only the sign convention, not the substance (re-estimating with an operational theme as reference leaves satisfaction strongly positive and preserves the ordering). Four estimation choices follow.

First, the specification is associative; θ is a topic-model posterior used as a regressor, and we read. β_k as a predicted rating difference, not a causal effect. Second, because a 0-to-1 proportion and a within-simplex scale raw coefficients "hold others fixed" shift is not literal, we report effects as the predicted rating change for an interquartile shift in a theme's prevalence, and we report the within-R^2 (the variance θ explains after app and month are partialled out). Third, with only six apps, app-level clustering has too few clusters; our primary inference is a wild cluster bootstrap over the six apps using Webb (2023) six-point weights (which give finer resolution than the 2^6 Rademacher sign patterns), and we also enumerate all 64 Rademacher patterns for an exact discrete p-value (Cameron et al., 2008; MacKinnon & Webb, 2016). We treat the conventional cluster-robust and app-month-clustered t-statistics as anti-conservative bounds, not as the primary inference, and we do not report bootstrap p-values below the $1/64$ discrete floor as if they were finer. Fourth, to address the same-text hazard directly, the cross-text test splits each app-month's reviews at random into halves A and B, builds theme prevalence from A, and predicts the ratings of the disjoint reviews in B; if A's theme content predicts B's ratings, the association is not a within-review tautologytwo procedural notes. The Webb wild bootstrap imposes the null that a theme's coefficient is zero, redraws the six cluster-level residuals from the six-point distribution, and re-estimates the coefficient and its cluster-robust t-statistic over 999 replications; the exact version instead enumerates all 64 sign assignments, so the reported p-value cannot fall below the genuine discrete floor. The seed-bootstrap refits the whole LDA-then-regression pipeline under eight seeds, aligns each refit's topics to the base solution by maximum posterior correlation, and reports the standard deviation of each theme's interquartile effect across seeds; it asks not whether the topics are identical run to run (they are not) but whether the rating associations they imply are stable.

IV. RESULTS

A) The Sample and its themes

Table 1 shows a sample weighted toward the four high-volume apps, with LinkAja thinner and Sakuku very thin, and a mean rating (3.5) below lifetime averagesa consequence of sampling a recent window. The rating distribution is sharply bimodal, as app-store ratings usually are. Ratings differ across apps, with the lowest for OVO (which is ~27% of the analysis sample) and the highest for the standalone GoPay app and DANA; this cross-app variation is what app fixed effects absorb.

Table 1: Sample Description by App (Main Analysis Window)

App	Type	N	Mean rating	SD	%1★	%5★	Months
DANA	standalone wallet	10,000	3.95	1.61	19.4	65.5	2025-11–2026-05
GoPay	standalone wallet	9,915	4.20	1.47	15.0	72.9	2025-11–2026-05

Gojek	embedded superapp	9,864	3.57	1.80	29.0	57.6	2025-11–2026-05
LinkAja	standalone wallet	2,829	3.57	1.78	28.6	55.5	2025-11–2026-05
OVO	standalone wallet	10,111	2.29	1.73	59.4	25.6	2025-11–2026-05
Sakuku	standalone wallet	110	2.75	1.81	47.3	32.7	2025-11–2026-05
All		42,829	3.50	1.81	30.7	55.3	2025-11–2026-05

Notes: *Main window = balanced months in which all apps are observed. Star rating is the Google Play 1–5 score.*

The ten themes (Table 2) are split into one positive theme and nine failure-or-function themes, each mapped to an e-S-QUAL dimension. The most prevalent (21%) is a general satisfaction theme (efficiency; "good," "easy," "fast," "safe," "satisfied"), our reference. The rest cover fulfillment (top-up, transfer, balance; fees and deductions), system availability (crashes and errors; account access and login; updates and upgrades), responsiveness (customer service), a reliability-decline theme, and two themes specific to the super-app (ride-hailing-and-lending and food-delivery), which lie outside payment service quality. No standalone cashback or promotion theme emerged at any topic count; Section 4.6 measures promotion language directly rather than inferring from absence.

Table 2: Latent Themes in E-Wallet Reviews, Mapped to E-S-QUAL Service-Quality Dimensions

Theme	Top terms (stemmed)	e-S-QUAL dimension	Prevalence	N dom.
Top-up, transfer & balance	saldo; masuk; uang; transaksi; top; bayar; transfer; potong; kirim; beli; hasil; bank	Fulfillment / Efficiency	0.111	3,920
Customer service & responsiveness	baik; masalah; buruk; kendala; layan; respon; sekali; chat; ganggu; sama; mohon; selalu	Responsiveness / Contact (recovery)	0.074	1,585
Fees & admin charges	aja; potong; admin; apa; pake; banyak; biaya; mau; banget; kena; lah; buat	Fulfillment (billing transparency)	0.104	3,238
Ease of use & satisfaction	sangat; bantu; bagus; mudah; cepat; transaksi; sekali; aman; puas; banget; lebih; mantap	Efficiency (reference)	0.207	8,571
Ride-hailing & lending (super-app)	makin; driver; selalu; jauh; mantap; moga; susah; banyak; jalan; lebih; jemput; tolak	Non-payment (super-app, n/a)	0.077	2,211
Account access & login	akun; tiba; hilang; nomor; saldo; login; ganti; email; mau; masuk; kode; cara	System availability / Security	0.078	2,195
App crashes & errors	buka; sering; eror; error; mulu; bagus; lot; bgt; susah; oke; padahal; update	System availability	0.085	2,379
Food delivery & waiting (super-app)	driver; lama; jam; pesan; nunggu; makan; banget; order; mau; gofood; dapet; cancel	Non-payment (super-app, n/a)	0.087	2,824
Updates & account upgrades	mau; baru; terus; udah; bintang; upgrade; cicil; fitur; gagal; coba; aja; malah	System availability / Fulfillment	0.102	3,271
Declining reliability / disappointment	hari; pakai; lama; kecewa; udah; pinjam; bayar; mau; transaksi; sekarang; padahal; sering	Fulfillment / Reliability	0.075	1,946

Notes: *LDA themes from stemmed Indonesian review text. e-S-QUAL dimensions per Parasuraman, Zeithaml & Malhotra (2005). Prevalence = mean topic proportion; N dom. = reviews where the theme is modal.*

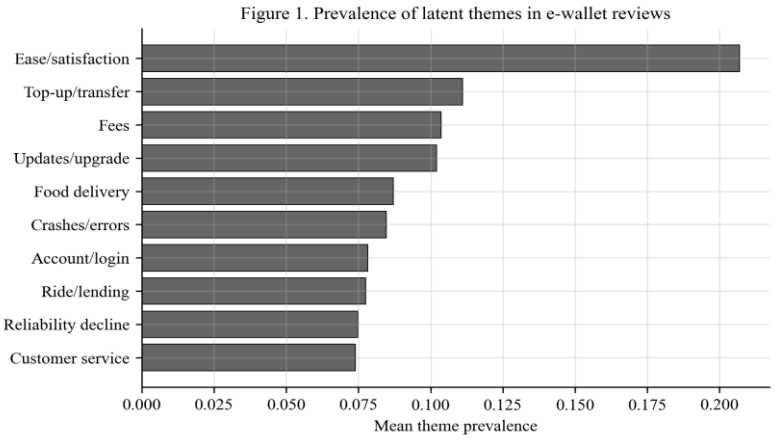

Figure 1. Latent Theme Prevalence in E-Wallet Reviews

The themes repay a closer look, since they answer RQ1. Five sit in the fulfillment and system-availability families that the e-S-QUAL framework treats as core. The top-up, transfer, and balance theme ("saldo," "transfer," "kirim," "potong") is the

operational heart of a wallet; the fees-and-deductions theme ("admin," "biaya," "kena") gathers money leaving the balance without a clear reason; the account-and-login theme ("akun," "login," "kode," "hilang") is about getting in at all; the crashes-and-errors theme ("eror," "lemot," "gabisa") is technical failure; and the updates-and-upgrades theme ("update," "upgrade," "cicil," "gagal") mixes version churn with the verification and pay-later flows users are pushed through. A responsiveness theme ("respon," "layan," "chat," "kendala") captures whether anyone answers when something breaks, and a reliability-decline theme ("kecewa," "padahal," "sering") is the longitudinal voice of a user whose trust has eroded. Anchoring the positive end is the satisfaction theme ("bagus," "mudah," "cepat," "aman," "puas," "mantap").

B) Sentiment check

The tuned InSet score correlates with the star at a Spearman's rho of 0.45 (Table 3), rising across the five rating levels. Sentiment and star agree most at the extremes and blur in the middle, and many positive-rated reviews register as neutral once ambiguous words are removed. This is enough to confirm the text carries valence; the heavier inferential work is carried by the rating itself, modeled directly.

Table 3: Validation of Inset Lexicon Sentiment Against Star Ratings

Metric	Value
N reviews	69,006
Lexicon hit rate	71.0%
Spearman ρ (Sentiment, rating)	0.452***
Pearson r (norm. sentiment, rating)	0.380***
3-class agreement	69.8%
Polarity agreement (pos vs neg)	75.5%

Notes: *InSet terms with unambiguous polarity; *** p<0.001. Agreement on lexicon-bearing reviews.*

C) Which dimensions track the rating

Table 4 and Figure 2 report the estimates. In terms of satisfaction, every failure dimension is associated with a lower rating. The interpretable magnitudes are modest: a 25th-to-75th percentile shift in a theme's prevalence is associated with about −0.14 stars for top-up and transfer problems, −0.14 for fees, −0.12 for updates and upgrades, −0.085 for food delivery, −0.08 for account and login, −0.07 for crashes, −0.05 for customer service, and −0.03 for ride/lending. Theme content explains a within-R^2 of 0.33 (after app and month are removed; total R^2 0.44). The rankings by association strength and by raw prevalence are only moderately correlated (rank correlation 0.62): food delivery, for instance, is of middling prevalence but among the strongest correlates, so the rating-weighted picture is not just a restatement of what is talked about most. The standalone-wallet estimates are nearly identical to the pooled ones for the payment themes (e.g., top-up −3.84 vs −3.89 raw), and the unfiltered full-sample estimates track the filtered ones closely (coefficient correlation 0.97), so neither pooling nor the articulability filter drives the result.

Table 4: Theme Prevalence and Star Rating (Review-Level OLS, App & Month FE)

Theme	Coef.	SE (app×month)	SE (two-way)	Std. coef.
Top-up, transfer & balance	-3.887***	(0.064)	(0.090)	-0.411
Customer service & responsiveness	-2.562***	(0.095)	(0.171)	-0.185
Fees & admin charges	-3.856***	(0.181)	(0.452)	-0.384
Ride-hailing & lending (super-app)	-1.378***	(0.259)	(0.658)	-0.114
Account access & login	-3.750***	(0.082)	(0.138)	-0.323
App crashes & errors	-3.174***	(0.191)	(0.416)	-0.272
Food delivery & waiting (super-app)	-4.093***	(0.211)	(0.543)	-0.400
Updates & account upgrades	-3.083***	(0.144)	(0.359)	-0.309
Declining reliability/disappointment	-3.687***	(0.092)	(0.208)	-0.286

Notes: *Associative, not causal. DV = 1–5 star rating. Regressors = theme prevalence (reference theme omitted). SE clustered by app×month (primary) and two-way by app & month. Std. coef. from z-scored variables. * p<0.10, ** p<0.05, *** p<0.01.*

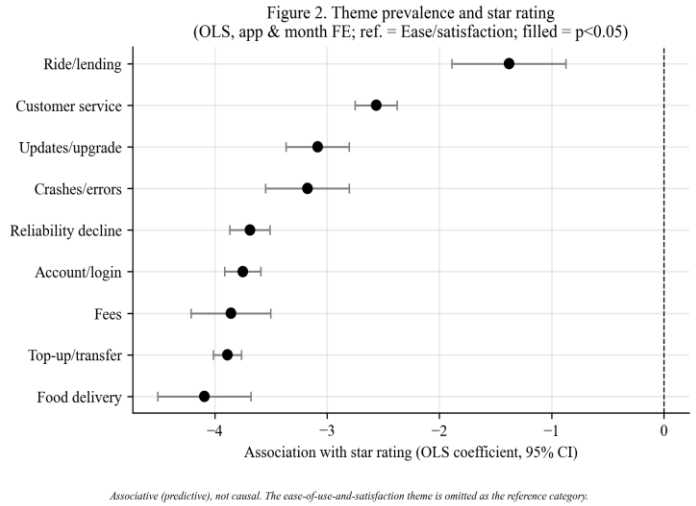

Figure 2. Theme-Rating Associations with App and Month Fixed Effects (Associative, Not Causal; Bars Are 95% Confidence Intervals)

The within-R^2 Of 0.33 is worth dwelling on: after differences between apps and between months are removed, theme content alone accounts for a third of the remaining variance in the star, which is high for review-level data and consistent with the rating being, in large part, a summary of which service-quality experience the review describes. The moderate rank correlation between association strength and prevalence (0.62) is the practical payoff of the regression over a simple frequency count: a team reading only "what is complained about most" would over-weight high-volume but rating-neutral chatter and under-weight dimensions like food delivery that are less frequent yet more tightly tied to the star.

D) Robustness and falsification

Because the specification is associative and shares text between regressor and outcome, the robustness work carries the argument.

The association is not a within-review tautology (cross-text). Building theme prevalence from a random half of an app-month's reviews and predicting the ratings of the disjoint other half, eight of nine themes remain significant predictors (account/login, crashes, food delivery, and customer service at $p < 0.001$; fees, reliability decline, top-up, and updates at $p < 0.05$; only ride/lending is insignificant). Because the regressor here comes from *different* reviews than the outcome, this is the test that most directly rebuts the same-text concern, and it holds.

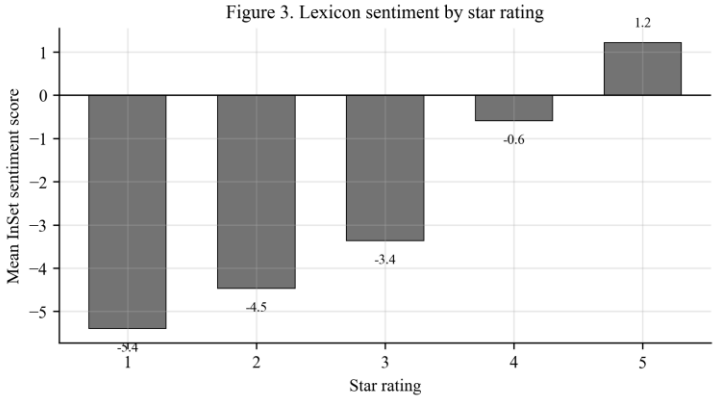

Figure 3. Lexicon Sentiment Against the Star Rating

It is not merely an effect on vocabulary (valence-stripped). Refitting the topic model on text with every InSet sentiment word removed, the explained variance barely moves (within the pooled model, total R^2 0.438 to 0.423) and all nine themes remain significant; the coefficient spread widens rather than collapses. This rules out an affect-word tautology, though, because event-denoting words such as "gagal" or "eror" are not all sentiment words; it does not by itself rule out referential coupling, which is why the cross-text test above is the load-bearing one. The identity of the failing theme is itself informative: ride/lending, the one super-app theme among the nine, is also the only one whose cross-text coefficient is insignificant and wrong-signed.

That the lone non-payment theme is the lone failure is reassuring rather than troubling it is the theme least likely to reflect a genuine wallet service-quality experience, and the cross-text design correctly declines to certify it.

Inference holds under the few-cluster correction. Under the Webb six-point wild cluster bootstrap over six apps, eight of nine themes are significant at 5% (top-up and account/login $p \approx 0.002$; customer service and fees 0.009; reliability 0.011; updates 0.013; food 0.032; crashes 0.037), with only ride/lending marginal (0.105). Exact enumeration of the 64 Rademacher patterns places the strongest themes at the discrete floor ($1/64 \approx 0.016$) rather than at the misleadingly fine values conventional formulas produce. All nine survive Benjamini-Hochberg and Romano-Wolf corrections in the conventional specification (Table 5).

Table 5: Multiple-Testing–Robust Inference For Theme→Rating Associations

Theme	Coef.	t	p (cluster)	p (BH-FDR)	p (perm.)	p (Romano–Wolf)
Top-up, transfer & balance	-3.887	-85.12	0.000	0.000	0.001	0.001
Customer service & responsiveness	-2.562	-41.48	0.000	0.000	0.001	0.001
Fees & admin charges	-3.856	-79.99	0.000	0.000	0.001	0.001
Ride-hailing & lending (super-app)	-1.378	-23.99	0.000	0.000	0.001	0.001
Account access & login	-3.750	-69.31	0.000	0.000	0.001	0.001
App crashes & errors	-3.174	-59.33	0.000	0.000	0.001	0.001
Food delivery & waiting (super-app)	-4.093	-76.27	0.000	0.000	0.001	0.001
Updates & account upgrades	-3.083	-65.78	0.000	0.000	0.001	0.001
Declining reliability/disappointment	-3.687	-64.15	0.000	0.000	0.001	0.001

Notes: *Coef./t from FWL-residualized model. p (cluster) = app×month clustered; BH-FDR = Benjamini–Hochberg; perm. = permutation (B=2000); Romano–Wolf = stepdown max-t controlling FWER.*

The broad pattern is robust, but the fine ranking is seed-sensitive. A label-permutation placebo collapses the largest association to near zero (max $|\text{coef}| \approx 0.12$). Dropping each app in turn, equal-weighting apps so OVO's 27% share cannot dominate, and dropping the crash-spike months (February-March 2026) all leave the coefficients close to baseline (e.g., top-up −3.87 to −4.17 across these). However, a seed-bootstrap that refits the topic model under eight seeds and re-estimates the regression is sobering: the *sign* of most associations is stable, but the interquartile magnitudes carry seed standard deviations of 0.04-0.13 of the same order as the effects themselves and the set of three most-negative themes overlaps the base solution only 1.4 of 3 on average. We therefore claim the *dimensional* result (failure dimensions below the satisfaction threshold, cross-text-validated) and decline to claim a precise rank-ordering among the failure dimensions.

Table 6: Few-Cluster Inference and Cross-Text Validation by Theme

Theme	Coef (pooled)	Coef (standalone)	p (Webb WCB, 6 cl.)	p (exact 2^6)	Cross-text coef	Cross-text p
Top-up, transfer & balance	-3.887	-3.835	0.002***	0.016	-2.71	0.004***
Customer service & responsiveness	-2.562	-2.585	0.009***	0.016	-7.66	0.000***
Fees & admin charges	-3.856	-3.734	0.009***	0.000	-6.08	0.001***
Ride-hailing & lending (super-app)	-1.378	-0.402	0.105	0.094	3.07	0.307
Account access & login	-3.750	-3.641	0.002***	0.031	-8.58	0.000***
App crashes & errors	-3.174	-3.273	0.037**	0.031	-5.44	0.000***
Food delivery & waiting (super-app)	-4.093	-2.222	0.032**	0.016	-10.49	0.000***
Updates & account upgrades	-3.083	-2.965	0.013**	0.031	-3.27	0.022**
Declining reliability/disappointment	-3.687	-3.550	0.011**	0.062	-6.95	0.002***

Notes: *Pooled = six apps; standalone = five wallets. Webb WCB = six-point wild cluster bootstrap over six apps (B=999); exact = enumeration of all 2^6 Rademacher sign patterns (discrete floor 1/64=0.016). Cross-text: theme prevalence from disjoint reviews in the same app-month predicting held-out ratings. * p<0.10, ** p<0.05, *** p<0.01.*

This split between a stable qualitative pattern and an unstable fine ranking is itself worth recording. Applied service-quality work that uses topic models rarely refits the model under several seeds before reading coefficients off a single solution; when we do, the ordinal claim that interests managers most *which* failure hurts most is the fragile one, while the categorical claim that operational failures sit well below the satisfaction baseline is what the data support. Reporting both, rather than the one seed that yields the tidiest ranking, is the honest summary and the one we would defend before a skeptical referee.

E) The super-app comparison

Removing Gojek and running solely on standalone, the food-delivery coefficient falls from about −4.09 to −2.22, and the ride/lending coefficient from −1.38 to −0.40, while the payment themes remain unchanged. That the two super-app themes attenuate while payment themes do not is what should happen if rides and food are experiences that the super-app container imports into the review stream rather than properties of the wallet. We therefore treat standalone wallets as the primary and use the super-app themes for comparison.

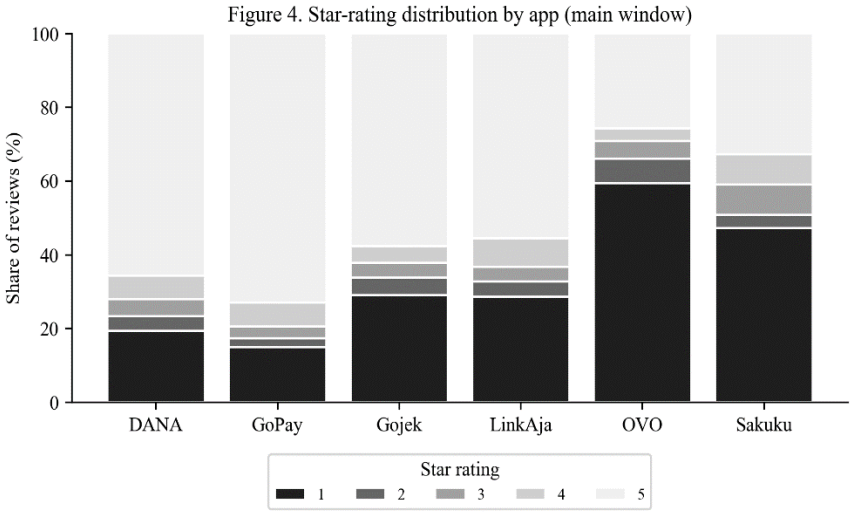

Figure 4. Star-rating distribution by app.

F) How common is promotion talk?

Because no cashback theme emerged and topic models cannot surface rare topics, we measured promotion language directly using a dictionary (cashback, promo, diskon, voucher, kupon, gratis, bonus, and related terms). Promotion terms appear in 4.2% of in-window reviews, and their share declines with the rating, from 6.5% in one-star to 2.6% in five-star reviews. Promotion is thus a minor, complaint-tilted presence in written reviews, not a dominant theme. We are careful not to read "no topic" as "promotions do not matter": satisfied promotion users may not write, and our filter removes many short positive reviews where a happy cashback mention would live.

Table 7: Robustness Diagnostics Summary

Diagnostic	Value
Within-R^2 (theta after app+month FE)	0.3341
Valence-stripped LDA: R^2 orig -> stripped	0.438 -> 0.423 (9/9 sig)
Cross-text: themes predicting held-out ratings	8/9
Unfiltered vs filtered coef correlation	0.97
Seed-bootstrap: top-3 negative overlap (8 seeds)	1.38/3 (fine rank seed-sensitive)
Horse race: rank-corr(	coef
Sensitivity: equal-weight apps & drop Feb-Mar	coefficients within ~5-10% of baseline
Promotion terms: overall / by star 1->5	4.15% / 6.46-6.24-5.98-3.8-2.59%
Topic stability (best-match Jaccard, 3 seeds)	0.311

Notes: *Cross-text and seed-bootstrap from the panel-revision battery; valence-stripped, promotion dictionary, and token survival on the in-window Sample.*

G) Over time

Across the seven months, the associations are stable, and theme prevalence shifts only modestly month to month, with a visible rise in the crashes theme in February 2026 (which the drop-period check above shows does not drive the result). Seven months is too short a time frame for long-run dynamics, so we report within-window stability rather than a trend.

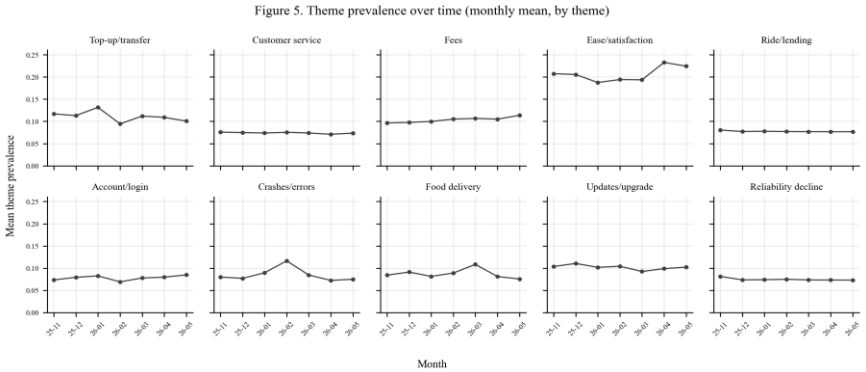

Figure 5. Theme Prevalence over Time (Small Multiples)

V. DISCUSSION

A) What the pattern means

Among Indonesian e-wallet users who write substantive reviews, low ratings are associated with e-service-quality failures rather than promotions. Fulfillment failures (money that does not move, money quietly deducted), system-availability failures (lost access, crashes), and weak responsiveness are the dimensions that travel with low ratings; a satisfaction theme anchors the top. These correlations sit inside the continuance literature even though the data and method differ: confirmation of expectations and perceived usefulness are the engines of satisfaction (Cao et al., 2018; Humbani & Wiese, 2019; Zhou, 2012), and service-quality reliability is exactly where expectations in a payment app are confirmed or violated. The contribution is not to overturn that literature but to show, in behavioral text, cross-text-validated, and across structurally different apps, which e-S-QUAL dimensions written dissatisfaction loads onto and that promotion is not among them.

Two features distinguish this from the typical single-app review study. First, the dimensional pattern holds across six structurally different apps bank-neutral, telco, ride-hailing, bank-issued, and a super-app so it is a property of the category, not of one product's defect profile; a single-app study could not separate the two. Second, the honest limit is as informative as the finding: that the fine ordering among failure dimensions shifts with the topic-model seed tells product teams and researchers to act on the *dimension* fulfillment and system availability dominate written dissatisfaction rather than on a spurious decimal-place ranking of which exact theme is worst. A literature that often reports precise topic rankings from a single LDA solution should treat those rankings with the same caution.

B) Contribution

The knowledge claim is that, for this market and medium, written dissatisfaction is an e-service-quality phenomenon concentrated in fulfillment and system availability, robust to who writes and which app, and not an artifact of measuring theme and rating from the same text. We do not claim methodological novelty in the components lexicon sentiment, LDA, and fixed-effects regression are standard but the cross-text validation, the seed-bootstrap accounting for a topic-model-generated regressor, and the few-cluster Webb bootstrap are the transferable methodological pieces, applied here to a comparative Indonesian corpus that the continuance literature has studied mostly through surveys (Cao et al., 2018; Furinto et al., 2023; Widodo et al., 2019). The comparative design also yields a second result: the container in which a wallet lives changes its review stream. Hence, pooling a super-app with standalone wallets imports rating consequences that a payments team does not control.

C) Implications

For product and operations teams, the e-S-QUAL ordering is usable: at the margins this market occupies, investment in fulfillment reliability (top-ups, transfers, balance accuracy), transparent fees, and account recovery should move the rating more than investment in promotions. For platform strategists weighing a standalone wallet against an embedded one, the super-app result is a caution. For supervision, the salience of fee and fulfillment complaints is suggestive rather than actionable: public review streams are an exploratory signal of where disclosed terms and lived experience may diverge, but they would need validation against formal complaint data before any supervisory use; we do not propose them as a monitoring instrument here.

The e-S-QUAL framing makes the practical message portable. Because the themes are read as service-quality dimensions rather than app-specific topics, a team at any wallet can ask the same question of its own review stream which dimension is bleeding rating and benchmark against the category pattern reported here. Dimension-level guidance is also more durable than a feature backlog: fulfillment and system availability are stable constructs, whereas the specific bug or feature that instantiates them changes from release to release.

VI. LIMITATIONS AND FUTURE WORK

The design is associative; although the cross-text test rebuts a within-review tautology and the valence-stripped test rules out affect-word coupling, we make no causal claim. The fine ranking among failure dimensions is sensitive to the topic-model seed, so we claim the dimensional pattern, not a precise ordering; an ensembled or seeded-topic model is a natural next step. The corpus is doubly selected into writing at all, which favors the very satisfied and the very frustrated, and into the articulability filter, which removes more positive than negative reviews so "dissatisfaction is operational" should be read as "among users who write articulable reviews"; we report the unfiltered Sample as a check, but the qualifier is real. Coverage is recent and uneven (seven balanced months; the densest apps are reachable only a limited distance back), so the evidence is within-window. Sentiment rests on a tuned lexicon; a contextual model such as IndoBERT (Koto et al., 2020) would be the natural alternative measure, and the seed instability of the fine theme ranking which we report rather than hide is itself a caution: studies that lean on a single topic solution to rank themes may report orderings that would not survive a refit, and our seed-bootstrap suggests treating any such ranking as indicative. The few-cluster setting (six apps) bounds the power of any inference regardless of the bootstrap, so a wider roster of apps would tighten the conclusions as much as a longer panel. One app, Sakuku, is thin and contributes little; dropping it leaves the result unchanged. Google Play reviews are a self-selected slice of users, silent on those who never write. An out-of-text criterion beyond the corpus such as next-month app ratings, developer incident logs, or complaint records would strengthen the predictive validity of the cross-text evidence.

VII. CONCLUSION

The star rating signals dissatisfaction but does not disclose its source. By mining tens of thousands of Indonesian e-wallet reviews and connecting discovered themes to the star with fixed effects, e-S-QUAL framing, and a demanding falsification battery, we show that for users who write, dissatisfaction is an e-service-quality phenomenon: failures of fulfillment, system availability, and responsiveness are the dimensions most associated with low ratings, while promotions are a minor, complaint-tilted presence. The association is predictive, not causal, but it is not a relabeling of complaint words theme content from other reviews predicts a review's rating and it is stable across apps, periods, and the writing filter, with the honest caveat that the fine ordering among failure dimensions depends on the topic-model solution. For a market that spent years competing on subsidies, written dissatisfaction now tracks the unglamorous reliability of the payment service itself.

VIII. REFERENCES

[1] Aryanti, F. A. D., Luthfiarta, A., & Soeroso, D. A. I. (2025). Aspect-based sentiment analysis with LDA and the IndoBERT algorithm on the mental health app Riliv. *JOURNAL OF APPLIED INFORMATICS AND COMPUTING*. https://doi.org/10.30871/jaic.v9i2.8958

[2] Asri, Y., Kuswardani, D., Suliyanti, W. N., Manullang, Y. O., & Ansyari, A. R. (2025). Sentiment analysis based on Indonesian language lexicon and IndoBERT on user reviews of the PLN mobile application. *Indonesian Journal of Electrical Engineering and Computer Science*. https://doi.org/10.11591/ijeecs.v38.i1.pp677-688

[3] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*. https://doi.org/10.1145/2133806.2133826

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*. https://doi.org/10.5555/944919.944937

[5] Çallı, L. (2022). Exploring mobile banking adoption and service quality features through user-generated content: The application of a topic modeling approach to Google Play Store reviews. *International Journal of Bank Marketing*. https://doi.org/10.1108/ijbm-08-2022-0351

[6] Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*. https://doi.org/10.1162/rest.90.3.414

[7] Cao, X., Yu, L., Liu, Z., Gong, M., & Luqman, A. (2018). Understanding mobile payment users' continuance intention: A trust transfer perspective. *Internet Research*. https://doi.org/10.1108/intr-11-2016-0359

[8] Chen, X., & Li, S. (2016). Understanding continuance intention of mobile payment services: An empirical study. *Journal of Computer Information Systems*. https://doi.org/10.1080/08874417.2016.1180649

[9] Egami, N., Fong, C., Grimmer, J., Roberts, M. E., & Stewart, B. (2022). How to make causal inferences using texts. *Science Advances*. https://doi.org/10.1126/sciadv.abg2652

[10] Fainusa, A. F., Nurcahyo, R., & Dachyar, M. (2019). Conceptual framework for digital wallet user satisfaction. *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*. https://doi.org/10.1109/icetas48360.2019.9117285

[11] Furinto, A., Tamara, D., Hartono, D., & Simamora, E. (2023). *Continuation use of the digital wallet using the extended ECM model in Indonesia*. https://doi.org/10.46254/au01.20220079

[12] Genc-Nayebi, N., & Abran, A. (2016). A systematic literature review: Opinion mining studies from mobile app store user reviews. *Journal of Systems and Software*. https://doi.org/10.1016/j.jss.2016.11.027

[13] Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*. https://doi.org/10.1257/jel.20181020

[14] Grimmer, J., & Stewart, B. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*. https://doi.org/10.1093/pan/mps028

[15] Guo, Y., Barnes, S. J., & Jia, Q. (2016). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation. *Tourism Management*. https://doi.org/10.1016/j.tourman.2016.09.009

[16] Guzmán, E., & Maalej, W. (2014). *How do users like this feature? A fine-grained sentiment analysis of app reviews*. https://doi.org/10.1109/re.2014.6912257

[17] Harman, M., Jia, Y., & Zhang, Y. (2012). *App store mining and analysis: MSR for app stores*. https://doi.org/10.1109/msr.2012.6224306

[18] Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. https://doi.org/10.1145/1014052.1014073

[19] Humbani, M., & Wiese, M. (2019). An integrated framework for the adoption and continuance intention to use mobile payment apps. *International Journal of Bank Marketing*. https://doi.org/10.1108/ijbm-03-2018-0072

[20] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2018). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, and a survey. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-018-6894-4

[21] Jha, N., & Mahmoud, A. (2019). Mining non-functional requirements from app store reviews. *Empirical Software Engineering*. https://doi.org/10.1007/s10664-019-09716-7

[22] Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP*. https://doi.org/10.18653/v1/2020.coling-main.66

[23] Koto, F., & Rahmaningtyas, G. Y. (2017). *Inset lexicon: Evaluation of a word list for indonesian sentiment analysis in microblogs*. https://doi.org/10.1109/ialp.2017.8300625

[24] Kumar, A., Adlakaha, A., & Mukherjee, K. (2018). The effect of perceived security and grievance redressal on continuance intention to use m-wallets in a developing country. *International Journal of Bank Marketing*. https://doi.org/10.1108/ijbm-04-2017-0077

[25] Kumar, A., Chakraborty, S., & Bala, P. K. (2023). Text mining approach to explore determinants of grocery mobile app satisfaction using online customer reviews. *Journal of Retailing and Consumer Services*. https://doi.org/10.1016/j.jretconser.2023.103363

[26] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*. https://doi.org/10.1038/44565

[27] Leem, B., & Eum, S.-W. (2021). Using text mining to measure mobile banking service quality. *Industrial Management & Data Systems*. https://doi.org/10.1108/imds-09-2020-0545

[28] Luiz, W., Viegas, F., Alencar, R. O. de, Mourão, F., Salles, T., Carvalho, D. B. F., Gonçalves, M. A., & Rocha, L. (2018). *A feature-oriented sentiment rating for mobile app reviews*. https://doi.org/10.1145/3178876.3186168

[29] MacKinnon, J. G., & Webb, M. D. (2016). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*. https://doi.org/10.1002/jae.2508

[30] Mimno, D., Wallach, H., Talley, E. M., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models*ScholarWorks@UMassAmherst (University of Massachusetts Amherst)*.

[31] [31] Nayebi, M., Cho, H., & Ruhe, G. (2018). App store mining is not enough for app improvement. *Empirical Software Engineering*. https://doi.org/10.1007/s10664-018-9601-1

[32] Oh, Y. K., & Kim, J.-M. (2022). What improves customer satisfaction in mobile banking apps? An application of text mining analysis. *ASIA MARKETING JOURNAL*. https://doi.org/10.53728/2765-6500.1581

[33] Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. Now Publishers, Inc. eBooks. https://doi.org/10.1561/9781601981516

[34] Paramita, E. D., & Cahyadi, E. R. (2024). The determinants of behavioral intention and QRIS use behavior as a digital payment method using the extended UTAUT model. *Indonesian Journal of Business and Entrepreneurship*. https://doi.org/10.17358/ijbe.10.1.132

[35] Parasuraman, A., Zeithaml, V. A., & Malhotra, A. (2005). E-s-QUAL. *Journal of Service Research*. https://doi.org/10.1177/1094670504271156

[36] Pranatawijaya, V. H., Sari, N. N. K., Rahman, R. A., Christian, E., & Geges, S. (2024). Unveiling user sentiment: Aspect-based analysis and topic modeling of ride-hailing and Google Play app reviews. *Journal of Information Systems Engineering and Business Intelligence*. https://doi.org/10.20473/jisebi.10.3.328-339

[37] Prasetyo, I. D., & Irawan, H. (2025). Assessing m-government service quality: A transformer-based sentiment and topic modeling analysis of Indonesia's police super app. *2025 13th International Conference on Cyber and IT Service Management (CITSM)*. https://doi.org/10.1109/citsm67730.2025.11291473

[38] Rachman, A., Julianti, N., & Arkoyah, S. (2024). Challenges and opportunities for QRIS implementation as a digital payment system in Indonesia. *EkBis Jurnal Ekonomi Dan Bisnis*. https://doi.org/10.14421/ekbis.2024.8.1.2134

[39] Röder, M., Both, A., & Hinneburg, A. (2015). *Exploring the space of topic coherence measures*. https://doi.org/10.1145/2684822.2685324

[40] Sihaloho, J. E., Ramadani, A., & Rahmayanti, S. (2020). Implementasi sistem pembayaran Quick Response Indonesia Standard untuk perkembangan UMKM di Medan. *Jurnal Manajemen Bisnis*. https://doi.org/10.38043/jmb.v17i2.2384

[41] Soelasih, Y., & Sumani, S. (2022). Factors influencing millennials' intention to continue using digital wallets in Indonesia. *Binus Business Review*. https://doi.org/10.21512/bbr.v13i3.8561

[42] Sulistiyani, D., Nurchayati, D., Nurchayati, D., & D, N. D. H. (2024). User experience of mobile banking application in Indonesia: New technology of banking. *GLOBAL BUSINESS & FINANCE REVIEW*. https://doi.org/10.17549/gbfr.2024.29.2.127

[43] Uncovska, M., Freitag, B., Meister, S., & Fehring, L. (2023). Rating analysis and BERTopic modeling of consumer versus regulated mHealth app reviews in Germany. *Npj Digital Medicine*. https://doi.org/10.1038/s41746-023-00862-3

[44] Webb, M. D. (2023). Reworking wild bootstrap-based inference for clustered errors. *Canadian Journal of Economics/Revue Canadienne d Économique*. https://doi.org/10.1111/caje.12661

[45] Widodo, M., Irawan, M. I., & Sukmono, R. A. (2019). Extending UTAUT2 to explore digital wallet adoption in Indonesia. *2019 International Conference on Information and Communications Technology (ICOIACT)*. https://doi.org/10.1109/icoiact46704.2019.8938415

[46] Widyanto, H. A., Kusumawardani, K. A., & Yohanes, H. (2021). Safety first: Extending UTAUT to better predict mobile payment adoption by incorporating perceived security, perceived risk, and trust. *Journal of Science and Technology Policy Management*. https://doi.org/10.1108/jstpm-03-2020-0058

[47] Zhou, T. (2012). An empirical examination of continuance intention of mobile payment services. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2012.10.034