Journal of Economics, Entrepreneurship, Management Business and Accounting, 2026, 4(4),1145-1182.
https://doi.org/10.61255/jeemba.v4i4.1356

Evaluating M-Pajak Service through Human-Validated IndoRoBERTa Analysis of Google Play Reviews

Hana Tamara Putri* & Mufidah
Universitas Batanghari, Jambi, 36122, Indonesia

ABSTRACT

Purpose – This study evaluates user-perceived experience in Indonesia's M-Pajak mobile tax application using Google Play reviews and identifies temporal, sentiment, and lexical patterns in users' expressed evaluations.
Design/methodology/approach – The study analyzed a relevance-ranked, retrievable corpus scraped on 24 May 2026. Of 7,009 retrieved reviews, 6,980 valid reviews from 2021–2026 were retained. IndoRoBERTa sentiment labels were evaluated against 1,500 reviews selected through proportionate year-stratified sampling and manually adjudicated by two annotators. The analysis combined temporal ratings, temporal sentiment, rating–sentiment alignment, and transparent bigram-to-dimension coding.
Findings/Results – Human annotators achieved 97.20% agreement (Cohen's kappa = 0.9132), while IndoRoBERTa achieved 94.20% agreement with the human gold standard (kappa = 0.8243). Within the retrieved corpus, 78.81% of reviews were classified as negative, and later-year review shares were increasingly concentrated in low ratings and negative sentiment. These patterns describe expressed review-based evaluation and do not establish that overall application quality or population-level satisfaction necessarily declined. Authentication, OTP, login, NPWP, EFIN, payment, and system-error barriers dominated negative expressions.
Originality/Value – The study offers a human-validated, multi-layer review-analysis framework for evaluating mobile public services, makes the lexical grouping procedure auditable, and translates recurrent complaints into operational service-improvement priorities.

ARTICLE INFO

Keywords:
M-Pajak, Mobile Tax Application, IndoRoBERTa, Sentiment Analysis, Digital Tax Administration

Article Information:
Received: 30/04/2026
Revised: 30/06/2026
Accepted: 10/07/2026

ISSN:
2985-3168 (Online)
2985-3222 (Print)

*Corresponding Author at:
Universitas Batanghari, Jl. Slamet Riyadi, Sungai Putri, Jambi 36122, Indonesia.
E-mail address: hanatamarap@gmail.com (Hana Tamara Putri)

1. Introduction

Digital transformation in public administration has shifted the evaluation of government services from institution-centered digitization toward citizen-centered service delivery. Twizeyimana & Andersson (2019) frame e-government value as a public-value issue, not merely a technical implementation outcome. In a similar vein, Dechamps et al. (2025) argue that citizen-centric digital government requires closer attention to how public services are experienced, interpreted, and evaluated by users. This perspective is particularly relevant for mobile public service delivery, where citizens interact with government through everyday digital interfaces rather than through conventional administrative counters.

Mobile government applications have become important service channels because they extend public services into portable, always-accessible environments. Sharma et al. (2018) show that user perspectives are central to understanding mobile government adoption, while Wang & Teo (2020) demonstrate that online service quality and perceived value are closely linked to mobile government success. These studies suggest that mobile public service applications should be evaluated not only by their availability, but also by whether users perceive them as reliable, accessible, and useful in practice.

Digital tax administration represents a particularly important domain of mobile public service delivery because taxpayers use these systems to fulfil obligatory administrative tasks. Saptono et al. (2023) demonstrate that e-tax system quality is associated with user satisfaction and tax compliance intention, indicating that perceived service quality matters in digital tax environments. Wicaksono et al. (2021) further show that taxpayers' opinions can reveal practical improvement areas in Indonesia's digital tax filing system. In this sense, the evaluation of mobile tax services needs to account for user-perceived experience, especially when the service involves authentication, identity verification, account access, and transaction-related procedures.

M-Pajak is an official mobile tax application associated with Indonesia's Directorate General of Taxes and is distributed through Google Play under the package identifier id.go.pajak.djp. The Google Play listing describes M-Pajak as an official mobile application developed by Indonesia’s Directorate General of Taxes to support taxpayers in fulfilling tax obligations independently, quickly, and securely, with features including Coretax account activation, digital certificate services, location-based access to tax offices, and regulatory information (Google Play, 2026). These developments position M-Pajak as a visible citizen-facing interface within Indonesia's mobile tax administration ecosystem.

For evaluating such an application, app-store reviews offer a useful form of review-based evidence because they contain both explicit star ratings and free-text user feedback. Nakamura et al. (2022) emphasize that app-store reviews can reveal user experience factors that are difficult to observe from technical metrics alone. Ramzy & Ibrahim (2024) also demonstrate the value of large-scale app-review sentiment analysis for understanding user satisfaction in public-service-related mobile applications. Nevertheless, star ratings and textual reviews should not be treated as fully interchangeable indicators: a rating provides an ordinal summary, whereas review text may describe specific usability problems, access barriers, frustrations, or appreciation in greater detail.

2. Literature Review & Research Development

The existing M-Pajak literature provides several useful but still fragmented insights. Putra & Fathurrahman (2022) examine quality-improvement priorities for Indonesian mobile tax service applications using a Delphi approach. Setiyani et al. (2022) focus on M-Tax/M-Pajak adoption and tax compliance through the lens of Diffusion of Innovation Theory. Sidiq et al. (2024) discuss challenges in mobile application development for Tax Administration 3.0, whereas Budianto et al. (2025) analyze M-Pajak Google Play reviews using XGBoost. These studies establish the relevance of M-Pajak as a research object, but they do not yet provide an integrated framework that combines longitudinal rating dynamics, validated transformer-based sentiment classification, rating-sentiment alignment, and lexical experience dimension grouping.

Transformer-based natural language processing provides a methodological pathway for addressing this gap because it can capture contextual meaning in informal and user-generated text. Liu et al. (2019) introduced RoBERTa as a robustly optimized transformer architecture, and Koto et al. (2020) demonstrate the importance of Indonesian language resources for contextual NLP tasks. The IndoRoBERTa model used in the present study is a RoBERTa-based Indonesian sentiment classifier made available through the Hugging Face model repository and trained for Indonesian sentiment classification (Wongso, 2023). Because Google Play reviews often contain spelling variation, abbreviations, code-mixed expressions, and non-standard grammar, contextual models are appropriate for deriving model-generated sentiment labels from Indonesian user reviews.

However, the use of transformer-based sentiment labels in public-service evaluation should be accompanied by validation and interpretability. Model outputs are analytical classifications rather than direct representations of ground truth; therefore, human validation is necessary to examine whether the classifier aligns with expert judgment. Moreover, sentiment labels alone do not explain what users actually discuss. For this reason, the present study links validated IndoRoBERTa-based sentiment classification with rating analysis and bigram-based lexical pattern extraction, allowing positive and negative user-perceived experience to be interpreted through both aggregate distributions and recurring textual expressions.

Against this background, the present study positions Google Play reviews of M-Pajak as a longitudinal record of user-perceived experience and expressed evaluation. It examines explicit star ratings, IndoRoBERTa-generated sentiment labels, the alignment between ratings and textual sentiment, and the lexical patterns through which positive and negative experiences are expressed. Table 1. positions the study against prior M-Pajak and adjacent digital public service research, highlighting how the present work extends existing knowledge through a validated, multi-layer computational text analysis design.

Table 1. Positioning of the Present Study Against Prior M-Pajak Research

Author(s) and year	Research context	Data source	Analytical method	Main focus or limitation	Positioning of the present study
Putra and Fathurrahman (2022)	Indonesian mobile tax service applications	Expert input and Delphi rounds	Delphi study	Prioritizes mobile tax service quality improvement themes but does not analyze public	Adds Google Play review evidence, temporal rating analysis, validated sentiment

Author(s) and year	Research context	Data source	Analytical method	Main focus or limitation	Positioning of the present study
				app reviews, ratings, or temporal user evaluations.	classification, and lexical interpretation of user-perceived experience.
Setiyani et al. (2022)	M-Tax/M-Pajak adoption in Indonesia	Taxpayer survey data	Diffusion of Innovation Theory	Focuses on adoption and compliance-related perception rather than unsolicited review text or app-store rating dynamics.	Shifts attention from adoption intention to observed review-based evaluation in a real mobile tax service environment.
Sidiq et al. (2024)	M-Pajak development challenges and Tax Administration 3.0	Literature review and interviews	Descriptive qualitative analysis	Provides a development-side perspective on mobile application challenges rather than citizen-side evaluation from platform reviews.	Complements development-oriented analysis with citizen-facing evidence derived from public reviews and validated sentiment labels.
Budianto et al. (2025)	M-Pajak Google Play reviews	App-store review text	KDD pipeline and XGBoost	Classifies criticism in M-Pajak reviews but does not integrate human-validated transformer classification, rating-sentiment alignment, or longitudinal interpretation.	Extends direct M-Pajak review mining through validated IndoRoBERTa classification and multi-layer analysis linking ratings, sentiment, and lexical dimensions.
Wicaksono et al. (2021)	Digital tax filing in Indonesia	Individual taxpayers' opinions	Exploratory qualitative study	Examines browser-based e-filing improvement rather than mobile app reviews and app-store analytics.	Moves the inquiry into the mobile tax application context using naturally occurring Google Play review data.

Author(s) and year	Research context	Data source	Analytical method	Main focus or limitation	Positioning of the present study
Saptono et al. (2023)	E-tax system quality and tax compliance intention in Indonesia	Survey of tax-system users	PLS-SEM and mediation analysis	Explains how perceived e-tax system quality relates to satisfaction and compliance intention, but not how a specific mobile app is discussed in public review ecosystems.	Provides review-based evidence from M-Pajak as a concrete mobile tax-service interface.
Sharma et al. (2018)	Mobile government service adoption	Survey of mobile governmen t users	SEM and neural-network modelling	Develops an adoption-oriented explanation but does not use app-store reviews or textual complaint evidence.	Examines post-use expressed evaluation from public reviews rather than only adoption intention.
Wang & Teo (2020)	Mobile government success	Survey of mobile police users	Informatio n systems success model	Highlights service quality and satisfaction but does not mine user review texts or compare ratings with sentiment labels.	Operationalizes service-quality concerns through user-generated review text and rating-sentiment alignment.

Source: Authors' computation and analysis (2026).

Taken together, the literature indicates several unresolved gaps. First, M-Pajak research remains limited and is dominated by adoption, development, or improvement-oriented studies rather than integrated user-evaluation research. Second, Indonesian digital tax studies have generally relied on surveys, interviews, or qualitative opinions, whereas Google Play reviews provide naturally occurring user expressions at scale. Third, app-review research shows that ratings and textual comments can contain different evaluative signals, but rating-sentiment alignment remains underexamined in the M-Pajak context. Fourth, although transformer-based Indonesian sentiment classification provides contextual sensitivity, its use in mobile public service evaluation is stronger when the model is validated against human judgment and then connected to interpretable lexical dimensions.

The novelty of the present study lies in its validated and multi-layered review-based framework. Rather than relying only on ratings, conventional sentiment classification, or adoption surveys, the study combines temporal rating dynamics, IndoRoBERTa-based sentiment distribution, human validation, rating-sentiment alignment, and bigram-based lexical interpretation. This design enables the study to identify not only whether expressed

evaluations are positive or negative, but also how users describe the operational conditions that shape their positive and negative user-perceived experience.

Because this study is descriptive and exploratory, it develops research questions rather than directional causal hypotheses. The analysis addresses the following questions:

RQ1. How did M-Pajak user ratings change during 2021–2026?

RQ2. How were IndoRoBERTa-based sentiment labels distributed over time?

RQ3. To what extent did star ratings align with textual sentiment?

RQ4. Which lexical patterns characterized positive and negative user-perceived experiences?

In response to these gaps, this study investigates user evaluations of the Indonesian M-Pajak mobile application using Google Play reviews as review-based evidence. Specifically, it examines the temporal dynamics of user ratings, the yearly distribution of model-generated sentiment labels, the degree of alignment between star ratings and validated sentiment classifications, and the lexical patterns through which positive and negative user-perceived experience is expressed. The study contributes to the literature by extending M-Pajak research toward a citizen-facing analysis of expressed evaluation, integrating validated IndoRoBERTa sentiment classification with rating and lexical analysis, and translating recurrent bigram expressions into broader experience dimensions. The next section details the data source, cleaning procedure, IndoRoBERTa classification workflow, human validation protocol, and analytical framework used to connect ratings, sentiment labels, and lexical patterns.

3. Methodology

This study adopts a computational text analysis design to examine user evaluations and sentiment patterns in Google Play reviews of the M-Pajak mobile application. App-store reviews have been widely treated as valuable sources of user feedback for opinion mining and software-related analysis, as shown in systematic reviews by Genc-Nayebi & Abran (2017) and Dąbrowski et al. (2022). The methodological design combines data scraping, data cleaning, transformer-based sentiment classification, human validation, rating-based temporal analysis, rating-sentiment alignment, and lexical pattern analysis. This multi-stage approach enables the study to examine not only the distribution of ratings and model-generated sentiment labels, but also the recurring lexical expressions that underlie positive and negative user experiences.

The analysis is structured to support the empirical results presented in the subsequent section. First, review ratings are examined longitudinally to identify temporal shifts in explicit user evaluation. Second, IndoRoBERTa-based sentiment labels are analyzed over time to capture the semantic polarity expressed in review texts. Third, rating-sentiment alignment is assessed because star ratings and textual sentiment may provide related but not fully identical evaluative signals, a methodological issue highlighted by Al-Natour & Turetken (2020). Finally, bigram-based lexical patterns are extracted and grouped into broader experience dimensions to identify the main sources of satisfaction and dissatisfaction.

3.1 Data Collection

The data used in this study were collected from publicly available Google Play reviews of the M-Pajak mobile tax application. The source application was identified using the package identifier id.go.pajak.djp, and the scraping session was completed on 24 May 2026. The data source was the Google Play application page at https://play.google.com/store/apps/details?id=id.go.pajak.djp&hl=en-US. Google Play reviews were used because they combine explicit star ratings with free-text feedback, enabling

numerical and semantic indicators of expressed user evaluation to be examined jointly (Dąbrowski et al., 2022; Genc-Nayebi & Abran, 2017).

The scraping process was conducted in a Python environment using the google-play-scraper library with lang = id, country = id, Sort.MOST_RELEVANT, and filter_score_with = None. The request limit was set to 500,000 to avoid imposing an artificial cap, but the platform–library interaction returned 7,009 records. Sort.MOST_RELEVANT was used in the original collection because the study was designed to capture salient, publicly visible review narratives across rating levels and historical years rather than to reconstruct a complete chronological stream. Yearly grouping was based on each review timestamp, not on retrieval order. Nevertheless, relevance ranking can prioritize reviews according to Google Play's opaque ranking criteria and may therefore affect corpus composition. Accordingly, the temporal analysis is interpreted only as a comparison within the retrievable relevance-ranked corpus available under this configuration on the scraping date. It is not presented as a census of all M-Pajak reviews ever posted, a population estimate of all users, or definitive evidence of application deterioration. A direct MOST_RELEVANT-versus-NEWEST sensitivity comparison is recommended for future replication.

The extracted records contained review text, rating score, review timestamp, application-version information, reply-related metadata, and user display fields returned by the library. The analysis used review text, rating score, timestamp-derived year, and model-generated sentiment labels. User-identifying fields were excluded from interpretation and reporting. The resulting dataset should therefore be understood as the retrievable review corpus under the stated date and configuration, rather than as an independently verified archive of every review ever submitted to Google Play.

Table 2. Summary of Data Collection Procedure

Component	Description
Application	M-Pajak
Google Play package identifier	id.go.pajak.djp
Data source	Google Play review page of the M-Pajak application
Scraping library	google-play-scraper
Language and country settings	lang = id; country = id
Sorting setting	Sort.MOST_RELEVANT
Rating filter	filter_score_with = None (all rating scores included)
Requested review count	500,000 reviews
Initial retrieved records	7,009 reviews
Observation period in cleaned dataset	2021–2026
Scraping date	24 May 2026
Corpus interpretation	Retrievable relevance-ranked corpus under the stated configuration; not a verified census of all reviews ever posted
Temporal-use boundary	Year comparisons use review timestamps and describe within-corpus review composition only

Source: Authors' computation and analysis (2026).

3.2 Data Cleaning and Screening

After data retrieval, the dataset underwent a structured cleaning and screening process to ensure that the analytical corpus contained valid rating scores and meaningful textual content. Text-preprocessing decisions can affect downstream sentiment-analysis results, as demonstrated by Symeonidis et al. (2018); therefore, the score variable was first converted into a numeric format, and only scores within the valid Google Play rating scale of 1 to 5 were retained. Review content was then converted into string format and stripped of leading and trailing whitespace before textual validity checks were applied.

The cleaning procedure applied three main content-based screening criteria. In sentiment-analysis pipelines for noisy user-generated text, preprocessing and normalization are commonly used to reduce non-informative variation before classification and lexical analysis (Duwairi & El-Orfali, 2014). First, review text was required to be non-empty. Second, reviews consisting entirely of numbers were excluded because they do not provide interpretable textual evidence for sentiment or lexical analysis. Third, reviews consisting only of emoji, emoticons, symbols, whitespace, or punctuation were removed because they do not contain sufficient lexical information for text-based analysis. Emoji-only reviews were detected using a regular-expression-based procedure that removed emoji characters, symbols, whitespace, and punctuation; reviews were classified as emoji-only when no meaningful textual content remained after this procedure.

The initial dataset consisted of 7,009 records. After applying the cleaning criteria, 6,980 reviews were retained for analysis and 29 reviews were removed. All removed records were classified as emoji- or emoticon-only content. No records were removed because of invalid rating scores, empty content, or numeric-only content. The cleaning summary is presented in Table 3.

Table 3. Summary of Data Cleaning Results

Cleaning Indicator	Frequency
Initial retrieved records	7,009
Records after cleaning	6,980
Removed records	29
Invalid score	0
Empty content	0
Numeric-only content	0
Emoji/emoticon-only content	29

Source: Authors' computation and analysis (2026).

The cleaned dataset covered reviews from 2021 to 2026. The distribution of reviews across years was uneven, reflecting natural variation in user review activity, user engagement, and application-related experience over time. No artificial balancing procedure was applied because preserving the natural distribution of reviews was important for maintaining the ecological validity of the temporal analysis. The final yearly distribution of the cleaned dataset is presented in Table 4.

Table 4. Yearly Distribution of Reviews after Cleaning

Year	Frequency	Percentage (%)
2021	202	2.89
2022	549	7.87
2023	1,087	15.57
2024	1,720	24.64
2025	2,216	31.75
2026	1,206	17.28
Total	6,980	100.00

Source: Authors' computation and analysis (2026).

Because the 2026 observations may represent an incomplete observation year, temporal findings involving 2026 were interpreted with caution. This consideration is particularly relevant for year-by-year comparisons of rating composition and sentiment distribution.

3.3 Transformer-Based Sentiment Classification

Sentiment classification was conducted using a transformer-based sentiment classifier for Indonesian text, namely IndoRoBERTa. The IndoRoBERTa checkpoint used in the analysis was w11wo/indonesian-roberta-base-sentiment-classifier, which is documented in the Hugging Face model card by Wongso (2023). This model was selected because transformer-based and deep-learning text-classification approaches are designed to represent contextual and semantic relationships in text (Minaee et al., 2021), including spelling variation, abbreviated expressions, and non-standard grammatical patterns commonly found in Google Play reviews.

The model was implemented using the Hugging Face Transformers library in a Python environment. GPU acceleration was used when available; otherwise, inference was performed on CPU. For GPU-based inference, half-precision computation was used to improve processing efficiency, while full precision was used for CPU-based processing. Reviews were processed in batches of 32 texts, and text inputs were truncated to a maximum sequence length of 512 tokens to comply with transformer input constraints. A chunk-based processing function was used to ensure reproducibility and scalability, with a chunk size of 10,000 observations.

Each review was classified into one of three sentiment categories: positive, neutral, or negative, following the label structure of the selected Indonesian RoBERTa sentiment classifier (Wongso, 2023). IndoRoBERTa outputs were normalized into lowercase standardized sentiment labels. The prediction confidence scores generated by the model were retained in the analytical dataset together with the standardized labels. These model-generated labels were treated as analytical classifications rather than as definitive ground-truth labels for the full corpus.

Table 5. Summary of Sentiment Classification Configuration

Component	Specification
Sentiment model	IndoRoBERTa
IndoRoBERTa checkpoint	w11wo/indonesian-roberta-base-sentiment-classifier
Implementation library	Hugging Face Transformers
Sentiment categories	Positive, neutral, and negative

Component	Specification
Batch size	32 reviews per inference batch
Chunk size	10,000 observations
Maximum sequence length	512 tokens
Hardware configuration	GPU when available; CPU otherwise
Output retained	Standardized sentiment label and prediction score

Source: Authors' computation and analysis (2026).

3.4 Human Validation and Model Reliability Assessment

To strengthen the credibility of the transformer-based sentiment classification, the study used a manually annotated validation sample of 1,500 reviews. The sample was selected through proportionate stratified sampling by review year so that each year contributed approximately the same share as in the cleaned corpus: 43 reviews from 2021, 118 from 2022, 234 from 2023, 370 from 2024, 476 from 2025, and 259 from 2026. Reviews were then selected within each yearly stratum using a random-sampling procedure. The sample was not balanced by rating or anticipated sentiment because preserving the natural review distribution was important for evaluating the model under realistic class imbalance. This design also ensured that the validation exercise covered the complete observation period rather than being dominated exclusively by the largest annual strata.

Two Indonesian-speaking academic researchers familiar with Indonesian tax-administration terminology and the interpretation of user-generated service feedback served as annotators. Before full annotation, they received a written codebook, discussed constructed examples, and completed a calibration exercise outside the 1,500-review evaluation sample. The annotators then worked independently and were not shown IndoRoBERTa predictions during coding. The codebook defined positive sentiment as predominantly favorable evaluation, successful task completion, appreciation, or endorsement; neutral sentiment as factual description, information seeking, or a statement without a clearly dominant evaluative orientation; and negative sentiment as complaint, failure, difficulty, frustration, or adverse service evaluation. For mixed reviews, annotators assigned the label corresponding to the dominant evaluative meaning; where positive and negative content remained genuinely balanced, the review was coded as neutral. Context, negation, colloquial Indonesian, spelling variation, and sarcasm cues were considered. After independent coding, disagreements were reviewed against the codebook and resolved through consensus discussion. The consensus label was treated as the human gold standard.

Inter-annotator reliability was assessed using percentage agreement and Cohen's kappa (Cohen, 1960). Agreement between the adjudicated human gold standard and IndoRoBERTa was evaluated using overall agreement, Cohen's kappa, class-level precision, recall, F1-score, confusion-matrix analysis, and year-specific agreement. These complementary measures were retained because accuracy alone can be misleading under class imbalance (Sokolova & Lapalme, 2009; Hicks et al., 2022). The validated IndoRoBERTa labels were subsequently used for temporal sentiment, rating–sentiment alignment, and lexical analyses, while remaining explicitly treated as analytical classifications rather than definitive ground truth.

Table 6. Sentiment Annotation Rules and Constructed Examples

Label/rule	Operational definition	Constructed example
Positive	Predominantly favorable evaluation, appreciation, successful completion, usefulness, or endorsement.	"The application is easy to use and makes tax reporting faster."
Neutral	Factual statement, information request, or no clearly dominant evaluative orientation.	"How can I obtain an EFIN through this application?"
Negative	Complaint, failure, difficulty, frustration, inability to complete a task, or adverse service evaluation.	"The OTP does not arrive, so I cannot log in."
Mixed-review rule	Assign the dominant evaluative meaning; use neutral only when favorable and unfavorable content remain genuinely balanced.	"The features are useful, but login repeatedly fails." → Negative because the operational failure dominates.

Source: Authors' annotation protocol; examples are constructed and are not verbatim user reviews.

3.5 Analytical Framework

The study followed a multi-layer analytical framework designed to connect explicit user ratings, model-generated sentiment labels, and lexical expressions of user experience. App-review research increasingly combines multiple analytical lenses to extract richer evidence from user feedback (Dąbrowski et al., 2022), while systematic mapping work on mobile-app user experience highlights the value of linking review evidence to broader experiential factors (Nakamura et al., 2022). The analytical sequence consisted of five main stages: temporal rating analysis, temporal sentiment analysis, rating-sentiment alignment, positive lexical pattern analysis, and negative lexical pattern analysis.

First, temporal rating analysis was conducted by cross-tabulating rating scores by year. This analysis provided a descriptive baseline for identifying changes in explicit user evaluations over time. Second, temporal sentiment analysis examined the yearly distribution of IndoRoBERTa-based sentiment labels, allowing the study to determine whether textual sentiment showed a similar temporal pattern to star ratings. Third, rating-sentiment alignment was examined by cross-tabulating sentiment labels across rating scores. This stage was important because, as Al-Natour & Turetken (2020) demonstrate, star ratings and sentiment-analysis outputs can be related indicators but may not fully align in consumer-review settings. Fourth, lexical pattern analysis was conducted on reviews classified as positive by the validated IndoRoBERTa model. Bigram terms were extracted to identify recurring expressions associated with favorable user evaluations, reflecting the usefulness of textual features and n-gram-style evidence in app-review analysis (Maalej et al., 2016). Fifth, the same lexical pattern procedure was applied to reviews classified as negative. Negative bigram terms were used to identify recurring expressions associated with dissatisfaction, access barriers, administrative

burden, and system-related frustration. The overall analytical workflow is summarized below in Table 7.

Table 7. Analytical Workflow of the Study

Stage	Analytical Component	Purpose
1	Temporal rating analysis	To examine yearly changes in explicit star ratings.
2	Temporal sentiment analysis	To analyze yearly sentiment distributions based on labels generated by the validated IndoRoBERTa model.
3	Rating–sentiment alignment	To assess whether numerical ratings and textual sentiment convey consistent evaluative signals.
4	Positive lexical pattern analysis	To identify recurring bigram expressions associated with positive user evaluations.
5	Negative lexical pattern analysis	To identify recurring bigram expressions associated with negative user evaluations.
6	Experience dimension grouping	To consolidate related bigram terms into broader positive and negative experience dimensions.

Source: Authors' computation and analysis (2026).

3.6 Lexical Pattern Extraction and Experience Dimension Grouping

Lexical pattern analysis was conducted to identify recurring expressions underlying positive and negative user evaluations. Before bigram extraction, the text was lowercased and URLs, numbers, punctuation, excessive whitespace, common stopwords, and highly generic non-informative terms were removed. Bigrams were extracted separately from reviews classified as positive and negative by the validated IndoRoBERTa model. Frequency tables and dimension-level bar charts were treated as the primary analytical evidence. Word clouds were retained only as supplementary visual summaries and were not used to determine rankings, dimension frequencies, or substantive conclusions.

The 30 most frequent positive bigrams and 30 most frequent negative bigrams were ranked by occurrence. Each term was interpreted within the sentiment-labelled reviews in which it appeared, because expressions such as "bayar pajak" or "live chat" are not intrinsically positive or negative. Their analytical polarity derives from the review context, consistent with the principle that sentiment depends on contextual usage rather than isolated lexical forms (Nandwani & Verma, 2021).

Experience-dimension grouping followed a transparent two-stage coding procedure. One coder assigned each of the 60 ranked bigrams to an operationally defined dimension according to its dominant experiential meaning. A second coder independently reviewed the proposed assignments and the surrounding review contexts. Differences were discussed until a consensus assignment was reached. Because the published results report final consensus coding rather than two retained independent dimension-label sets, a separate kappa statistic was not calculated for this stage. The complete bigram-to-dimension mapping and frequency subtotals are reported in Tables 20a and 20b, allowing readers to audit how the dimension totals were produced.

The coding dimensions were defined before final tabulation and included usability/accessibility, appreciation, helpfulness/public value, support/account assistance, tax-

service convenience, and perceived security for positive reviews; and authentication/OTP/login, payment/administrative burden, NPWP/account data, system error/service frustration, and EFIN recovery for negative reviews. Each bigram was assigned to one dimension only, and the reported dimension frequency equals the sum of the frequencies of its assigned bigrams. This one-to-one assignment prevents double counting and makes the aggregation reproducible from Tables 20a and 20b.

3.7 Ethical and Privacy Considerations

This study used publicly available Google Play review data and analyzed the information at an aggregate level. Even when online-platform data are publicly accessible, ethics scholarship cautions that researchers should consider contextual privacy, traceability, and potential identifiability risks (Gliniecka, 2023). User-identifying fields, including user names and profile images, were not used in the analytical interpretation and were not reported in the manuscript. No attempt was made to identify, profile, or contact individual users. The analysis focused on review text, rating scores, timestamps, model-generated sentiment labels, and aggregate lexical patterns, in line with recommendations that studies using online or social-media data report privacy-protection practices transparently (Zhang et al., 2024).

Because the study relied on publicly available, non-interventional online review data, individual informed consent was not required. Nevertheless, ethical discussions of online-data research emphasize that public availability does not remove researcher responsibility for privacy, confidentiality, and appropriate data handling (Hunter et al., 2018; Samuel & Buchanan, 2020). Therefore, the study minimized privacy risks by avoiding the disclosure of personally identifiable information and by presenting findings only in aggregate statistical, temporal, and lexical forms. Verbatim review quotations were also avoided to reduce the risk that individual user posts could be traced back through search engines or platform searches (Zhang et al., 2024).

4. Results and Discussion

4.1 Human Validation and IndoRoBERTa Reliability Assessment

Human validation assessed whether IndoRoBERTa labels were consistent with human judgment. The 1,500-review sample was selected through proportionate stratification by year and independently annotated by two trained Indonesian-speaking researchers using the codebook summarized in Table 5a. The year-specific allocations were 43, 118, 234, 370, 476, and 259 reviews for 2021–2026, respectively. Disagreements were resolved by consensus, and the adjudicated labels served as the human gold standard.

The inter-annotator reliability results indicate that the human annotation process was highly consistent. Out of 1,500 validation samples, the two annotators assigned identical labels to 1,458 reviews, corresponding to an agreement rate of 97.20%. The Cohen's Kappa value was 0.9132, which falls into the category of almost perfect agreement. This result suggests that the human labeling process was sufficiently reliable and that the adjudicated labels provide a stable basis for evaluating the model-generated sentiment labels.

Table 8. Summary of Human Annotation Reliability and IndoRoBERTa Agreement

Comparison	N	Agreement (f)	Agreement (%)	Cohen's Kappa	Interpretation
Annotator 1 vs Annotator 2	1500	1458	97.20	0.9132	Almost perfect agreement
Human Gold Standard vs IndoRoBERTa	1500	1413	94.20	0.8243	Almost perfect agreement

Source: Authors' computation and analysis (2026).

The distribution of the human gold standard labels shows that the validation dataset was dominated by negative sentiment. Specifically, 1,218 reviews were classified as negative, representing 81.20% of the validation sample. Meanwhile, 118 reviews were labeled as neutral, accounting for 7.87%, and 164 reviews were labeled as positive, accounting for 10.93%. This distribution confirms the presence of class imbalance in the validation data. Therefore, the evaluation of IndoRoBERTa should not rely solely on accuracy, but should also consider macro-averaged and weighted performance metrics as well as Cohen's Kappa.

When the IndoRoBERTa labels were compared with the human gold standard, the model demonstrated a very high level of agreement. IndoRoBERTa produced 1,413 labels that matched the adjudicated human labels out of 1,500 cases, yielding an overall agreement rate of 94.20%. The Cohen's Kappa value was 0.8243, indicating almost perfect agreement between IndoRoBERTa and human judgment. This finding provides strong empirical support for using IndoRoBERTa as the primary sentiment classification model in this study.

Table 9. Distribution of Human Gold Standard Labels

Human Label	Frequency	Percentage (%)
Negative	1218	81.20
Neutral	118	7.87
Positive	164	10.93
Total	1500	100.00

Source: Authors' computation and analysis (2026).

The classification metrics further confirm the robustness of IndoRoBERTa. The model achieved an accuracy of 0.9420, a macro precision of 0.8653, a macro recall of 0.8866, and a macro F1-score of 0.8722. In addition, the weighted precision, weighted recall, and weighted F1-score were 0.9485, 0.9420, and 0.9443, respectively. The high weighted F1-score indicates that the model performed strongly under an imbalanced label distribution, while the macro F1-score suggests that the model maintained a reasonably balanced performance across sentiment categories.

At the class level, IndoRoBERTa performed particularly well in identifying negative and positive sentiment. For the negative class, the model obtained a precision of 0.9742, recall of 0.9622, and F1-score of 0.9682. For the positive class, the model achieved a precision of 0.9732, recall of 0.8841, and F1-score of 0.9265. The neutral class was comparatively more challenging, with a precision of 0.6486, recall of 0.8136, and F1-score of 0.7218. This pattern indicates that although the model was able to capture most neutral reviews, some reviews predicted as neutral by the model were judged differently by human annotators.

Table 10. Class-Level Performance of IndoRoBERTa against the Human Gold Standard

Label	Precision	Recall	F1-score	Support
Negative	0.9742	0.9622	0.9682	1218
Neutral	0.6486	0.8136	0.7218	118
Positive	0.9732	0.8841	0.9265	164

Source: Authors' computation and analysis (2026).

Table 11. Overall IndoRoBERTa Evaluation Metrics

Metric	Value
Accuracy	0.9420
Macro Precision	0.8653
Macro Recall	0.8866
Macro F1-score	0.8722
Weighted Precision	0.9485
Weighted Recall	0.9420
Weighted F1-score	0.9443
Cohen's Kappa	0.8243

Source: Authors' computation and analysis (2026).

The confusion matrix provides a more detailed explanation of the model's classification behavior. Among the 1,218 reviews labeled as negative by the human gold standard, 1,172 were also classified as negative by IndoRoBERTa, while 42 were classified as neutral and 4 as positive. Among the 118 human-labeled neutral reviews, 96 were correctly classified as neutral, whereas 22 were classified as negative. Among the 164 human-labeled positive reviews, 145 were correctly classified as positive, while 9 were classified as negative and 10 as neutral. These results indicate that the main classification challenges occurred around the boundary between neutral and the two polarized sentiment classes.

The yearly evaluation also shows that IndoRoBERTa maintained a stable level of agreement with the human gold standard across different review years. The Cohen's Kappa values ranged from 0.7313 to 0.9233. The highest agreement was observed in 2021, followed by 2022, 2023, and 2024, all of which reached almost perfect agreement. Although the Kappa values for 2025 and 2026 were lower, both still indicated substantial agreement. This temporal pattern suggests that IndoRoBERTa's performance was not confined to a single period, but remained consistently strong across the validation sample.

Table 12. Confusion Matrix of Human Gold Standard and IndoRoBERTa

Human Label	IndoRoBERTa Negative	IndoRoBERTa Neutral	IndoRoBERTa Positive	Total
Human Negative	1172	42	4	1218
Human Neutral	22	96	0	118
Human Positive	9	10	145	164
Total	1203	148	149	1500

Source: Authors' computation and analysis (2026).

Table 13. Yearly Agreement between Human Gold Standard and IndoRoBERTa

Year	N	Agreement (f)	Agreement (%)	Cohen's Kappa	Interpretation
2021	43	41	95.35	0.9233	Almost perfect agreement
2022	118	112	94.92	0.9079	Almost perfect agreement
2023	234	226	96.58	0.8960	Almost perfect agreement
2024	370	351	94.86	0.8189	Almost perfect agreement
2025	476	440	92.44	0.7313	Substantial agreement
2026	259	243	93.82	0.7854	Substantial agreement

Source: Authors' computation and analysis (2026).

Overall, the human validation results provide strong methodological support for the use of IndoRoBERTa in the subsequent sentiment analysis. The high inter-annotator reliability confirms the credibility of the human gold standard, while the strong agreement between IndoRoBERTa and the adjudicated labels demonstrates that the model closely approximates expert human judgment. Accordingly, IndoRoBERTa can be considered a reliable model for generating sentiment labels in this study.

4.2 Temporal Dynamics of User Ratings

Having established the reliability of the human validation process and the robustness of IndoRoBERTa-based sentiment classification, the analysis proceeds to examine the temporal dynamics of explicit user evaluations. Table 13 reports the yearly distribution of star ratings from 2021 to 2026, thereby providing an initial descriptive basis for understanding how users evaluated the M-Pajak application over time.

Table 14. Distribution of Review Ratings by Year

Year	Rating 1		Rating 2		Rating 3		Rating 4		Rating 5		Total	
	f	%	f	%	f	%	f	%	f	%	f	%
2021	57	28.22	14	6.93	13	6.44	16	7.92	102	50.50	202	100.00
2022	288	52.46	32	5.83	35	6.38	14	2.55	180	32.79	549	100.00
2023	792	72.86	55	5.06	35	3.22	26	2.39	179	16.47	1087	100.00
2024	1329	77.27	61	3.55	37	2.15	32	1.86	261	15.17	1720	100.00
2025	1763	79.56	79	3.56	42	1.90	36	1.62	296	13.36	2216	100.00
2026	995	82.50	45	3.73	15	1.24	16	1.33	135	11.19	1206	100.00
Total	5224	74.84	286	4.10	177	2.54	140	2.01	1153	16.52	6980	100.00

Source: Authors' computation and analysis (2026).

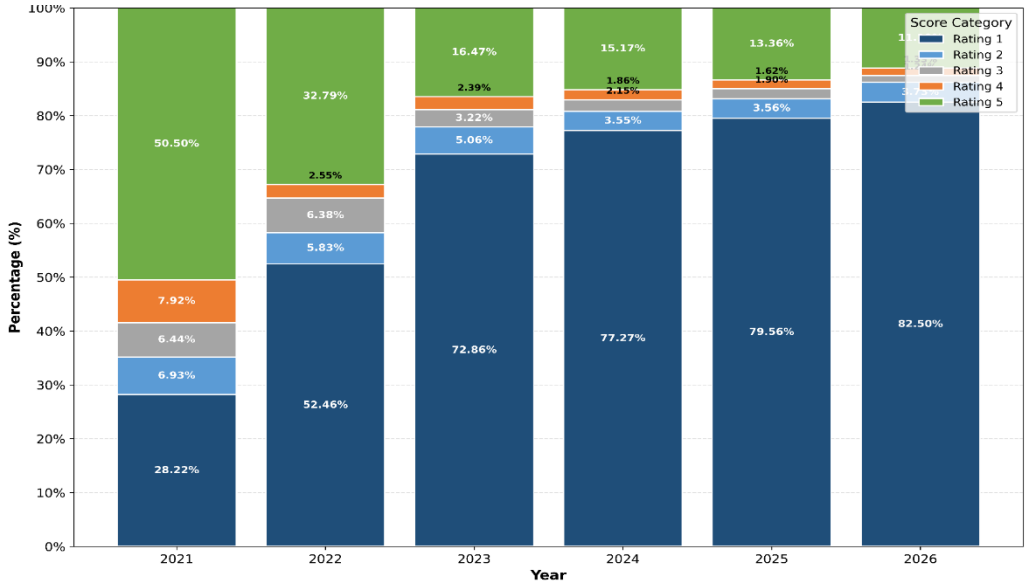
The distribution of review ratings indicates a pronounced dominance of low ratings across the full observation period. Out of 6,980 reviews, 5,224 reviews were assigned a one-star rating, representing 74.84% of all observations. By contrast, five-star ratings accounted for 1,153 reviews, or 16.52% of the dataset, while two-, three-, and four-star ratings collectively represented only 8.65%. This distribution suggests that user evaluations were not evenly spread across the rating scale; rather, they were heavily concentrated at the lowest end of the rating spectrum.

Within the retrieved relevance-ranked corpus, the yearly composition of ratings shifted markedly. In 2021, one-star ratings accounted for 28.22% of reviews, while five-star ratings represented 50.50%, indicating that early user evaluations were comparatively more favorable. However, the proportion of one-star ratings increased sharply to 52.46% in 2022, 72.86% in 2023, 77.27% in 2024, 79.56% in 2025, and 82.50% in 2026. At the same time, five-star ratings declined from 50.50% in 2021 to 11.19% in 2026. This inverse movement suggests a marked shift in expressed user evaluations, from a relatively more favorable pattern in the early period toward a predominantly negative evaluative pattern in later years.

Figure 1 visualizes this temporal shift more clearly by displaying the trajectory of each rating category over time. The figure is particularly useful because it highlights the widening gap between one-star and five-star ratings, while also showing that intermediate ratings remained relatively marginal throughout the observation period. Thus, the graphical evidence reinforces the tabular finding that dissatisfaction became increasingly dominant in user evaluations.

Overall, the rating-based analysis shows that later-year observations in the retrieved corpus contained a larger share of low ratings. This pattern should be described as a shift in expressed review-based evaluation, not as proof that the application necessarily became worse or that satisfaction declined among all M-Pajak users. App-store reviews are self-selected, the MOST_RELEVANT ranking may affect which reviews were retrievable, review behavior can intensify around updates or service disruptions, and the 2026 data cover only the collection window available on 24 May 2026. Star ratings also provide only a coarse ordinal indicator; the next section therefore examines review-text sentiment.

Figure 1. Distribution of User Ratings Over Time



Source: Authors' visualization (2026).

4.3 Temporal Distribution of IndoRoBERTa-Based Sentiment Labels

While rating scores provide explicit numerical evaluations, textual reviews may contain more nuanced expressions of satisfaction, frustration, ambiguity, or mixed experience. Therefore, Table 14 extends the temporal analysis by presenting the yearly distribution of sentiment labels generated using the validated IndoRoBERTa model.

Table 15. Distribution of Sentiment Labels by Year

Year	Negative		Neutral		Positive		Total	
	f	%	f	%	f	%	f	%
2021	94	46.53	51	25.25	57	28.22	202	100.00
2022	346	63.02	87	15.85	116	21.13	549	100.00
2023	847	77.92	112	10.30	128	11.78	1087	100.00
2024	1380	80.23	147	8.55	193	11.22	1720	100.00
2025	1817	81.99	189	8.53	210	9.48	2216	100.00
2026	1017	84.33	88	7.30	101	8.37	1206	100.00
Total	5501	78.81	674	9.66	805	11.53	6980	100.00

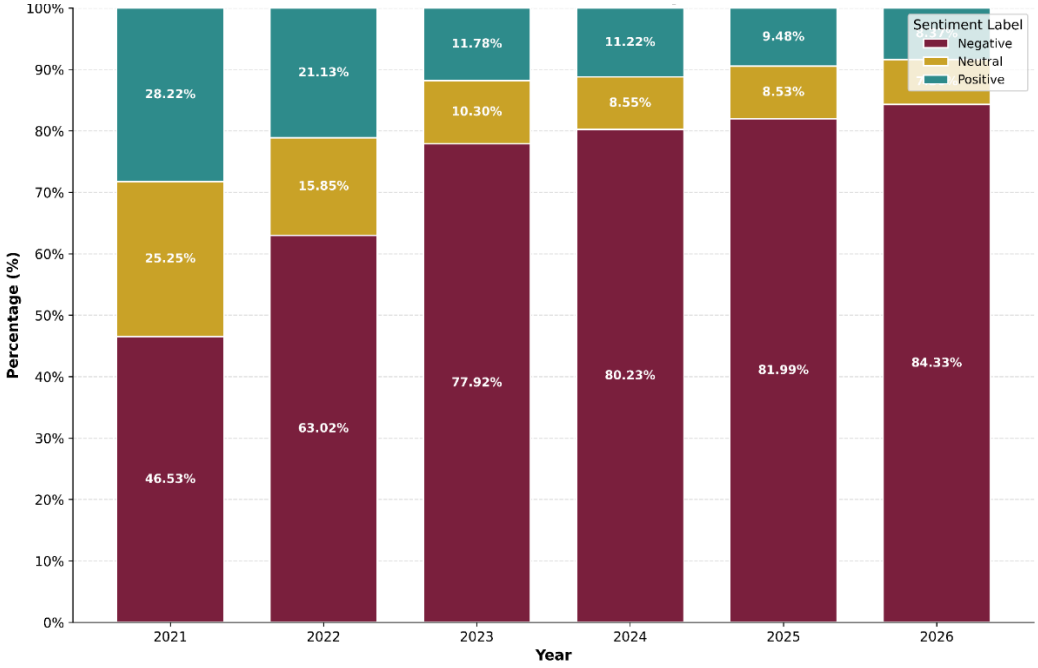
Source: Authors' computation and analysis (2026).

The sentiment distribution confirms that negative sentiment was the dominant polarity in the dataset. Across all years, 5,501 reviews were classified as negative, accounting for 78.81% of the total sample. Positive sentiment comprised 805 reviews, or 11.53%, while neutral sentiment accounted for 674 reviews, or 9.66%. This pattern is highly consistent with the rating distribution reported in Table 13, where one-star ratings also constituted the majority of user evaluations.

Within the retrieved corpus, the annual share of reviews classified as negative increased from 46.53% in 2021 to 84.33% in the available 2026 observations, while positive and neutral shares decreased. This is a descriptive change in the composition of reviews captured under the selected retrieval configuration. It may reflect a combination of user self-selection, complaint-oriented posting, application updates, temporary service disruptions, changes in the active user base, and ranking effects. It should not be interpreted as a causal estimate of application-quality decline or as a representative trend for all users.

Figure 2 complements Table 14 by illustrating the changing composition of sentiment labels within the retrieved corpus. The figure is used descriptively to show relative annual shares; it does not establish the cause of the observed changes and should be read together with the corpus-boundary and incomplete-2026 cautions.

Figure 2. Distribution of Sentiment Labels Over Time



Source: Authors' visualization (2026).

Taken together, the rating and sentiment analyses show a convergent shift toward a larger proportion of negative expressed evaluations within the retrieved review corpus. The convergence strengthens the description of the review stream, but it does not convert the data into a representative user-satisfaction survey or establish that the application itself deteriorated. Because ratings and text sentiment are related but not identical, the next section examines their alignment directly.

4.4 Rating–Sentiment Alignment Based on IndoRoBERTa Classification

Table 15 examines the relationship between star ratings and IndoRoBERTa-based sentiment labels. This analysis is important because star ratings represent explicit user scoring, whereas sentiment labels represent the semantic orientation of the written review. Assessing their alignment therefore helps determine whether numerical ratings and textual sentiment convey consistent signals of user experience.

Table 16. Distribution of Sentiment Labels Across Ratings

Rating	Negative		Neutral		Positive		Total	
	f	%	f	%	f	%	f	%
1	4851	92.86	313	5.99	60	1.15	5224	100.00
2	245	85.66	33	11.54	8	2.80	286	100.00
3	115	64.97	43	24.29	19	10.73	177	100.00
4	52	37.14	33	23.57	55	39.29	140	100.00
5	238	20.64	252	21.86	663	57.50	1153	100.00
Total	5501	78.81	674	9.66	805	11.53	6980	100.00

Source: Authors' computation and analysis (2026).

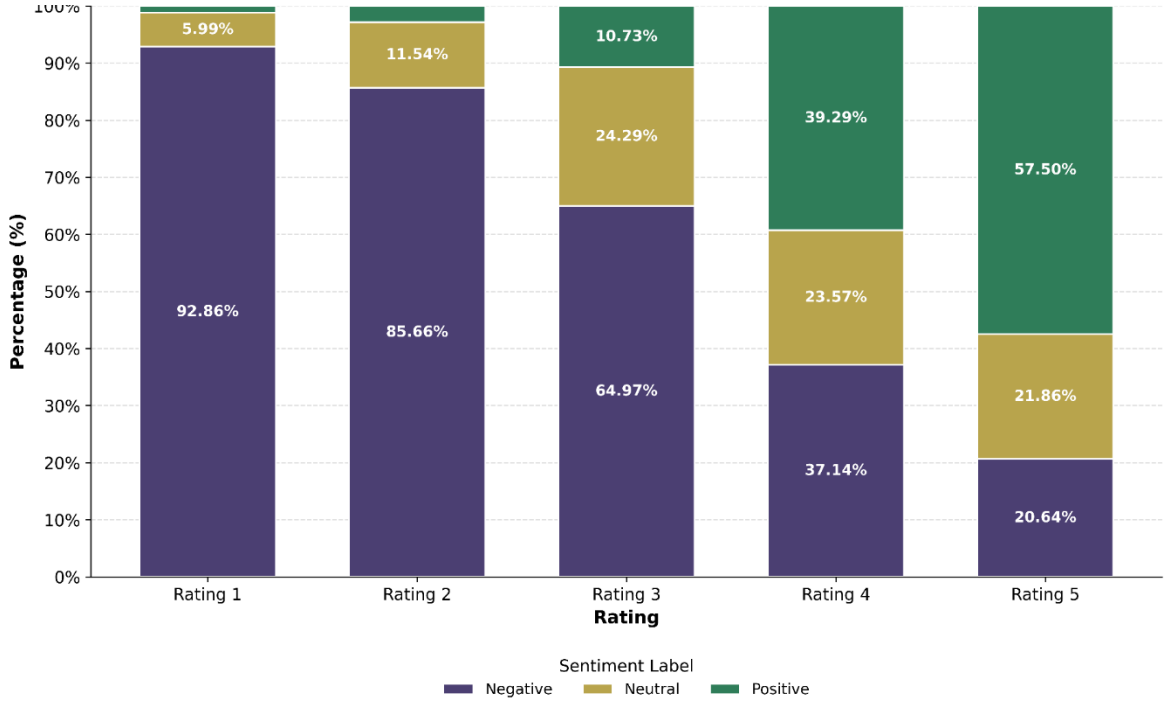
The results show strong alignment at the lower end of the rating scale. Among one-star reviews, 4,851 out of 5,224 reviews were classified as negative, corresponding to 92.86%. Similarly, 85.66% of two-star reviews were classified as negative. These results indicate that low ratings were highly consistent with negative textual sentiment, confirming that users who assigned low scores generally articulated dissatisfaction in their written comments as well.

The alignment becomes less straightforward for middle and high ratings. Three-star reviews were still dominated by negative sentiment, with 64.97% classified as negative, 24.29% as neutral, and only 10.73% as positive. Four-star reviews showed a more balanced pattern, with 39.29% classified as positive, 37.14% as negative, and 23.57% as neutral. Although five-star reviews were mostly positive, the alignment was not complete: 57.50% of five-star reviews were classified as positive, while 20.64% were classified as negative and 21.86% as neutral. This suggests that higher star ratings do not always correspond to uniformly positive textual content.

Figure 3 provides a visual representation of these alignment patterns by showing how sentiment categories are distributed across rating levels. The figure clarifies two key findings. First, low ratings are strongly associated with negative sentiment. Second, higher ratings contain more heterogeneous textual sentiment, indicating that some users may provide favorable scores while still expressing reservations, suggestions, or partial dissatisfaction in their comments.

These findings demonstrate that sentiment classification adds analytical value beyond the use of rating data alone. Ratings provide an efficient summary of user evaluation, but IndoRoBERTa-based sentiment labels capture the semantic tone embedded in review texts. Having established both the temporal dominance of negative evaluations and the partial alignment between ratings and textual sentiment, the analysis next turns to lexical patterns to identify the specific themes that shape positive and negative user experiences.

Figure 3. Distribution of Sentiment Labels Across Ratings



Source: Authors' visualization (2026).

4.5 Lexical Patterns in Positive Reviews

After examining aggregate sentiment distributions and their alignment with rating scores, the next analytical step is to identify the lexical patterns underlying positive user experiences. Table 16 presents the top 30 bigram terms appearing in reviews classified as positive based on labels generated by the validated IndoRoBERTa model.

Table 17. Top 30 Bigram Terms in Positive Reviews Based on Labels Generated by the Validated IndoRoBERTa Model

Rank	Term	Frequency
1	good job	7
2	mudah digunakan	6
3	membantu wajib	5
4	the best	5
5	cepat mudah	4
6	semoga kedepannya	4
7	live chat	4
8	membayar pajak	3
9	semoga membantu	3
10	secara online	3
11	mudah diakses	3
12	mudah aman	3
13	mudah akses	3
14	mendapatkan efin	3
15	dikirim email	3
16	membantu masyarakat	3
17	pajak membantu	3
18	good apk	3
19	membantu mudah	3
20	pelaporan pajak	3
21	yg membantu	3
22	pelayanan pajak	3
23	mudah dipahami	2
24	efin mudah	2
25	mudah bgt	2
26	beri bintang	2
27	lapor pajak	2
28	proses cepat	2
29	bermanfaat bagi	2
30	lancar kendala	2

Source: Authors' computation and analysis (2026).

The positive bigram distribution suggests that favorable reviews were primarily associated with ease of use, perceived helpfulness, and general appreciation. The most frequent positive terms include "good job", "mudah digunakan", "membantu wajib", "the best", and "cepat mudah". Although the absolute frequencies of these terms are relatively modest, their semantic content indicates that users valued the application when it was perceived as simple, accessible, useful, and supportive of tax-related activities.

Several positive bigrams also reflect the public service function of the application. Terms such as "membantu masyarakat", "pajak membantu", "pelaporan pajak", "pelayanan pajak", and "lapor pajak" indicate that some users associated positive experiences with the broader utility of digital tax administration. In addition, phrases such as "mendapatkan efin", "dikirim email", and "live chat" suggest that account-related assistance and support channels contributed to positive perceptions when they functioned effectively.

To move beyond individual bigram expressions, the positive bigrams were grouped into broader experience dimensions, as summarized in Table 17. The final assignment of every ranked positive bigram is shown in Table 20a. The one-to-one coding rule ensures that the dimension totals can be reconstructed directly from the listed term frequencies.

Table 18. Frequency Distribution of Positive Experience Dimensions Based on Grouped Bigram Term Occurrences

Dimension	Frequency	Percentage (%)
Ease of Use and Accessibility	23	23.96
General Appreciation and Endorsement	21	21.88
Helpfulness and Public Value	19	19.79
Support and Account Assistance	15	15.62
Tax Service and Reporting Convenience	15	15.62
Trust and Perceived Security	3	3.12
Total	96	100.00

Source: Authors' computation and analysis (2026).

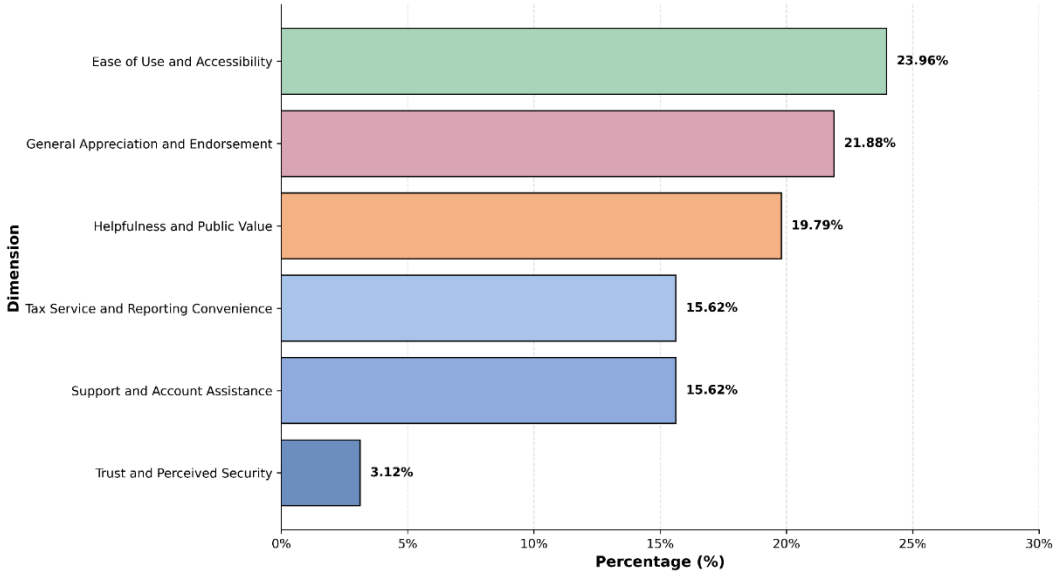
As shown in Table 17, Ease of Use and Accessibility represents the largest positive experience dimension, accounting for 23.96% of grouped positive bigram occurrences. This is followed by General Appreciation and Endorsement at 21.88% and Helpfulness and Public Value at 19.79%. These results indicate that positive user experience was mainly constructed around practical usability and perceived usefulness rather than around security-related reassurance. Given the relatively low absolute frequencies of the positive bigrams, these patterns should be interpreted as indicative lexical signals rather than exhaustive representations of all positive user experiences. This limitation also provides a useful comparative baseline for the next section, which examines the more prominent lexical patterns found in negative reviews.

Figure 4 is retained only as a supplementary visual overview of positive bigrams. The analytical interpretation relies on the ranked frequencies in Table 16 and the structured dimension frequencies in Table 17 and Figure 5. The word cloud is not used for quantitative comparison because font size and spatial arrangement are less precise than tabular or bar-chart displays.

Figure 4. Supplementary Word Cloud of Bigram Terms in Positive Reviews

Source: Authors' visualization (2026).

Figure 5. Dominant Positive Experience Dimensions in Positive Reviews



Source: Authors' visualization (2026).

4.6 Lexical Patterns in Negative Reviews

To complement the analysis of positive reviews, Table 18 presents the top 30 bigram terms in negative reviews based on validated IndoRoBERTa labels. Compared with the positive bigram distribution, the negative terms show substantially higher absolute frequencies. Although this pattern partly reflects the larger number of reviews classified as negative, the repeated occurrence of access-, verification-, and transaction-related terms indicates that negative experiences were concentrated around recurring operational issues.

Table 19. Top 30 Bigram Terms in Negative Reviews Based on Labels Generated by the Validated IndoRoBERTa Model

Rank	Term	Frequency
1	bayar pajak	317
2	kode verifikasi	142
3	minta ampun	106
4	kode otp	92
5	login susah	89
6	log in	78
7	daftar npwp	68
8	gak masuk	59
9	live chat	58
10	ga login	57
11	susah login	55
12	no npwp	55
13	minta kode	48
14	minta efin	48
15	verifikasi email	48
16	susahnya minta	47
17	verifikasi gagal	45
18	lupa password	44
19	lupa sandi	41
20	npwp online	40
21	ga masuk	40
22	eror mulu	33
23	efin gagal	33
24	nomor npwp	33
25	kirim kode	32
26	pajak ribet	32
27	gak login	31
28	daftar online	29
29	error mulu	28
30	daftar susah	28

Source: Authors' computation and analysis (2026).

Before interpreting these terms, it is important to note that their polarity is derived from the sentiment label of the review context rather than from the intrinsic sentiment of the bigram terms themselves. For example, terms such as "bayar pajak" and "live chat" are not inherently negative; however, their repeated occurrence within negative reviews indicates that they were

frequently embedded in complaint-oriented contexts. The most frequent negative bigram is "bayar pajak", with 317 occurrences, followed by "kode verifikasi" with 142 occurrences, "minta ampun" with 106 occurrences, "kode otp" with 92 occurrences, and "login susah" with 89 occurrences. These terms suggest that negative user experiences were strongly associated with core access and transaction processes, particularly verification codes, OTP mechanisms, login difficulty, and tax payment activities. The recurrence of terms such as "gak masuk", "ga login", "susah login", "verifikasi gagal", "lupa password", and "lupa sandi" further indicates that authentication and access barriers were central sources of dissatisfaction.

The lexical evidence also shows that administrative identity and account-related issues contributed substantially to negative experiences. Terms such as "daftar npwp", "no npwp", "npwp online", "nomor npwp", "daftar online", and "daftar susah" indicate that users encountered difficulties related to NPWP registration, identification, or account linkage. In addition, EFIN-related terms such as "minta efin" and "efin gagal" point to recovery and verification problems, while expressions such as "eror mulu", "error mulu", and "pajak ribet" reflect broader frustration with system reliability and administrative burden.

To synthesize these recurring negative expressions, the negative bigrams were grouped into five experience dimensions, as summarized in Table 19. The final assignment of every ranked negative bigram is shown in Table 20b. The one-to-one assignment rule prevents double counting and makes the dimension totals auditable.

Table 20. Frequency Distribution of Negative Experience Dimensions Based on Grouped Bigram Term Occurrences

Dimension	Frequency	Percentage (%)
Authentication, OTP, and Login Barriers	901	48.55
Tax Payment and Administrative Burden	349	18.80
NPWP Registration and Account Data Issues	253	13.63
System Error and Service Frustration	225	12.12
EFIN Request and Recovery Problems	128	6.90
Total	1856	100.00

Source: Authors' computation and analysis (2026).

Table 21. Positive Bigram-to-Dimension Coding Map

Dimension	Assigned positive bigrams (frequency)	Subtotal
Ease of Use and Accessibility	mudah digunakan (6); cepat mudah (4); secara online (3); mudah diakses (3); mudah akses (3); mudah dipahami (2); mudah bgt (2)	23
General Appreciation and Endorsement	good job (7); the best (5); semoga kedepannya (4); good apk (3); beri bintang (2)	21
Helpfulness and Public Value	membantu wajib (5); membantu masyarakat (3); pajak membantu (3); membantu mudah (3); yg	19

Dimension	Assigned positive bigrams (frequency)	Subtotal
	membantu (3); bermanfaat bagi (2)	
Support and Account Assistance	live chat (4); semoga membantu (3); mendapatkan efin (3); dikirim email (3); efin mudah (2)	15
Tax Service and Reporting Convenience	membayar pajak (3); pelaporan pajak (3); pelayanan pajak (3); lapor pajak (2); proses cepat (2); lancar kendala (2)	15
Trust and Perceived Security	mudah aman (3)	3
Total	All 30 positive bigrams	96

Source: Authors' consensus coding of the ranked positive bigrams (2026).

Table 22. Negative Bigram-to-Dimension Coding Map

Dimension	Assigned negative bigrams (frequency)	Subtotal
Authentication, OTP, and Login Barriers	kode verifikasi (142); kode otp (92); login susah (89); log in (78); gak masuk (59); ga login (57); susah login (55); minta kode (48); verifikasi email (48); verifikasi gagal (45); lupa password (44); lupa sandi (41); ga masuk (40); kirim kode (32); gak login (31)	901
Tax Payment and Administrative Burden	bayar pajak (317); pajak ribet (32)	349
NPWP Registration and Account Data Issues	daftar npwp (68); no npwp (55); npwp online (40); nomor npwp (33); daftar online (29); daftar susah (28)	253
System Error and Service Frustration	minta ampun (106); live chat (58); eror mulu (33); error mulu (28)	225
EFIN Request and Recovery Problems	minta efin (48); susahnya minta (47); efin gagal (33)	128
Total	All 30 negative bigrams	1,856

Source: Authors' consensus coding of the ranked negative bigrams (2026).

Table 22 shows that Authentication, OTP, and Login Barriers constitute the largest negative experience dimension, accounting for 901 occurrences or 48.55% of grouped negative bigram occurrences. This finding indicates that nearly half of the recurring negative lexical patterns were related to users' inability to access, verify, or authenticate their accounts. The second-largest dimension is Tax Payment and Administrative Burden, with 349 occurrences or 18.80%, followed by NPWP Registration and Account Data Issues with 253 occurrences or 13.63%.

System Error and Service Frustration accounts for 12.12%, while EFIN Request and Recovery Problems accounts for 6.90%.

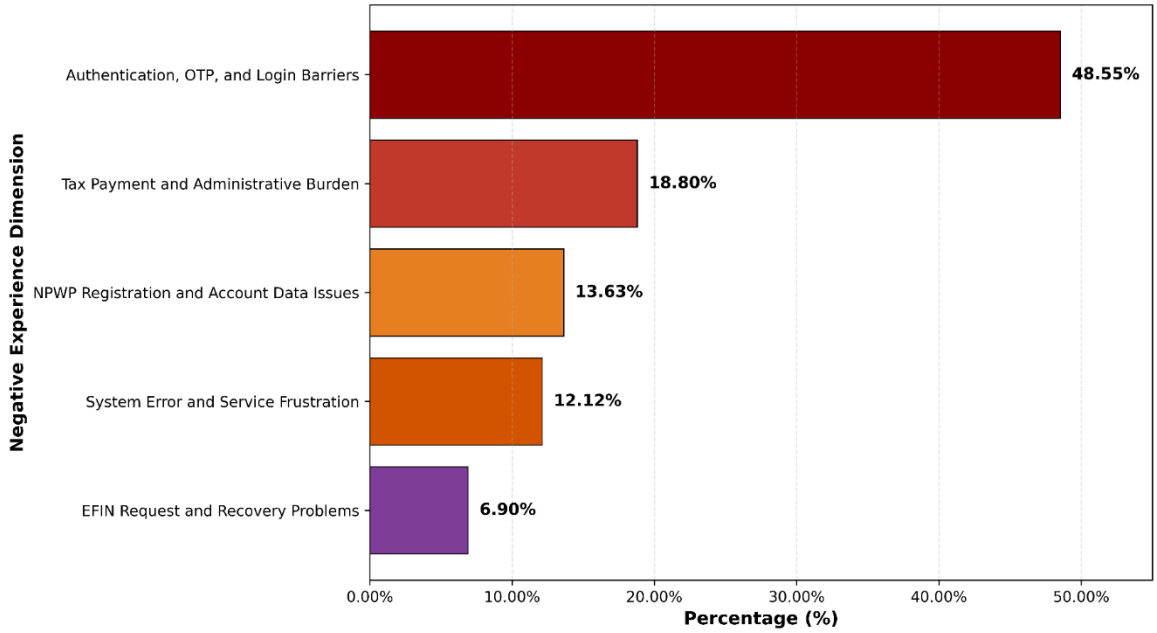
Figure 6 is retained only as a supplementary lexical overview. The substantive interpretation is based on the ranked frequencies in Table 18, the auditable coding map in Table 20b, and the dimension frequencies in Table 19 and Figure 7. This ordering prevents visually salient word-cloud terms from being treated as more analytically precise than the structured evidence.

Overall, the negative lexical analysis indicates that user dissatisfaction was not merely a generalized expression of poor evaluation, but was anchored in specific operational obstacles. The most critical barriers occurred at the entry point of digital service use, particularly login, OTP, verification, and account recovery processes. This suggests that improving authentication reliability, simplifying account access, strengthening NPWP and EFIN workflows, and reducing transaction-related friction may be central to improving perceived service quality in the M-Pajak application.

Figure 6. Supplementary Word Cloud of Bigram Terms in Negative Reviews

Source: Authors' visualization (2026).

Figure 7. Dominant Negative Experience Dimensions in Negative Reviews



Source: Authors' visualization (2026).

In summary, the results reveal a coherent pattern across rating distributions, IndoRoBERTa-based sentiment classifications, rating–sentiment alignment, and lexical analysis. Within the relevance-ranked retrievable corpus, later-year observations contained higher shares of low ratings and negative sentiment, and this compositional shift was reflected in both explicit ratings and review-text polarity. The lexical analysis further indicates that positive experiences were mainly associated with ease of use, perceived helpfulness, and service convenience, whereas negative experiences were concentrated around authentication, verification, login, NPWP, EFIN, system error, and tax payment-related barriers. These findings provide an empirical basis for the subsequent discussion, particularly in relation to digital service reliability, user accessibility, and administrative usability in mobile tax service delivery.

4.7 Integrated Discussion

4.7.1 Reliability of Human-Validated Transformer-Based Sentiment Classification

The validation results indicate that the sentiment classification pipeline is sufficiently reliable for substantive interpretation of the M-Pajak review corpus. In annotation-based computational linguistics, the credibility of automated analysis depends not only on model architecture but also on the reliability of the human-labelled reference data used to evaluate it (Artstein & Poesio, 2008). In this study, the high inter-annotator agreement and Cohen's Kappa value show that the human gold standard was internally consistent, while the strong agreement between IndoRoBERTa and the adjudicated labels supports the use of model-generated sentiment labels in the subsequent temporal, alignment, and lexical analyses.

The role of Cohen's Kappa is particularly important because it adjusts agreement for chance and is widely used for nominal classification tasks (Cohen, 1960). The high agreement obtained in this study therefore strengthens the interpretation that IndoRoBERTa was not merely reproducing majority-class tendencies, but approximating human judgement with a substantial degree of consistency. At the same time, the lower performance for the neutral class should be interpreted carefully. Neutral reviews often contain ambiguous, short, or mixed statements, and such class-level variation is common in sentiment-analysis settings where

precision, recall, F1-score, and confusion matrices provide more informative evidence than accuracy alone (Sokolova & Lapalme, 2009).

Methodologically, the results also reinforce the value of transformer-based classification for Indonesian user-generated text. Transformer models are designed to capture contextual relationships in language, and RoBERTa-style architectures improve contextual text representation through robust pretraining procedures (Liu et al., 2019). The use of an Indonesian RoBERTa sentiment classifier documented by Wongso (2023) is therefore appropriate for review texts that include informal expressions, abbreviated forms, and context-dependent polarity. Nevertheless, the validated labels should be understood as analytical classifications rather than definitive ground truth; this caution is consistent with broader evaluation guidance that classification metrics must be interpreted in relation to the task, label distribution, and intended use of the model (Hicks et al., 2022).

4.7.2 User Evaluations of M-Pajak as a Mobile Public Service Application

The temporal results suggest that the retrieved M-Pajak review stream contained increasingly large shares of negative expressed evaluation across the observed years. This pattern matters because mobile government applications are citizen-facing interfaces through which accessibility and service reliability are experienced (Twizeyimana & Andersson, 2019; Wang & Teo, 2020). However, the result is evidence about the composition of a self-selected, relevance-ranked review stream. It is not evidence that all users became less satisfied or that the application necessarily worsened over time.

The interpretation is especially relevant in digital tax administration because users depend on the application for obligatory administrative tasks. Poor experiences may therefore generate intense complaint behavior. Prior research links e-tax system quality with satisfaction and compliance intention (Saptono et al., 2023), but the present descriptive design does not test that causal pathway. The observed shifts may also coincide with updates, outages, workflow changes, or changes in who chose to review the application. Accordingly, the results identify review-based service-friction signals rather than a causal decline in satisfaction or compliance.

This finding extends earlier M-Pajak and Indonesian tax-technology studies by shifting attention from adoption and development perspectives toward expressed user evaluation. The evidence is observational, platform-specific, and bounded by the selected scraping configuration, but it provides a useful citizen-facing signal for service monitoring. Its strongest use is issue prioritization and hypothesis generation for subsequent internal diagnostics, usability testing, and representative user research—not population-level generalization.

4.7.3 Rating-Sentiment Alignment and the Limits of Star Ratings

The rating-sentiment alignment results show that numerical ratings and textual sentiment are strongly related but not interchangeable. This distinction is important because star ratings provide a compact ordinal summary, whereas review text contains the semantic content through which users explain their experiences. Al-Natour & Turetken (2020) show that star ratings and sentiment analysis can provide different but complementary views of consumer reviews. In the M-Pajak data, the strong concentration of negative sentiment among one-star and two-star reviews confirms that low ratings generally corresponded with negative textual evaluation.

However, the alignment became less straightforward among middle and high ratings. The presence of negative and neutral sentiment within four-star and five-star reviews indicates that some users may assign relatively favorable scores while still describing operational problems, mixed experiences, or partial satisfaction. This finding supports the methodological

decision to examine both rating distribution and sentiment labels. App-review studies have emphasized that user reviews contain rich issue-related information that cannot be fully captured by aggregate ratings alone (Genc-Nayebi & Abran, 2017). In software-engineering research, app reviews have also been used to identify recurring complaints, feature requests, and experience problems that are hidden behind numerical summaries (Dąbrowski et al., 2022).

The implication is that rating-based monitoring is useful but insufficient for public-service evaluation. Ratings identify the direction and intensity of explicit evaluation, while validated sentiment labels clarify the semantic orientation of the written review. In this study, the alignment analysis functions as a bridge between the temporal rating analysis and the lexical analyses. It justifies the move from broad distributional patterns to the specific terms and experience dimensions that explain why users evaluated the application positively or negatively.

4.7.4 Positive Experience Patterns: Ease of Use, Helpfulness, and Service Convenience

The positive lexical patterns show that favorable evaluations of M-Pajak were mainly associated with ease of use, accessibility, helpfulness, public value, support, reporting convenience, and perceived security. These dimensions are consistent with mobile government literature, which emphasizes that users evaluate public applications through service quality, usability, information quality, and perceived value (Wang & Teo, 2020). The prominence of expressions such as ease of use and quick access suggests that positive experiences emerged when users perceived the application as reducing rather than increasing the administrative effort required to interact with the tax authority.

The positive dimensions also show that some users interpreted M-Pajak through its public-service function. Terms associated with helping taxpayers, supporting reporting, and facilitating tax-related tasks indicate that favorable reviews were not limited to generic app appreciation. This is consistent with the public-value view of e-government, where digital services are evaluated in terms of their ability to support meaningful citizen-government interactions (Twizeyimana & Andersson, 2019). In the tax context, perceived usefulness and system quality are particularly important because digital platforms mediate required interactions between taxpayers and government agencies (Saptono et al., 2023).

At the same time, the positive lexical results should be interpreted proportionally. Positive bigram frequencies were much lower than negative bigram frequencies, reflecting the overall imbalance of sentiment in the corpus. Therefore, the positive patterns should not be read as evidence that favorable experiences dominated the application. Instead, they identify the conditions under which users were more likely to express appreciation: when the application was perceived as accessible, supportive, and practically useful for tax-related tasks. These conditions provide a constructive contrast to the negative patterns discussed below.

4.7.5 Negative Experience Patterns: Authentication, Login, OTP, NPWP, EFIN, Tax Payment, and System Errors

The negative lexical patterns indicate that dissatisfaction with M-Pajak was concentrated around specific operational barriers rather than around vague or general dislike. The most dominant dimension concerned authentication, OTP, and login barriers. This finding is substantively important because authentication is the entry point of digital public service use. If users cannot log in, verify their identity, or receive OTP codes reliably, then downstream services such as reporting, payment, account management, and information access become difficult or impossible to complete.

The second layer of negative experience involved tax payment and administrative burden, followed by NPWP registration and account-data issues, system errors, and EFIN recovery problems. These dimensions suggest that the most visible service frictions occurred along an access-identity-transaction chain: users first faced entry barriers, then encountered account or identity inconsistencies, and finally reported difficulties with payment, recovery, or system stability. Research on digital tax services emphasizes that system quality is closely connected to user satisfaction and compliance-related intention (Saptono et al., 2023). The M-Pajak findings therefore show why mobile tax applications must be evaluated not only as information tools but also as transaction-intensive public service infrastructures.

The lexical evidence also aligns with app-review mining literature, which treats recurring review terms as signals of concrete user problems. Maalej et al. (2016) show that app reviews can be automatically classified to identify issue types, while Dąbrowski et al. (2022) argue that app-review analysis can inform software-engineering and maintenance priorities. In the present study, recurring expressions such as login difficulty, verification failure, NPWP-related problems, EFIN requests, payment burden, and repeated errors indicate areas where service improvement may be most visible to users. These findings should not be interpreted as direct technical diagnoses, but they provide empirically grounded signals for prioritizing investigation by the service provider.

The context-dependent interpretation of negative bigrams is also important. Terms such as 'bayar pajak' or 'live chat' are not intrinsically negative; they become analytically negative because they occur in reviews classified as negative by the validated IndoRoBERTa model. This supports the methodological decision to interpret lexical patterns within sentiment-labelled review contexts. Sentiment-analysis scholarship cautions that polarity depends on semantic usage and context rather than on isolated lexical items alone (Nandwani & Verma, 2021). Consequently, the negative dimension analysis should be understood as a contextual synthesis of recurring complaint expressions, not as a dictionary-based interpretation of individual words.

4.7.6 Theoretical, Methodological, and Practical Implications

Theoretically, this study contributes to mobile government and digital tax administration research by positioning mobile tax applications as experience-intensive public service interfaces. Much of the existing discussion on e-government value emphasizes service delivery, citizen orientation, and public value (Twizeyimana & Andersson, 2019). The M-Pajak findings add that public value in mobile tax services is affected not only by availability but also by the reliability of access, identity verification, account linkage, and transaction support. In this sense, the study extends mobile public service evaluation from adoption-oriented questions to experience-oriented evidence derived from user-generated reviews.

Methodologically, the study demonstrates the value of integrating multiple forms of review-based evidence. Rating analysis captures explicit scoring, transformer-based sentiment classification captures the semantic orientation of review text, rating-sentiment alignment tests the relationship between these two signals, and lexical pattern grouping translates recurring expressions into interpretable experience dimensions. This multi-layer approach responds to calls in app-review research for methods that move beyond raw review counts or isolated sentiment scores (Genc-Nayebi & Abran, 2017). It also aligns with research showing that app reviews can support issue identification, maintenance decisions, and user-experience analysis when analyzed systematically (Nakamura et al., 2022).

Practically, the findings can be translated into a staged service-improvement agenda. Authentication and OTP reliability should be addressed first because access failure blocks all downstream functions. Login recovery, NPWP and EFIN integration, payment-error handling, in-app error communication, and support responsiveness should then be treated as connected parts of the same user journey rather than isolated features. Table 21 converts the review-based signals into operational actions and monitoring indicators. These actions should be validated through internal logs, incident records, usability testing, and service-process analysis before implementation.

App-store reviews can also be incorporated into a continuous post-release feedback process. The Directorate General of Taxes could establish pre-update and post-update monitoring windows, track the share and frequency of recurring complaint bigrams, review sentiment and rating composition after major releases, and connect review patterns with operational incident data. Reviews should be treated as an early-warning and issue-discovery channel rather than as a representative satisfaction survey. Combining review signals with completion rates, authentication logs, support tickets, and user testing would provide a more reliable basis for prioritizing corrective action.

Table 23. Operational Service-Improvement Priorities for M-Pajak

Service priority	Operational actions	Monitoring indicators
OTP delivery reliability	Audit delivery gateways; provide timed resend and alternative verification routes; display delivery-status feedback.	OTP delivery success rate; median delivery time; resend failure rate; related review frequency.
Login and account recovery	Simplify recovery steps; unify password/sandi terminology; provide guided recovery and clear lockout instructions.	Recovery completion rate; failed-login rate; median recovery time; repeat-login complaint share.
NPWP and EFIN workflows	Validate data synchronization; explain mismatch causes; integrate NPWP registration and EFIN request/recovery guidance.	Data-mismatch rate; workflow completion rate; EFIN resolution time; NPWP/EFIN complaint frequency.
Payment-error handling	Preserve transaction state; provide clear failure codes; explain retry, reversal, and escalation procedures.	Payment success rate; unresolved-error rate; duplicate-attempt rate; payment-related complaint share.
In-app error communication	Replace generic error messages with task-specific explanations, next steps, and traceable incident codes.	Share of errors with actionable messages; repeat error rate; support contacts per error code.
Live chat and support responsiveness	Set service-level targets; route tax/account issues to	First-response time; resolution time; first-contact resolution; reopened-case

Service priority	Operational actions	Monitoring indicators
	appropriate specialists; provide case tracking.	rate; support-related sentiment.
Post-update review monitoring	Compare 7-day and 30-day pre/post-update review patterns; link review spikes with release and incident logs.	Change in low-rating share; negative-sentiment share; top complaint bigrams; incident-to-review time alignment.

Source: Authors' recommendations based on the review-based findings (2026).

5. Conclusion and Recommendations

This study evaluated expressed user-perceived experience in Indonesia's M-Pajak mobile tax application through a human-validated IndoRoBERTa analysis of Google Play reviews. The study combined temporal rating analysis, transformer-based sentiment classification, rating–sentiment alignment, and auditable lexical-dimension coding. The final corpus contained 6,980 valid reviews retrieved under the stated MOST_RELEVANT configuration on 24 May 2026. The findings therefore characterize this retrievable review corpus and should not be generalized automatically to all M-Pajak users or all reviews ever posted.

The findings show that the use of human validation strengthened the methodological credibility of the sentiment classification process. The high agreement between human annotators and the strong alignment between the adjudicated human labels and IndoRoBERTa outputs support the use of the model-generated labels for subsequent analysis. At the same time, the study treats these labels as analytical classifications rather than definitive ground truth, particularly because neutral and context-dependent reviews remain more difficult to classify than clearly positive or negative expressions.

Substantively, the retrieved corpus was dominated by low ratings and negative sentiment labels. Later-year observations also contained larger negative shares than earlier observations. These findings indicate a shift in expressed review-based evaluation within the corpus, not definitive evidence that the application became worse or that population-level satisfaction declined. Rating–sentiment alignment further showed that star ratings and textual sentiment are closely related but not interchangeable, confirming the value of analyzing both structured scores and unstructured feedback.

The lexical analysis identifies two contrasting patterns of user experience. Positive reviews were mainly associated with ease of use, accessibility, helpfulness, service convenience, support, and perceived security. These findings suggest that users valued M-Pajak when it reduced administrative effort and supported tax-related tasks effectively. In contrast, negative reviews were concentrated around authentication, OTP, and login barriers, followed by tax payment and administrative burden, NPWP registration and account data issues, system errors, and EFIN recovery problems. These patterns indicate that access reliability, identity verification, account integration, and transaction stability are central to user-perceived experience in mobile tax service delivery.

The study contributes to digital government and digital tax administration research in three main ways. First, it positions M-Pajak as a citizen-facing mobile public service interface whose performance can be evaluated through review-based evidence. Second, it demonstrates a methodological approach that integrates validated Indonesian-language transformer classification with rating analysis, sentiment alignment, and lexical experience dimension

grouping. Third, it provides practical evidence for service improvement by identifying the application functions and service touchpoints most frequently associated with user frustration or appreciation.

For service improvement, the Directorate General of Taxes and M-Pajak developers should apply the operational priorities summarized in Table 21: improve OTP delivery reliability, simplify login recovery, integrate NPWP and EFIN workflows, strengthen payment-error handling, provide actionable in-app error messages, improve live-chat responsiveness, and monitor review patterns after each major update. Review-derived priorities should be verified against operational data and user testing before system changes are finalized.

6. Limitations and Future Research

Several limitations should be considered. First, Google Play reviews are self-selected expressions and do not represent the full population of M-Pajak users. Second, the corpus was retrieved using Sort.MOST_RELEVANT, whose opaque ranking may prioritize particular reviews and alter the balance of years, ratings, or issue types compared with Sort.NEWEST. The analysis therefore describes the retrievable relevance-ranked corpus rather than a complete chronological census. Third, the 2026 observations cover only the period available up to the 24 May 2026 scraping date and should not be treated as a complete year. Fourth, update events, temporary service disruptions, complaint behavior, and changes in the reviewing population may contribute to temporal differences; the study does not identify their causal effects.

Fifth, full-corpus sentiment labels were generated by a validated IndoRoBERTa model and remain analytical classifications rather than definitive ground truth, particularly for neutral or mixed reviews. Sixth, the human-validation protocol used a naturally imbalanced, year-stratified sample; performance could differ in a sentiment-balanced sample or in rare ambiguous cases. Seventh, the lexical analysis focused on frequent bigrams and one-to-one consensus grouping. Although Tables 20a and 20b improve transparency, the approach may miss sarcasm, longer narratives, cross-sentence context, and less frequent but consequential issues. Finally, the study is descriptive and exploratory and does not infer causal relationships between application updates, tax policy, system incidents, and review patterns.

Future research should replicate the scraping with Sort.NEWEST and compare yearly rating, sentiment, and lexical distributions with the MOST_RELEVANT corpus as a sensitivity analysis. Studies could also link review timestamps to application release notes, system-incident logs, and institutional responses; apply aspect-based sentiment analysis or change-point detection; and combine app reviews with representative taxpayer surveys, interviews, and usability tests. Such triangulation would help distinguish persistent service-design problems from temporary incidents, platform-ranking effects, and self-selection in review posting.

Overall, M-Pajak review evidence is most informative when ratings, validated sentiment labels, and transparent lexical coding are interpreted jointly and within clearly stated corpus boundaries. The study identifies accessible, helpful, and convenient service conditions in positive reviews and recurrent authentication, OTP, login, NPWP, EFIN, payment, and system-error barriers in negative reviews. These patterns provide a structured basis for investigation and service prioritization, while the cautious interpretation prevents review-stream changes from being overstated as definitive population-level satisfaction decline.

Declarations

Ethics Statement

This study used publicly available Google Play reviews and analyzed the data at an aggregate level. User-identifying information, including user names and profile images, was excluded from analytical interpretation and was not reported in the manuscript. No attempt was made to identify individual users. Because the study used publicly available, non-interventional online review data, individual informed consent was not required.

Data and Code Availability Statement

The analysis code and a de-identified processed dataset will be made available through the corresponding author's GitHub repository upon publication, subject to Google Play platform restrictions and privacy considerations. User-identifying fields, including user names and profile images, will not be included in the shared dataset.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare no conflict of interest.

Declaration of AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used ChatGPT and Jenni AI for language refinement, drafting assistance, and editing support. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the manuscript.

References

Al-Natour, S., & Turetken, O. (2020). A comparative assessment of sentiment analysis and star ratings for consumer reviews. *International Journal of Information Management*, *54*, 102132. https://doi.org/10.1016/j.ijinfomgt.2020.102132

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Budianto, T. A. C., Dermawan, B. A., & Jajuli, M. (2025). Analisis Sentimen Ulasan Aplikasi M-Pajak pada Google Play Store Menggunakan XGBoost. *Jurnal Sistem Informasi Dan Aplikasi*, *3*(1), 11–25. https://doi.org/10.52958/jsia.v3i1.12440

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Dąbrowski, J., Letier, E., Perini, A., & Susi, A. (2022). Analysing app reviews for software engineering: a systematic literature review. *Empirical Software Engineering*, *27*(2), 43. https://doi.org/10.1007/s10664-021-10065-7

Dechamps, S., Simonofski, A., & Burnay, C. (2025). Citizen-centricity in digital government: A theoretical and empirical typology. *Government Information Quarterly*, *42*(1), 102005. https://doi.org/10.1016/j.giq.2024.102005

Duwairi, R., & El-Orfali, M. (2014). A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *Journal of Information Science*, *40*(4). https://doi.org/10.1177/0165551514534143

Genc-Nayebi, N., & Abran, A. (2017). A systematic literature review: Opinion mining studies from mobile app store user reviews. *Journal of Systems and Software*, *125*, 207–219. https://doi.org/10.1016/j.jss.2016.11.027

Gliniecka, Martyna. (2023). The Ethics of Publicly Available Data Research: A Situated Ethics Framework for Reddit. *Social Media + Society*, *9*(3), 20563051231192020. https://doi.org/10.1177/20563051231192021

Google Play, G. P. (2026). *M-Pajak [Mobile application listing]*. https://play.google.com/store/apps/details?id=id.go.pajak.djp&hl=en-US

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, *12*(1), 5979. https://doi.org/10.1038/s41598-022-09954-8

Hunter, R. F., Gough, A., O'Kane, N., McKeown, G., Fitzpatrick, A., Walker, T., McKinley, M., Lee, M., & Kee, F. (2018). Ethical Issues in Social Media Research for Public Health. *American Journal of Public Health*, *108*(3), 343–348. https://doi.org/10.2105/AJPH.2017.304249

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 757–770). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.66

Leem, B.-H., & Eum, S.-W. (2021). Using text mining to measure mobile banking service quality. *Industrial Management & Data Systems*, *121*(5), 993–1007. https://doi.org/10.1108/IMDS-09-2020-0545

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://doi.org/10.48550/arXiv.1907.11692

Maalej, W., Kurtanović, Z., Nabil, H., & Stanik, C. (2016). On the automatic classification of app reviews. *Requirements Engineering*, *21*(3), 311–331. https://doi.org/10.1007/s00766-016-0251-9

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning--based Text Classification: A Comprehensive Review. *ACM Comput. Surv.*, *54*(3). https://doi.org/10.1145/3439726

Nakamura, W. T., de Oliveira, E. C., de Oliveira, E. H. T., Redmiles, D., & Conte, T. (2022). What factors affect the UX in mobile apps? A systematic mapping study on the analysis of app store reviews. *Journal of Systems and Software*, *193*, 111462. https://doi.org/10.1016/j.jss.2022.111462

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1), 81. https://doi.org/10.1007/s13278-021-00776-6

Putra, A., & Fathurrahman, R. (2022). Improving the Quality of the Mobile Tax Service Apps in Indonesia: A Delphi Study. *Budapest International Research and Critics Institute-Journal (BIRCI-Journal)*, *5*(2), 8319–8330. https://doi.org/10.33258/birci.v5i2.4614

Ramzy, M., & Ibrahim, B. (2024). User satisfaction with Arabic COVID-19 apps: Sentiment analysis of users' reviews using machine learning techniques. *Information Processing & Management*, *61*(3), 103644. https://doi.org/10.1016/j.ipm.2024.103644

Samuel, Gabrielle, & Buchanan, Elizabeth. (2020). Guest Editorial: Ethical Issues in Social Media Research. *Journal of Empirical Research on Human Research Ethics*, *15*(1–2), 3–11. https://doi.org/10.1177/1556264619901215

Saptono, P. B., Hodžić, S., Khozen, I., Mahmud, G., Pratiwi, I., Purwanto, D., Aditama, M. A., Haq, N., & Khodijah, S. (2023). Quality of E-Tax System and Tax Compliance Intention: The Mediating Role of User Satisfaction. *Informatics*, *10*(1). https://doi.org/10.3390/informatics10010022

Setiyani, L., Indahsari, A. N., Monica, S., Poerwati, A. E., Ezrafel, A., & Rustam, R. (2022). Analysis of M-Tax Mobile Application Adoption on Tax Compliance in Indonesia Using Diffusion of Innovation Theory (DIT). *Central European Management Journal*, *30*(4), 1202–1212. https://doi.org/10.57030/23364890.cemj.30.4.121

Sharma, S. K., Al-Badi, A., Rana, N. P., & Al-Azizi, L. (2018). Mobile applications in government services (mG-App) from user's perspectives: A predictive modelling approach. *Government Information Quarterly*, *35*(4), 557–568. https://doi.org/10.1016/j.giq.2018.07.002

Sidiq, D., Raharjo, T., & Trisnawaty, N. W. (2024). Towards Tax Administration 3.0: Bracing the Challenges in Mobile Application Development. *Jurnal Informatika Ekonomi Bisnis*, *6*(2), 410–417. https://doi.org/10.37034/infeb.v6i2.915

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, *110*, 298–310. https://doi.org/10.1016/j.eswa.2018.06.022

Twizeyimana, J. D., & Andersson, A. (2019). The public value of E-Government – A literature review. *Government Information Quarterly*, *36*(2), 167–178. https://doi.org/10.1016/j.giq.2019.01.001

Verkijika, S. F., & Neneh, B. N. (2021). Standing up for or against: A text-mining study on the recommendation of mobile payment apps. *Journal of Retailing and Consumer Services*, *63*, 102743. https://doi.org/10.1016/j.jretconser.2021.102743

Wang, C., & Teo, T. S. H. (2020). Online service quality and perceived value in mobile government success: An empirical study of mobile police in China. *International Journal of Information Management*, *52*, 102076. https://doi.org/10.1016/j.ijinfomgt.2020.102076

Wicaksono, P. T., Tjen, C., & Indriani, V. (2021). Improving the tax e-filing system in Indonesia: An exploration of individual taxpayers' opinions. *Jurnal Akuntansi Dan Auditing Indonesia*, *25*(2). https://doi.org/10.20885/jaai.vol25.iss2.art4

Wongso, W. (2023). *indonesian-roberta-base-sentiment-classifier (Revision e402e46)*. Hugging Face. https://doi.org/10.57967/hf/0644

Yi, J., & Oh, Y. K. (2022). The informational value of multi-attribute online consumer reviews: A text mining approach. *Journal of Retailing and Consumer Services*, *65*, 102519. https://doi.org/10.1016/j.jretconser.2021.102519

Zhang, Y., Fu, J., Lai, J., Deng, S., Guo, Z., Zhong, C., Tang, J., Cao, W., & Wu, Y. (2024). Reporting of Ethical Considerations in Qualitative Research Utilizing Social Media Data on Public Health Care: Scoping Review. *J Med Internet Res*, *26*, e51496. https://doi.org/10.2196/51496