

From user reviews to actionable insights: Machine learning prediction and thematic analysis of mobile learning satisfaction

Majdi Abdellatif^{a*}

^aArab Open University, Faculty of Computer Studies, Department of Computer Science, Riyadh, Saudi Arabia

CHRONICLE

Received August 25, 2025
Received in revised format
December 15, 2025
Accepted December 22 2025
Available online
December 28 2025

Keywords:

User satisfaction
Thematic analysis
App reviews
Machine-learning
Mobile apps
Duolingo

ABSTRACT

The widespread adoption of mobile applications has led to a substantial increase in user feedback. Systematically extracting actionable insights from large volumes of unstructured user feedback presents a significant analytical challenge, particularly in mobile learning (m-learning) environments. This study addresses this gap through a mixed-method analysis of 36,625 Duolingo application reviews by integrating machine learning and thematic analysis. The primary objectives are to determine the predictability of user satisfaction from review text and to identify specific thematic factors that underpin user sentiment. Multiple machine learning classifiers were trained using the Bag-of-Words (BoW) model and transformer-based document embeddings to predict the binary satisfaction sentiment label. The results demonstrated that user satisfaction was highly predictable, with logistic regression and neural network models achieving accuracy exceeding 90% (AUC up to 0.968, MCC up to 0.818). Thematic analysis revealed a clear bifurcation in sentiment; core pedagogical features (e.g., internationalization, personalization, and usability) were the primary drivers of satisfaction, whereas significant dissatisfaction stemmed from technical instability (e.g., bugs and performance issues) and monetization strategies (e.g., pricing and ad intrusiveness). This mixed-method provides an interpretable satisfaction-prediction framework and delivers evidence-based recommendations for app developers aiming to maintain competitiveness through continuous user-centered enhancements.

© 2026 by the authors; licensee Growing Science, Canada.

1. Introduction

The limitations of traditional, classroom-based learning in fostering essential 21st-century skills, such as personalized learning and critical thinking, have prompted a transition toward more flexible educational models. Consequently, m-learning has emerged as a dominant paradigm in education and training. This shift has been driven by the widespread use of smartphones and further accelerated by the COVID-19 pandemic (Qazi et al., 2024). This expansion is reflected in the global mobile learning (m-learning) market, which is projected to reach \$77.4 billion by 2025, and in reports indicating that 67% of U.S. companies have adopted mobile training for employees.

With the growth of m-learning applications and competitive market landscape, user satisfaction has become a critical performance metric that directly influences the success and sustainability of m-learning platforms. Therefore, understanding the specific factors driving user satisfaction is essential for both pedagogical advancement and platform development (Grljević et al., 2025). Studies indicate that high satisfaction correlates with platform loyalty, retention, and positive word-of-mouth promotion, all of which contribute significantly to organizational growth (Singh & Suri, 2022). User satisfaction is tangibly expressed in app store ratings and reviews, which directly affect application visibility, download rates, and revenue (Sandy et al., 2025). Therefore, app store reviews represent a valuable, high-volume, and real-time source of user feedback. Each review provides both qualitative text and quantitative rating, offering a dual lens through which user sentiments can be assessed. An aggregated analysis of these data can reveal common pain points and drivers of satisfaction, providing actionable insights for developers and researchers, particularly in response to software updates, feature releases, and policy changes.

* Corresponding author

E-mail address: m.mohammed@arabou.edu.sa (M. Abdellatif)

Although app reviews offer a rich dataset, their vast volume and unstructured nature pose significant analytical challenges that necessitate robust computational methods for evaluation (Singh & Suri, 2022). Recent studies have employed machine learning (ML) algorithms to address this issue. For example, (Sandy et al., 2025) utilized two ML algorithms to analyze 110,668 Duolingo reviews, achieving a 90% accuracy rate, and identified content-related themes (74.2%) as the primary drivers of user satisfaction. However, the broadness of this dominant category risks obscuring specific, actionable user pain points, limiting guidance for targeted platform enhancement. Similarly, (Kong et al., 2024) found a moderate correlation between review sentiments and star ratings (Pearson's $r = 0.43$). However, the model's low adjusted R-squared value (0.391) indicates that over 60% of the variance in user ratings remained unexplained by the topics identified.

This study addresses these gaps by employing a dual approach that combines ML classification with an in-depth thematic analysis to evaluate the drivers of user satisfaction in the Duolingo mobile application. The methodological novelty of this study lies in the complementary integration of machine learning and thematic analysis, rather than in their individual applications. Machine learning provides scalable, data-driven sentiment classification and quantitative validation of user satisfaction, while thematic analysis offers qualitative context and interpretable insights into the underlying satisfaction drivers. This hybrid approach establishes an interpretable satisfaction-prediction framework that yields actionable insights beyond what either method alone can achieve. Specifically, ML addresses what users feel or express, while thematic analysis explains why they feel that way.

As a globally recognized m-learning application, Duolingo has over 500 million active users and relies heavily on artificial intelligence (AI), making it an ideal case study. Beyond its scale, Duolingo's pedagogical engine is fundamentally AI-driven. Its proprietary Birdbrain model estimates both (i) the difficulty of each exercise and (ii) a learner's current proficiency, and updates these estimates after every interaction (Bicknell et al., 2023). Birdbrain enables real-time personalization at massive scale (≈ 1 billion exercises completed per day) by selecting items that keep learners in an optimal challenge zone. When a learner incorrectly answers an exercise, the system simultaneously decreases the learner's estimated ability for that skill domain while increasing the exercise's difficulty rating. More recently, Duolingo integrated OpenAI's GPT-4 to power conversational tools (Roleplay and Explain My Answer) that provide contextual dialogue practice and fine-grained grammatical feedback (Castro et al., 2023; Team, 2023). Taken together, these systems position Duolingo as a scalable, AI-driven tutor that aspires to approximate the individualized benefits highlighted by Bloom's 2-sigma problem (BLOOM, 1984).

Identifying the factors that contribute to user satisfaction on such prominent platforms provides valuable insights for educators, developers, and researchers in the m-learning sector. To achieve these goals, this research pursued the following objectives:

1. Systematic collection, preprocessing, and annotation of a large-scale corpus of user reviews for Duolingo from Google Play and iOS app stores.
2. We evaluated and compared the performance of multiple ML classifiers in predicting user satisfaction using both Bag-of-Words (BoW) and transformer-based text vectorization methods.
3. A rigorous data-driven thematic analysis was conducted to identify key qualitative themes underlying user satisfaction and dissatisfaction with Duolingo.

The structure of this study is as follows: Section 2 provides a literature review; Section 3 describes the data and methodology; and Section 4 presents and discusses the results in relation to the existing literature. Finally, Section 5 concludes the paper and suggests directions for future research.

2. Literature Review

Traditional methods, which primarily include interviews (Jeno et al., 2022), surveys (García De Blanes Sebastián et al., 2025) and questionnaires (Kim & Ong, 2005) exhibit several notable limitations when applied to rapidly evolving mobile learning (m-learning) applications. These methods are often described as "one-off, speed-based, and norm-referenced," offering a static snapshot that fails to capture authentic, longitudinal, and context-rich insights into users' actual interactions with an application over time.

2.1 ML Techniques for Evaluating Satisfaction

Modern ML approaches provide quantitative measures of user satisfaction by analyzing app reviews, social media platforms, and community forums. Kong et al., (2024) analyzed 7,292 reviews from Duolingo users by using sentiment analysis to assess user satisfaction. Their results indicated that sentiment scores were correlated with star ratings (Pearson's $r = 0.43$), confirming that a more positive language in reviews corresponded to higher satisfaction ratings. However, correlation alone does not establish causation, and an adjusted R-squared value of 0.391 suggests that over 60% of variability in user ratings was due to factors not captured or analyzed in their study. Similarly, (Sandy et al., 2025) analyzed over 100,000 Duolingo reviews from Google Play, using logistic regression and Naïve Bayes classifiers. These models achieved an accuracy of 90% in classifying

user feedback into four primary themes: content, instruction, performance, and user interface/user experience (UI/UX). The analysis revealed content quality was the primary driver of user satisfaction, accounting for 74.2% of all reviews. Nevertheless, standard ML models, although potentially accurate, often function as "black boxes" and fail to provide interpretable rationales for their predictions. Beyond sentiment analysis, explainable machine learning (XAI) methods are increasingly being employed to predict and identify drivers of satisfaction. Darko et al. (Darko et al., 2024) applied advanced natural language processing (NLP) and six ML algorithms to analyze 17,717 user-generated reviews and social media posts for mental health applications. The authors used explainability tools, such as SHAP and LIME, to clarify model decisions, highlighting the features (topics, sentiments, and keywords) that most influenced user satisfaction. The integration of XAI tools significantly addresses the "black-box" criticisms of ML models, enhancing the interpretability and trustworthiness of their findings.

2.2 Thematic Analysis for Qualitative Satisfaction Insights

While ML provides broad patterns and metrics, thematic analysis deepens understanding by exploring qualitative data, thus elucidating why users experience satisfaction or dissatisfaction. Ahmed *et al.* (Ahmed et al., 2021), conducted a thematic analysis of over 205,000 user reviews of mental health chatbot applications. They identified distinct satisfaction drivers by categorizing reviews according to their star ratings. Positive reviews have highlighted usability and emotional support, including factors such as effective consultation and empathetic, friend-like interactions. Conversely, negative reviews revealed key drivers of dissatisfaction, including technical issues, privacy concerns, and repetitive content.

Similarly, (Pikhart et al., 2024) conducted a questionnaire-based study involving 148 university students across Iraq, Taiwan, and the Czech Republic. Participants used tools such as Duolingo and ChatGPT, and a thematic analysis of their qualitative feedback revealed distinct satisfaction factors. Positive feedback emphasizes ease of use, efficiency in skill development (e.g., vocabulary acquisition and speaking practice), and accessibility. Conversely, dissatisfaction themes included connectivity problems, limited content variety, and repetitive exercises, which negatively affected user engagement and depth of learning. In summary, user satisfaction with Duolingo is shaped by the multifaceted interplay between design, functionality, and learning effectiveness. Table 1 provides a synthesized overview of the key elements that influence user satisfaction with Duolingo, as identified in recent studies.

Table 1
Key Drivers Influencing User Satisfaction in Duolingo

Satisfaction drivers	Description and example feature in context of Duolingo	Synthesis of Key Findings from Literature
Gamification and engagement	Streak count, XP points, badges, leaderboards, and rewards enhancing motivation	Gamification consistently emerges as primary drivers of satisfaction in language apps (Kamsik et al., 2023; Kong et al., 2024; Zhang & Pan, 2024). Studies highlight that features such as points and leaderboards increase user engagement and learning persistence, but can also risk superficial engagement if not combined with meaningful learning design. Kong et al. (Kong et al., 2024) specifically link features like points and leaderboards to enhanced user motivation and engagement.
Usability and UI design	Intuitive interface, ease of navigation, consistent design	An intuitive interface is crucial for user satisfaction. Studies show that poor UI/UX is a frequent source of negative reviews, whereas positive feedback often highlights ease of navigation and clear design (Erdođdu et al., 2024; Y. Qi & Xu, 2024). However, excessive simplification may limit opportunities for deeper learner interaction.
Technical aspects	Performance (Fast lesson loading times) and stability (synchronization of progress across devices).	App performance, including stability and speed, is a fundamental factor of user satisfaction and trust. Frequent crashes, lags, or bugs reduce user confidence, while smooth performance enhances engagement and encourages long-term usage (Erdođdu et al., 2024; Ulfiah et al., 2025).
User Experience	Ease of use, and the core loop of short, manageable lessons combined with instant feedback.	The overall user experience is a holistic measure combining usability, content quality, and performance. Positive experiences are linked to learners' willingness to recommend the app and to sustain engagement (de Araújo & Eddine, 2020; Y. Qi & Xu, 2024).
Personalization and adaptation	Algorithm adjusting question difficulty and mistakes review sessions	Adaptive learning algorithms that tailor content to the user's skill level are critical for maintaining learner satisfaction. Personalized feedback mechanisms are frequently highlighted as a key feature in positive app reviews (Changala et al., 2024; Pearlin & Gandhi, 2024; Saniyah, 2023).
Content quality and variety	Comprehensive lessons, interactive exercises (listening, speaking, matching).	Users express higher satisfaction when content is comprehensive and varied, especially interactive and culturally relevant exercises. Content depth and authenticity are recurrent themes in user reviews (Hawa & Roslani, 2024; Oshadi & Thelijagoda, 2022; Sandy et al., 2025).
Accessibility and Mobility	Offline access and availability on iOS, Android, and web.	The "anytime, anywhere" nature of mobile learning is a core advantage. Features like offline access and cross-platform compatibility significantly enhance the app's utility and user satisfaction (Kamsik et al., 2023; Sakkir & Syamsuddin, 2023). This feature makes the app adaptable to learners' daily routines.
User feedback and iteration	user-initiated bug reports, regular platform updates, and community interaction	Studies stress that Apps that incorporate user feedback into iterative development processes consistently receive more favorable evaluations. (de Araújo & Eddine, 2020; Y. Qi & Xu, 2024).

2.4 AI-Enhanced Personalization in Duolingo: Birdbrain and GPT-4

Duolingo’s AI stack combines Birdbrain, a dual-estimation personalization engine, with GPT-4-based tutoring to approximate individualized instruction at scale (Bicknell et al., 2023; Team, 2023). Operating at a massive scale with over a billion daily user interactions, Birdbrain employs a logistic, Item Response Theory (IRT)-inspired formulation to simultaneously model both exercise difficulty and learner ability after every interaction.

This dual-estimation mechanism allows for a continuous feedback loop where every user response recalibrates the system, ensuring the session generator delivers exercises precisely calibrated to the learner’s current state. Complementing Birdbrain’s adaptive item selection, Duolingo integrates OpenAI’s GPT-4 (Duolingo Max) to extend personalization to conversational practice and pedagogical feedback (Team, 2023). Features like “Explain My Answer” provide context-aware grammatical feedback, while “Roleplay” enables open-ended conversational practice, effectively tightening the feedback loop and simulating human-like dialogue. Together, these AI components transform Duolingo from a gamified content platform into a scalable AI tutor that approximates individualized instruction through intelligent automation.

Around this core, Duolingo employs a four-stage AI integration model, a pipeline that transforms the entire educational content lifecycle: (1) curriculum design, where large language models analyze pedagogical frameworks to structure lesson sequences; (2) content creation, in which AI generates and filters exercise variants for quality and difficulty; (3) interactive exercise generation, which produces multiple versions of an exercise to prevent repetition; and (4) real-time lesson personalization, where the Birdbrain algorithm constructs individualized lessons using exercises at the optimal difficulty for each learner.

3. Data and Methods

This study aimed to identify the factors influencing Duolingo user satisfaction by integrating ML with a thematic analysis. The authors collected, prepared, and annotated user reviews; applied dual feature extraction techniques; evaluated multiple classification algorithms; and conducted a thematic analysis. Each phase is described in detail below:

3.1 Data Collection

User reviews were collected using the Appbot tool¹ on March 1, 2025. Initially, 2,265,543 Google Play and 336,758 iOS reviews from the previous year were identified. Table 2 presents the sentiment distribution statistics. To comply with data ethics policies, data collection was restricted to reviews submitted within the last three months, resulting in 50,875 Google Play and 29,289 iOS reviews. After filtering for English-language reviews only, the final corpus consisted of 19,289 Google Play and 20,475 iOS reviews. Each review included metadata (review ID, geography, application version, rating, date, author, subject, content, sentiment, and language).

Table 2

Sentiment distribution form March 1, 2024 to March 1, 2025

Google Play		iOS Store	Google Play		iOS store
Star	Review	Review	Sentiment	Review	Review
1	1772349	210919	Positive	1849709	213956
2	242885	45174	Neutral	150971	32517
3	89677	26307	Mixed	114204	25030
4	43789	15215	Negative	150659	65255
5	116843	39143			

3.2 Text Preprocessing Pipeline

A systematic NLP pipeline was implemented using the Akkio platform², Python NLTK, and Excel to ensure analytical consistency.

1. Normalization: Removed punctuation, special characters, emojis, numbers, and excess whitespace.
2. Redundancy reduction: Standardized repeated characters (e.g., "loooove" became "love") and colloquial expressions.
3. Linguistic filtering: Excluded stop words and high-frequency terms to isolate meaningful semantic content.
4. Case standardization: Converted all text to lowercase.
5. Deduplication: Eliminated duplicate reviews.
6. Length threshold: Discarded reviews containing fewer than four words.

¹ App review & ratings analysis for mobile teams,” Appbot. Accessed: May 12, 2025. [Online]. Available: <https://appbot.co/>

² <https://app.akkio.com/login>

This preprocessing resulted in 18,100 analyzable Google Play reviews and 18,525 analyzable iOS reviews, thus preserving the original meanings.

3.3 Sentiment Annotation Framework

A hybrid annotation protocol was used to ensure robust sentiment labeling:

- **Baseline labels:** Appbot assigned initial sentiment (positive, neutral, negative, and mixed) using star-rating heuristics and automated analyses, aligning with the established literature (McIlroy et al., 2016; Oyeode et al., 2020).
- **Bias mitigation:** Sentiment labels were cross-checked with Akkio's generative AI tool to enhance reliability, achieving 65% agreement (23,806 reviews). Reviews with discrepancies (12,819 reviews, 35%) were excluded from subsequent analyses. Reviews categorized as neutral or mixed during the initial collection (Table 2) were intentionally excluded from this analysis as they did not align with the study's objective of predicting and analyzing satisfaction and dissatisfaction drivers. This conservative approach prioritized the creation of a high-confidence, reliably labeled dataset to minimize noise and label bias, ensuring a more robust foundation for model training and evaluation.

3.4 Data Vectorization

Text was transformed into numerical representations using:

- **Bag-of-words (BOW):** Generated term-frequency vectors highlighting frequent word usage, offering a transparent baseline.
- **Transformer-based embeddings** capture the semantic and contextual relationships within dense vector spaces, thereby enabling the detection of nuanced thematic patterns.

This dual-strategy vectorization enabled a comparative evaluation of the feature-space effectiveness.

3.4 Sentiment Classification

Binary sentiment classification was performed using seven ML algorithms: Multinomial Naïve bayes (MNB), Stochastic Gradient Descent (SGD), Logistic Regression (LR), Random Forest (RF), k-Nearest Neighbor (kNN), Support Vector Machine (SVM), and Neural Networks. From the 65% re-annotated dataset (23,806 reviews), 11,656 reviews were allocated for model training. To ensure robust model training and mitigate issues related to data imbalance and noise in user-generated content, two complementary strategies were employed. First, during the preprocessing stage, Cleaning Under-Sampling (CUS) was applied to address both class imbalance and noisy observations in the training data. Unlike simple random under sampling, which discards majority-class samples indiscriminately, CUS strategically removes those samples considered ambiguous (close to decision boundaries) or redundant (overly easy to classify). This pruning process not only balances the class distribution but also improves data quality by filtering mislabeled or uninformative samples. Consequently, classifiers are trained on a cleaner, more representative dataset, which enhances generalization performance (He & Garcia, 2009; Kang et al., 2017; Xiao et al., 2019).

Second, stratified cross-validation was employed to preserve class distribution across all training and validation folds. Unlike random partitioning, which can inadvertently underrepresent minority classes, stratification guarantees that each fold maintains the same class ratio as the full dataset, thereby reducing sampling bias during evaluation. This technique is widely recognized for improving the reliability of performance estimates, particularly in imbalanced datasets (C. Qi et al., 2019).

3.5 Model Configuration and Hyperparameter Tuning

To achieve optimal performance and mitigate overfitting, a systematic hyperparameter optimization (HPO) procedure was applied to the top-performing models. Initially, all candidate models were benchmarked using default hyperparameters to establish baseline performance across the two vectorization techniques. Thereafter, LR and NN, identified as the best performers, were subjected to Bayesian optimization to maximize the F1-score on a validation set. For LR, the hyperparameters tuned included the regularization strength (inverse parameter C), and regularization penalty type (l2 vs l1). For NN, the optimization space included: number of hidden layers (1–3), units per layer (32, 64, 128), dropout rate ranging between 0.2–0.5, and learning rate (0.001 or 0.0001). The final models reported in Tables 3 and 4 reflect these optimized configurations and demonstrate superior generalization compared to their default-parameter counterparts (Shankar et al., 2020; Yang & Shami, 2020).

3.6 Machine-Assisted Thematic Analysis (MATA)

To identify the underlying drivers of satisfaction, a Machine-Assisted Thematic Analysis (MATA) was employed. MATA integrates the computational capacity of unsupervised machine learning, which detects latent patterns within large-scale

textual data, with the researcher’s expertise in interpreting and contextualizing the identified themes. This integration ensures that the resulting themes are both computationally robust and qualitatively meaningful. The initial phase of topic identification employed the Akkio AI platform to perform unsupervised clustering. This platform provides a no-code AI solution that democratizes data analysis and predictive modeling for business users by identifying topics based on word frequency and contextual relationships. The computational procedure generated 70 preliminary clusters, each representing a potential topic within the dataset. These machine-generated clusters served as the foundational coding framework for subsequent qualitative analysis. While this approach enabled efficient thematic discovery, the no-code nature of the platform-imposed limitations on the level of control over clustering algorithms and model parameters. Consequently, the analysis emphasized interpreting machine-generated clusters rather than fine-tuning the underlying process. This reflects a methodological trade-off between the rapid, high-level insights offered by automated tools and the deeper, customizable analyses afforded by traditional research methods, a consideration critical for reproducibility and scholarly rigor in applied machine learning research.

The machine-generated clusters were not accepted at face value. To ensure analytical rigor, Braun and Clarke’s six-phase Reflexive Thematic Analysis (RTA) framework (Byrne, 2022) was applied. This involved iterative procedures of familiarization (manual review of sample reviews within each cluster), reviewing and refining themes (merging semantically overlapping clusters and splitting heterogeneous ones), and generating and naming themes (assigning concise, descriptive labels that captured the central organizing concept).

4. Results & Discussion

The results indicate that automated text classification can accurately predict user satisfaction based on review content, whereas qualitative thematic analysis provides contextual insights into why certain reviews are positive or negative. This section organizes the findings into three parts: first, an analysis of the geographic distribution of user reviews to understand regional usage patterns; second, a detailed evaluation of predictive models developed to classify user satisfaction; and third, thematic analysis results identifying the features and issues that most influence user satisfaction.

4.1 Geographic Distribution of User Review

Granular geographical insights were derived from the iOS App Store, as this source provided country-specific review data, a feature absent from the Google Play dataset. The distribution revealed a significant concentration of reviews from English-speaking regions, justifying the focus of this study on English-language feedback. This approach aligns with prior studies analyzing user feedback within specific international contexts, such as the Czech Republic, Taiwan, Iraq (Pikhart et al., 2024), Turkey (Solmaz, 2025), and China (Yao, 2024). As shown in Fig. 1, the United States contributed 11,207 reviews (52.3% of the iOS dataset), over four times the volume of the second-ranking country, the United Kingdom (2,681 reviews). The top 20 countries accounted for 89.1% of all reviews, highlighting Duolingo’s strong presence in North America and Western Europe. This pattern is partially consistent with prior research (Shortt et al., 2023), which also identified the United States as having the highest representation of user reviews, likely reflecting the app’s origins and the establishment of a user base in its home market. Recognizing this predominantly Western user demographic is important when interpreting thematic analysis results, because the identified satisfaction drivers may reflect regional contexts.

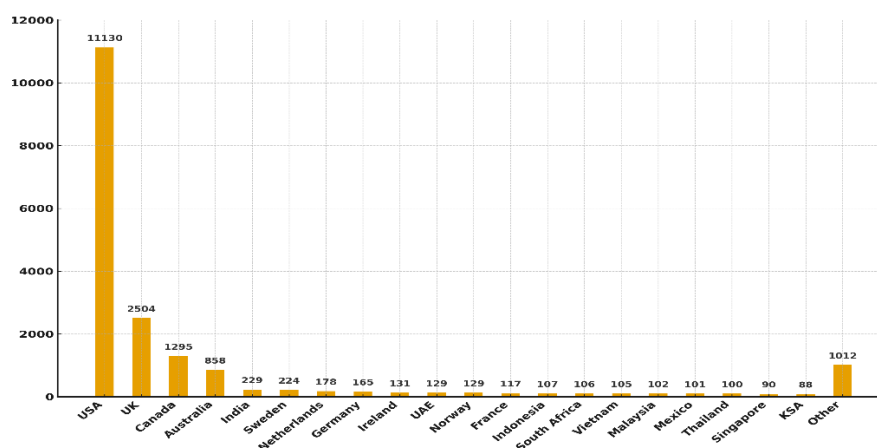


Fig. 1. The top 20 countries with the highest user review volumes derived from the iOS App Store

4.2 Prediction Model Performance

Seven distinct classification models were evaluated using two text vectorization techniques: the traditional BoW model and modern document embeddings. The results summarized in Table 3 and Table 4 demonstrate that user satisfaction is highly predictable from the review content. The performance of these algorithms was evaluated using

six standard metrics: Area Under the Curve (AUC), Classification Accuracy (CA), F1-Score (F1), Precision (Prec), Recall, and Matthews Correlation Coefficient (MCC). Several models achieved strong performance, with LR and neural networks attaining accuracies (CA) above 90%, correctly classifying approximately nine out of ten reviews. For instance, the LR model achieved an AUC of 0.968 and an MCC of 0.812, indicating excellent discrimination and reliability. This substantially exceeds the baseline expectations and aligns with previous studies (Sandy et al., 2025), which similarly found LR to be a top performer (AUC \approx 0.812). Comparative research further indicates that LR provides a robust performance comparable to that of neural networks across sentiment classification tasks. For instance, prior studies reported nearly identical accuracies for LR and recurrent neural networks (0.99 vs. 0.98) in Amazon reviews (Ashbaugh & Zhang, 2024), highlighting the efficacy of linear models for text classification tasks with informative features.

A comparative analysis of the two vectorization techniques provides additional insights. Using the BoW representation, linear models (LR and SGD) demonstrated exceptional performance, suggesting that simple word-frequency features contain strong sentiment signals. Conversely, distance-based algorithms (kNN and SVM) performed poorly with BoW (e.g., SVM's AUC was only 0.606), a result anticipated owing to their reliance on the Euclidean distance in high-dimensional, sparse term vectors. Switching to dense document embeddings markedly improved the performance of these models (kNN's AUC improved from 0.799 to 0.931; SVM's from 0.606 to 0.862), because embeddings better captured the semantic context. Notably, the neural network emerged as the top performer with embeddings (AUC 0.968, MCC 0.818), capitalizing on richer contextual information. LR model consistently achieved high performance (AUC \approx 0.96), underscoring the predictive power of high-impact keywords (e.g., "bug", "good", "love") for effective linear classification. This result provides a foundation for the satisfaction-prediction framework.

Table 3

Model Performance using Bag-of-Words (BOW)

Model	AUC	CA	F1	Prec	Recall	MCC
LR	0.968	0.906	0.906	0.906	0.906	0.812
RF	0.937	0.881	0.881	0.881	0.881	0.762
MNB	0.958	0.844	0.842	0.865	0.844	0.709
NN	0.948	0.885	0.885	0.886	0.885	0.771
kNN	0.799	0.719	0.716	0.732	0.719	0.451
SGD	0.919	0.919	0.919	0.920	0.919	0.840
SVM	0.606	0.528	0.403	0.678	0.528	0.142

Table 4

Models Performance using Document Embeddings

Model	AUC	CA	F1	Prec	Recall	MCC
LR	0.966	0.907	0.907	0.907	0.907	0.815
RF	0.922	0.845	0.845	0.845	0.845	0.690
MNB	0.910	0.832	0.832	0.832	0.832	0.663
NN	0.968	0.909	0.909	0.909	0.909	0.818
kNN	0.931	0.870	0.870	0.872	0.870	0.742
SGD	0.884	0.884	0.884	0.884	0.884	0.768
SVM	0.862	0.777	0.777	0.779	0.777	0.556

4.3 Thematic Analysis of user satisfaction

The hybrid thematic analysis consolidated the initial 70 machine-generated clusters into a coherent set of themes (Table 5). This process translated high-volume computational groupings into human-interpretable patterns that represent the core aspects of user feedback on mobile learning satisfaction and provides an interpretable satisfaction-prediction. The frequency is defined as the aggregate number of times a specific theme was identified across all user reviews. The sentiment percentage is calculated as the ratio of either positive or negative mentions to the combined total of positive and negative mentions for that theme or topic. The analysis revealed a bifurcated user experience: core educational functionalities (internationalization and personalization) generated overwhelmingly positive sentiments, whereas significant dissatisfaction emerged regarding technical quality and monetization strategies. Each thematic cluster is detailed below and contextualized with relevant literature.

1) Global adoption

This cluster addresses Duolingo's core strategic strengths, and highlights its alignment with international markets.

Internationalization (N=7,966): This theme captures reviews of Duolingo's language offerings, cultural appropriateness, and overall content quality. The overwhelmingly positive

Table 5

Theme and the corresponding frequency in user reviews

Cluster	Theme	Negative	Neutral	Positive	Frequency	sentiment
Global adoption	Internationalization	1189	145	6632	7966	84.80% ● ↑
	personalization	485	70	2445	3000	83.70% ● ↑
	Social & community	41	0	51	93	54.84% ● ↑
Usability	UX Design	843	46	1259	2148	59.90% ● ↑
	Gamification	383	17	645	1045	62.74% ● ↑
Content Quality and Multimedia	Curriculum depth	197	20	1235	1452	86.24% ● ↑
	Content repetition	527	26	578	1131	52.31% ● ↑
	Feature requests	466	59	670	1195	58.98% ● ↑
	Audio	200	21	193	414	50.89% ● ↓
	Video	52	6	90	148	63.38% ● ↑
	Image	41	3	43	87	51.19% ● ↑
Stability and Reliability	Bugs	2068	81	595	2744	77.66% ● ↓
	Performance	1348	38	851	2237	61.30% ● ↓
	Updates	352	25	307	684	53.41% ● ↓
	Devices	97	1	26	124	78.86% ● ↓
	Connectivity	75	2	58	135	56.39% ● ↓
Monetization Strategy	Pricing & payment	1439	68	736	2243	66.16% ● ↓
	Advertising	775	43	480	1298	61.75% ● ↓
Trust and Control	Customer Support	681	29	856	1566	55.69% ● ↑
	Notifications	343	13	124	480	73.45% ● ↓
	Security & security	298	7	129	434	69.7% ● ↓
	Login & Sign Up	82	0	20	102	80.39% ● ↓

sentiment (84%) underscored that internationalization was the cornerstone of Duolingo's mission to democratize education. For example, native English speakers can select more than 40 courses, whereas native Arabic speakers have access to only nine courses. This disparity is expected given that English is the most commonly studied languages in 122 countries, reflecting the high global demand. Given Duolingo's \approx 40 million monthly active users and more than 500 million registered users (Portnoff et al., 2021), this sentiment further validates the application's global reach and cultural adaptability supported by an extensive and engaged learner base. These findings are consistent with prior works (Hawa & Roslaini, 2024; Kamsik et al., 2023) that demonstrated Duolingo's effectiveness across diverse cultural contexts. From a software-engineering standpoint, the results indicated the successful implementation of both user-interface localization (L10N) and deep content-level internationalization (I18N).

Personalization (N=3,000): This theme describes how well Duolingo tailored instruction to the needs and goals individual learners. Reviews in this category spanned use cases ranging from primary school pupils (Solmaz, 2025) to university-level test-takers (Isaacs et al., 2023; Yao, 2024), frequently requesting adjustable difficulty levels or personalized course recommendations. The high positive sentiment (83.7%) supports research showing that adaptive features benefit various learner profiles (Portnoff et al., 2021). A recent systematic review of mobile applications reported that adaptive systems are generally more effective. Interestingly, this strong approval contrasts with scholarly critiques that the app lacks depth for advanced learners (Kong et al., 2024; Solmaz, 2025; Ulfiah et al., 2025). Nevertheless, for most users, the existing level of personalization and content breadth appear to be adequate.

Social and community (93): This theme encompasses reviews of social features such as progress-sharing and friend tracking. Despite the small sample, sentiment was split (54.8% positive; 45.2% negative). The negative subset highlights persistent localization challenges (Isaacs et al., 2023; Shortt et al., 2023; Yao, 2024), whereas the positive subset confirms that Duolingo's gamified social elements (e.g., leaderboards) promote community-driven engagement (Ashilah et al., 2025; Shortt et al., 2023; Yao, 2024).

2) Usability:

This cluster examines how interface design and accessibility shape overall user experience and learning effectiveness.

UX Design (N=2,148): Reviews of this theme addresses navigation, aesthetics, and intuitiveness. Positive sentiment (59.9%) reinforces literature praising Duolingo's visual appeal (Soad et al., 2016). Empirical studies have further validated these

strengths, reporting high System Usability Scale scores (90 / 100) (Kong et al., 2024), and demonstrated a clear correlation between interface quality and user satisfaction (Mohtar et al., 2023). Nevertheless, consistent with other usability investigations (Erdoğan et al., 2024; Soyupak & İpek, 2024), a subset of users perceived the interface as lacking innovation or freshness, thereby indicating the need for continuous refinement and periodic UI updates to sustain engagement.

Gamification (N=1,045): This theme encompasses reviews that discuss gamified elements such as points, streaks, and leaderboards. A majority of these reviews (62.7%) were positive, reinforcing the view that gamification serves as a key driver of motivation and engagement in m-learning applications (Kong et al., 2024; Solmaz, 2025), (Bitrián et al., 2021; Mohtar et al., 2023; Shortt et al., 2023). Conversely, the remaining negative feedback aligns with prior research (Kong et al., 2024; Pearlin & Gandhi, 2024), indicating that gamified mechanics, while initially motivating, may become distracting and ultimately diminish engagement, thus raising concerns about their contribution to deeper, meaningful language acquisition (Rehman & Iqbal, 2024). These findings imply that, although gamification is beneficial, its implementation must be carefully balanced to avoid undermining long-term educational outcomes.

3) Content Quality and Variety

This cluster summarizes users' perspectives on learning outcomes, content breadth, and media elements.

Curriculum depth (N=1,452): This theme captures users' perceptions of the learning curve, lesson difficulty, and content depth. An overwhelmingly positive sentiment (86.2%) emerged, indicating that Duolingo maintains learners within an optimal challenge zone by offering content that is demanding yet manageable (Shortt et al., 2023). This result corroborates the efficacy of Duolingo's adaptive-learning engine, Birdbrain (Bicknell et al., 2023), which tailors instruction to individual proficiency levels. This result corroborates the efficacy of Duolingo's adaptive-learning engine, Birdbrain (Bicknell et al., 2023), which uses an IRT-inspired logistic model to align item difficulty with a learner's evolving proficiency. Although several scholars have criticized the platform for its limited depth or variety at advanced stages (Kong et al., 2024; Pikhart et al., 2024; Rehman & Iqbal, 2024; Shortt et al., 2023; Solmaz, 2025, 2025; Ulfiah et al., 2025), this study suggests that such limitations do not significantly affect the overall satisfaction.

Content repetition (N=1,131): Comments on this theme highlight lessons and exercise repetitions. Sentiment was evenly split (52.3% positive, 47.7% negative). Positive reviewers viewed repetition as beneficial (Azhima & Halim, 2024), whereas others reported boredom or reduced critical thinking (Erdoğan et al., 2024; Solmaz, 2025).

Feature Requests (N=1,195): This theme captures requests for new functions and enhancements. Approximately 59% of these reviews expressed desire for more in-depth grammatical explanations (Solmaz, 2025), improved cultural contextualization of lessons (Ashilah et al., 2025), and diverse technical refinements (Kong et al., 2024). Such demands align with the suggestions of previous studies (Kong et al., 2024; Soad et al., 2016). From a software engineering perspective, such feedback is a valuable resource for requirements engineering, enabling developers to prioritize features that most benefit the user base.

Multimedia (audio (N=414), video (N=148) and images (N=87)): This composite theme addresses user evaluations of audio, video, and image elements, which are essential for effective pedagogy (Kong et al., 2024; Soad et al., 2016; Solmaz, 2025). Sentiment was mixed, with a weighted positivity of 46.9%. Video content received favorable ratings (63.4%), images were perceived neutrally (51.2%), and the audio component exhibited a slightly negative skew (50.9%). This outcome was unexpected because audio functionality is fundamental to language learning, and prior literature generally praises Duolingo's audio quality (Bicknell et al., 2023; Oyebo et al., 2020; Solmaz, 2025). From a software-engineering perspective, shortcomings in the core audio medium undermine the pedagogical value of supplementary media such as videos and images. Addressing these deficiencies will require rigorous cross-device testing and enhanced quality-assurance procedures to ensure consistent audio performance across platforms.

4) Stability and Reliability

This cluster addresses critical non-functional requirements that affect user satisfaction and sustained engagement with the Duolingo application.

Bugs (N=2,744): Reviews of this theme report functional errors, application crashes, glitches, and other disruptive behaviors. These issues were among the most frequently reported and carried a pronounced negative sentiment (77.7% dissatisfaction), underscoring application reliability as a significant pain point affecting user satisfaction. This finding empirically confirms prior observations by Kong et al. (Kong et al., 2024), who identified "bugs, crashes, and app freezes" as primary technical problems that negatively influence user experiences. From a software engineering perspective, such widespread bug reports indicate potential weaknesses in Duolingo's regression testing and software release process. Failure to effectively manage this technical debt poses a substantial risk to sustained user satisfaction, as persistent technical issues constitute a major source of user frustration (Kong et al., 2024).

Performance (N=2,237): This theme includes user complaints regarding the application speed, responsiveness, loading times, and resource consumption. A substantial proportion (61.3%) of these reviews expressed negative sentiments, frequently mentioning slow load times and excessive battery or data usage. Such performance concerns echo the findings from previous studies on mobile learning applications (Erdođdu et al., 2024). From a software performance engineering perspective, these issues pose critical problems owing to Duolingo's reliance on frequent, and short learning sessions. A high latency or lag directly interrupts the learning flow, thereby diminishing the gamified engagement of the application. Performance problems are particularly severe for users with lower-end devices or unstable network conditions (Ulfiah et al., 2025), highlighting the importance of optimization for consistent user experience across diverse contexts.

Updates (N=684): This theme encompasses the user reviews of the application updates and feature modifications. A slightly negative sentiment (53.4%) indicated possible issues in Duolingo's change management and deployment processes. This finding aligns with prior research (de Araújo & Eddine, 2020; Y. Qi & Xu, 2024), suggesting that updates, although intended to resolve technical issues, may inadvertently introduce new bugs or unpopular alterations. These findings underscore that delicate balance developers must strike between rapid deployment and stability assurance during each release cycle.

Connectivity (N=135): This theme includes reviews concerning Internet connectivity requirements, offline functionality, and progress synchronization. Although relatively few in number, these reviews predominantly expressed negative sentiments (56.4%), emphasizing the necessity for robust offline functionality, as previously highlighted (de Araújo & Eddine, 2020; Kamsik et al., 2023; Sakkir & Syamsuddin, 2023). For a mobile application explicitly designed for on-the-go learning, the absence of reliable offline functionality constitutes a significant usability issue, undermining Duolingo's core value proposition of anytime, anywhere practice.

Devices (N=124): Reviews within this theme addresses compatibility issues associated with specific devices, operating system versions, or device-specific errors. A notably high negative sentiment (78.9%) confirmed findings from prior research (Erdođdu et al., 2024; Y. Qi & Xu, 2024) indicating that technical issues often vary by platform or device. Although Duolingo's multi-platform availability (iOS, Android, and Web) is advantageous, the findings highlight the critical need for comprehensive testing strategies to ensure consistent user experience across fragmented hardware ecosystems.

5) Monetization Strategy

This cluster addresses the pronounced user dissatisfaction related to Duolingo's monetization approach, a prevalent challenge in freemium mobile applications.

Pricing and payment (N=2,243): This theme combines user opinions regarding premium subscription costs, the perceived value of paid features, and the transaction process, including billing, refunds, and subscription management. A significant majority (66%) expressed dissatisfaction, indicating that users frequently perceived the premium version's benefits as insufficient to justify its cost or found the payment process further eroded their trust. This dissatisfaction highlights the fact that user concerns extend beyond pricing alone, encompassing the design and clarity of transaction procedures. These results align with those of prior studies that addressed the cost sensitivity in freemium applications (Ashilah et al., 2025).

Advertising (N=1,298): This theme includes reviews concerning advertisement frequency, intrusiveness, and relevance in Duolingo's free version. A substantial proportion of users (61.7%) expressed negative sentiments, suggesting that current advertising practices negatively impacted their learning experience. For educational applications that require focused attention, intrusive advertisements can disrupt concentration, undermine educational objectives, and potentially transform a key revenue source into a driver of user attrition. However, the existing literature presents mixed perspectives; while some studies (Foulds et al., 2021) support this finding, others argue that advertisements alone are insufficient to motivate users to upgrade to premium versions (Ashilah et al., 2025).

6) Trust and Control

This cluster pertains to user experience dimensions related to customer support, data privacy and security, and user autonomy in managing app interactions.

Customer Support (N=1,566): This theme captures user experiences with Duolingo's customer support channels and services, particularly their responsiveness and effectiveness in issue resolution. Sentiments were mixed, with approximately 55.7% positive and 44.3% negative, highlighting the challenges of providing scalable, and effective customer support to a large user base. Prior research affirms that accessible support mechanisms significantly contribute to user satisfaction and continuous improvement (Pearlin & Gandhi, 2024; Qi & Xu, 2024). However, the current findings suggest that numerous users still perceive Duolingo's support as insufficient to address their inquiries adequately, indicating potential gaps in the quality of customer service.

Notifications (N=480): This theme includes user reviews addressing Duolingo's reminders and alerts (e.g., daily practice prompts and streak notifications). A considerable majority (73.4%) expressed negative sentiments, suggesting that notifications intended to enhance user re-engagement frequently produced the opposite effect. Users often described these notifications as intrusive or irritating rather than motivational. This finding is somewhat unexpected, as gamification elements such as streak reminders typically aim to foster user engagement; however, in this case, they appear to induce frustration. These results highlight the importance of implementing personalized user-configurable notification settings that respects individual autonomy and preferences.

Privacy and Security (N=434): This theme covers user concerns regarding Duolingo's data collection practices, account permissions, and user information security. Although mentioned with moderate frequency, a significantly negative sentiment (69.7%) indicates potential reputational risks. This finding aligns with prior research (Soad et al., 2016) that identified system-level security weaknesses and raised questions about responsible handling of extensive learner data collected by Duolingo's "Birdbrain" AI engine (Bicknell et al., 2023). These results suggest that Duolingo's communication regarding data practices might insufficiently reassure some users, emphasizing the need for enhanced transparency and user-centric control to foster and sustain trust.

Login and Sign-Up (N=102): This theme addresses user issues encountered during account creation, login, or password recovery processes. An exceptionally high negative sentiment (80.4%) is particularly concerning given the fundamental nature of these functionalities. Such dissatisfaction presents a significant barrier to initial user acquisition and re-engagement. From the learners' perspective, friction at this entry stage disrupts the initiation of the learning journey, causing immediate dissatisfaction and increasing the likelihood of app abandonment.

Neutral Sentiment: Although neutral sentiment appears less frequently, its interpretive value should not be underestimated. These comments typically offer conditional feedback (e.g., "works fine, but..."), revealing how effectively the application meets diverse user needs across varying contexts. From a software engineering perspective, neutral reviews can provide valuable insights into the application quality and maturity. For instance, the 81 neutral comments related to "Bugs" may reflect awareness of non-critical issues that do not impair core functionality but could be prioritized in the development backlog to refine the user experience. Similarly, the 145 neutral observations on internationalization likely reflect requests for additional language support or regional adaptations rather than criticism of existing offerings. In essence, neutral reviews often point to incremental improvements that developers can prioritize to elevate user experiences from "okay" to "great."

5. Conclusion And Future Work

This study combined ML methods with thematic analysis to create a robust and interpretable model for evaluating m-learning satisfaction, as validated through user reviews of Duolingo. The findings demonstrated that, while quantitative models can accurately predict user sentiment, qualitative methods are essential for uncovering the contextual factors underlying these predictions. The primary contributions of this study are twofold. Theoretically, it provides a reproducible framework connecting natural-language user feedback with measurable satisfaction indicators. Practically, the study offers developers, researchers, and educators a proactive diagnostic tool to identify latent dissatisfaction drivers in m-learning apps before they negatively impact user engagement. For researchers, the interpretable satisfaction-prediction framework enables a systematic way to identify and track sources of user dissatisfaction in real-time. For educators and learning designers, the findings highlight that the success of an m-learning platform is not solely determined by its pedagogical features but also by its technical stability and monetization model. This approach enables data-driven, cost-effective iterations that can help maintain learner retention and enhance the educational impact. The practical implications for Duolingo and similar platform developers are evident from the findings. The thematic analysis (Table 5) reveals a critical disconnect between user satisfaction and the platform's non-functional requirements and monetization strategies. Integrating these findings with platform architecture clarifies why high satisfaction is observed around curriculum depth and personalization. Birdbrain's continual parameter updates and lesson selection plausibly sustain optimal difficulty, while GPT-4 features (Roleplay; Explain My Answer) provide contextual, responsive feedback that mimics human tutoring in micro-interactions. This synergy strengthens the interpretation that Duolingo's AI stack contributes materially to engagement and progression, rather than acting as superficial gamification alone.

Despite the robustness of the findings, this study is not without limitations. The exclusive focus on Duolingo constrains the immediate generalizability of the findings to other mobile learning applications, particularly those with distinct user demographics, pedagogical approaches, or interaction paradigms. However, this limitation should be contextualized within the broader methodological contributions of this research. The platform-specific findings such as user dissatisfaction with Duolingo's monetization strategies, gamification elements, or technical implementation issues are inherently bound to this particular application. These contextual insights, while valuable for understanding Duolingo's user experience, cannot be directly extrapolated to other m-learning platforms without empirical validation. Different applications may exhibit sources of user friction, user expectations, and satisfaction drivers that are not captured in the current analysis. Nevertheless, the core methodological framework presented in this study demonstrates significant potential for cross-platform applicability. The

interpretable satisfaction-prediction framework represents a generalizable approach for extracting actionable insights from large-scale unstructured user feedback. This methodology can be adapted to analyze user reviews from diverse m-learning, Future research should validate these methodological contributions by applying the framework to multiple m-learning platforms simultaneously, enabling comparative analysis of satisfaction factors across different application types, target demographics, and learning objectives. Such multi-platform studies would strengthen the external validity of the findings and provide more robust insights into universal versus platform-specific factors influencing mobile learning satisfaction. Additionally, although large, the dataset predominantly consisted of reviews from English-speaking users, potentially restricting the applicability of the identified themes to Duolingo's broader global audience. To address the limitations of this study, future work should incorporate more diverse datasets and conduct longitudinal analyses to track how user satisfaction evolves, particularly in response to major feature updates and policy changes. To complement this approach, comparative studies across various language-learning platforms are required to distinguish platform-specific issues from universal challenges in the educational technology domain. Ultimately, the most critical next step is to investigate the direct link between user satisfaction and learning outcomes. By integrating these findings with educational data mining, researchers can determine whether positive and negative user feedback measurably impacts engagement and learning success.

Acknowledgment

The author thanks the Arab Open University for supporting this work

References

- Ahmed, A., Aziz, S., Khalifa, M., Shah, U., Hassan, A., Abd-Alrazaq, A., & Househ, M. (2021). *Thematic Analysis on User Reviews for Depression and Anxiety Chatbot A pps: Machine Learning Approach (Preprint)*. Crossref. <https://doi.org/10.2196/preprints.27654>
- Ashbaugh, L., & Zhang, Y. (2024). A Comparative Study of Sentiment Analysis on Customer Reviews Using Machine Learning and Deep Learning. *Computers*, 13(12), 340. <https://doi.org/doi.org/10.20944/preprints202411.0741.v1>
- Ashilah, F. D., Efendi, N. H., Havara, Y. F., Handayani, P. W., & Harahap, N. C. (2025). An Analysis of Factors Impacting Users' Choice of Freemium or Premium Services in a Mobile-Assisted Language Learning App. *Electronic Journal of E-Learning*, 23(1), Article 1. <https://doi.org/10.34190/ejel.23.1.3894>
- Azhima, F., & Halim, A. (2024). The Integration of Duolingo in Classroom Setting: A Case Study of its Impact on English Language Learning. *Riwayat: Educational Journal of History and Humanities*, 7(3), 877–890. <https://doi.org/doi.org/10.24815/jr.v7i3.39326>
- Bicknell, K., Brust, C., & Settles, B. (2023). How Duolingo's AI Learns what you Need to Learn: The language-learning app tries to emulate a great human tutor. *IEEE Spectrum*, 60(3), 28–33. <https://doi.org/10.1109/MSPEC.2023.10061631>
- Bitrián, P., Buil, I., & Catalán, S. (2021). Enhancing user engagement: The role of gamification in mobile apps. *Journal of Business Research*, 132, 170–185. <https://doi.org/doi.org/10.1016/j.jbusres.2021.04.028>
- BLOOM, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, 13(6), 4–16. <https://doi.org/10.3102/0013189X013006004>
- Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, 56(3), 1391–1412. <https://doi.org/10.1007/s11135-021-01182-y>
- Castro, E., Saurabh, S., & Catherine, M. (2023). *How Artificial Intelligence Can Personalize Education* [Online post]. IEEE Spectrum. <https://spectrum.ieee.org/how-ai-can-personalizeeducation> (2023)
- Changala, R., Borde, A., Subhashini, R., Pathak, P., Rao, V. S., & Bala, B. K. (2024). Sentiment Analysis in Mobile Language Learning Apps Utilizing LSTM-GRU for Enhanced User Engagement and Personalized Feedback. *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 1–7. <https://doi.org/DOI:%252010.1109/ICEEICT61591.2024.10718406>
- Darko, A. P., Antwi, C. O., Adjei, K., Zhang, B., & Ren, J. (2024). Predicting determinants influencing user satisfaction with mental health app: An explainable machine learning approach based on unstructured data. *Expert Systems with Applications*, 249, 123647. <https://doi.org/doi.org/10.1016/j.eswa.2024.123647>
- de Araújo, P. A. M., & Eddine, E. A. C. (2020). *The reviews of users of the Duolingo application: Usability and objectivity in the learning process*. 8(09). <https://doi.org/DOI:%2520https://doi.org/10.29121/granthaalayah.v8.i9.2020.1326>
- Erdoğan, F., Kırdar, K., Eski, F., & Okumuş, E. (2024). Usability Analysis of Mobile Applications Used for Foreign Language Learning. *International Journal of Pioneering Technology and Engineering*, 3(02), 47–52. <https://doi.org/doi.org/10.56158/jpte.2024.90.3.02>
- Foulds, O., Azzopardi, L., & Halvey, M. (2021). Investigating the Influence of Ads on User Search Performance, Behaviour, and Experience during Information Seeking. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 107–117. <https://doi.org/10.1145/3406522.3446024>
- García De Blanes Sebastián, M., Sarmiento Guede, J. R., Azuara Grande, A., & Filipe, A. F. (2025). UTAUT-2 predictors and satisfaction: Implications for mobile-learning adoption among university students. *Education and Information Technologies*, 30(3), 3201–3237. <https://doi.org/10.1007/s10639-024-12927-1>
- Grljević, O., Marić, M., & Božić, R. (2025). Exploring Mobile Application User Experience Through Topic Modeling. *Sustainability (2071-1050)*, 17(3). <https://doi.org/10.3390/su17031109>

- Hawa, Z. L., & Roslaini, R. (2024). STUDENTS'PERCEPTION TOWARD USING DUOLINGO FOR LEARNING ENGLISH. *KLASIKAL: JOURNAL OF EDUCATION, LANGUAGE TEACHING AND SCIENCE*, 6(3), 604–611. <https://doi.org/doi.org/10.52208/klasikal.v6i3.1213>
- He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Isaacs, T., Hu, R., Trenkic, D., & Varga, J. (2023). Examining the predictive validity of the Duolingo English Test: Evidence from a major UK university. *Language Testing*, 40(3), 748–770. <https://doi.org/10.1177/02655322231158550>
- Jeno, L. M., Egelanddal, K., & Grytnes, J.-A. (2022). A qualitative investigation of psychological need-satisfying experiences of a mobile learning application: A Self-Determination Theory approach. *Computers and Education Open*, 3, 100108. <https://doi.org/10.1016/j.caeo.2022.100108>
- Kamsik, A. E., Daud, A., & Masyhur, M. (2023). Students' perception on the use of the Duolingo application as a medium for developing university-level English language skills. *Journal of English Language Teaching and Learning (Jette)*, 5(1), 1–19. <https://doi.org/doi.org/10.18860/jette.v5i2.23667>
- Kang, Q., Chen, X., Li, S., & Zhou, M. (2017). A Noise-Filtered Under-Sampling Scheme for Imbalanced Classification. *IEEE Transactions on Cybernetics*, 47(12), 4263–4274. <https://doi.org/10.1109/TCYB.2016.2606104>
- Kim, G.-M., & Ong, S. M. (2005). An exploratory study of factors influencing m-learning success. *Journal of Computer Information Systems*, 46(1), 92–97. <https://doi.org/doi.org/10.1080/08874417.2005.11645872>
- Kong, L., Koh, D. H., & Antonenko, P. (2024). Evaluating User Experience in E-Learning Platforms: An NLP-Enhanced Analysis of Duolingo Reviews. *International Journal of Human-Computer Interaction*, 1–14. <https://doi.org/10.1080/10447318.2024.2427361>
- McIlroy, S., Ali, N., Khalid, H., & E. Hassan, A. (2016). Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. *Empirical Software Engineering*, 21(3), 1067–1106. <https://doi.org/10.1007/s10664-015-9375-7>
- Mohtar, S., Jomhari, N., Omar, N. A., Mustafa, M. B. P., & Yusoff, Z. M. (2023). The usability evaluation on mobile learning apps with gamification for middle-aged women. *Education and Information Technologies*, 28(1), 1189–1210. <https://doi.org/10.1007/s10639-022-11232-z>
- Oshadi, D. M. K., & Thelijjagoda, S. (2022). AppGuider: Feature Comparison System using Neural Network with FastText and Aspect-based Sentiment Analysis on Play Store User Reviews. *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, 1148–1155. <https://doi.org/DOI:%252010.1109/ICOSEC54921.2022.9952093>
- Oyebode, O., Alqahtani, F., & Orji, R. (2020). Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews. *IEEE Access*, 8, 111141–111158. <https://doi.org/10.1109/ACCESS.2020.3002176>
- Pearlin, E., & Gandhi, S. M. G. (2024). Enhancing User Behavior Analysis in Mobile Language Learning Apps Through Gamification and AI Integration: A Transformer-Based Deep Learning Approach. *2024 International Conference on Data Science and Network Security (ICDSNS)*, 1–6. <https://doi.org/DOI:%252010.1109/ICDSNS62112.2024.10690955>
- Pikhart, M., Klimova, B., & Al-Obaydi, L. H. (2024). Exploring university students' preferences and satisfaction in utilizing digital tools for foreign language learning. *Frontiers in Education*, 9, 1412377. <https://doi.org/doi.org/10.3389/feduc.2024.1412377>
- Portnoff, L., Gustafson, E., Rollinson, J., & Bicknell, K. (2021). Methods for Language Learning Assessment at Scale: Duolingo Case Study. *International Educational Data Mining Society*.
- Qazi, A., Qazi, J., Naseer, K., Hasan, N., Hardaker, G., & Bao, D. (2024). M-Learning in education during COVID-19: A systematic review of sentiment, challenges, and opportunities. *Heliyon*, 10(12). <https://doi.org/10.1016/j.heliyon.2024.e32638>
- Qi, C., Diao, J., & Qiu, L. (2019). On Estimating Model in Feature Selection With Cross-Validation. *IEEE Access*, 7, 33454–33463. <https://doi.org/10.1109/ACCESS.2019.2892062>
- Qi, Y., & Xu, R. (2024). Research on user interface design and interaction experience: A case study from “Duolingo” Platform. *EAI Endorsed Transactions on Scalable Information Systems*, 11(5), 61. <https://doi.org/doi.org/10.4108/eetsis.5461>
- Rehman, Z., & Iqbal, A. (2024). Investigating the Challenges of Using Duolingo for Language Learning: A Simplified Review. *Contemporary Journal of Social Science Review*, 2(04), 863–878.
- Sakkir, G., & Syamsuddin, N. A. (2023). Students' perceptions of Duolingo Mobile assisted language learning (MALL) in learning English vocabulary. *EduLine: Journal of Education and Learning Innovation*, 3(3), 381–388. <https://doi.org/doi.org/10.35877/454RI.eduline1970>
- Sandy, T. A., Ghufuron, A., & Muhtadi, A. (2025). Text Classification of Duolingo Reviews on Google Play: Insights for Enhancing M-Learning Applications. *International Journal of Interactive Mobile Technologies*, 19(7). <https://doi.org/10.3991/ijim.v19i07.52891>
- Saniyah, S. M. (2023). Duolingo and Learner Autonomy: Investigating The Role of Personalization and Gamification in Promoting Self-directed Language Learning. *ENJEL: English Journal of Education and Literature*, 2(02), 141–147. <https://doi.org/doi.org/10.30599/enjel.v2i02.529>
- Shankar, K., Zhang, Y., Liu, Y., Wu, L., & Chen, C.-H. (2020). Hyperparameter Tuning Deep Learning for Diabetic Retinopathy Fundus Image Classification. *IEEE Access*, 8, 118164–118173. <https://doi.org/10.1109/ACCESS.2020.3005152>

- Shortt, M., Tilak, S., Kuznetcova, I., Martens, B., & Akinkuolie, B. (2023). Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning*, 36(3), 517–554. <https://doi.org/10.1080/09588221.2021.1933540>
- Singh, Y., & Suri, P. K. (2022). An empirical analysis of mobile learning app usage experience. *Technology in Society*, 68, 101929. <https://doi.org/doi.org/10.1016/j.techsoc.2022.101929>
- Soad, G. W., Duarte Filho, N. F., & Barbosa, E. F. (2016). Quality evaluation of mobile learning applications. *2016 IEEE Frontiers in Education Conference (FIE)*, 1–8. <https://doi.org/10.1109/FIE.2016.7757540>
- Solmaz, O. (2025). Impacts of digital applications on emergent multilinguals' language learning experiences: The case of Duolingo. *Education and Information Technologies*, 30(7), 9185–9214. <https://doi.org/10.1007/s10639-024-13185-x>
- Soyupak, O., & İpek, H. (2024). INVESTIGATION OF THE USABILITY AND USER EXPERIENCE OF MOBILE LANGUAGE LEARNING APPLICATIONS: BUSUU, DUOLINGO, AND MEMRISE. *Turkish Online Journal of Design Art and Communication*, 14(4), 840–855. <https://doi.org/doi.org/10.7456/tojdac.1510008>
- Team, D. (2023, March 14). *Introducing Duolingo Max, a learning experience powered by GPT-4*. Duolingo Blog. <https://blog.duolingo.com/duolingo-max/>
- Ulfiah, U., Rasyadan, M. F. R., Utami, W. T., Sunardi, S., & Murad, D. F. (2025). Impact of usability on continuance usage intention in language learning apps with gamification features. *Bulletin of Electrical Engineering and Informatics*, 14(1), 696–705. <https://doi.org/doi.org/10.11591/eei.v14i1.8023>
- Xiao, Z., Wang, L., & Du, J. Y. (2019). Improving the Performance of Sentiment Classification on Imbalanced Datasets With Transfer Learning. *IEEE Access*, 7, 28281–28290. <https://doi.org/10.1109/ACCESS.2019.2892094>
- Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415, 295–316. <https://doi.org/10.1016/j.neucom.2020.07.061>
- Yao, D. (2024). Does perceived test fairness affect test preparation?—A case study of Duolingo English Test. *Heliyon*, 10(23). <https://doi.org/doi.org/10.1016/j.heliyon.2024.e40579>
- Zhang, Y., & Pan, W. (2024). A UTAUT2 Model Expansion: Investigating the Effect of Interactive Learning Environment and Gamification on Duolingo User Base in China. *International Journal of Human–Computer Interaction*, 1–15. <https://doi.org/10.1080/10447318.2024.2426736>



© 2026 by the authors; licensee Growing Science, Canada. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).