J-INTECH (Journal of Information and Technology)
Accredited Sinta 4 Ministry of Higher Education, Science and Technology Republic of Indonesia SK No. 10/C/C3/DT.05.00/2025
E-ISSN: 2580-720X || P-ISSN: 2303-1425

Quality Analysis of an Interactive Programming Learning Platform Based on ISO/IEC 25010 Using a String-Matching Approach on User Reviews

Mochamad Chandra Saputra[1], Satrio Agung Wicaksono[2], Satrio Hadi Wijoyo[3], Prasetya Naufal Rahmandita[4], Buce Trias Hanggara [5]

[1,2,3,4,5] *University of Brawijaya, Faculty of Computer Science, Departemen of Information System, Jl. Veteran No.10-11, Ketawanggede, Kec. Lowokwaru, Kota Malang, Jawa Timur 65145, Indonesia*

Keywords

E-learning; Efficiency; ISO/IEC 25010; Perceived User Experience; Quality in Use; Satisfaction.

Corresponding Author:
andra@ub.ac.id

Abstract

This study aims to analyze user perceptions of Quality in Use of the Khan Academy e-learning platform, focusing on two key characteristics defined in the ISO/IEC 25010 standard: satisfaction and efficiency. Based on the analysis of user review data, there is a clear difference in volume between the two aspects: 539 reviews (65%) reflected the satisfaction aspect, while only 290 reviews (35%) were related to efficiency. This indicates that users are approximately 1.86 times more likely to comment on satisfaction than on efficiency. For satisfaction, most of reviews were positive (420 reviews, or 77.9%), while 119 reviews (22.1%) expressed negative sentiments. These results suggest that most users are satisfied with their experience using Khan Academy, particularly due to factors such as flexible access time, user convenience, and the wide availability of learning materials. In contrast, the efficiency-related reviews exhibited a more even distribution, with 154 positive reviews (53.1%) and 136 negative reviews (46.9%). This closer balance indicates that while some users appreciate the platform's performance, others report encountering technical issues, including slow access speeds, navigation difficulties, and system instability. Overall, user perception of the Khan Academy e-learning system is generally positive, especially regarding satisfaction. However, the findings also underscore the importance of addressing technical performance challenges to improve efficiency and ensure a seamless learning experience. These insights provide a valuable basis for the development of user-centered e-learning systems and contribute to the evaluation of system quality from the Quality in Use perspective.

1. Introduction

In today's digital era, applications have become an inseparable part of modern human life. Users from diverse backgrounds rely on applications for various purposes, including entertainment, communication, productivity, and education. In this context, maintaining application quality is a crucial aspect to ensure continued relevance and competitiveness in the eyes of users [1]. Application quality is not solely assessed based on technical features; it is also determined by how users perceive its usefulness, ease of use, and the overall satisfaction they derive from using it a concept known as quality-in-use.

Educational applications constitute a critical component of the ongoing digital transformation. As the demand for remote learning continues to rise, e-learning platforms have undergone rapid development and emerged as indispensable tools in contemporary education. E-learning is defined as a technology-mediated learning process that transcends the constraints of time and space [2]. One prominent example of such an application is Khan Academy, which offers a comprehensive array of interactive learning materials, including instructional videos, practice exercises, and articles. The platform is accessible via both mobile and desktop devices; however, feature availability may vary depending on the device used [3].

The adoption of e-learning platforms has exhibited a marked upward trajectory. Statistical data on online education reveal a dramatic increase in course enrollments, particularly in the aftermath of the COVID-19 pandemic, with millions of new users registering annually[4]. Empirical studies have indicated that digital learning environments contribute positively to various educational outcomes, including enhanced academic performance, improved time efficiency, and elevated levels of learner satisfaction [5]. These findings underscore the growing relevance and pedagogical value of educational applications, affirming their potential as effective and scalable instruments for delivering high-quality learning experiences.

In the development of high-quality educational applications, understanding user feedback is a critical factor. One commonly employed approach involves the analysis of user reviews available on application distribution platforms such as Google Play. These reviews provide valuable insights into users' perceptions, experiences, and expectations regarding the application. However, the manual analysis of thousands of review entries is both time-consuming and resource-intensive, rendering it impractical for large-scale evaluation. Consequently, there is a growing need for automated text analysis methods, with text mining emerging as a prominent and effective approach for extracting meaningful patterns from unstructured user-generated content [6].

Text mining enables researchers to extract meaningful information from large collections of unstructured textual data, such as user-generated reviews. One of its core techniques is sentiment analysis, which aims to classify user opinions into predefined sentiment categories typically positive, negative, or neutral based on the semantic content of the text [7]. For instance, a user comment such as "terrible on desktop and not easy to get to on phone" reflects dissatisfaction related to the satisfaction attribute, whereas a statement like "Downloading is so unstable, please make it resumable" indicates a concern associated with the efficiency aspect of the system. Such classification facilitates a more nuanced understanding of user experiences and allows developers to identify specific areas for improvement within the application.

The evaluation of software quality from the user's perspective can be systematically conducted using the ISO/IEC 25010:2011 standard framework, specifically the Quality-in-Use model. This model assesses quality based on five core characteristics: Effectiveness, Efficiency, Satisfaction, Freedom from Risk, and Context Coverage [8]. Although sentiment analysis and keyword extraction have been used in software quality research, there is a lack of studies that systematically map user feedback to ISO/IEC 25010 Quality In Use Characteristics using data-driven methods. Even fewer studies combine string matching algorithms with semantic keyword extraction to automatically infer Quality In Use attributes from natural language reviews in educational platforms. Given its direct emphasis on end-user perceptions and needs, this framework is particularly well-suited for evaluating the user experience of educational applications such as Khan Academy. By focusing on how effectively and efficiently users can achieve their learning goals while maintaining satisfaction, minimizing

risks, and ensuring adaptability across varying contexts the Quality-in-Use model provides a comprehensive foundation for user-centered quality assessment.

Considering the preceding discussion, this study tries to conduct a comprehensive evaluation of the quality-in-use of the Khan Academy educational application by leveraging user-generated reviews from the Google Play Store. The methodological framework employed combines text mining and sentiment analysis techniques, underpinned by the ISO/IEC 25010 quality model. The findings of this research are anticipated to yield strategic insights and evidence-based recommendations for application developers, with the aim of enhancing software quality and fostering sustained user satisfaction over time.

The International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) are two preeminent international bodies responsible for the development and maintenance of global standards. The ISO/IEC standards encompass a broad spectrum of technical domains, including software engineering. Among these, ISO/IEC 25010 serves as a contemporary software quality model, superseding the earlier ISO/IEC 9126 standard [8].

The ISO/IEC 25010 framework comprises two primary components: Product Quality and Quality in Use. This study principally focuses on the Quality in Use dimension, which emphasizes the end-user perspective in assessing software quality based on actual user experience. According to ISO/IEC 25010 (2021), Quality in Use is delineated by five fundamental characteristics, as detailed in Table 1. These characteristics constitute the cornerstone for evaluating software quality from the user's standpoint, as they directly reflect user experience and satisfaction [9].

Table. 1. Quality In Use Characteristics

No	Characteristics	Definition
1	Effectiveness	The degree to which users can accurately and completely achieve specified goals within a defined context of use, as delineated in (ISO 9241-11)
2	Efficiency	The level of resources expended such as time, effort, and cost in relation to the accuracy and completeness with which users attain their objectives (ISO 9241-11)
3	Freedom From Risk	The extent to which a system or product minimizes potential adverse impacts on financial assets, human life, health, or environmental safety.
4	Satisfaction	The measure of fulfillment of user needs and expectations when interacting with a product or system in a particular context of use.
5	Context Coverage	The capability of a product or system to maintain efficiency, freedom from risk, and user satisfaction across both anticipated and unforeseen usage scenarios effectiveness.

Text mining constitutes a systematic process for extracting meaningful information from unstructured textual data by leveraging statistical methods, linguistic analysis, and machine learning techniques. This approach has been extensively applied across various domains, including user review processing, public opinion analysis, and data-driven decision-making [10].

Within the context of software systems, text mining proves instrumental in elucidating user needs and perceptions through the analysis of reviews posted on digital platforms. By transforming vast volumes of textual data into actionable insights, this technique enables researchers and practitioners to better comprehend user feedback and inform subsequent development and improvement efforts [11].

String matching refers to the computational process of identifying the presence of one or more text patterns within a given string or document. Among the various algorithms developed for this purpose, the Jaro-Winkler algorithm is widely recognized for its efficacy in handling typographical errors and enabling approximate string matching with tolerance for minor discrepancies [12].

The Jaro-Winkler algorithm quantifies the similarity between two strings by considering the number of matching characters and the number of transpositions. This characteristic renders the algorithm particularly suitable for analyzing user-generated comments, which often contain spelling variations or inconsistent terminology while conveying semantically similar content [13]. It has found extensive application in domains such as record linkage, data deduplication, and recommendation systems.

In this study, the Jaro-Winkler algorithm is employed to facilitate the alignment of user comments with specific Quality-in-Use attributes, such as efficiency or satisfaction, despite the diversity in linguistic style and orthographic variations. The Jaro distance (dj) between two strings namely, the user comment and the quality characteristic is computed using the following formula:

$$\frac{1}{3}\cdot\left(\frac{|\,w'_1\,|}{|\,w'_1\,|}+\frac{|\,w'_2\,|}{|\,w'_2\,|}+\frac{|\,w'_1\,|-T(w'_1,w'_2)}{|\,w'_1\,|}\right) \quad (1)$$

Where :

$□_1$, $□_2$ = comparing the words

$□'_1$ = a series of characters from w1, found in w2

$□'_2$ = a series of characters from w2, found in w1

Tw'$_1$,w'$_2$ = = □□□□□□□ □□ □□□□□□□□□□□□□□ □h □□□□□□□□□ □'$_1$ □□□ □'$_2$/2

Subsequently, an analysis is conducted on the pair of text strings, denoted as w1 and w2, wherein comprises characters that bear similarity to those present in w2, and vice versa. A character "a" from w1 is considered similar to the corresponding character in w2 if it occurs at the same or nearly the same position within w2. Two characters from w1 and w2 are regarded as a "match" if the positional distance between them does not exceed a specified threshold, as formalized by the following equation:

$$\left(\frac{\max|s_1|,|s_2|}{2}\right)-1 \quad (2)$$

2. Research Methodology

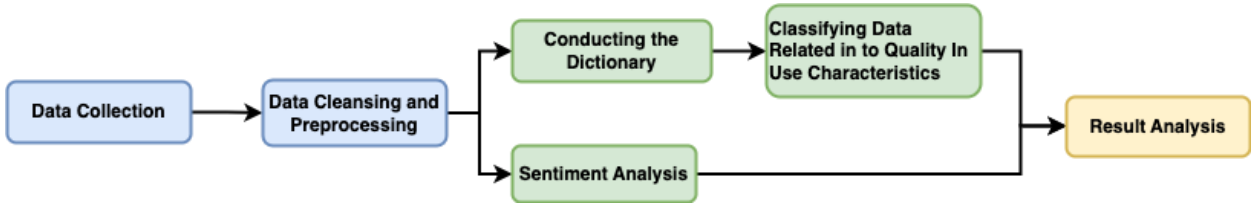

Figure 1. Research Methodology

Data collection for this study was conducted using web scraping techniques, enabling the extraction of user reviews from the application platform. The data were filtered based on criteria including year, month, date,application rating, and reviews written in English, covering a temporal range from January 1, 2021, to December 31, 2022.

The data preprocessing phase involved duplicate elimination and handling of missing entries. Duplicate reviews were identified and removed such that only one instance of each review was retained to avoid redundancy. Similarly, reviews with missing or empty content were excluded from the dataset. Prior to cleaning, user reviews underwent normalization procedures, including conversion to lowercase and removal of punctuation, symbols, and emojis, to ensure consistency and facilitate accurate analysis. Furthermore, a minimum character length threshold of 20 alphabetic characters was applied, based on the rationale that shorter reviews often lack sufficient semantic content for meaningful analysis in this context.

For dictionary construction, an extensive literature review was performed to identify relevant keywords pertaining to the Quality in Use attributes under investigation, specifically satisfaction and efficiency. The resulting lexicon comprised synonyms, formal definitions of the respective quality aspects, and terminologies derived from academic journals and authoritative English dictionaries.

Review classification was performed using a string matching approach whereby extracted keywords from user reviews were compared against the pre-established dictionary. This facilitated the assignment of labels to reviews corresponding to the defined Quality in Use aspects, namely satisfaction and efficiency. The workflow of the string matching process is illustrated in Figure 2.

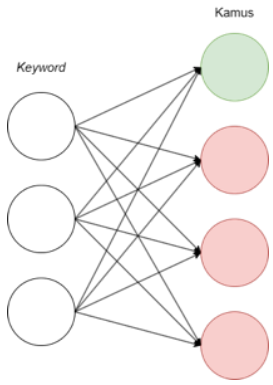

Figure 2. Illustration of String Matching

Related with the schema presented in Figure 2, each keyword identified within user review sentences was systematically matched against a curated lexicon constructed by the researchers. Upon detecting congruent keywords, their occurrences were aggregated to quantify the prevalence of specific Quality in Use attributes namely satisfaction or efficiency within the textual data.

The subsequent analytical procedure involved synthesizing the classification outcomes with sentiment polarity metrics, encompassing positive and negative sentiments, pertinent to each Quality in Use dimension examined in this research. The principal objective of this analysis was to elucidate the relationship between user sentiment distributions and the corresponding Quality in Use attributes, thereby enabling the identification of attributes that elicit the most substantial positive or negative user feedback. The insights derived from this process are intended to inform strategic recommendations aimed at enhancing the Quality in Use dimensions most critically evaluated by users.

3. Result and Discussion

In this study, data collection was conducted using the Google Colab platform, employing web scraping techniques. The researcher utilized the google-play-scraper library to extract user review data from the Khan Academy application available on the Google Play Store. Data retrieval was performed by inputting the official application URL corresponding to the research object. To ensure temporal relevance and clarity of scope, the dataset was restricted to the period between January 1, 2021, and December 31, 2022, resulting in a total of 5,803 user review entries. An exemplar of the collected data is presented in Table 2. The attributes gathered during this process included the review submission date, author name, star rating, and review content.

The selection of the 2021–2022 timeframe was predicated on its representation of a critical phase characterized by the intensive development and utilization of educational applications, particularly in the aftermath of the COVID-19 pandemic, which significantly accelerated the adoption of online learning platforms. During this interval, a substantial number of active users provided feedback concerning their experiences with Khan Academy, addressing both functional performance and user satisfaction dimensions. Consequently, the review data from this period are considered both relevant and representative for evaluating Quality in Use

characteristics in accordance with the ISO/IEC 25010 standard.

Tabel 2. Example of Data Collecting

Author	Rating	Review
Piyush Kumar. 26	5	Awesome amazing thanks for your hard work thankyou
V7I - TheSeventhSector	4	I love this on desktop but on mobile it has a massive problem imo. On desktop it seems that all your lessons are in order in relation to each other but on mobile it's not like this, they don't seem to be in any order at all. If this could be fixed it would be very helpful and make the app much more useful
Jitender Sharma	4	Very nice app
Allan Tyler	5	I love, Love, LOVE this app. I wanted to spend some time on an app that was not just scrolling though social fluff or popping coloured dots, but wanted bite-sized engagement. I wanted to refresh S advance school mathematics skills S knowledge. This offers well illustrated, well-explained, clear-to-follow, short tutorials, demonstrations, examples, practice exercises, review exercises, summaries, and self-assessments. Excellent! I highly recommend this, S I applaud its founder, creators, funders!
Dileep Patel	5	Good
jagat suman	5	✌
Hamza Daali	5	Wonderful app
Rima Rai	1	This is the most horrific, horrible app for learning. Pls believe me . I have downloaded as the stars and the review were very good but😔😔 it was a useless app pls don't waste net in it

An extensive data preprocessing phase was undertaken to guarantee the integrity and analytical viability of the dataset. The initial corpus consisted of 5,803 user review entries, within which redundancies and superfluous elements including punctuation marks, numerical characters, emojis, and special symbols were identified. The preprocessing pipeline commenced with text normalization through conversion of all alphabetic characters to lowercase. Thereafter, extraneous characters were systematically removed using Python's Regular Expression (Regex) libraries.

Deduplication was then performed, resulting in the elimination of 1,059 duplicated records and yielding a corpus of 4,744 unique reviews. Subsequent filtering addressed the removal of empty or semantically insufficient entries that persisted post-deduplication. Specifically, reviews comprising fewer than 20 alphabetic characters were excluded on the basis that their brevity undermined their informational contribution to the analytical process. Illustrative examples of such excluded reviews include succinct phrases like "very good app" or "please fix bug." This criterion led to the removal of an additional 929 entries, culminating in a refined dataset of 3,815 substantive reviews.

Moreover, the preprocessing entailed the exclusion of all numeric and non-alphabetic characters, thereby restricting the textual data to alphabetic characters (a–z) exclusively. This constraint was essential to comply with the methodological requirements of subsequent text mining and natural language processing analyses.

The data cleansing process is imperative to ensure that the dataset utilized for analysis is thoroughly purged of noise, which could otherwise compromise the accuracy of the results. Additionally, this process facilitates more efficient and expedited computational procedures. The outlined preprocessing steps align with contemporary research methodologies in the domain of text processing, which emphasize the critical importance of rigorous data cleaning prior to conducting further analytical operations [14][15][16].

The next phase in text data preprocessing encompasses several critical steps: punctuation removal, stopword elimination, lemmatization, and tokenization. Initially, punctuation removal involves the elimination of

punctuation marks such as periods, commas, exclamation points, question marks, and other symbols that lack significant semantic value within the context of sentiment analysis or text classification. This step aims to simplify the textual data and mitigate noise that could adversely affect analytical outcomes. Following this, stopword removal is performed to discard frequently occurring words that convey minimal informational value, such as "the," "is," "at," "and," among others. In the Indonesian language context, common stopwords include terms like "dan," "yang," and "adalah." The objective of this step is to reduce the dimensionality of the data while preserving only meaningful words pertinent to the analysis.

The next step is lemmatization, which entails transforming each word into its canonical or base form, known as the lemma. For instance, words such as "running," "ran," and "runs" are normalized to "run." Unlike stemming, which often crudely truncates words, lemmatization leverages grammatical rules and linguistic dictionaries to yield more accurate and contextually appropriate results. Finally, tokenization is conducted, which segments the text into discrete units or tokens, typically individual words. These tokens subsequently serve as inputs for downstream analytical processes, including classification or string matching. Collectively, these preprocessing steps facilitate the construction of a cleaner and more structured textual representation, thereby enabling analytical models to more accurately comprehend the context and content of each user review, as illustrated in Table 3 [14][16].

Table 3. Result of data preprocessing

Before	After
get free course transition unlike without time limit absolutely stunning think it d free trial abt week completely free thing need internet connection ur do thank u wonderful	['get', 'free', 'course', 'transition', 'unlike', 'without', 'time', 'limit', 'absolutely', 'stunning', 'think', 'it', 'd', 'free', 'trial', 'abt', 'week', 'completely', 'free', 'thing', 'need', 'internet', 'connection', 'ur', 'do', 'thank', 'u', 'wonderful']
amazing well everything really helpful study lesson interesting well explain free simply amazing	['amazing', 'well', 'everything', 'really', 'helpful', 'study', 'lesson', 'interesting', 'well', 'explain', 'free', 'simply', 'amazing']
one good learn tool student super effective course perfect bug sentence hint button multiple choice appear page go blank reason fix reloading progress remove leave frustrated need fix good overall	['one', 'good', 'learn', 'tool', 'student', 'super', 'effective', 'course', 'perfect', 'bug', 'sentence', 'hint', 'button', 'multiple', 'choice', 'appear', 'page', 'go', 'blank', 'reason', 'fix', 'reloading', 'progress', 'remove', 'leave', 'frustrated', 'need', 'fix', 'good', 'overall']

The development of the dictionary was conducted to enable the systematic classification of Quality in Use (QinU) characteristics, as delineated in the ISO/IEC 25010:2011 standard. This process comprised three principal stages: (1) extraction of terminology from the formal definitions of each QinU attribute, (2) synthesis of insights from prior empirical and theoretical literature on software quality evaluation, and (3) lexical expansion through the inclusion of semantically equivalent terms identified using authoritative language resources, including academic dictionaries and thesauri.

For instance, from the definition of the Satisfaction attribute "the degree to which user needs are satisfied when a product or system is used in a specified context of use" the term satisfied was extracted as a primary lexical indicator. This term, along with related expressions such as happy, pleased, and enjoyed, was systematically included in the lexicon to ensure comprehensive coverage of linguistic variations encountered in user-generated reviews. A similar procedure was applied to the Efficiency attribute, where keywords such as fast, responsive, and smooth were selected to reflect user perceptions related to performance and usability. The resulting dictionary was employed as a core analytical tool to annotate and classify user-generated textual data into predefined Quality in Use categories as shown in Table 4. This dictionary driven annotation supports the automation of sentiment mapping and contributes to enhancing the objectivity and reproducibility of the evaluation process.

Table 4. Dictionary for satisfaction and efficiency

No	Characteristics	Terms
1	*Satisfaction*	*usefulness, easy, simple, improve, enhance, clear, straightforward, smooth, boost, complex, complicated, useless, difficult, frustrate, pleasure, attractive, engaging, appeal, visual, interface, admirable, bored, pleased, pleasant, satisfied, excited, fun, delightful, happy, sad, upset, depress, awful, displeasing, ugly, terrible, confidence, expectation, trusted, navigate, faith, certainty, belief, doubt, misgiving, skepticism, wariness, disappoint, comfort, enjoy, satisfaction, tense, amenity, convenient, relax, relief, cozy, calm, nervous, strain, irritating, painful*
2	*Efficiency*	*resource, goal, lagging, acurate, efficiency, waste, time, reasonable, glitch, fast, cost, accessibility, compatible, overload, availability, background, capacity, slow, expertise, lack, ability, utilize, period, screen, skillfulness, ineptness, target, relevant, hours, load, properly*

The labeling of Quality in Use aspects in this study adheres to the ISO/IEC 25010 standard, with a particular focus on two primary characteristics: efficiency and satisfaction. The labeling process involves keyword extraction utilizing the KeyBERT algorithm applied to each user review. The extraction results in a ranked list of keywords, where higher scores indicate a stronger relevance to the review content.

Following the keyword extraction, a matching procedure is conducted using a pre-constructed lexicon, which was developed based on formal definitions and previous studies related to the efficiency and satisfaction aspects. This matching process identifies the predominant Quality in Use characteristic reflected in each review. The classification is determined based on predefined rules that associate specific keywords with either the efficiency or satisfaction dimension.

1. If the satisfaction score is greater than the efficiency score, the review is classified under the satisfaction aspect.
2. If the efficiency score is greater than the satisfaction score, the review is classified under the efficiency aspect.
3. If both scores are equal, the review is excluded from classification under either aspect.

This approach aims to automatically and objectively categorize user reviews based on the relevance of their content to the targeted Quality in Use characteristics.

Table 5. Example of using the Quality in Use Aspect Rule

No	Review	Keyword	Score	Label
1	Very well designed *study material, easy to grasp masters you to understand concepts smoothly. You can learn according to your pace*	['study', 'easy', 'concepts', 'understand', *'material', 'pace', 'smoothly', 'designed', 'grasp', 'well', 'masters', 'according']*	Score satisfaction: 2 (*easy = easy* => 1.0, *smoothly = smooth* => 0.91) Score efficiency: 0	Satisfaction
2	Good app but I tried many times but I wasn't able to update my info in my account.	['account', 'update', 'info', 'good', 'many' 'tried', 'able', 'times']	Score satisfaction: 0 Score efficiency: 1 (times = time => 0.93)	Efficiency
3	It is very depressing *that I*	['video', 'screen',	Score	Undifine

	have to quit full screen and tap next video When I am in a learning streak it's just mood killer put a next video button on the media player plzz	*'depressing', 'next', 'quit', 'mood', 'media', 'streak', 'button', 'tap', 'learning', 'plzz' 'player']*	satisfaction: 1 (depressing = depress => 0.9) Score efficiency: 1 (screen = screen => 1.0)

The objective of the labeling process based on the Quality in Use aspects defined in the ISO/IEC 25010 standard is to classify user reviews of software applications into software quality categories from the user's perspective, specifically focusing on the efficiency and satisfaction characteristics. This process is carried out by extracting keywords from each review using algorithms such as KeyBERT, followed by matching those keywords against a pre-constructed lexicon derived from formal definitions and relevant literature on Quality in Use. Through this method, each review can be systematically identified in terms of its inclination toward one of the specified quality aspects.

The benefits of this labeling process are substantial, both in practical and academic contexts. Practically, it provides direct insights from real user experiences, enabling application developers to understand and enhance software quality based on actual user evaluations. Furthermore, this approach supports the automation of software quality assessment, which traditionally relied on manual analysis, thereby increasing both efficiency and objectivity. From an academic standpoint, this study contributes to the development of more advanced evaluation models that leverage textual and sentiment-based user data, and it offers a valuable reference for strategic decision-making aimed at improving future digital products.

In the process of labeling user sentiment toward application quality, this study utilizes the Transformers library from Hugging Face, employing the DistilBERT model to automate sentiment classification on user review data. Prior to classification, the textual data underwent lemmatization to enhance the accuracy of the sentiment analysis. The implementation results indicate that, out of the total number of reviews analyzed, 896 reviews were identified as having negative sentiment and 2,919 reviews as having positive sentiment, after excluding entries with undefined sentiment labels.

The sentiment-labeled data were further analyzed based on the Quality in Use aspects efficiency and satisfaction. For the efficiency aspect, 136 reviews were classified as negative and 154 as positive. For the satisfaction aspect, 119 reviews exhibited negative sentiment, while 420 reviews were categorized as positive. These findings illustrate the distribution of user perceptions across the two targeted quality aspects.

In the evaluation phase, the performance of the automated classification model was assessed by comparing its output with expert-labeled data, focusing specifically on the satisfaction and efficiency aspects of Quality in Use. A total of 300 reviews were used for this evaluation, consisting of 150 reviews for each aspect. Reviews that could not be assigned to either of the two aspects were excluded from the evaluation metrics.

In the process of labeling user sentiment toward application quality, this study utilizes the Transformers library from Hugging Face, employing the DistilBERT model to automate sentiment classification on user review data. Prior to classification, the textual data underwent lemmatization to enhance the accuracy of the sentiment analysis. The implementation results indicate that, out of the total number of reviews analyzed, 896 reviews were identified as having negative sentiment and 2,919 reviews as having positive sentiment, after excluding entries with undefined sentiment labels.

The evaluation was conducted using a confusion matrix, as presented in Table 6. The results indicate that, for the actual efficiency class, 98 reviews were correctly predicted as efficiency (True Positive), 5 reviews were incorrectly classified as satisfaction, and 47 reviews could not be classified. In the case of the actual satisfaction class, 112 reviews were correctly predicted, 9 were misclassified as efficiency, and 29 reviews remained

undefined.

Table 6. Confusion Matrix for Quality in Use Characteristics

		Prediction		
		Efficiency	*Satisfaction*	Tidak Terdefinisi
Actual	*Efficiency*	98	5	47
	Satisfaction	9	112	29
	Undefine	0	0	0

For each Quality in Use aspect efficiency and satisfaction confusion matrices were independently computed. Specifically, for the efficiency aspect, the confusion matrix yielded the following values: True Positives (TP) = 98, False Positives (FP) = 9, False Negatives (FN) = 52, and True Negatives (TN) = 141. In the case of the satisfaction aspect, the corresponding values were: TP = 112, FP = 5, FN = 38, and TN = 145.

Subsequently, these values served as the basis for calculating key classification performance metrics, namely precision, recall, F1-score, and accuracy, to rigorously evaluate the effectiveness of the sentiment classification model with respect to each targeted aspect.

Table 7. Confusion matrix Quality in Use Characteristics

	Precision	*Recall*	*F1-score*	*Accuracy*
Satisfaction	96%	75%	84%	70%
Efficiency	92%	65%	76%	

The result of evaluation shown that, with respect to the satisfaction aspect, the classification system attained a precision of 96%, recall of 75%, F1-score of 84%, and an overall accuracy of 70%. In relation to the efficiency aspect, the system achieved a precision of 92%, recall of 65%, and F1-score of 76%. Given the balanced nature of the dataset employed, the reported accuracy of 70% can be considered a reasonably representative indicator of the automated labeling system's effectiveness.

In analyzing user feedback based on the ISO/IEC 25010 Quality in Use model, it is important to recognize that different reviews often reflect different quality dimensions, even when they discuss the same platform. To illustrate this, the study includes representative examples of user reviews that capture the contrast between Satisfaction and Efficiency. For example, some users express positive emotions related to their experience, such as enjoying the interface, feeling motivated by progress, or appreciating the overall design. These types of comments are typically associated with Satisfaction, as they reflect personal enjoyment, comfort, or emotional responses to the system.

On the other hand, other users focus on how smoothly or quickly they can complete tasks mentioning issues like slow loading times, delays in execution, or difficulty navigating. These concerns relate more directly to Efficiency, as they involve the resources (time, effort) required to achieve specific goals. By presenting these examples, the study highlights how users evaluate software quality from different angles. This distinction is important for ensuring that automated analysis methods such as keyword extraction and string matching can accurately classify user feedback according to the correct Quality in Use categories.

The relatively lower recall values reveal that the system exhibits limitations in comprehensively identifying all relevant instances within the respective classes. This underscores the necessity for continued refinement and further evaluation to enhance the model's sensitivity, particularly toward more nuanced or implicit expressions of the targeted Quality in Use aspects within textual reviews. In the domain of text classification, recall is a critical metric, as inadequate detection of pertinent categories can adversely affect the overall analytical quality.

The integration of precision and recall metrics into the F1-score offers a balanced and robust measure of classification performance, which is especially pertinent when evaluating models on datasets with relatively uniform class distributions, as observed in this research.

The evaluation of the automated sentiment classification model's performance was conducted by comparing its outputs with expert-annotated labels. This assessment employed a confusion matrix constructed from a dataset comprising 300 reviews, consisting of 178 positive and 122 negative instances, which had also been utilized in the Quality in Use aspect classification.

The confusion matrix results for sentiment classification reveal that, of the 178 positive reviews, 153 were correctly classified as positive (True Positives), while 25 were erroneously classified as negative (False Negatives). Conversely, among the 122 negative reviews, 99 were accurately identified as negative (True Negatives), and 23 were incorrectly labeled as positive (False Positives).

Key evaluation metrics including precision, recall, F1-score, and accuracy were computed. For the negative class, the precision attained was 80%, recall was 81%, and the F1-score was 80%. The positive class demonstrated a precision of 87%, recall of 86%, and an F1-score of 86%. The overall classification accuracy across both classes was 84%, as detailed in Table 8.

Table 8. Confusion Matrix for User Sentiment

	Precision	Recall	F1-score	Accuracy
Negative	80%	81%	80%	84%
Positive	87%	86%	86%	84%

The findings show that the sentiment labeling system exhibits satisfactory and balanced performance in accurately classifying both positive and negative user reviews. However, to ensure robustness and reliability, further validation is required, particularly with larger datasets and across diverse application domains.

An analysis of user perceptions toward e-learning platforms reveals a marked predominance of the satisfaction dimension relative to efficiency, as reflected by the unequal distribution of labeled data 539 instances associated with satisfaction compared to 290 for efficiency, as shown in Figure 3. This imbalance suggests that users predominantly focus on their affective experiences and overall satisfaction with the e-learning system, rather than on its technical efficiency or system performance. Such insights underscore the importance of prioritizing user satisfaction in the development and evaluation of e-learning solutions.

Users generally perceive significant benefits from the flexibility, ease of access, and convenience offered by online learning, which directly contribute to an enhanced positive perception of the satisfaction dimension [17][18]. Conversely, the comparatively lower volume of data associated with the efficiency aspect suggests the presence of limitations related to access speed, interface intuitiveness, or system stability. These issues may not have been explicitly articulated by users but have the potential to adversely affect their learning experience.

This imbalance may also reflect a tendency for users to provide more immediate affective feedback rather than technical critique, or it may indicate a limited user capacity to specifically identify and articulate efficiency-related problems. Such findings highlight the need for more targeted investigations into system performance factors to complement user satisfaction assessments in the evaluation of e-learning platforms.

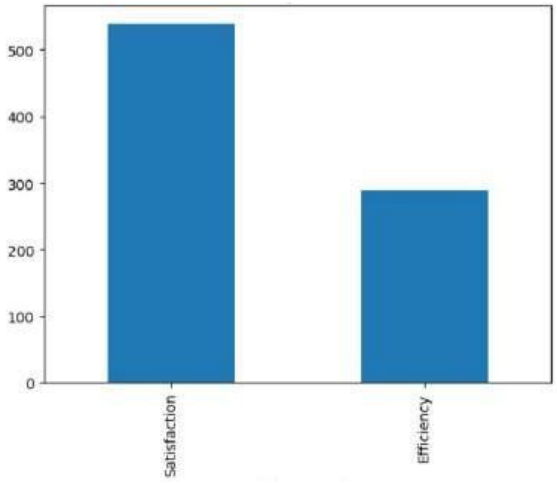

Figure 3. Results of clasifying reviews based on aspects of Quality in Use satisfaction and Efficiency

Consequently, despite the predominantly positive user perceptions regarding satisfaction with e-learning platforms, it is imperative for system developers to equally prioritize the enhancement of technical efficiency. Addressing this dimension is crucial to ensure that the learning experience is not only engaging and satisfactory but also effective, seamless, and devoid of operational hindrances.

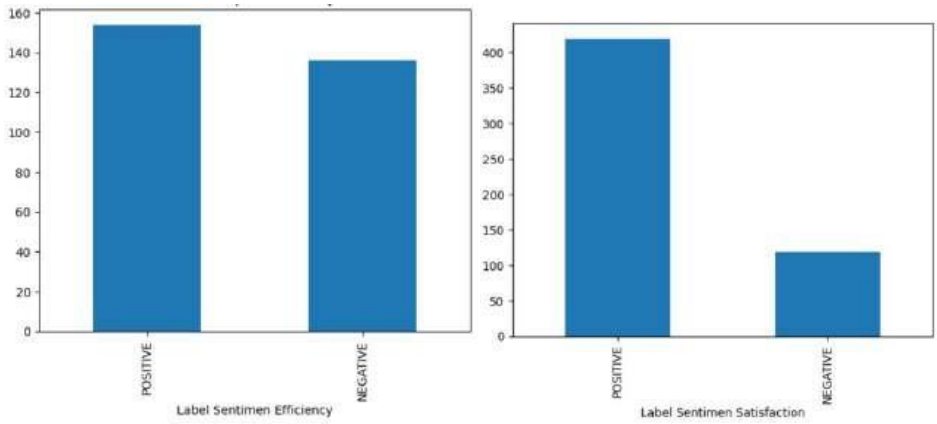

Figure 4. Result of Sentiment based on user reviews for the aspects of Quality in Use satisfaction and Efficiency

The analysis of user perceptions regarding the e-learning platform Khan Academy, as depicted in Figure 4, indicates a predominant inclination toward positive sentiment, especially in the satisfaction dimension. Within the analyzed dataset, 420 reviews were classified as positive and 119 as negative with respect to satisfaction. This distribution reflects that the majority of users express favorable evaluations of their experiences with the platform, highlighting factors such as convenience, temporal flexibility, and comprehensive access to learning resources [19].

The efficiency dimension demonstrates a comparatively balanced sentiment distribution, comprising 154 positive and 136 negative reviews, as illustrated in Figure 4. This suggests that while a substantial portion of users perceive the Khan Academy system as efficient characterized by the prompt and accessible delivery of educational content there remains a notable segment of users who identify limitations in technical performance, including access speed, navigational ease, and system stability.

Perceptions of efficiency within e-learning environments are strongly mediated by the quality of user interface

design and the seamlessness of technical operations during platform interaction. Deficiencies in these areas may negatively impact user satisfaction and pose challenges to sustained platform adoption and long-term engagement [20].

The distribution of this data reflects that, although users generally hold positive perceptions of their e-learning experiences particularly from an emotional or satisfaction perspective there remains a critical need to address technical improvements, such as efficiency. Such enhancements are essential for sustaining and advancing the success of e-learning platforms in the future. In particular, from a system development standpoint, comprehensive testing and quality assurance practices should be prioritized to ensure the robustness and reliability of the platform [21].

4. Conclusion

Based on the analysis of user perceptions toward the Khan Academy e-learning application, the system quality from the perspective of Quality in Use demonstrates a notably positive trend, particularly regarding the characteristics of satisfaction and efficiency. These perceptions directly reflect the experiences of users engaged in the learning process through the platform. While user feedback indicates a generally favorable attitude toward comfort and overall satisfaction, system developers are advised to prioritize improvements in technical quality, specifically in terms of usage efficiency. Achieving a balanced integration of both functional and non-functional aspects is critical to ensuring the long-term sustainability and acceptance of e-learning systems, especially in the current digital era where user experience serves as a key determinant of success.

5. References

[1] M. Ahmed and M. S. Fiaz, "Evaluating User Experience of Educational Mobile Applications: An Empirical Study," *Int J Hum Comput Interact*, vol. 38, no. 3, pp. 245–260, 2022.

[2] S. M. Mousavi and others, "Emerging Trends in E-Learning: Challenges and Solutions," *Educ Inf Technol (Dordr)*, vol. 26, pp. 2897–2913, 2021.

[3] Google Play Store, "Khan Academy – User Reviews," 2023.

[4] Class Central, "The State of MOOCs 2023," 2023.

[5] Y. Zhao, H. Xu, and X. Wang, "Investigating the Influence of E-Learning on Student Performance in Higher Education," *Comput Educ*, vol. 168, p. 104211, 2021.

[6] A. Kumar and G. Harit, "Text Mining Techniques for Analyzing User Reviews in Educational Apps," *Procedia Comput Sci*, vol. 199, pp. 865–872, 2022.

[7] A. Hassan, M. Ahmad, and S. Saeed, "Sentiment Analysis of User Reviews Using NLP and Machine Learning: A Review," *Journal of King Saud University – Computer and Information Sciences*, 2021.

[8] International Organization for Standardization, "ISO/IEC 25010:2011 Systems and software engineering – System and software quality models," 2021, *ISO, Geneva*.

[9] J. Ming, J. Wu, and Y. Li, "Evaluating Software Quality-in-Use with User Reviews: A Multi-Perspective Approach," *Journal of Systems and Software*, vol. 188, p. 111277, 2022, doi: 10.1016/j.jss.2022.111277.

[10] J. D. Silva, A. Carvalho, and L. Lima, "Using Text Mining to Analyze User Feedback in Software Applications," *Inf Process Manag*, vol. 58, no. 5, p. 102679, 2021, doi: 10.1016/j.ipm.2021.102679.

[11] Y. Yin, S. Lin, and X. Zhang, "Text Mining in Smart Education: Applications, Challenges, and Research Opportunities," *IEEE Access*, vol. 8, pp. 164301–164317, 2020, doi: 10.1109/ACCESS.2020.3022105.

[12] S. Budhrani, A. Thomas, and M. George, "Comparative Analysis of String Matching Algorithms in Big Data Environment," *Procedia Comput Sci*, vol. 185, pp. 92–99, 2021, doi: 10.1016/j.procs.2021.05.010.

[13] A. Kumar, M. Singh, and Z. Khan, "String Matching Techniques for Noisy Text Data: Applications in Real-World NLP Tasks," *Journal of Information Science and Engineering*, vol. 39, no. 2, pp. 145–159, 2023.

[14] H. Kaur and A. Sharma, "Data Preprocessing Techniques for Text Classification: A Review," *International Journal of Computer Sciences and Engineering*, vol. 10, no. 1, pp. 23–30, 2022, doi: 10.26438/ijcse/v10i1.2330.

[15] R. Gupta, P. Sharma, and R. Kumar, "A Review on Text Preprocessing Techniques in Sentiment Analysis," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, pp. 10657–10666, 2021.

[16] S. Albahli, I. Ahmad, and A. Hussain, "Preprocessing Techniques in Sentiment Analysis: A Comparative Review," *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 2, pp. 2049–2059, 2023, doi: 10.3233/JIFS-223207.

[17] D. Al-Fraihat, M. Joy, and J. Sinclair, "Evaluating E-learning systems success: An empirical study," *Comput Human Behav*, vol. 102, pp. 67–86, 2020, doi: 10.1016/j.chb.2019.08.004.

[18] L. Indriani and A. R. Jatmiko, "Analisis Kepuasan Pengguna Website PPDB Online Dengan Penerapan Metode Webqual 4.0 dan IPA (Studi SMK Negeri 1 Labuan Bajo)," *J-INTECH*, vol. 12, no. 02, pp. 207–218, Dec. 2024, doi: 10.32664/j-intech.v12i02.1282.

[19] L. A. Hussein and M. F. Hilmi, "The influence of convenience on the usage of Learning Management System," *Electronic Journal of e-Learning*, vol. 19, no. 6, 2021.

[20] R. M. Yılmaz, F. G. K. Yılmaz, H. T. Öztürk, and B. Sezer, "Investigating the effect of e-learning systems on academic achievement and attitudes: A meta-analysis," *Educational Technology Research and Development*, vol. 69, no. 6, pp. 3345–3371, 2021, doi: 10.1007/s11423-021-10045-6.

[21] M. C. Saputra and T. Katayama, "Proposal of a Method to Measure Test Suite Quality Attributes for White-Box Testing," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021, doi: 10.14569/IJACSA.2021.0120535.