Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 7 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i07.52891

PAPER

Text Classification of Duolingo Reviews on Google Play: Insights for Enhancing M-Learning Applications

Teguh Arie Sandy(✉), Anik Ghufron, Ali Muhtadi, Pujiriyanto

Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

teguhariesandy@uny.ac.id

ABSTRACT

As digital education tools gained prominence, user feedback played a crucial role in refining and personalizing learning experiences. This study analyzed over 100,000 Google Play reviews of the Duolingo language-learning app, using text classification techniques to extract key insights into user sentiment and preferences. By employing natural language processing (NLP) methods, specifically logistic regression and Naive Bayes classifiers, the study categorized feedback into four primary themes: content, instruction, performance, and user interface and user experience (UI/UX). Logistic regression achieved an AUC score of 0.812, precision of 0.904, recall of 0.900, and F1-score of 0.894, while Naive Bayes achieved an AUC score of 0.806, precision of 0.904, recall of 0.900, and F1-score of 0.894. Both models demonstrated an accuracy rate of 90%. The results indicated that content was the most significant concern for users, comprising 74.2% of all reviews, followed by instructional feedback (14%) and performance issues (9.1%). This analysis provided valuable insights for developers aiming to enhance Duolingo's user experience by addressing content quality, improving pedagogical approaches, resolving technical issues, and refining the user interface. The findings also contributed to the broader field of educational technology by demonstrating the application of machine learning techniques in understanding user feedback at scale.

KEYWORDS

Duolingo, mobile learning, user feedback analysis, data mining, sentiment analysis

1 INTRODUCTION

In the evolving landscape of language learning technologies, Duolingo emerged as a leading platform, boasting millions of active users worldwide. As the demand for digital education tools increased, the significance of user feedback in shaping these technologies became paramount. User reviews, particularly on platforms like Google Play, served not only as direct feedback to developers but also as indicators of user satisfaction and areas needing improvement [1], [2]. Given the volume of data generated, manually analyzing these reviews could have been more practical,

Sandy, T.A., Ghufron, A., Muhtadi, A., Pujiriyanto. (2025). Text Classification of Duolingo Reviews on Google Play: Insights for Enhancing M-Learning Applications. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(7), pp. 206–223. https://doi.org/10.3991/ijim.v19i07.52891

Article submitted 2024-10-14. Revision uploaded 2025-01-04. Final acceptance 2025-01-06.

necessitating automated methods to harness valuable insights from large-scale user feedback [3].

The field of natural language processing (NLP) provides the necessary tools to address this challenge through text classification techniques. These techniques allow for the systematic categorization of text data, enabling the automation of user feedback analysis on a scale that manual methods could not achieve [4]. This study applied text classification methods to analyze user feedback from Google Play reviews of the Duolingo app, aiming to identify prevailing sentiments and concerns among users. The insights derived guided improvements in in-app features and user experience, potentially increasing user engagement and satisfaction [5], [6].

Prior research on text classification demonstrated its effectiveness across various domains, including sentiment analysis, topic modeling, and customer feedback processing. Studies such as those by Pang and Lee [5] and Liu [6] laid the foundation for sentiment analysis using machine learning techniques, showcasing their capability to extract sentiments from product reviews and social media data. However, the application of these methodologies specifically to educational technology apps like Duolingo remained underexplored [2], [7]. This study sought to fill this gap by focusing on the categorization of feedback into distinct themes, such as app content, usability, instructional quality, and overall performance, providing a nuanced understanding of user experiences.

The complexity of user feedback, encompassing a range of languages and informal styles, posed unique challenges for text classification [8]. This necessitated the use of robust NLP techniques to preprocess and transform raw data into a format suitable for analysis. In this study, preprocessing steps such as tokenization, stemming, and the elimination of stopwords were employed to refine the text data [9]. Subsequently, the Bag of Words model was used for vectorization, transforming text into numerical data suitable for machine learning models [10].

For the classification task, two widely used algorithms were selected due to their proven effectiveness and efficiency in handling textual data: logistic regression and Naive Bayes. Logistic regression, a predictive analysis algorithm, was well-suited for binary and multiclass classification problems and was known for its simplicity and interpretability [11]. Naive Bayes, on the other hand, offered advantages in terms of computational efficiency and performance with smaller datasets, making it ideal for preliminary analysis phases [4]. The comparative analysis of these models provided insights into their suitability for application in user feedback analysis within the context of language learning apps [12].

Model performance was evaluated using several metrics, including accuracy, precision, recall, F1 score, and the Matthews correlation coefficient (MCC). These metrics provided a comprehensive view of model effectiveness, capturing both the accuracy and quality of the classifications. This dual approach ensured that the models not only accurately categorized feedback but also maintained high standards of quality in the classifications, which was crucial for practical applications such as informing app development and enhancing user experience [13].

In summary, this study leveraged advanced text classification techniques to analyze user feedback from Duolingo's Google Play reviews. By systematically categorizing feedback into thematic areas, the study uncovered actionable insights that informed improvements in app functionality and user interface design. This approach not only enhanced understanding of user satisfaction drivers but also contributed to the broader field of educational technology by demonstrating the application of NLP techniques in real-world scenarios. The outcomes of this study provided a blueprint for other developers in the digital education space, highlighting the importance of user-centered design and continuous feedback integration in the development process.

2 LITERATURE REVIEW

2.1 Text classification and natural language processing

Text classification is a fundamental aspect of NLP. It aims to categorize text into predefined categories, facilitating the efficient handling of large text datasets. A pivotal study by Aggarwal and Zhai [3] provides a comprehensive overview of the NLP field, discussing various text classification techniques that have been adapted to analyze sentiments, sort emails, and automate customer service responses. In educational technology, these techniques help identify themes and sentiments in user feedback, which is critical for app development and user experience enhancement.

2.2 Sentiment analysis in user feedback

Sentiment analysis deals explicitly with identifying the polarity of text data—whether the expressed opinion in the text is positive, negative, or neutral. Pang and Lee [14] and Liu [6] have extensively explored sentiment analysis, establishing foundational models that employ both supervised and unsupervised learning techniques. These methodologies have been pivotal in parsing complex user-generated data, such as product reviews and social media posts [15], [16]. Recent advancements in sentiment analysis include the use of data augmentation to address issues such as class imbalance and limited labeled data, as demonstrated by Abonizio et al. [17], who found that augmenting datasets significantly improves classifier robustness.

Furthermore, pre-processing techniques, including lemmatization and noise reduction, have been shown to enhance model accuracy on social media datasets [18]. Advanced machine learning approaches, such as the SRNN-MAFM model by Jain et al. [19], have achieved high accuracy in sentiment classification tasks, emphasizing the role of deep learning and attention mechanisms. These innovations not only improve the reliability of sentiment analysis but also broaden its applications, such as analyzing consumer emotions to predict product success or failure, as explored by Thelwall et al. [20].

2.3 Application in educational technology

Despite extensive studies in other domains, the application of text classification in educational technology, particularly in user feedback for language learning apps like Duolingo, still needs to be explored. Research by Kukulska-Hulme [2] and Viberg and Grönlund [21] highlights the growing importance of immediate feedback in educational settings. Still, it lacks depth in analyzing user feedback through automated text classification. Alonso et al. [22] demonstrated the potential of multi-label classification for analyzing feedback in online education systems, which could be extended to language learning applications. Similarly, Mekala et al. [23] introduced a deep-learning pipeline for classifying user feedback with high accuracy, showcasing the effectiveness of artificial intelligence in low-resource environments.

Further advancements include sentiment analysis to enhance educational decision-making, as highlighted by Shaik et al. [24], who developed a framework for processing student feedback and categorizing it into actionable insights. Lee et al. [25] addressed challenges in educational text classification, particularly data scarcity, by using cross-encoding augmentation techniques, proving effective in multi-label scenarios.

Additionally, Rogers et al. [26] conducted a systematic review of real-time text classification methods, emphasizing the importance of natural language processing techniques in handling user-generated content, which could inform the design of text classification systems in educational apps. These findings underscore the untapped potential of text classification to enhance the understanding of user interactions and satisfaction in educational technology. In parallel, the integration of mobile technologies such as cloud computing and augmented reality (AR) has significantly transformed the delivery and structure of learning experiences. According to Papadakis et al. [27], the synergy between cloud technologies and AR offers immense potential for increasing interactivity in educational tools, particularly in mobile learning environments. These technologies provide a solid foundation for immediate feedback, which is crucial for the effectiveness of educational apps. However, Papadakis et al. [27] also point out that users face challenges when evaluating the appropriateness and educational value of these apps, especially in early childhood education.

On the other hand, research by Papadakis et al. [28] stresses the growing importance of educational apps for young children. Still, it identifies a gap in understanding parents' perceptions of the educational value of these tools. Parents often struggle to assess the effectiveness of educational apps due to the vast variety of app types and content. Both studies emphasize that while immediate feedback plays a key role in learning, there is still a gap in systematically understanding user experiences through tools like automated text classification to improve these apps further.

2.4 Challenges in user feedback analysis

User feedback presents unique challenges due to the informal and diverse language used, the presence of multilingual text, and the subjective interpretation of sentiments. Boiy and Moens [29] highlight the difficulty of accurately interpreting the context and sentiment of user comments, which can significantly affect classification accuracy. This challenge is compounded in multilingual settings, where a single platform like Duolingo receives feedback in numerous languages, as discussed by Cruz et al. [30], who emphasize the need for practical multilingual sentiment analysis tools.

2.5 Machine learning techniques in text classification

Logistic regression and Naive Bayes have been frequently used among machine learning models due to their efficiency and effectiveness in binary and multiclass classification tasks. Jurafsky and Martin [31] describe these models' applicability in NLP, underscoring their ability to handle large datasets with high-dimensional features, as is typical in text data. Their adaptability to various text classification scenarios is evidenced in their widespread use across different domains, from spam detection to sentiment analysis.

"Among the machine learning models, logistic regression and Naive Bayes have been frequently used due to their efficiency and effectiveness in binary and multiclass classification tasks. Jurafsky and Martin [31] describe these models' applicability in NLP, underscoring their ability to handle large datasets with high dimensional features, as is typical in text data. The adaptability of logistic regression and Naive Bayes to various text classification scenarios is evidenced in their widespread use across different domains, from spam detection to sentiment analysis. For instance, in the context of cultural tourism management, as explored in [32], data mining

techniques like these play a key role in efficiently processing large sets of textual data to derive insights. Additionally, mobile learning apps often rely on similar algorithms for personalized content delivery and user engagement analysis [33].

2.6 Comparative studies and model effectiveness

Comparative studies, such as those by Zhang et al. [10], have evaluated the performance of different text classification algorithms, including logistic regression and Naive Bayes. These studies typically assess model accuracy, precision, recall, and F1 scores to determine suitability for specific applications. However, such comparative analyses are rare in educational technology applications, particularly in studies that focus on user feedback in language learning platforms.

Comparative studies, such as those by Zhang et al. [10] have evaluated different text classification algorithms' performance, including logistic regression and Naive Bayes. These studies typically assess model accuracy, precision, recall, and F1 scores to determine suitability for specific applications. However, there needs to be more such comparative analyses in educational technology applications, particularly in studies that focus on user feedback in language learning platforms. For example, sentiment analysis techniques and intense learning approaches have been effectively applied in evaluating user feedback on platforms, such as in the case of hotel performance evaluation through social media data [34]. Additionally, research on Arabic text sentiment analysis has also shown promise in predicting student attitudes, which could be adapted for language learning platforms to understand user sentiment and engagement better [35].

2.7 Gaps in the literature

The literature review reveals a significant gap in comprehensive studies that apply advanced text classification techniques to user feedback in educational technologies, specifically language learning apps. There remains considerable scope for research on using NLP to enhance user experience and app functionality based on systematic feedback analysis. Moreover, the multilingual nature of user feedback and the informal use of language represent areas requiring further exploration and methodological innovation.

In conclusion, while NLP and text classification have significant applications across various domains, their potential in educational technology, especially in analyzing and leveraging user feedback for language learning platforms like Duolingo, has yet to be fully realized. This study aims to bridge this gap, applying rigorous text classification methodologies to provide actionable insights into user preferences and pain points, thereby informing targeted enhancements in the Duolingo app.

3 METHODS

This study employed a systematic methodology to analyze user feedback from Google Play reviews of the Duolingo application using text classification techniques. The framework, as illustrated in Figure 1, consisted of five key stages: data preparation, preprocessing, modeling, visualization, and evaluation. The process began with data preparation, where user reviews were collected from the Google

Play Store through web crawling. These reviews were then labeled to create a dataset suitable for analysis.

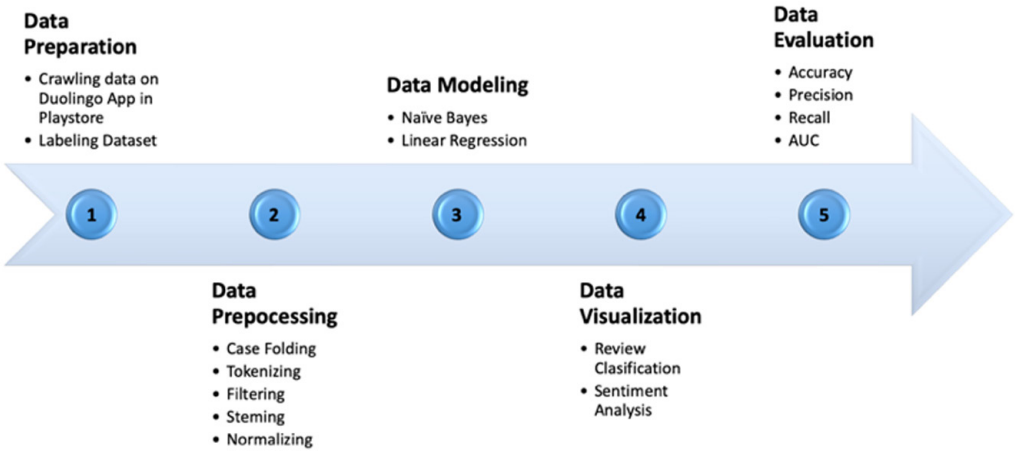

Fig. 1. Workflow of the methodology in the Duolingo analysis

In the data preprocessing stage, the raw data was refined into a structured format using NLP techniques. This involved case folding, which converted all text to lowercase for uniformity, tokenizing to split the text into individual words, and filtering to remove unnecessary characters, punctuation, and stop words. Additionally, stemming and normalizing were performed to reduce words to their root forms and standardize the text for consistency.

The preprocessed data was then used in data modeling, employing two machine learning algorithms: Naïve Bayes and logistic regression. Naïve Bayes was chosen for its efficiency in text classification tasks, particularly with smaller datasets, while logistic regression was selected for its simplicity and effectiveness in binary and multiclass classification.

The categorized feedback was further explored in the data visualization stage, where insights were derived by grouping reviews into predefined themes and conducting sentiment analysis to identify positive, negative, or neutral sentiments in the user feedback. Finally, the performance of the classification models was assessed in the data evaluation stage using metrics such as Accuracy, Precision, Recall, F1 Score, and AUC (Area Under the Curve) to ensure robust and reliable results.

The dataset consists of user reviews of the Duolingo app sourced from the Google Play Store from January 2020 to December 2022. This robust sample captures a wide range of user interactions and sentiments during this period [36]. Reviews were extracted using a web scraping tool that was compliant with Google's API policies to ensure ethical data collection. The final dataset comprises over 100,000 reviews (110,668 total), each containing a star rating and text comment [37].

Preprocessing the textual data is crucial for preparing it for analysis. The preprocessing steps include text cleaning, tokenization, stop word removal, and stemming. Text cleaning involves removing HTML tags, memorable characters, and numbers using regular expressions, which are standard in NLP tasks [3]. Tokenization was then performed using the Natural Language Toolkit (NLTK) library in Orange 3, splitting the text into individual words or tokens for a more straightforward analysis [38]. Common words that do not contribute to sentiment analysis, such as "and" or "the," were removed using a customized list suitable for the context of language learning feedback [29]. Lastly, stemming was applied to reduce words to their base or root

form, consolidating variations like “learn,” “learning,” and “learned” into a single word, which is essential for standardizing the text [39].

To perform a practical evaluation of machine learning models [3], 10-fold cross-validation is a commonly used technique that ensures the model is appropriately validated and evaluated for its generalizability. This method involves dividing the dataset into ten equal parts, or folds, and using each part in turn for testing while the remaining nine folds are used for training. By repeating this process for all folds, we obtain a more accurate and reliable measure of model performance, reducing the likelihood of overfitting [29], [38], [39].

In Orange3, the first step in this process is to load the dataset using the File widget. This widget allows users to import various dataset formats (such as CSV or Excel) into the project. Once the dataset is loaded, preprocessing steps can be performed if necessary. For instance, when dealing with text data, the Text widget can be used to clean the data by removing HTML tags, tokenizing the text, removing stop words, and applying stemming. These preprocessing tasks are essential for preparing the text data, especially for sentiment analysis and text classification. Non-text data can be preprocessed using the Preprocess widget, which allows for normalization, feature selection, and other data preparation techniques.

After preprocessing, the next crucial step is applying 10-fold cross-validation, which can be accomplished by using the Cross Validation widget in Orange3. This widget divides the data into ten equal parts and applies the cross-validation process, ensuring that each data point is used for both training and testing. This helps in evaluating how well the model performs across different data subsets. The Test and Score widget is then used to assess the performance of the model by providing key metrics such as accuracy, precision, recall, and F1-Score. These metrics give insights into the model’s ability to classify the data and handle challenges such as class imbalance.

In this evaluation, logistic regression and Naïve Bayes models are often selected for text classification tasks due to their effectiveness. logistic regression predicts the probability of a categorical dependent variable, making it suitable for classifying user feedback based on text features. Naïve Bayes, a probabilistic model base’ on Bayes’ Theorem, assumes independence among predictors and is particularly efficient in handling large datasets, making it ideal for text classification. Both models can be connected to the Cross-Validation widget for 10-fold cross-validation in Orange3.

The results of the cross-validation process are displayed in the Test and Score widget, which shows the performance metrics for each fold and the average performance across all folds. This provides a holistic view of’ the model’s performance, offering more reliable results than a single training-test split. The use of 10-fold cross-validation ensures a more stable and accurate evaluation, improving the likelihood of selecting a robust model that can generalize well to unseen data.

In conclusion, 10-fold cross-validation in Orange3 provides a robust framework for evaluating machine learning models, offering reliable and repeatable results. This method reduces bias and overfitting, ensuring that the model selected is effective across different data subsets. The process is especially beneficial for text classification tasks, as it ensures that models like logistic regression and Naïve Bayes are thoroughly tested on various portions of the data before final evaluation [10], [13], [31].

For reference, Figure 2 illustrates Orange3’s workflow, highlighting the key steps in the data preprocessing, model training, and evaluation process. This flowchart helps visually guide users through the steps involved in building and evaluating a machine-learning model in Orange3.

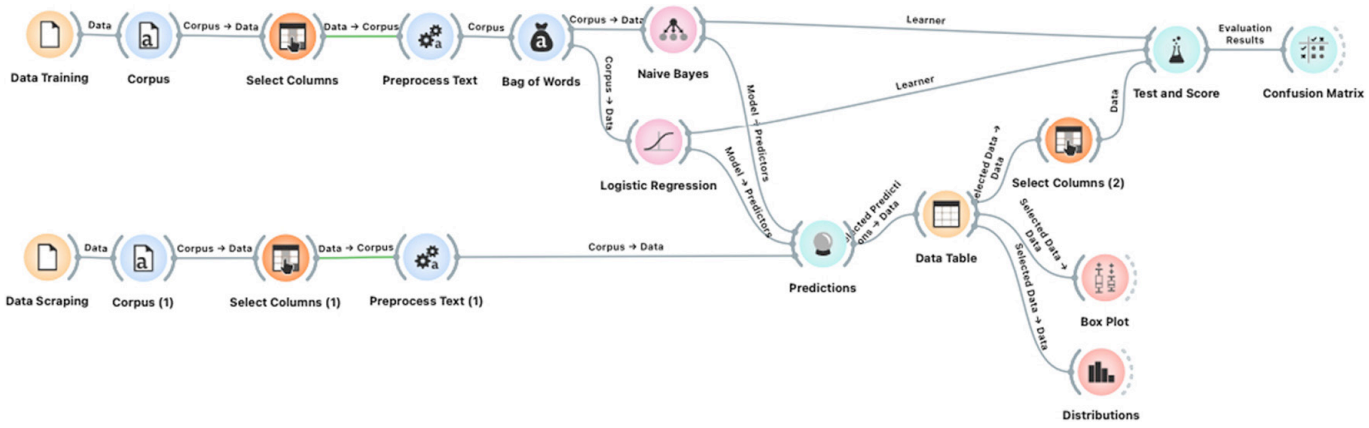

Fig. 2. Orange3 workflow analysis

Model performance was evaluated using several key metrics to assess the accuracy and quality of predictions. Accuracy measures the overall correctness of the model in predicting the feedback category, while precision and recall evaluate the accuracy of optimistic predictions and the ability to find all relevant instances in the dataset, respectively [39]. The F1 score, a harmonic mean of precision and recall, provided a balanced metric that was handy for datasets with uneven class distributions [29]. Additionally, the MCC was used to provide a more informative measure than accuracy alone, accounting for true and false positives and negatives, especially when dealing with imbalanced datasets [40]. These metrics offer a comprehensive assessment of each model's effectiveness in accurately classifying user feedback, making them essential for selecting the most suitable model for ongoing analysis [41].

4 RESULT

The analysis of user feedback from Duolingo using text classification provided comprehensive insights into how users perceive and interact with the app. By categorizing reviews into four distinct themes—content, instructional, performance, and user interface and user experience (UI/UX)—we gained a clearer understanding of the areas that matter most to users and how they relate to their overall satisfaction with the app. Each of these categories provides valuable insights into potential areas of improvement and user expectations.

4.1 Review classification

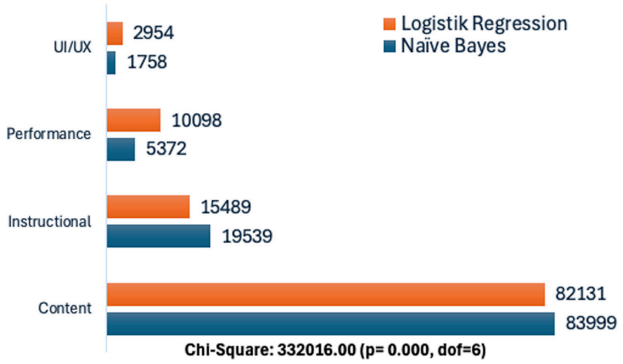

Fig. 3. Naïve Bayes and logistic regression review classification result

Figure 3 presents a comparative analysis of user feedback on the Duolingo application, categorized into four primary themes: content, instructional, performance, and UI/UX. The chart illustrates the distribution of reviews processed using two text classification models—Logistic regression and Naïve Bayes. Content dominates as the most significant category, with 82,131 reviews classified by logistic regression and 83,999 by Naïve Bayes, highlighting its importance to user satisfaction. Instructional feedback follows as the second most discussed theme, with logistic regression identifying 15,489 reviews and Naïve Bayes categorizing 19,539. Performance-related feedback and UI/UX concerns show lower review counts yet remain critical for enhancing the app's overall functionality and user experience.

The inclusion of a Chi-square test result ($\chi^2 = 332016.00$, $p = 0.000$, $dof = 9$) emphasizes the statistical significance of the classification results, validating the reliability of the analysis. This figure underscores the effectiveness of using logistic regression and Naïve Bayes models to categorize user feedback systematically and reveals actionable insights into areas for improvement. Developers can leverage these findings to prioritize enhancements in content delivery and instructional design while addressing technical and interface-related challenges to meet user expectations effectively.

The majority of user reviews focused on content, which made up an overwhelming 74.2% of the total feedback. Logistic regression classified 82,131 reviews, while Naïve Bayes categorized 83,999 reviews in this theme. This heavy emphasis on content indicates that Duolingo's users are most concerned with the quality and breadth of the language courses offered. Content-related reviews likely include feedback on the variety of lessons, the depth of language coverage, and the inclusion of features such as vocabulary-building exercises, grammar instruction, and conversational practice. Given that content is the cornerstone of any educational app, this result is not surprising. However, it suggests that Duolingo must continuously innovate and expand its language offerings to maintain user satisfaction. Users may be requesting new language courses, more advanced lessons, or updates to existing content to keep pace with evolving language learning trends.

The Instructional category, comprising 14% of the total feedback, highlights pedagogical aspects of the app. Logistic regression classified 15,489 reviews, and Naïve Bayes identified 19,539 reviews under this category. This includes feedback on the effectiveness of teaching methods, the structure of lessons, and whether the app supports users' long-term language acquisition goals. Users may have also commented on the app's gamification elements, such as streaks, rewards, and quizzes, and whether these features are beneficial or distracting from real learning. The significant percentage of instructional feedback shows that users care not only about what they are learning but also about how it is being taught. Balancing engaging features and meaningful learning outcomes is critical for the app's success, especially for retaining serious language learners over time.

Performance, covering technical issues such as bugs, glitches, app crashes, and other related problems, accounted for 9.1% of the feedback. Logistic regression classified 10,098 reviews, while Naïve Bayes identified 5,372 reviews in this theme. Although this represents a smaller portion of total feedback compared to content and instruction, it remains crucial. Technical issues can significantly disrupt the user experience, especially in a learning context where consistency and smooth functionality are essential. Performance-related feedback might also address app responsiveness, download times, or compatibility with specific devices and operating systems. Duolingo's developers can leverage this data to resolve performance bottlenecks, ensuring the app runs efficiently across a range of devices and network conditions.

Addressing these issues would likely result in higher user satisfaction and improved retention rates, especially among users who have experienced technical difficulties.

Lastly, UI/UX was the smallest category, with 2.7% of the feedback. Logistic regression identified 2,954 reviews, while Naïve Bayes classified 1,758 reviews under this theme. Although a relatively minor focus, this category should not be overlooked. Reviews in this category likely discuss the app's design, layout, ease of navigation, and overall usability. A well-designed UI/UX is essential for keeping users engaged, particularly in a learning app where intuitive navigation can enhance the learning experience by reducing cognitive load. Even with strong content and instruction, negative feedback related to design or navigation could undermine the app's overall effectiveness. Continuous improvement in UI/UX would ensure that users can seamlessly access content and make the learning process more enjoyable.

4.2 Sentiment analysis

The sentiment analysis, which was conducted by evaluating the star ratings accompanying user reviews, provided a clear indication of overall user satisfaction in Figure 4. The vast majority of reviews were overwhelmingly positive, with **82.9%** (91,755 reviews) awarding Duolingo a **5-star** rating. An additional **10.4%** (11,538 reviews) were **4-star** ratings, leading to a combined **93.3%** of the reviews reflecting positive user sentiment. This high level of satisfaction indicates that most users find Duolingo to be a highly effective and enjoyable platform for language learning. Users who rated the app highly may have been satisfied with the content variety, lesson structure, or the app's engaging gamification features.

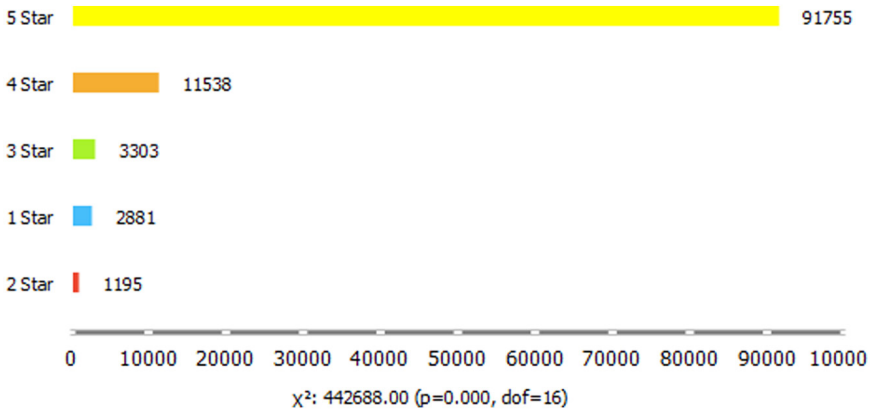

Fig. 4. Star rating review classification result

In contrast, only a tiny fraction of users provided negative feedback, with **1-star** and **2-star** reviews comprising **2.6%** (2,881 reviews) and **1.1%** (1,195 reviews), respectively. These lower-rated reviews likely reflect users who encountered significant frustrations with the app, whether due to technical problems, dissatisfaction with content, or issues with the app's teaching methodology. While the percentage of negative feedback is small compared to the overall positive sentiment, it still represents a critical opportunity for improvement. Addressing these negative reviews, especially those that focus on performance or content concerns, could help Duolingo enhance the app and reduce user churn.

4.3 Model performance

Table 1. Model performance

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.812	0.9	0.894	0.904	0.9	0.759
Naive Bayes	0.806	0.9	0.894	0.904	0.9	0.759

Two machine learning models—logistic regression and Naive Bayes—were tested to classify user feedback into four categories: Content, Instruction, Performance, and UI/UX. Both models demonstrated high performance, achieving similar results across multiple metrics, including accuracy, F1 score, precision, recall, and MCC. The results, illustrated in Table 1, highlight the robustness of both approaches for this task.

Logistic regression achieved an AUC score of 0.812, alongside an accuracy of 0.900 and an F1 score of 0.894. Its precision and recall were 0.904 and 0.900, respectively, with an MCC value of 0.759, reflecting its ability to balance predictions effectively. Naive Bayes, on the other hand, showed a slightly lower AUC score of 0.806 but matched logistic regression in terms of accuracy (0.900), F1 score (0.894), precision (0.904), recall (0.900), and MCC (0.759).

The findings indicate that both models are highly effective for categorizing user reviews into predefined categories. However, logistic regression displayed a marginal advantage in distinguishing between categories due to its slightly higher AUC score. This minor distinction makes logistic regression a more favorable choice for real-world applications where fine-grained distinctions in classification are critical. Nevertheless, given the comparable overall performance, both models can be confidently deployed for similar feedback classification tasks. Figure 4 provides a detailed breakdown of these metrics, emphasizing the practical utility of both models.

4.4 Confusion matrices

The confusion matrices for both logistic regression and Naive Bayes provided further insights into the classification performance. Both models performed exceptionally well in classifying **Content** and **Instruction** reviews, with a high number of correct classifications. However, there needs to be more misclassification between **performance** and **content** reviews, as users may need more clarification on technical issues related to their experience of the content itself. Additionally, **UI/UX** reviews were occasionally misclassified, likely due to the relatively smaller sample size in this category.

The confusion matrices for both models are displayed in Tables 2 and 3.

Table 2. Logistic regression confusion matrix

	Content	Instructional	Predicted Performance	UI/UX	Sum
Actual Content	77350	1031	2193	266	80840
Actual Instructional	1221	18087	1938	825	22071
Actual Performance	286	658	4951	34	5929
Actual UI/UX	87	45	41	1655	1828
Actual SUM	82131	15489	10098	2954	110668

Table 3. Naïve Bayes confusion matrix

	Content	Instructional	Predicted Performance	UI/UX	Sum
Actual Content	79563	1876	337	355	82131
Actual Instructional	1119	14063	214	93	15489
Actual Performance	2581	2704	4764	45	10094
Actual UI/UX	736	896	57	1265	2954
Actual SUM	83999	19539	5372	1758	110668

The confusion matrices for logistic regression (refer to Table 2) and Naïve Bayes (refer to Table 3) offer a detailed comparison of how each model performed in classifying user reviews into four distinct categories: Content, Instruction, Performance, and UI/UX. While both models showed strong performance across the board, key differences emerged in how they handled specific categories.

For the Content category, both models exhibited strong performance, with 79,563 reviews correctly classified under Content by Naïve Bayes and 77,350 correctly classified by logistic regression. This demonstrates that both models excelled in identifying Content-related feedback, which makes sense given that this category constitutes a substantial portion of the dataset (approximately 74.2%). The slight edge held by Naïve Bayes in terms of correct classifications can be attributed to its slightly better handling of the Content category.

When it comes to Instructional, both models performed robustly, though there were some differences in the number of correct predictions. Logistic regression accurately classified 18,087 Instruction-related reviews, while Naïve Bayes predicted 14,063 correctly. The difference of around 4,000 reviews suggests that logistic regression may be more adept at handling the structured nature of Instructional feedback. Nonetheless, both models performed well in predicting Instruction-related reviews, suggesting they are both effective at capturing feedback related to pedagogical methods.

The Performance category posed a more significant challenge for both models. logistic regression correctly classified 4,951 performance reviews, while Naïve Bayes classified 4,764 correctly. While the number of misclassifications was similar across both models, the more significant issue here is the frequent confusion between Performance and Content. Both models misclassified 2,581 performance reviews as Content, highlighting the difficulty in distinguishing reviews that address technical performance issues alongside content-related feedback. Additionally, a notable number of performance reviews (2,704) were misclassified as Instruction, indicating that both models struggled with reviews that touch on performance aspects but also include instructional feedback. This overlap between categories presents a challenge for models attempting to differentiate between them.

For the UI/UX category, both models achieved similar accuracy, correctly predicting 1,265 UI/UX reviews. However, there were significant misclassifications, particularly where UI/UX feedback was misclassified as either Content or Instruction—896 reviews were misclassified as Content, and 736 as Instruction. This suggests that both models faced difficulty distinguishing between UI/UX-related comments and those related to content or instructional design. This overlap is likely due to the nature of UI/UX feedback, which often involves comments about design elements that are closely linked to the content or instructional aspects of the user experience.

In conclusion, while both logistic regression and Naïve Bayes performed similarly, logistic regression showed a slight advantage in accurately classifying more nuanced

categories, especially Performance and Content. However, both models still demonstrated room for improvement, particularly in reducing misclassifications between categories that are frequently discussed together, such as Content and Performance. To further improve classification accuracy, it may be beneficial to explore more advanced feature extraction techniques, such as word embeddings, which could help capture deeper contextual relationships within the text.

5 DISCUSSION

5.1 Interpretation of results

The data-driven insights derived from analyzing Duolingo's Google Play reviews highlight distinct areas where the app can enhance its user experience and functionality. The predominant focus on content-related feedback, which comprises a significant 74.2% of all reviews, clearly indicates the users' deep interest in the quality and diversity of educational material offered [36]. This insight is crucial for Duolingo, suggesting that continued investment in and innovation of its language courses can substantially enrich user satisfaction and learning outcomes [1].

The next significant area of feedback is Instruction, accounting for 14.0% of the reviews. This reflects users' concerns with Duolingo's pedagogical approaches, including the effectiveness of its teaching methodologies and the structure of its lessons [42], [43], [44]. Given that instructional strategies directly influence learning efficacy, Duolingo should consider exploring varied instructional designs that cater to different learning styles and preferences to optimize educational impact [21].

Though less frequent, performance-related feedback is equally important as it addresses the technical aspects of the app. Any technical glitches or performance issues can severely disrupt the learning process, leading to frustration and potential user dropout. Therefore, maintaining a smooth, bug-free user experience is imperative for user retention and satisfaction.

Lastly, the most minor yet significant category of UI/UX highlights the importance of intuitive design and user-friendly interfaces in educational app [31]. A well-crafted user interface can enhance the learning experience by minimizing distractions and making navigation effortless, allowing users to focus more on learning than on figuring out app functionalities.

5.2 Model comparison

This study compares the performance of logistic regression and Naive Bayes models. Both are highly effective, with logistic regression slightly edging out Naive Bayes in terms of the area under the curve (AUC). While both models are capable of handling text classification tasks efficiently, logistic regression might provide a slight advantage in more complex or nuanced classification scenarios due to its robustness in handling diverse data distributions.

However, the choice between these models can depend on specific requirements, such as the need for quick preliminary insights, for which Naive Bayes could be preferred due to its computational efficiency, or scenarios requiring detailed probability outputs for each class, for which logistic regression would be more appropriate due to its interpretative ease [29].

5.3 Limitations and challenges

This study acknowledges several limitations that need to be considered. One primary limitation is the reliance on user-generated content from Google Play, which may include biased or non-representative samples of the global user base [1]. Additionally, the inherent noise and subjectivity in user reviews can introduce variability that affects the accuracy of the text classification [45]. The use of the Bag of Words model for feature extraction also presents a challenge, as it disregards the context and order of words, potentially leading to a loss of deeper semantic meaning. This limitation could be addressed in future studies by adopting more advanced NLP techniques, such as word embeddings or contextual models, which offer a richer understanding of language nuances [46]. Furthermore, focusing solely on English-language reviews may exclude insights from non-English-speaking users, potentially skewing the analysis towards English-centric user experiences [30]. To address this, it is essential to integrate multilingual text analysis to capture a broader spectrum of user feedback [29].

5.4 Future directions

To enhance the comprehensiveness and accuracy of future analyses, incorporating a wider range of data sources, such as reviews from other app stores and direct feedback mechanisms within the app, would be beneficial [41]. Employing advanced NLP techniques, such as deep learning models that understand the linguistic context better than traditional models, could significantly improve the classification and analysis of the user feedback [13]. A longitudinal approach to analyzing user feedback could also provide dynamic insights into how user sentiments evolve in response to changes and updates in the app, helping Duolingo align its offerings more closely with user needs over time [6], [37]. Additionally, future research could incorporate multilingual feedback analysis using multilingual NLP techniques, such as mBERT or XLM-R, and explore the use of word embeddings or deep learning models like LSTM or Transformer to capture more complex contextual relationships within the text. Furthermore, including user demographic analysis could provide valuable insights into specific preferences or needs of particular user groups, leading to more targeted improvements for the application or platform being studied.

6 CONCLUSION

This study underscores the effectiveness of using text classification techniques, such as logistic regression and Naive Bayes, to analyze large-scale user feedback for educational technology platforms such as Duolingo. By categorizing over 100,000 Google Play reviews, we identified four key areas of user concern: Content, Instruction, Performance, and UI/UX. Content, making up the majority of the feedback, highlights the importance of continuously expanding and enhancing Duolingo's language offerings to meet user expectations. Instructional feedback also plays a significant role, suggesting that users are looking for more diverse and effective teaching methods. While the app generally performs well, technical issues and performance-related complaints still hinder the user experience, and attention to these areas is critical for retaining users and improving satisfaction.

The confusion matrices for both models revealed areas where classification could be improved, particularly in distinguishing between closely related categories

such as Performance and Content. While both logistic regression and Naive Bayes performed well, future work could explore more sophisticated techniques, like word embeddings or deep learning models, to capture the context and semantic meaning of user feedback more accurately.

Furthermore, expanding the analysis to include reviews in multiple languages would provide a more comprehensive view of Duolingo's global user base and offer deeper insights into the experiences of non-English speakers. Overall, this study demonstrates the value of applying machine learning techniques to user feedback analysis and provides a practical framework for improving educational technologies based on real user experiences. These insights will guide future improvements in both content offerings and technical performance, which will be crucial for Duolingo's continued success.

7 REFERENCES

[1] H. Ahmed, "Duolingo as a bilingual learning app: A case study," *Arab World English Journal (AWEJ)*, vol. 7, no. 2, pp. 255–267, 2016. https://doi.org/10.24093/awej/vol7no2.17

[2] A. Kukulska-Hulme, "How should the higher education workforce adapt to advancements in technology for teaching and learning?" *The Internet and Higher Education*, vol. 15, no. 4, pp. 247–254, 2012. https://doi.org/10.1016/j.iheduc.2011.12.002

[3] C. C. Aggarwal and C. Zhai, Eds., *Mining Text Data*. Boston, MA: Springer US, 2012. https://doi.org/10.1007/978-1-4614-3223-4

[4] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual web texts," *Inf. Retrieval*, vol. 12, pp. 526–558, 2009. https://doi.org/10.1007/s10791-008-9070-z

[5] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundation and Trends in Information Retrieval*, vol. 2, nos. 1–2, pp. 1–135, 2008. https://doi.org/10.1561/1500000011

[6] B. Liu, *Sentiment Analysis and Opinion Mining*, in Synthesis Lectures on Human Language Technologies. Cham: Springer International Publishing, 2012. https://doi.org/10.1007/978-3-031-02145-9

[7] O. Viberg and Å. Grönlund, "Cross-cultural analysis of users' attitudes toward the use of mobile devices in second and foreign language learning in higher education: A case from Sweden and China," *Computers & Education*, vol. 69, pp. 169–180, 2013. https://doi.org/10.1016/j.compedu.2013.07.014

[8] F. L. Cruz, J. A. Troyano, B. Pontes, and F. J. Ortega, "Building layered, multilingual sentiment lexicons at synset and lemma levels," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5984–5994, 2014. https://doi.org/10.1016/j.eswa.2014.04.005

[9] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Beijing: O'Reilly, 2009.

[10] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *Int. J. Mach. Learn. and Cyber.*, vol. 1, pp. 43–52, 2010. https://doi.org/10.1007/s13042-010-0001-0

[11] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed., 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[12] V. Bonta, N. Kumaresh, and N. Janardhan, "A comprehensive study on Lexicon based approaches for sentiment analysis," *Asian Journal of Computer Science and Technology (AJCST)*, vol. 8, no. S2, pp. 1–6, 2019. https://doi.org/10.51983/ajcst-2019.8.S2.2037

[13] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information Fusion*, vol. 36, pp. 10–25, 2017. https://doi.org/10.1016/j.inffus.2016.10.004

[14] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundation and Trends in Information Retrieval*, vol. 2, nos. 1–2, pp. 1–135, 2008. https://doi.org/10.1561/1500000011

[15] N. Garg and K. Sharma, "Text pre-processing of multilingual for sentiment analysis based on social network data," *International Journal of Election and Computer Engineering (IJECE)*, vol. 12, no. 1, pp. 776–784, 2022. https://doi.org/10.11591/ijece.v12i1.pp776-784

[16] P. Kumawat and R. Dey, "A review on sentiment analysis," *International Research Journal of Computer Science (IRJCS)*, vol. 9, no. 4, pp. 52–56, 2022. https://doi.org/10.26562/irjcs.2021.v0904.001

[17] H. Q. Abonizio, E. C. Paraiso, and S. Barbon, "Toward text data augmentation for sentiment analysis," *IEEE Transaction Artificial Intelligence*, vol. 3, no. 5, pp. 657–668, 2022. https://doi.org/10.1109/TAI.2021.3114390

[18] M. A. Palomino and F. Aider, "Evaluating the effectiveness of text pre-processing in sentiment analysis," *Applied Sciences*, vol. 12, no. 17, p. 8765, 2022. https://doi.org/10.3390/app12178765

[19] V. Jain, A. K. Saxena, A. Senthil, A. Jain, and A. Jain, "Cyber-bullying detection in social media platform using machine learning," in *2021 10th International Conference on System Modeling and Advancement in Research Trends (SMART)*, 2021, pp. 401–405. https://doi.org/10.1109/SMART52563.2021.9676194

[20] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in Twitter events," *J. Am. Soc. Inf. Sci.*, vol. 62, no. 2, pp. 406–418, 2011. https://doi.org/10.1002/asi.21462

[21] O. Viberg and Å. Grönlund, "Cross-cultural analysis of users' attitudes toward the use of mobile devices in second and foreign language learning in higher education: A case from Sweden and China," *Computers & Education*, vol. 69, pp. 169–180, 2013. https://doi.org/10.1016/j.compedu.2013.07.014

[22] D. Ruiz Alonso, C. Zepeda Cortés, H. Castillo Zacatelco, J. L. Carballido Carranza, and J. L. García Cué, "Multi-label classification of feedbacks," *Journal of Intelligent & Fuzzy Systems*, vol. 42, no. 5, pp. 4337–4343, 2022. https://doi.org/10.3233/JIFS-219224

[23] R. R. Mekala, A. Irfan, E. C. Groen, A. Porter, and M. Lindvall, "Classifying user requirements from online feedback in small dataset environments using deep learning," in *2021 IEEE 29th International Requirements Engineering Conference (RE)*, Notre Dame, IN, USA, 2021, pp. 139–149. https://doi.org/10.1109/RE51729.2021.00020

[24] T. Shaik, X. Tao, C. Dann, C. Quadrelli, Y. Li, and S. O'Neill, "Educational decision support system adopting sentiment analysis on student feedback," in *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Niagara Falls, ON, Canada, 2022, pp. 377–383. https://doi.org/10.1109/WI-IAT55865.2022.00062

[25] H. S. Lee *et al.*, "Cross encoding as augmentation: Towards effective educational text classification," *arXiv preprint arXiv:2305.18977*, 2023. https://doi.org/10.48550/ARXIV.2305.18977

[26] D. Rogers, A. Preece, M. Innes, and I. Spasic, "Real-time text classification of user-generated content on social media: Systematic review," *IEEE Trans. Computational. Social Systems*, vol. 9, no. 4, pp. 1154–1166, 2022. https://doi.org/10.1109/TCSS.2021.3120138

[27] J. Vaiopoulou, S. Papadakis, E. Sifaki, D. Stamovlasis, and M. Kalogiannakis, "Parents' perceptions of educational apps use for kindergarten children: Development and validation of a new instrument (PEAU-p) and exploration of parents' profiles," *Behavioral Sciences*, vol. 11, no. 6, p. 82, 2021. https://doi.org/10.3390/bs11060082

[28] S. Papadakis *et al.*, "Unlocking the power of synergy: The joint force of cloud technologies and augmented reality in education," Криворізький державний педагогічний університет, 2023. https://doi.org/10.31812/123456789/7399

[29] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual web texts," *Inf. Retrieval*, vol. 12, pp. 526–558, 2009. https://doi.org/10.1007/s10791-008-9070-z

[30] F. L. Cruz, J. A. Troyano, B. Pontes, and F. J. Ortega, "Building layered, multilingual sentiment lexicons at synset and lemma levels," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5984–5994, 2014. https://doi.org/10.1016/j.eswa.2014.04.005

[31] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed., 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[32] T. Yuensuk, P. Limpinan, W. Nuankaew, and P. Nuankaew, "Information systems for cultural tourism management using text analytics and data mining techniques," *International Journal of Interactive Mobile Technology (iJIM)*, vol. 16, no. 9, pp. 146–163, 2022. https://doi.org/10.3991/ijim.v16i09.30439

[33] S. Okuboyejo and O. Koyejo, "Examining users' concerns while using mobile learning apps," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 15, no. 15, pp. 47–58, 2021. https://doi.org/10.3991/ijim.v15i15.22345

[34] R. A. Hameed, W. J. Abed, and A. T. Sadiq, "Evaluation of hotel performance with sentiment analysis by deep learning techniques," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 9, pp. 70–87, 2023. https://doi.org/10.3991/ijim.v17i09.38755

[35] E. Alshdaifat, A. Al-shdaifat, and A. Alsarhan, "Machine learning algorithms for attitude prediction from Arabic text: Detecting student attitude towards online learning," *International Journal of Interactive Mobile Technology (iJIM)*, vol. 18, no. 12, pp. 42–56, 2024. https://doi.org/10.3991/ijim.v18i12.47197

[36] I. Garcia, "Learning a language for free while translating the web. Does Duolingo work?" *International Journal of English Linguistics (IJEL)*, vol. 3, no. 1, pp. 19–25, 2013. https://doi.org/10.5539/ijel.v3n1p19

[37] N. Altrabsheh, M. Cocea, and S. Fallahkhair, "Sentiment analysis: Towards a tool for analysing real-time students feedback," in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, Limassol, Cyprus, 2014, pp. 419–423. https://doi.org/10.1109/ICTAI.2014.70

[38] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Beijing: O'Reilly, 2009.

[39] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014. https://doi.org/10.48550/ARXIV.1408.5882

[40] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019. https://doi.org/10.3390/info10040150

[41] B. Settles, C. Brust, E. Gustafson, M. Hagiwara, and N. Madnani, "Second language acquisition modeling," in *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana, Association for Computational Linguistics, 2018, pp. 56–65. https://doi.org/10.18653/v1/W18-0506

[42] Z. Li, C. J. Bonk, and C. Zhou, "Supporting learners self-management for self-directed language learning: A study within Duolingo," *Interactive Technology and Soft Education*, vol. 21, no. 3, pp. 381–402, 2024. https://doi.org/10.1108/ITSE-05-2023-0093

[43] P. Munday, "The case for using Duolingo as part of the language classroom experience," *Revista Iberoamericana de Educación a Distancia (RIED)*, vol. 19, no. 1, pp. 83–101, 2015. https://doi.org/10.5944/ried.19.1.14581

[44] M. Nushi and M. Eqbali, "Duolingo: A mobile application to assist second language learning," *Teaching English with Technology*, vol. 17, pp. 89–98, 2017.

[45] Y. Choi and H. Lee, "Data properties and the performance of sentiment classification for electronic commerce applications," *Inf. Syst. Front.*, vol. 19, pp. 993–1012, 2017. https://doi.org/10.1007/s10796-017-9741-7

[46] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014. https://doi.org/10.48550/ARXIV.1409.0473

8 AUTHORS

Teguh Arie Sandy is a doctoral candidate in Education at Yogyakarta State University. His research focuses on the development of teaching materials, multimedia, e-learning, mobile learning, social media analysis, and sentiment analysis. Currently, he is interested in literature research, including bibliometric analysis, systematic reviews, and meta-analyses (E-mail: teguhariesandy@uny.ac.id).

Anik Ghufron is a Professor and a lecturer in the Department of Curriculum and Educational Technology, Faculty of Education and Psychology, Yogyakarta State University.

Ali Muhtadi is a Professor and lecturer in the Department of Curriculum and Educational Technology, Faculty of Education and Psychology, Yogyakarta State University.

Pujiriyanto holds a Doctorate and is a lecturer in the Department of Curriculum and Educational Technology, Faculty of Education and Psychology, Yogyakarta State University.