JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6015

SENTIMENT ANALYSIS OF PLAYER FEEDBACK IN ALGORUN: A STUDY OF DEEP LEARNING MODELS FOR GAME-BASED LEARNING

Rio Andriyat Krisdiawan[1*]; Nur Alamsyah[2]; Tito Sugiharto[3]

Information Technology[1,3]
Universitas Kuningan, Kuningan, Indonesia[1,3]
www.uniku.ac.id[1,3]
rioandriyat@uniku.ac.id[1*], tito@uniku.ac.id[3]

Technology and Informatics[2]
Universitas Informatika dan Bisnis Indonesia[2]
www.unibi.ac.id[2]
nuralamsyah@unibi.ac.id[2]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *AlgoRun: Coding Game is a game-based learning application aimed at teaching computational thinking (CT) concepts such as variables, conditions, loops, and functions. Evaluating user feedback in such educational games is challenging, as traditional sentiment analysis techniques often overlook nuanced responses. Despite its potential to inform content improvements, sentiment analysis in game-based learning remains underexplored. This study compares the performance of deep learning models—DNN, CNN, RNN with LSTM, and Bidirectional LSTM—for sentiment classification of AlgoRun user reviews, using TF-IDF and word embeddings as feature extraction methods. A total of 1,440 reviews were scraped from the Google Play Store, translated, and preprocessed using data preparation techniques (dropna, fillna), text preprocessing (case folding, cleaning, tokenization, stopword removal, stemming), and feature extraction (TF-IDF and word embeddings). The dataset was labeled into negative, neutral, and positive classes, and split 80% for training and 20% for testing. Among the tested models, the DNN with TF-IDF achieved the highest accuracy of 98.86%, followed by CNN with Word Embeddings (96.97%), Bidirectional LSTM (96.59%), and RNN with LSTM (92.42%). The DNN also showed stable performance and convergence at the 10th epoch, outperforming other models in precision, recall, and F1-score. These results suggest that DNN with TF-IDF is highly effective for sentiment classification in the context of game-based learning. The findings offer useful guidance for developers to adapt content and enhance game quality based on user feedback. This research also contributes to the growing body of literature on leveraging sentiment analysis to optimize educational applications.*

Keywords: *computational thinking, deep learning models, game-based learning, player feedback, sentiment analysis.*

Intisari— *AlgoRun: Coding Game adalah aplikasi pembelajaran berbasis permainan yang ditujukan untuk mengajarkan konsep pemikiran komputasional (CT) seperti variabel, kondisi, loop, dan fungsi. Mengevaluasi umpan balik pengguna dalam permainan edukasi semacam itu merupakan tantangan, karena teknik analisis sentimen tradisional sering kali mengabaikan respons yang bernuansa. Meskipun berpotensi untuk menginformasikan peningkatan konten, analisis sentimen dalam pembelajaran berbasis permainan masih kurang dieksplorasi. Studi ini membandingkan kinerja model pembelajaran mendalam—DNN, CNN, RNN dengan LSTM, dan Bidirectional LSTM—untuk klasifikasi sentimen ulasan pengguna AlgoRun, menggunakan TF-IDF dan penyematan kata sebagai metode ekstraksi fitur. Sebanyak 1.440 ulasan diambil dari Google Play Store, diterjemahkan, dan diproses terlebih dahulu menggunakan teknik persiapan data (dropna, fillna), pra proses teks (pelipatan huruf besar-kecil, pembersihan, tokenisasi, penghapusan stopword, stemming), dan ekstraksi fitur (TF-IDF dan penyematan kata). Kumpulan data diberi label menjadi kelas negatif, netral, dan*

positif, dan dibagi 80% untuk pelatihan dan 20% untuk pengujian. Di antara model yang diuji, DNN dengan TF-IDF mencapai akurasi tertinggi sebesar 98,86%, diikuti oleh CNN dengan Word Embeddings (96,97%), Bidirectional LSTM (96,59%), dan RNN dengan LSTM (92,42%). DNN juga menunjukkan kinerja dan konvergensi yang stabil pada epoch ke-10, mengungguli model lain dalam hal presisi, recall, dan skor F1. Hasil ini menunjukkan bahwa DNN dengan TF-IDF sangat efektif untuk klasifikasi sentimen dalam konteks pembelajaran berbasis permainan. Temuan ini menawarkan panduan yang berguna bagi pengembang untuk mengadaptasi konten dan meningkatkan kualitas permainan berdasarkan umpan balik pengguna. Penelitian ini juga berkontribusi pada semakin banyaknya literatur tentang pemanfaatan analisis sentimen untuk mengoptimalkan aplikasi pendidikan.

***Kata Kunci**: pemikiran komputasional, model pembelajaran mendalam, pembelajaran berbasis permainan, umpan balik pemain, analisis sentimen.*

INTRODUCTION

The integration of games into education in Indonesia has been actively supported by the Ministry of Education, Culture, Research, and Technology (Kemendikbudristek), which promotes the adoption of digital technology in the learning process. One notable initiative is the implementation of educational games designed to enhance 21st-century skills, such as creativity, critical thinking, communication, and collaboration [1]. Game-based learning platforms like AlgoRun hold significant potential to achieve these goals interactively, particularly by fostering computational thinking (CT) skills, which are essential for programming. CT encompasses cognitive and visual-spatial abilities that enable individuals to solve problems systematically [2], in both effective and efficient ways [3].

The Indonesian government also promotes the integration of technology in education through the Merdeka Belajar Kampus Merdeka (MBKM) policy, which emphasizes the use of games as effective learning tools. This policy aligns with research demonstrating that interactive technologies, such as game-based learning, can enhance student motivation and comprehension of complex concepts [4]. Game-based learning is widely recognized as an effective method for delivering educational materials in an interactive and engaging way [5]. It aims to enhance cognitive skills, perseverance [6], comprehension [7], problem-solving abilities, motivation [8], and assessment capabilities [9]. However, one of the key challenges in teaching computational thinking (CT) in Indonesia is the lack of appealing educational tools for students. AlgoRun, with its foundational programming elements such as variables, loops, conditions, and functions, presents a promising solution to this challenge by offering a more enjoyable and engaging learning experience compared to traditional, often monotonous methods.

Despite their potential, measuring the effectiveness and player engagement of these learning-based games remains a significant challenge [10]. A commonly used approach for understanding user experience is sentiment analysis of player feedback [11], [12]. However, traditional sentiment analysis techniques often fall short in capturing the complex nuances of user reviews, particularly within the context of educational games that encompass diverse linguistic and cultural backgrounds.

This study addresses these gaps by proposing a novel approach that integrates advanced deep learning models—DNN, CNN, RNN with LSTM, and Bidirectional LSTM—with dual feature extraction techniques (TF-IDF and Word Embeddings). Unlike prior studies that limit feature extraction to singular techniques, this approach leverages the strengths of both to capture document-level significance and contextual word relationships, enhancing sentiment classification accuracy. Furthermore, the systematic evaluation of model performance across diverse linguistic and cultural feedback datasets distinguishes this study from previous works.

The primary contributions of this research include:

1. Introducing a hybrid methodology combining TF-IDF and Word Embeddings to improve feature representation in sentiment analysis tasks.
2. Evaluating the effectiveness of various deep learning architectures for sentiment analysis within game-based learning platforms, filling a research gap in the educational domain.
3. Providing actionable insights for game developers to adaptively customize content and improve educational outcomes based on user sentiment.

In this context, the application of sentiment analysis to game-based learning remains underexplored, despite its potential to offer valuable insights for game developers in adaptively

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6015

customizing content and enhancing the quality of players' learning experiences. Deep learning-based sentiment analysis serves as an effective tool for understanding user responses to educational games, providing data-driven insights that can inform decision-making in both game development and educational policy.

Table 1. Literature Study of Sentiment Analysis

Author Name	Problem Researched	Method Used	Research Results
Chakraborty et al. (2022) [13]	Understand the audience's attitude towards the product and classify sentiment as positive, negative, or neutral.	BERT module for classification and fuzzy logic for sentiment analysis.	Excels in accuracy, precision, and non-parametric size.
Kaushik Dhola et al. (2021) [14]	Sentiment analysis using Support Vector Machine and Naive Bayes' Multinomial against deep learning methods such as BERT and LSTM	Comparison of SVM, Naive Bayes, Bert, LSTM methods.	Comparative performance analysis of classification algorithms
Romadhoni et al. (2022) [15]	Comparison of Naive Bayes and LSTM in sentiment analysis towards Permendikbud No.30. Dataset of 2765 tweets	Naive Bayes and LSTM	accuracy value of 77% for LSTM and 76% for Naïve Bayess
M.Totox et al. (2023) [16]	Amazon Product Analysis with LSTM and CNN	LSTM and CNN	LSTM accuracy is better than CNN with an epoch of 50
Nursina et al. (2023) [17]	Comparison of Naïve Bayes and K-Nearest Neighbor (KNN) for classifying airline passenger satisfaction.	Both algorithms were applied to passenger satisfaction survey data and evaluated based on accuracy.	Naïve Bayes outperformed KNN with 84.48% accuracy versus 65.38%.
Haifei Zhang et al. (2022) [18]	Sentiment analysis for Amazon product reviews, which highlights their relevance across various business and social domains.	RNN Model	85%, 70%, and 70% accuracy for each dataset

Source : (Research Result, 2024)

The literature review highlights that game-based learning not only enhances students' learning motivation [19], [20], but also imparts essential skills such as computational thinking. Research utilizing Naive Bayes [21] and LSTM methods to analyze sentiment towards Permendikbud No. 30 on a dataset of 2,765 tweets demonstrated that the LSTM model outperformed Naive Bayes, achieving an accuracy of 77% compared to 76%, with TF-IDF used as the text weighting method [15]. Another study by M. Totox et al. [16] applied the LSTM and CNN algorithms to analyze Amazon product reviews, revealing that LSTM achieved a slightly higher accuracy of 77.96% compared to CNN, which reached 77.97% after 50 epochs.

Research [18] employed RNN to analyze sentiment in Amazon product reviews, demonstrating its relevance across various business and social domains. The RNN achieved accuracy levels of 85%, 70%, and 70% on respective datasets, highlighting its competitiveness with contemporary models. Based on the literature and research review, four deep learning models were utilized: DNN with TF-IDF, CNN with word embeddings, RNN with LSTM, and Bidirectional LSTM, to analyze player review sentiments for AlgoRun. The selection of these methods was guided by each model's capability to process text and comprehend sentiment context. DNN and CNN are renowned for their efficiency in classifying large datasets, while LSTM and Bidirectional LSTM excel in capturing long-term dependencies in word sequences, which is essential for accurate sentiment analysis [22]. Research Objectives:

1. To test and compare the performance of deep learning models (DNN, CNN, RNN, LSTM, and Bidirectional LSTM) for sentiment classification using TF-IDF and Word Embeddings.
2. To identify the most effective deep learning algorithm for analyzing player feedback sentiment.
3. To provide actionable recommendations for game developers to customize content and adjust difficulty levels based on sentiment analysis results.
4. To contribute empirical data that supports the literature on game-based learning and Computational Thinking, particularly in the Indonesian context.

Through in-depth sentiment analysis, this research aims to provide valuable insights for game-based learning developers and education policymakers to design more adaptive learning tools that align with student needs and national education policies. The significance of this study lies in its ability to bridge the gap in understanding user

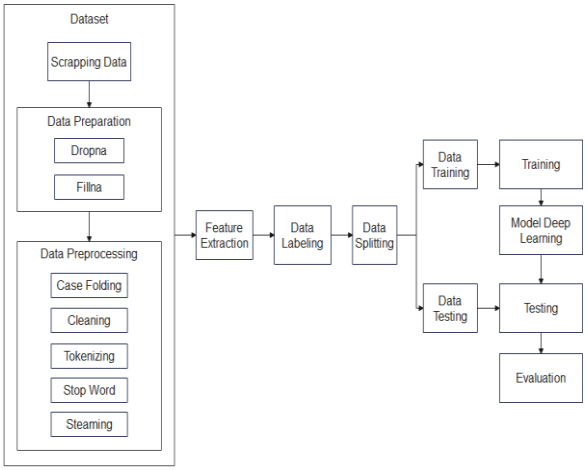
feedback within game-based learning environments, particularly in the context of developing computational thinking (CT) skills. While existing sentiment analysis research predominantly focuses on e-commerce and social media, educational applications remain underexplored. By addressing this gap, the study offers actionable insights for game developers to create more adaptive and engaging learning tools. Furthermore, the findings support national educational policies, such as Merdeka Belajar Kampus Merdeka (MBKM), by promoting innovative and technology-driven approaches to learning.

The novelty of this study lies in its methodological contributions and application in the context of game-based learning. Specifically, this research introduces a novel framework by integrating diverse feature extraction methods—TF-IDF for capturing document-level significance and word embeddings for contextual word relationships—with advanced deep learning models. Unlike previous studies, this research systematically evaluates multiple state-of-the-art models (DNN, CNN, RNN with LSTM, and Bidirectional LSTM) for sentiment analysis of player feedback within an educational game context, namely AlgoRun. This approach not only provides a comparative understanding of model performance and adaptability to sentiment analysis tasks but also highlights the suitability of different architectures for handling diverse language patterns and review complexities.

Furthermore, the study demonstrates how sentiment analysis results can be utilized to customize game content, offering actionable insights for game developers to align player feedback with iterative improvements. These findings contribute significantly to advancing the field of game-based learning by addressing the underexplored area of player sentiment in educational applications. Ultimately, this research provides practical implications for improving educational outcomes and enriching the learning experience for students while supporting broader educational policies aimed at fostering technology-driven learning.

MATERIALS AND METHODS

This research focuses on sentiment analysis of player reviews for the game-based learning AlgoRun, aiming to evaluate and compare the performance of several deep learning models. The models tested include RNN, LSTM, GRU, Bi-LSTM, and Bi-GRU to classify sentiments into positive, neutral, and negative categories based on user reviews. The study using a laptop equipped with an AMD FX-9830P RADEON R7 processor, 12 compute cores (4C+8G) at 3.00 GHz, and 16GB of RAM. Python programming on Google Colab was utilized for implementation. The research methodology involves several stages, as outlined in Figure 1.



Source : (Research Result, 2024)

Figure 1. Research Procedure

The research procedure illustrated in Figure 1 consists of several stages, including data scraping, data preparation, data preprocessing, feature extraction, data labeling, data splitting, classification, and evaluation. Each stage of the procedure is described in detail as follows:

A. Scraping Data

The initial stage of the research involves collecting data in the form of AlgoRun player reviews from the Google Play Store through a scraping process. This approach captures a wide range of sentiments from various languages and countries, ensuring a diverse and representative dataset. The scraping process is performed using a combination of relevant keywords and an appropriate API, specifically google_play_scraper. Examples of reviews scraped from the Play Store in multiple languages are presented in Table 2.

Table 2. Some Reviews of Scraping all languages

Create at	Username	Content
2024-09-02 19:57:25	Minenhle Khumalo	Made me think differently
2024-08-20 19:00:53	Not Thebest	Great app but I am stuck on conditionals 7 without hints
5/14/2024 4:26:09	Omer Rivlin	Very fun and challenging app for kids abd adults. Wonderful way to get intuition for programming!

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

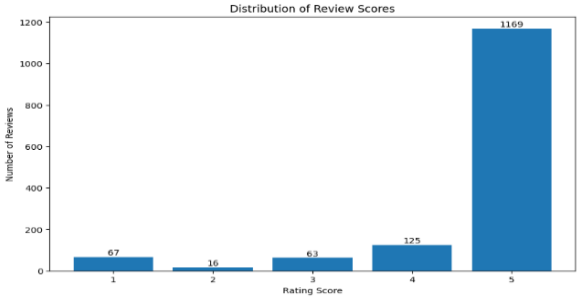
VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6015

Table 2. Some Reviews of Scraping all languages (Continue.)

Create at	Username	Content
5/19/2023 11:53:17	Daniel Rudki	Mantap gamenya coba tingkatin lagi gamenya pasti keren
3/27/2024 9:56:51	Doug Heffernan	Ein wunderbares Spiel um Algorithmen zu lernen. Sehr zu empfehlen

Source : (Research Result, 2024)

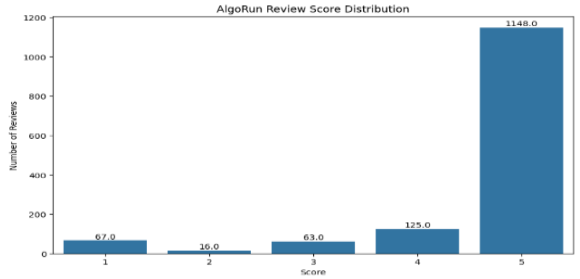
After data collection, each foreign-language review was automatically translated into Bahasa Indonesia using the Google Translate API to ensure consistency in language for sentiment analysis. Following the translation process, the dataset comprised 1,440 reviews, encompassing various scores, dates, and app versions. The distribution of the collected review ratings is depicted in Figure 2.



Source : (Research Result, 2024)
Figure 2. Data distribution based on review ratings

B. Data Preparation

After successfully retrieving and translating the data, the data preparation process was conducted to ensure data quality. The Dropna technique was employed to remove missing or empty data, while Fillna was utilized to fill missing values with appropriate replacements. As a result, the total number of reviews decreased from 1,440 to 1,419. This demonstrates the effectiveness of the Dropna and Fillna techniques in handling missing and incomplete data. The distribution of the prepared data is illustrated in Figure 3.



Source : (Research Result, 2024)
Figure 3. Data distribution based on review ratings after data preparation

The visualization in Figure 3 illustrates the distribution of user review scores for the AlgoRun - Coding Game app on the Google Play Store. The dataset includes 67 reviews with a rating of 1, 16 reviews with a rating of 2, 63 reviews with a rating of 3, 125 reviews with a rating of 4, and 1,148 reviews with a rating of 5. Notably, the number of reviews with a rating of 5 decreased from 1,169 to 1,148 after applying the Dropna and Fillna techniques, indicating that the data preparation stage was conducted effectively and in accordance with the procedure. Examples of the prepared data based on review ratings are presented in Table 3.

Table 3. Preparation result

Score	Content
1	AlgoRun bahkan tidak mau mulai. Tidak peduli b...
2	Awalnya mudah, tetapi begitu sampai pada fungs...
3	Saya terjebak di level 6 kondisional, bisakah ...
4	Permainan yang bagus
5	Membuatku berpikir secara berbeda

Source : (Research Result, 2024)

A general analysis of the data presented in Table 3 reveals that the content of each review aligns with the ratings, demonstrating consistency with the data preparation procedure.

C. Data Preprocessing

The preprocessing stage is conducted to clean and prepare text data before it is used in the model [23]. This step is crucial as effective preprocessing can significantly enhance the accuracy of the algorithm. The following steps are undertaken during this stage:

1. **Case Folding:** All text is converted to lowercase to eliminate discrepancies between uppercase and lowercase letters, ensuring uniformity in word representation.
2. **Cleaning**: Punctuation marks and non-alphanumeric special characters are removed to focus on meaningful words.
3. **Tokenization**: Sentences are broken down into individual words (tokens), enabling machine learning algorithms to process the text effectively.
4. **Stop word removal**: Common words that do not contribute significant meaning (stop words) are removed to enhance the quality of the analysis.
5. **Stemming**: Words are reduced to their root form, consolidating variations of words with the same meaning into a single entity.

The results of the process can be seen in table 4.

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i4.6015

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

Table 4. Data preprocessing process

1	2	3	4	5	6
content	lower_content	cleaned_content	tokenized_content	filtered_tokens	tokens_stemmed
Permainan yang bagus 🎯	permainan yang bagus 🎯	permainan yang bagus	[permainan, yang, bagus]	[permainan, bagus]	[permainan, bagu]
Membuatku berpikir secara berbeda	membuatku berpikir secara berbeda	membuatku berpikir secara berbeda	[membuatku, berpikir, secara, berbeda]	[membuatku, berpikir, berbeda]	[membuatku, berpikir, berbeda]
Aplikasi yang bagus tapi saya terjebak pada ko...	aplikasi yang bagus tapi saya terjebak pada ko...	aplikasi yang bagus tapi saya terjebak pada ko...	[aplikasi, yang, bagus, tapi, saya, terjebak, ...	[aplikasi, bagus, terjebak, kondisi, 7, petunjuk]	[aplikasi, bagu, terjebak, kondisi, 7, petunjuk]
Aplikasi bagus. Putra saya yang berusia 5 tahu...	aplikasi bagus. putra saya yang berusia 5 tahu...	aplikasi bagus putra saya yang berusia 5 tahun...	[aplikasi, bagus, putra, saya, yang, berusia, ...	[aplikasi, bagus, putra, berusia, 5, mengikuti]	[aplikasi, bagu, putra, berusia, 5, mengikuti]
Cemerlang dan sederhana	cemerlang dan sederhana	cemerlang dan sederhana	[cemerlang, dan, sederhana]	[cemerlang, sederhana]	[cemerlang, sederhana]
Luar biasa 👌	luar biasa 👌	luar biasa	[luar, biasa]	[]	[]

Source : (Research Result, 2024)

Table 4 is a very clear overview of each step in the data preprocessing process. Each row represents a review, and each column shows the results of each preprocessing step.

1. **Column Content**: is the raw, unprocessed review text. There is data that still has icons, words that have no meaning, and so on.
2. **Lower_content column**: at this stage, the text is converted into all lowercase letters.
3. **Cleaned_content column**: Punctuation marks and special characters are removed to help focus on words that actually have meaning.
4. **Tokenized_content column**: Sentences are broken down into individual words (tokens).
5. **Filtered_tokens column**: Common words that do not provide much information (stop words) such as “which”, “and”, “at” are removed.
6. **Tokens_stemmed column**: Words are converted into their base form (stemming). for example, the words “play” and “play” will become “play”.

D. Feature Extraction

The feature extraction process utilizes two primary methods: TF-IDF (Term Frequency-Inverse Document Frequency) and Word Embedding. These methods transform text from its original format into numerical representations suitable for processing by deep learning models. TF-IDF features are employed in the Simple Neural Network (DNN) to classify sentiments by identifying patterns in word distributions across documents, enabling effective sentiment prediction [24]. A total of 968 TF-IDF features were generated, as detailed in Table 5, which presents the terms along with their corresponding TF-IDF scores.

Table 5. Term TF-IDF

Word	Tfidf_Score
bagus	131.661664
yang	97.176446
saya	69.614948
untuk	69.351793
permainan	67.840589
sangat	65.052186
ini	58.377110
game	48.847225
dan	48.298544
menyenangkan	43.355419

Source : (Research Result, 2024)

Table 5 provides valuable insights into the most relevant words based on their TF-IDF scores. Words with the highest scores, such as "good," "game," and "play," highlight the primary topics discussed in the reviews of AlgoRun - Coding Game on the Google Play Store. A visualization of the TF-IDF results, illustrating word occurrences, is presented in Figure 4.

Source : (Research Result, 2024)

Figure 4. Visualization of TF-IDF Feature into Word Cloud.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 8. NO. 1 AGUSTUS 2022
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i2.XXXX

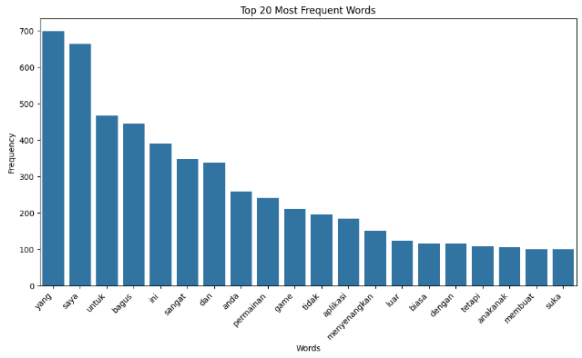
Word embeddings are continuous vector representations that capture the contextual meaning of words within a corpus. These embeddings are used as input for deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

Table 6. Results of Word Embedding

Content	Score
bagus	444
permainan	241
game	210
aplikasi	184
menyenangkan	151
anakanak	106
suka	99
level	97
pemrograman	91
logika	81
membantu	78
berfikir	76
algoritma	75
belajar	74
coding	70
menyukainya	69
terbaik	68
menarik	62
otak	56

Source : (Research Result, 2024)

The analysis of Table 6 reveals that words with the highest scores in the embedding process, such as "good," "game," and "play," highlight a dominant theme centered on users' positive experiences with the app or game. Additionally, words like "fun," "like," and "learn" emphasize the importance of both educational and entertainment aspects, which are primary concerns for users. The high scores of these terms suggest that positive and meaningful concepts related to learning and digital activities dominate the discussion, offering valuable insights into user perceptions and usage patterns. The Word Embedding features are visualized in Figure 5.



Source : (Research Result, 2024)
Figure 5. Word frequency analysis

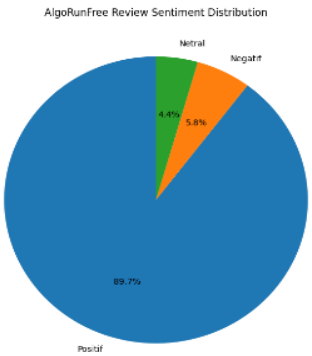
Figure 5 illustrates the word frequency analysis of AlgoRun player feedback, highlighting the dominance of words such as "good," "very," and "like," which reflect positive sentiment. However, the presence of terms like "no" and "difficult" suggests areas for potential improvement. Words such as "game" and "app" confirm that most feedback focuses on the core aspects of the AlgoRun game-based learning platform.

The frequency distribution of these words provides valuable insights for the further development of the AlgoRun platform. For instance, the term "difficult" suggests the need to adjust the complexity of certain learning materials, while terms like "fun" and "make" demonstrate that the game aspect of the platform has successfully engaged players. These insights can guide the creation of a roadmap for new features better aligned with player needs and preferences.

Moreover, the analysis of frequently occurring words offers valuable input for developing deep learning models. These keywords can be leveraged to train models for more accurate sentiment identification. By understanding the context in which these words appear, researchers can design more relevant features, ultimately enhancing the model's performance in analyzing player feedback.

E. Data Labeling

The data labeling process is carried out to label each data sample according to the sentiment contained in the review text.



Source : (Research Result, 2024)
Figure 6. AlgoRunFree Review Sentiment Distribution

Figure 6 is a visualization of the label distribution of the AlgoRunFree data. The labels used are positive, neutral, and negative based on predefined score values. The review score is given based on the user's assessment, where a high score (4-5) as positive, a neutral score (3) as neutral, and a low score (1-2) is categorized as negative.

F. Data Splitting

Once the data is labeled, the next step involves dividing it into two subsets: a training set and a testing set. In this study, 80% of the data is allocated

for training, while the remaining 20% is reserved for testing. The training set is used to train the deep learning models, while the testing set evaluates the model's performance on previously unseen data. After splitting the dataset of 1,419 reviews, 1,135 reviews were assigned to the training set, and 284 reviews were allocated to the testing set.

G. Classification Method

In the classification stage, four deep learning methods were used to predict sentiment from reviews. The specifications of the four models are shown in Table 7.

Table 7. Model specification

Model	Feature Extraction	Architecture
DNN with TF-IDF	TF-IDF	DNN
CNN with Word Embeddings	Word Embeddings	CNN
RNN with LSTM	Word Embeddings	LSTM
Bidirectional LSTM	Word Embeddings	Bidirectional LSTM

Source : (Research Result, 2024)

The first method is Simple Neural Network (DNN) which uses TF-IDF features to generate a numerical vector representation of the text. TF-IDF is used to calculate the importance of a word in a document relative to the rest of the document set, without considering word order or context. The DNN model utilizes a fully connected architecture. Each layer is defined as:

$$h^{(l)} = \sigma\left(W^{(l)}h^{(l-1)} + b^{(l)}\right) \quad (1)$$

where ($h^{(l)}$) represents the activations of the (l) −th layer, ($W^{(l)}$) is the weight matrix, ($b^{(l)}$) is the bias vector, and (σ)($b^{(l)}$) is the activation function (ReLU in this study). The output layer applies a softmax function to compute the probability distribution over sentiment classes.

Convolutional Neural Network (CNN) utilizes Word Embeddings to identify local patterns in the text, such as key phrases or word order, and predicts sentiment based on these features. The core operation of CNN involves convolutions, where filters are applied to extract meaningful features from the text:

$$c_{i,j} = \sum_{k=1}^{n}\sum_{l=1}^{m} w_{k,l}x_{i+k-1,j+l-1} + b \quad (2)$$

Here, ($w_{k,l}$)is the convolution filter, (x) is the input matrix, (b)is the bias, and ($c_{i,j}$) is the output of the convolution operation. Max-pooling is then applied to reduce dimensionality while preserving important features.

The third model is a Recurrent Neural Network (RNN) with an LSTM that is able to capture the context of the word order in the text, making it ideal for tasks involving temporal dependencies. LSTM, which is an advanced version of RNN, is used due to its ability to handle long sequences and avoid the vanishing gradient problem. LSTM is defined to capture long-term dependencies using gating mechanisms[25], [26]. The cell state (c_t) and hidden state (h_t) are updated as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (4)$$

$$\tilde{c_t} = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c_t} \quad (6)$$

$$h_t = o_t \odot \tanh(c_t) \quad (7)$$

Here, (f_t, i_t, o_t) are forget, input, and output gates, respectively. Finally, Bidirectional LSTM is applied to process text from two directions (forward and backward), allowing the model to understand the context better. With Bidirectional LSTM, the model can consider information from both sides of the text when making predictions, which is very useful in sentiment classification, especially when the words at the beginning and end of the text are related. The Bidirectional LSTM processes input sequences in both forward and backward directions to capture context from both ends:

$$h_t^{\text{bi}} = \text{concat}\left(h_t^{\text{forward}}, h_t^{\text{backward}}\right) \quad (8)$$

Bidirectional LSTM extends the standard LSTM model by processing input sequences in two directions: forward and backward. This dual processing allows the model to capture both past and future context at each time step in a sequence. In the forward pass, the hidden state $\left((h_t^{\text{forward}})\right)$is computed based on the current input $((x_t))$and the hidden state from the previous timestep$\left((h_{t-1}^{\text{forward}})\right)$. In the backward pass, the hidden state$\left((h_t^{\text{backward}})\right)$is determined using the current input$((x_t))$and the future hidden state$\left((h_{t+1}^{\text{backward}})\right)$. These forward and backward hidden states are then concatenated((h_t^{"{bi}} = "{concat}(h_t^{"{forward}}}, h_t^{"{backward}}})))to form a final comprehensive representation at each time step.

This bidirectional approach is particularly effective in sentiment analysis as it allows the model

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti
No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri

VOL. 8. NO. 1 AGUSTUS 2022
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i2.XXXX

to understand relationships between words, regardless of their position in the sentence. For example, in the phrase "not very good," the backward LSTM identifies the negation ("not"), while the forward LSTM captures the positive sentiment of the word "good." This dual processing ability enables Bidirectional LSTM to handle long dependencies and extract sentiments from complex expressions effectively. In this study, the Bidirectional LSTM demonstrated strong performance, achieving high accuracy and efficiently classifying player reviews by capturing nuanced contextual information. These four methods are integrated to provide a comprehensive framework for text-based sentiment classification, aiming to deliver accurate and efficient results.

RESULTS AND DISCUSSION

A. Result

This study reveals variations in the performance of the four deep learning models tested. Below are the detailed results of the sentiment evaluation, highlighting the performance of each model in classifying player reviews for the AlgoRun application.

1. Accuracy Testing

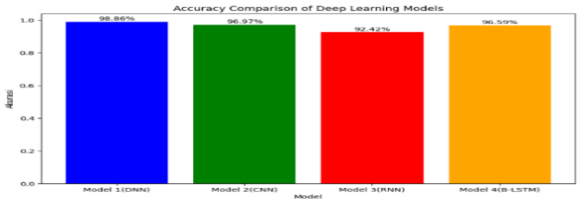
The accuracy testing results are presented in Table 8, with a visual representation shown in Figure 7. The findings highlight the varied performance of the four deep learning models. The DNN model with TF-IDF achieved the highest accuracy of 98.86%, outperforming the other models due to the simplicity and effectiveness of TF-IDF in handling reviews with common language patterns. The CNN model with Word Embeddings attained an accuracy of 96.97%, showcasing its ability to capture both local and semantic patterns in the text, although it was slightly less effective compared to DNN.

Table 8. Summary of accuracy metrics results

Model	Accuracy	Precision	Recall	F1-Score
DNN with TF-IDF	0.9886	0.9887	0.9886	0.9886
CNN with Word Embeddings	0.9697	0.9746	0.9697	0.9688
RNN with LSTM	0.9242	0.8841	0.9242	0.9002
Bidirectional LSTM	0.9659	0.9662	0.9659	0.9644

Source : (Research Result, 2024)

The Bidirectional LSTM achieved an accuracy of 96.59%, excelling at capturing the contextual relationships in word order, particularly in longer reviews. However, its performance slightly lagged behind that of DNN and CNN.



Source : (Research Result, 2024)

Figure 7. Visualization of summary of accuracy metric results

The RNN with LSTM recorded the lowest accuracy at 92.42%, suggesting that while it is effective in managing word order, it struggles to adapt to the diverse language patterns present in the reviews.

2. Epoch Value Testing

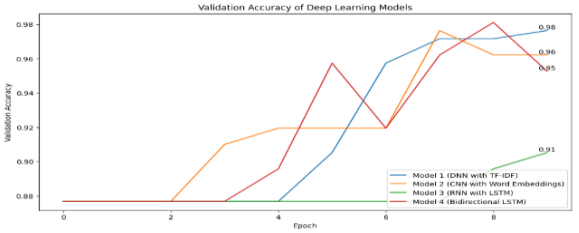
The accuracy per epoch test results are presented in Table 9. The table demonstrates that each of the four deep learning models used in the sentiment analysis of player feedback for AlgoRun exhibits a distinct pattern of performance improvement across epochs.

Table 9. Compare epoch value testing

Epoch	Model DNN	Model CNN	Model RNN	Model B-LSTM
Epoch 1	0.87677	0.87677	0.87677	0.87677
Epoch 2	0.87677	0.87677	0.87677	0.87677
Epoch 3	0.87677	0.87677	0.87677	0.87677
Epoch 4	0.87677	0.90995	0.87677	0.87677
Epoch 5	0.87677	0.91943	0.87677	0.89573
Epoch 6	0.90521	0.91943	0.87677	0.95734
Epoch 7	0.95734	0.91943	0.87677	0.91943
Epoch 8	0.97156	0.97630	0.87677	0.96208
Epoch 9	0.97156	0.96208	0.89573	0.98104
Epoch 10	0.97630	0.96208	0.90521	0.95260

Source : (Research Result, 2024)

The DNN model achieved the highest accuracy of 97.63% at the 10th epoch, highlighting its superior ability to utilize TF-IDF features for sentiment classification. The CNN model demonstrated consistent improvement starting from the 4th epoch and achieved an accuracy of 96.20% at the 10th epoch, showcasing its effectiveness in capturing spatial features through word embeddings. A visualization of these results is provided in Figure 8.



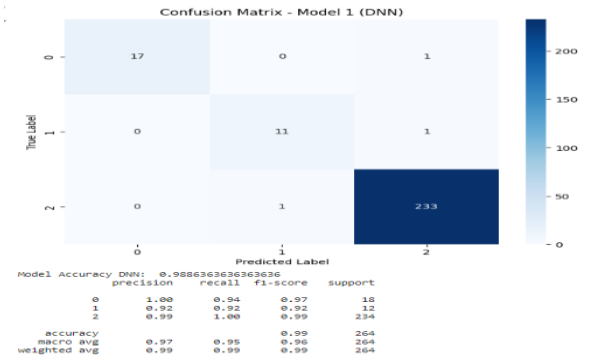
Source : (Research Result, 2024)

Figure 8. Compare epoch validation accuracy

The Bidirectional LSTM performed well, achieving an accuracy of 95.26% at the 10th epoch, demonstrating its strength in capturing the contextual relationships in word order within reviews. In contrast, the RNN with LSTM showed slower improvement, reaching a maximum accuracy of 90.52% at the 10th epoch, making it the lowest-performing model. Overall, these results indicate that DNN and CNN deliver superior performance in sentiment analysis, while Bidirectional LSTM remains robust in understanding contextual relationships between words.

3. Confusion Matrix

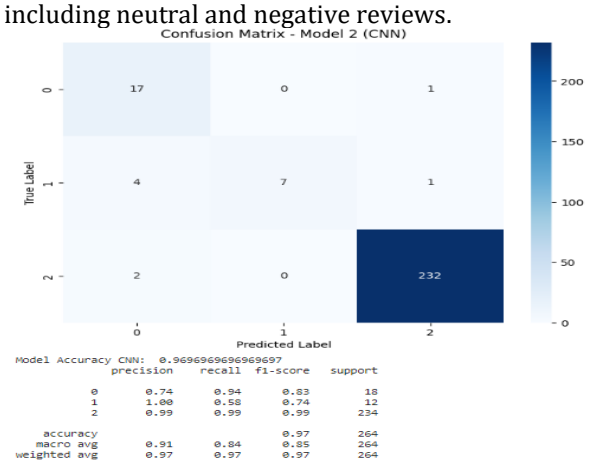
The confusion matrix is used to evaluate the performance of a classification model, providing a detailed overview of the model's ability to correctly and incorrectly predict class labels.



Source : (Research Result, 2024)

Figure 9. Confusion Matrix DNN

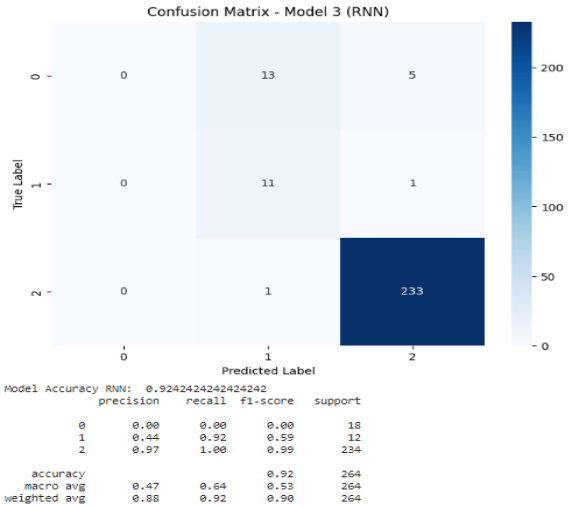
Based on the confusion matrix of the DNN (TF-IDF) model in Figure 9, the model achieved the highest accuracy of 98.86%, demonstrating exceptional performance in classifying positive reviews with near-perfect precision and recall. It also maintained stability across all classes, including neutral and negative reviews.



Source : (Research Result, 2024)

Figure 10. Confusion Matrix CNN

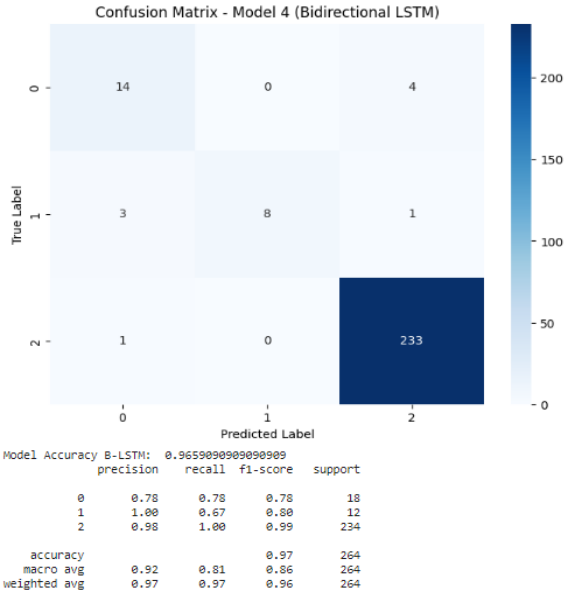
The CNN (Word Embeddings) model, as shown in Figure 10, achieved an accuracy of 96.97%. While it performed strongly in detecting positive reviews, it was slightly less effective than DNN in classifying neutral and negative reviews.



Source **:** (Research Result, 2024)

Figure 11. Confusion Matrix RNN

Based on the confusion matrix of Model 3 (RNN) in Figure 11, the model achieved an accuracy of 92.42%. It excelled in classifying class 2, with near-perfect precision (0.97) and recall (0.99). However, the model's performance was inconsistent across all classes. Notably, class 0 was not detected correctly, with both precision and recall at 0. Additionally, the model struggled with classifying class 1, achieving a precision of only 0.44. This performance indicates a strong bias toward the majority class (class 2), leading to an imbalance in classification accuracy across different classes.



Source : (Research Result, 2024)

Figure 12. Confusion Matrix B-LSTM

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti
No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri

VOL. 8. NO. 1 AGUSTUS 2022
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i2.XXXX

Based on the confusion matrix of Model 4 (Bidirectional LSTM) in Figure 12, the model achieved an accuracy of 96.59%. It performed exceptionally well in classifying class 2, with nearly perfect precision (0.98) and recall (1.00). Additionally, the model demonstrated solid performance on class 0, achieving a precision and recall of 0.78, although four instances were misclassified as class 2. However, the model's performance on class 1 was slightly weaker, with a precision of 1.00 but a recall of 0.67, indicating that the model struggled to consistently detect class 1.

Overall, the model exhibited stability across all classes, with macro average precision, recall, and F1-score exceeding 0.80. The weighted average precision and recall, both at 0.97, further highlight the model's ability to handle data imbalance effectively. All four models—DNN, CNN, RNN, and Bidirectional LSTM—performed well in the sentiment classification task, as evidenced by their high accuracy scores.

Table 10. Resume confusion matrix

Model	accuracy	Precision (average)	Recall (average)	F1-score (average)
DNN	High	High	High	High
CNN	High	High	Medium	Medium
RNN	Medium	Medium	Medium	Medium
B-LSTM	High	High	High	High

Source : (Research Result, 2024)

Based on the confusion matrix shown in Table 10, the models demonstrate a tendency to predict the majority class more accurately than the minority classes. This is evident from the lower recall values observed for the minority classes across all models.

B. Discussion

The results indicate that the DNN with TF-IDF excels in sentiment classification, achieving stable accuracy across all classes, making it the optimal choice for general language patterns. The CNN model demonstrates significant strength in capturing local features of the text, while the Bidirectional LSTM is particularly well-suited for handling sequential context in longer reviews. However, the RNN with LSTM, despite its strength in processing word order, shows a bias toward the majority class and is less effective for minority classes.

The DNN model with TF-IDF achieved the highest accuracy (98.86%) due to its ability to effectively capture global patterns in the text. TF-IDF, which emphasizes the importance of words based on their frequency across documents, proved to be highly effective for sentiment classification in this context. In contrast, the CNN model with word embeddings, while still achieving high accuracy (96.97%), excelled in capturing local patterns such as n-grams and phrases, making it suitable for reviews with more nuanced expressions.

The Bidirectional LSTM model, though slightly less accurate (96.59%), demonstrated its strength in handling sequential dependencies, particularly in longer reviews where context from both past and future words is crucial. However, the RNN with LSTM struggled with minority classes, likely due to its inability to handle imbalanced data effectively. This highlights the importance of choosing the right model based on the specific characteristics of the dataset.

The challenge of data imbalance is evident across all models, as reflected in their reduced performance on minority classes. Nonetheless, the bidirectional LSTM provides more balanced results compared to the RNN, indicating its relative effectiveness in addressing this issue.

CONCLUSION

This research evaluates the effectiveness of various deep learning models in sentiment analysis of player reviews for the game AlgoRun. The study utilizes TF-IDF and Word Embedding for feature extraction and applies four deep learning models—DNN with TF-IDF, CNN with Word Embeddings, RNN with LSTM, and Bidirectional LSTM—to determine the most effective model for classifying user sentiments into positive, neutral, and negative categories. The feature extraction process using TF-IDF and Word Embedding provided valuable insights into word frequency distributions, offering actionable data for game developers to enhance the AlgoRun platform and similar game-based learning applications. These insights support the customization of content and improvement of game quality to better align with user feedback.

At the model evaluation stage, the DNN model with TF-IDF features achieved the highest accuracy at 98.86%, followed by the CNN model, which recorded an accuracy of 96.97%. These findings highlight the superiority of DNN with TF-IDF in processing user review data within the context of game-based learning. While RNN and Bidirectional LSTM demonstrated comparatively lower performance, they still captured important aspects of sentiment analysis, particularly in understanding word order and sequential context. This research contributes significantly to the literature on game-based learning by identifying the most effective sentiment analysis model for application in educational games. Furthermore, it enhances understanding of player preferences and learning

experiences in the context of games, paving the way for future research in this domain.

For future research, exploring advanced transformer-based models such as BERT or GPT could improve sentiment analysis accuracy, especially in capturing deeper contextual relationships. Additionally, experiments with larger and more diverse datasets, including multi-language reviews, could validate the robustness of these models. Integrating real-time sentiment analysis into game-based learning platforms could also enable dynamic content adaptation, creating a more personalized and engaging learning experience for players.

REFERENCE

[1] K. R. dan T. (Kemendikbudristek) Kementerian Pendidikan, "Penggunaan teknologi digital dalam pembelajaran abad ke-21." Accessed: Sep. 28, 2024. [Online]. Available: https://kemendikbud.go.id

[2] R. Tariq, B. M. Aponte Babines, J. Ramirez, I. Alvarez-Icaza, and F. Naseer, "Computational thinking in STEM education: current state-of-the-art and future research directions," 2024, *Frontiers Media SA*. doi: 10.3389/fcomp.2024.1480404.

[3] L. K. Lee, T. K. Cheung, L. T. Ho, W. H. Yiu, and N. I. Wu, "Learning computational thinking through gamification and collaborative learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2020, pp. 339–349. doi: 10.1007/978-3-030-21562-0_28.

[4] L. M. Baso Intang Sappaile and B. F. F. A. S. M. B. M. Rudy Max Damara Gugat, "Dampak Penggunaan Pembelajaran Berbasis Game Terhadap Motivasi Dan Prestasi Belajar," *Jurnal Review Pendidikan dan Pengajaran*, vol. 7, no. 1, pp. 714–727, 2024.

[5] P. Mikrouli, K. Tzafilkou, and N. Protogeros, "Applications and Learning Outcomes of Game Based Learning in Education," *International Educational Review*, pp. 25–54, Mar. 2024, doi: 10.58693/ier.212.

[6] R. Israel-Fishelson and A. Hershkovitz, "Micro-persistence and difficulty in a game-based learning environment for computational thinking acquisition," *J Comput Assist Learn*, vol. 37, no. 3, pp. 839–850, Jun. 2021, doi: 10.1111/jcal.12527.

[7] M. Guenaga, A. Eguíluz, P. Garaizar, and J. Gibaja, "How do students develop computational thinking? Assessing early programmers in a maze-based online game," *Computer Science Education*, vol. 31, no. 2, pp. 259–289, Apr. 2021, doi: 10.1080/08993408.2021.1903248.

[8] H. Huang and Y. Li, "Exploring the Motivation of Livestreamed Users in Learning Computer Programming and Coding," *The Electronic Journal of e-Learning*, vol. 19, no. 5, pp. 363–375, 2021, [Online]. Available: www.ejel.org

[9] E. Rowe *et al.*, "Assessing implicit computational thinking in Zoombinis puzzle gameplay," *Comput Human Behav*, vol. 120, Jul. 2021, doi: 10.1016/j.chb.2021.106707.

[10] T. Guzsvinecz and J. Szűcs, "Length and sentiment analysis of reviews about top-level video game genres on the steam platform," *Comput Human Behav*, vol. 149, p. 107955, Dec. 2023, doi: 10.1016/j.chb.2023.107955.

[11] X. Li, Z. Zhang, and K. Stefanidis, "A Data-Driven Approach for Video Game Playability Analysis Based on Players' Reviews," *Information*, vol. 12, no. 3, p. 129, Mar. 2021, doi: 10.3390/info12030129.

[12] Y. Yu, D. T. Dinh, B. H. Nguyen, F. Yu, and V. N. Huynh, "Mining Insights From Esports Game Reviews With an Aspect-Based Sentiment Analysis Framework," *IEEE Access*, vol. 11, pp. 61161–61172, 2023, doi: 10.1109/ACCESS.2023.3285864.

[13] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A Three-Step Fuzzy-Based BERT Model for Sentiment Analysis," 2022, pp. 41–52. doi: 10.1007/978-981-19-0489-9_4.

[14] A. G. Ganie and S. Dadvandipour, "Traditional or deep learning for sentiment analysis : A review," *Multidiszciplináris tudományok*, vol. 12, no. 1, pp. 3–12, 2022, doi: 10.35925/j.multi.2022.1.1.

[15] R. G. A. Pratama and N. Cahyono, "Comparison Of Deep Learning Methods On Sentiment Analysis Using Word Embedding," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 1, pp. 1–8, Jul. 2024, doi: 10.33480/jitk.v10i1.5280.

[16] M. Totox and H. F. Pardede, "Exploring the Effectiveness of Deep Learning in Analyzing Review Sentiment," *JIKO (Jurnal Informatika dan Komputer)*, vol. 6, no. 2, Aug. 2023, doi: 10.33387/jiko.v6i2.6372.

[17] A. Nurdina and A. B. I. Puspita, "Naive Bayes and KNN for Airline Passenger Satisfaction Classification: Comparative Analysis," *Journal of Information System Exploration and Research*, vol. 1, no. 2, Jul. 2023, doi: 10.52465/joiser.v1i2.167.

[18] H. Zhang, J. Xu, L. Lei, Q. Jianlin, and R. Alshalabi, "A Sentiment Analysis Method Based on Bidirectional Long Short-Term Memory Networks," *Applied Mathematics and*

VOL. 8. NO. 1 AGUSTUS 2022
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i2.XXXX

Nonlinear Sciences, vol. 8, no. 1, pp. 55–68, Jan. 2023, doi: 10.2478/amns.2022.1.00015.

[19] I. Muhammad, F. A. Triansyah, A. Fahri, and A. Gunawan, "Analisis Bibliometrik: Penelitian Game-Based Learning pada Sekolah Menengah," 2023. [Online]. Available: https://jiped.org/index.php/JSP

[20] X. Rubio-Campillo, K. Marín-Rubio, and C. Corral-Vázquez, "La evaluación del Aprendizaje Basado en Juegos en contextos informales mediante ciencia de datos," *Revista Latinoamericana de Tecnología Educativa - RELATEC*, vol. 23, no. 2, pp. 9–26, Jul. 2024, doi: 10.17398/1695-288x.23.2.9.

[21] N. Punetha and G. Jain, "Bayesian game model based unsupervised sentiment analysis of product reviews," *Expert Syst Appl*, vol. 214, p. 119128, Mar. 2023, doi: 10.1016/j.eswa.2022.119128.

[22] M. Waqas and U. W. Humphries, "A critical review of RNN and LSTM variants in hydrological time series predictions," *MethodsX*, vol. 13, p. 102946, Dec. 2024, doi: 10.1016/j.mex.2024.102946.

[23] "Effects of Preprocessing on Text Classification in Balanced and Imbalanced Datasets," *KSII Transactions on Internet and Information Systems*, vol. 18, no. 3, Mar. 2024, doi: 10.3837/tiis.2024.03.004.

[24] Aji Gautama Putrada; Nur Alamsyah; Mohamad Nurkamal Fauzan, "BERT for Sentiment Analysis on Rotten Tomatoes Reviews," *2023 International Conference on Data Science and Its Applications (ICoDSA)*, 2023.

[25] N. Alamsyah, T. P. Yoga, and B. Budiman, "Improving Traffic Density Prediction Using Lstm With Parametric ReLU (PReLU) Activation," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 9, no. 2, pp. 154–160, Feb. 2024, doi: 10.33480/jitk.v9i2.5046.

[26] S. F. Pane, J. Ramdan, A. G. Putrada, M. N. Fauzan, R. M. Awangga, and N. Alamsyah, "A Hybrid CNN-LSTM Model With Word-Emoji Embedding For Improving The Twitter Sentiment Analysis on Indonesia's PPKM Policy," in *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, IEEE, Dec. 2022, pp. 51–56. doi: 10.1109/ICITISEE57756.2022.10057720.