Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

Comparison of XGBoost and LSTM in Knowledge Discovery for GrokAI Mobile Application Sentiment Analysis

Aliyananda Risyahputri, Dedy Kurniawan*, Ken Ditha Tania

Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Sriwijaya, Palembang, Indonesia
Email: [1]aliyanandar04@gmail.com, [2,*]dedykurniawan@unsri.ac.id, [3]kenya.tania@gmail.com
Correspondence Author Email: dedykurniawan@unsri.ac.id
Submitted: **05/11/2025**; Accepted: **06/12/2025**; Published**: 08/12/2025**

Abstract−Generative AI has provided real benefits in key sectors of the public sector. However, the rapid expansion of AI assistant services also raises concerns about whether newly released products can consistently meet user expectations, especially as negative experiences are increasingly expressed through public reviews. Its positive impacts encourage competitive rivalry among AI assistant product developers, including xAI, which also participates by formulating the Grok AI application. As a relatively new product with over 50 million downloads, GrokAI needs to perform an evaluation to maintain its competitiveness. This condition leads to the research goal of analyzing user sentiment toward GrokAI application through reviews on Google Play Store and comparing the performance of Machine Learning and Deep Learning classification models within the framework of Knowledge Discovery in Databases (KDD). This study uses 11,108 review data classified using the VADER Lexicon method, resulting in 7,633 positive reviews and 3,475 negative reviews. The data is then tested on XGBoost (Extreme Gradient Boosting) and LSTM (Long-Short Term Memory) models. The results show that the XGBoost model performs slightly better with an accuracy of 87.22%, compared to LSTM, which reaches 86.58%. However, both models exhibit significant performance disparities in classifying negative classes due to the extreme difference in data quantity. The knowledge discovery process reveals that the majority of positive sentiment appreciates the free access and general functions of the application. Meanwhile, negative sentiment focuses on complaints related to response time, output quality, and specific features such as image and voice. The main recommendation is to maintain the advantage of free access also improve features and processing logic to sustain loyalty and service quality. Future research is suggested to test models with more balanced data and optimize dataset cleaning to improve accuracy in minority classes.

Keywords: XGBoost; LSTM; Grok AI; Knowledge Discovery; Sentiment Analysis

1. INTRODUCTION

The contribution of artificial intelligence nowadays is not only centered on the technology sector but has also touched other critical lines within the public sector. This statement is proven by the emergence of artificial intelligence (AI) innovations that also serve as virtual assistants in problem solving, one of which is Generative AI, which significantly improves management decision accuracy through strong predictive capabilities and data analysis [1]. According to research by Albashrawi in 2025, Generative AI also substantially increases decision-making accuracy in the business and healthcare domains [2]. Furthermore, Generative AI has a considerable part in productivity. This is supported by Noy and Zhang study in 2023, which states that the use of Generative AI, specifically ChatGPT, can enhance productivity in writing tasks assigned to higher education professional staff [3]. Other research findings also indicate that AI use can increase worker productivity in problem solving, with an average increase of 15% above usual levels [4]. In addition to the lucrative results by its massive penetration, the dilemma arises to ensure that the innovations offered by Generative AI adapt well in the hands of its users. Over time, users' expectations for the application will also become more complex, so that if these expectations fail to be fulfilled, it can have an impact on reducing trust and desire to adopt the service. This is in line with research revealed by Huynh and Aichner in 2025, that trust in Generative AI is strongly driven by users' perceptions of fairness, social presence, and the interaction experience provided [5]. So, when these expectations are not met, users will tend to list negative evaluations that are ultimately reflected through app reviews.

The positive results brought by the existence of Generative AI in human life have driven software industry players to compete in launching AI assistant services with the latest versions. This competitive rivalry is evident with the emergence of similar products, such as ChatGPT developed by OpenAI, Gemini by Google, Perplexity, and other AI assistant products that can produce outputs in various fields using images, sounds, text, and videos [6]. As part of the evolution of Generative AI, Elon Musk's AI development company, xAI, also formulated Grok, a product initially aimed at real time data processing and humorous conversations on social media X [7]. After its first release in November 2023 [8], xAI officially expanded Grok's accessibility by launching a dedicated app for Android and iOS users. The transformation of Grok into an app introduces a new dynamic, as its relatively young status offers an important space for study within the context of Generative AI development. This hypothesis is supported by data indicating that Grok AI has gained over 50 million downloads and more than 1.43 million reviews on the Google Play Store since its initial release[9], making it an interesting object for further research.

The presence of these various products indicates that the trend of AI assistants based on chatbots and generative AI can lead users to become more selective in choosing services that can become market leaders. This condition indirectly encourages Grok AI, as a relatively new platform in the industry, to continuously evaluate its application performance to maintain its value and competitiveness with other similar applications. Therefore, sentiment analysis by grouping user comments and concerns about the application emerges as a concrete step to address this urgency. In

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

line with what was stated by Vadla in their 2024 research, one of the important things that industry sectors can consider recently is gathering user reviews to identify their expectations regarding product usage [10]. User and application interaction data is unavoidable because when people use social media platforms to share their thoughts, opinions, experiences, and feelings, a large portion of the data will be in unstructured text but it still contains insightful information [11]. These user review data can be accumulated through the Google Play Store or Apple Store and then processed using machine learning approaches to extract sentiments that represent users' feelings [12].

The user sentiment analysis approach based on review data has been applied in several previous studies. [13] highlights user sentiment for the Digital Identity application through 5,000 reviews on the Google Play Store from September 2022 to December 2023 using the K-Nearest Neighbors classification algorithm. The research results show that the KNN algorithm successfully classified with good accuracy, and user perceptions of the application tend to contain negative stigma. In the study conducted by Kulsum in the year 2022, the Support Vector Machine (SVM) algorithm was applied to reviews of the streaming app WeTV, dividing the testing scenarios into four based on the percentage of training and testing data. The third scenario produced the most excellent outcomes with 80% training and 20% testing data ratio, where positive user sentiment focused on the diversity of available series, and negative sentiment stemmed from user complaints about several app services [14]. Using two different methods and focusing on comparing algorithm performance, research in 2024 by Damayanti analyzed sentiment toward 1,000 user comments on the Alfagift app, a minimarket delivery service in Indonesia. In contrast to previous research, a lexicon-based method was used for data labeling prior to classification using SVM and Long-Short Term Memory (LSTM) algorithms. This experiment indicated that SVM's accuracy was slightly better than LSTM and suggested increasing the amount of data for future research [15].

Research by Setiawan and Nastiti in 2024 compares the performance of the XGBoost, SVM, and Extra Trees Classifier algorithms in classifying 50,000 user reviews of the DANA application. Equipped with the SMOTE oversampling technique, XGBoost and SVM achieved superior results with an accuracy score of 93%. It is recommended to gather a more diverse dataset and try other classification methods in future research [16]. Focusing on a more specific object of study related to Generative AI applications, Prasetyo in his study in the year 2025, used 800 social media review samples to develop a hybrid deep learning procedure utilizing CNN and Bi-LSTM for sentiment analysis of Indonesian language ChatGPT text. With preprocessing and Word2Vec CBOW word embeddings, this hybrid model achieved an accuracy of 95.24%, precision of 95.09%, recall of 95.15%, and F1 score of 95.99%. It is suggested to explore larger datasets and transfer learning to improve generalization in the future [17]. On the other hand, in 2024 Widaad and Anggraini analyzed user sentiment toward the ChatGPT application by comparing the performance of SVM and CNN algorithms. Based on an analysis of 155,529 reviews from the Google Play Store, which involved preprocessing steps such as emoji removal, case folding, and stemming, the Support Vector Machine (SVM) model achieved an accuracy of 85% (precision 0.83, recall 0.55, F1-score 0.58). Meanwhile, the Convolutional Neural Network (CNN) attained an accuracy of 84% (precision 0.68, recall 0.59, F1-score 0.62), with 75% of the reviews classified as positive. Future research is recommended to enhance application features based on user feedback [18].

Unlike previous research that generally focused on the classification of application sentiment using machine learning and deep learning models separately and focused on comparing algorithm performance only, this study will combine the comparison of the two learning models and add a knowledge discovery context to find sentiments that users often raise about applications. If the previous research focused on accuracy results, then *knowledge discovery* was applied as the final result of knowledge from the dataset in this study in order to reveal whether there are certain patterns based on words that appear that can be used as a basis for application improvement in the future. Although there have been previous studies that have analyzed opinions on applications such as Generative AI, in fact direct comparisons between XGBoost and LSTM in mobile-based Generative AI application objects are still rare, especially in Grok AI which has just transformed from just a feature to a separate application. In addition, the research space to utilize English-language reviews is also still open because it can capture universal perceptions from global users.

Based on this gap, this study is directed to apply the Knowledge Discovery in Database framework to Grok AI user reviews collected through the Google Play Store by testing XGBoost and LSTM in the sentiment classification process that has been categorized through the VADER (Valence Aware Dictionary and Sentiment Reasoning) Lexicon method. In addition to comparing the accuracy of the algorithm model, this research also aims to uncover word patterns that often appear as magic knowledge from the dataset so that it can be the basis for future application improvement. Through this approach, this research is expected to provide an overview of user perception while contributing to enriching the academic literature related to the comparative effectiveness of machine learning and deep learning methods in the context of AI-based application sentiment analysis.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The process in this study refers to the Knowledge Discovery in Database (KDD) framework. The KDD approach is considered advantageous for processing data to produce valid, up-to-date information that potentially has meaning based on patterns understood from large databases [19]. Additionally, it is stated that this research framework can

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

help systematically translate user sentiment based on data, making it easier to gain a deeper understanding of the impressions left [20]. Broadly speaking, the KDD stages are divided into several parts [21], namely the Extract, Transform, Load (ETL) process, which comprises collecting, integrating, cleansing, reducing, and transforming the data. These stages proceed to the main process, which is data mining, and end with data interpretation. In the context of sentiment analysis experiments, the ETL stage involves data cleaning steps such as stopword removal, tokenization, stemming, and data transformation into a format suitable for machine learning [22]. In previous research, the preprocessing or ETL stages are described along with emoji removal, case folding, and data normalization [18]. Regarding the transformation or restructuring stage, it will be divided into two methods suitable for each machine learning and deep learning algorithm. To illustrate more specifically, the KDD flow implemented in this study is as follows in Figure 1.

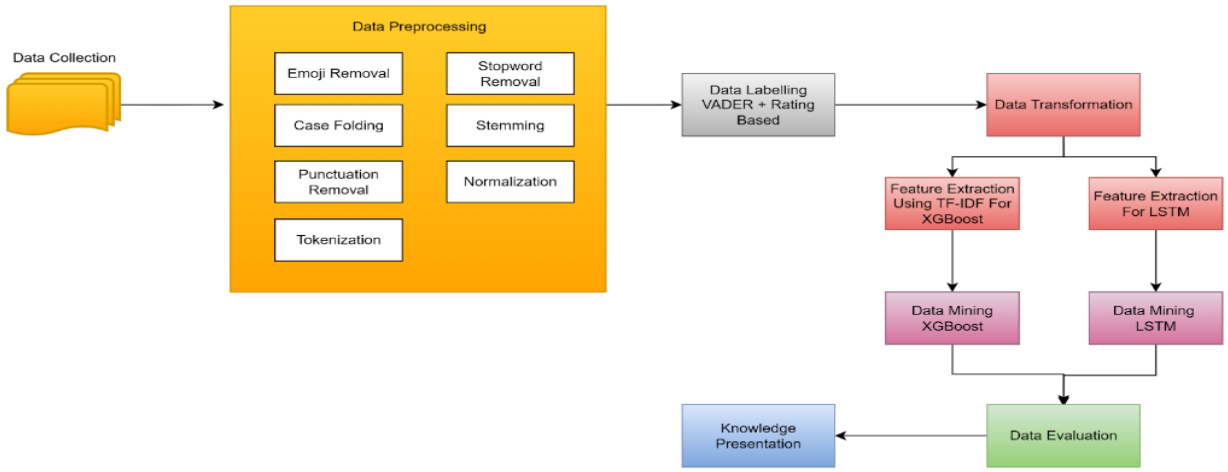

Figure 1. Research Flow Diagram

2.2 Data Collection

This research will focus on sentiment analysis of review data for the Grok AI application, totaling 11,245 user perceptions. The data acquired originates from the Google Play Store, comprising reviews written in English. The collection of reviews was conducted through data scraping performed on the Google Colaboratory platform utilizing Python as the programming language. The criteria applied during the review scraping process are as follows:

a. Reviews selected are in English.
b. Data retrieval was conducted randomly per rating with the “Newest” filter over a six month period, from April 2025 to October 2025.
c. The features or columns extracted for raw data are ‘score,’ representing the user rating of the application, and ‘content,’ which contains the review text expressed by the user.

The choice of an English reviews dataset is due to the limited number of Indonesian reviews, which only amounts to 2,463 reviews. In contrast, English language reviews are more diverse. Several studies mentioned in the previous chapter indicate that to achieve more stable model accuracy, a varied dataset is necessary during model training. Additionally, the decision to use English reviews is expected to yield sentiment outputs that reflect Grok AI users globally. The data preview is shown in Table 1 below.

Table 1. Data Collection Results

Content	Score
bad app 😠😠	1
worst	1
Extremely resource intensive. It's somewhat understandable because I am using a low end device. So it is to be somewhat expected, however, I still find it to be beyond excessively laggy and nearly unusable. I can't even type without grok temporarily freezing and gboard freezing after I type just a few words. In short, It works, just not very well.	1
…….	…..
very nice 👍	5
the best one out there ...no filters !!!	5
Fantastic app. I'm able to bring to life old photos of loved ones.	5

2.3 Data Preprocessing

Preprocessing is an important step that needs to be carried out in machine learning based experiments in general. Data that has been collected from certain sources is still partly raw, so it is considered necessary to go through quality

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

enhancement stages by adjusting structure and format [23]. This stems from the reality that low data quality has a crucial impression on the performance of learning and data mining conducted by algorithm models, as said by Ramirez-Gallego in [23]. Therefore, this study will implement a preprocessing phase as described below.

a. Emoji Removal
   The first step in the preprocessing process in this study is emoji removal, which is the stage of deleting emoticons or emojis found in review sentences. This aims to ensure that during data transformation, review data can be processed properly and focus on extracting meaning from the review sentences written by users.
b. Case Folding
   Case folding, also known as converting text to lowercase, is one of the basic preprocessing methods that functions to unify the representation of words. This process consolidates variations of identical words with different spellings, thereby reducing the number of features that need to be processed [24]. This approach has become a common practice in various Natural Language Processing (NLP) studies. In this research, words containing capital letters will be transformed into uniform lowercase format to improve the quality of the processed data.
c. Punctuation Removal
   Punctuation removal is a step aimed at minimizing data noise because most punctuation marks do not contribute to sentiment analysis [24]. Punctuation marks that are removed include exclamation marks (!), periods (.), commas (,), slashes (_), question marks (?), and others.
d. Stopword Removal
   After the sentence is cleaned of punctuation, the text will undergo stopword removal, where common functional words that do not provide relevant information in the main process will be eliminated. This removal aims to reduce data dimensionality because these words do not affect the sentiment polarity of reviews [25]. The stopwords to be removed include words like “which,” “and,” “in,” “is,” “the,” and similar words, leaving only words that are meaningful for sentiment analysis [26].
e. Stemming
   The final stage of data cleaning is stemming, also often called lemmatization, which is a method to reduce words to their root form by cutting affixes according to certain rules or utilizing a lexical database to find the base form of a word [27]. For instance, the root word "search" will be extracted from the term "searching”. This is done to maximize results when the data enters the data mining process.
f. Normalization
   The text that has gone through the previous six simplification stages will undergo normalization, which consolidates it into a format aligned with subsequent analysis processes. In this case, the final results of the stemming process will be combined with review score features in CSV format as before.

In addition to going through the above stages, review data also experience changes in the number of reviews across each score category. The decrease in the number of reviews across all scores is caused by the process of removing null (empty) values in columns. Overall, the total number of reviews before preprocessing was recorded at 11,245 data points, while after preprocessing it became 11,108 reviews. This indicates a reduction of 137 rows, which are considered anomalies. The results and stages of preprocessing that have been described will be presented in Table 2 below.

Table 2. Data Preprocessing Results

Preprocessing Stages	Results
Raw Data	It's very bad. After searching 3-5 message then I seen a pop-up indicator in my phone whenever I searched any messages? 😟
Emoji Removal	It's very bad. After searching 3-5 message then I seen a pop-up indicator in my phone whenever I searched any messages?
Case Folding	it's very bad. after searching 3-5 message then i seen a pop-up indicator in my phone whenever i searched any messages?
Punctuation Removal	its very bad after searching 35 message then i seen a popup indicator in my phone whenever i searched any messages
Stopword Removal	bad searching 35 message seen popup indicator phone whenever searched messages
Stemming / Lemmatization	bad search 35 message see popup indicator phone whenever search message

2.4 Data Labelling

In line with the ultimate goal of this research, which is to map perceptions based on reviews provided by users using sentiment analysis methods, a data labeling stage is necessary so that machine learning algorithm models can recognize reviews effectively. Data classification labeling is divided into two groups: ternary, which categorizes reviews as positive, negative, and neutral; and binary, which includes only positive and negative reviews [28]. The tendency to use one over the other depends on the analysis objectives of the study. For example, binary classification is often preferred because the presence of a neutral class is considered to reduce model accuracy. As stated by Valdivia in 2025 study by Tzimiris, neutral reviews are often discarded because they are considered to have ambiguous meaning, thus acting as noise that can lower the accuracy of binary based classification [29]. Therefore, this study

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

chooses to label reviews as positive and negative to produce more clearly segmented knowledge, which can provide more targeted input for application development and improvement. Based on these considerations, this experiment applies the VADER (Valence Aware Dictionary and Sentiment Reasoning) method, whose efficiency in automatically labeling data has been recognized [30]. VADER works with a lexicon-based approach that captures words in reviews and determines their polarity based on predefined values. Generally, VADER has a threshold that classifies data into positive, negative, and neutral labels, which is a score calculated based on the valence of each word in the lexicon, further called the compound score. If the score exceeds 0.05, it is categorized as positive, less than -0.05 it is classified as negative, and if the compound score falls between 0.05 and -0.05 it is considered neutral [31]. Since VADER adopts a ternary labeling as its default classification, but the research aims to focus on positive and negative reviews, this study will re-validate reviews labeled as neutral based on Google Play Store ratings. Neutral reviews with scores from 1 to 3 will be categorized as negative, while ratings of 4 and 5 will be considered positive reviews [32]

2.5 Data Transformation

Before being tested on the XGBoost and LSTM models, data that has undergone the labeling stage will move on to the preparation stage. For the XGBoost algorithm, the sentiment labels are first converted into numerical form, with 1 for positive and 0 for negative. Next, the 'cleaned_content' column containing review texts will be defined as features (X), and the 'label_numeric' sentiment column as the target (y) to be used for training the model. The model will then be trained and evaluated on the complete dataset that has been divided, with this study employing an 80:20 split, which means 80% training data and 20% testing data. Since machine learning cannot classify text-based data directly, the TF-IDF (Term Frequency-Inverse Document Frequency) approach will be used to convert the reviews into numerical representations, with training and testing data separated. The parameters used in TF-IDF are max_features = 5000, which means retaining the 5,000 most unique words with the highest TF-IDF values for machine learning model training. TF-IDF is a method implemented to recognize the most important word sequences in a dataset [33]. The TF-IDF weighting can be measured using the following techniques, using the overall amount of documents in the document compilation (N) and the number of documents containing the target word (n).

$$TF = \frac{(number\ of\ occurrences\ of\ a\ word\ in\ a\ document)}{(number\ of\ words\ in\ a\ document)} \quad (1)$$

$$IDF = log\frac{N}{n} \quad (2)$$

For the deep learning algorithm LSTM, the steps involved have slight differences in their procedures. The 'cleaned_content' and 'label_vader' columns from the labeling process will be converted into string data types and transformed into lists. This aims to ensure that the data can be properly converted into token form. Next, tokenization is performed using the Tokenizer() function from the Keras library. Tokenization is a typical method for breaking down a text into little components known as tokens. One of the simplest methods is word-based tokenization, where separation is done based on spaces. In this approach, each token generally represents a single word, although without additional preprocessing steps, tokens may still contain punctuation or irrelevant characters [34]. This process allows each word to be converted into a unique numerical index according to its order of appearance in the text dataset, so it can be processed by the LSTM model. After the tokenization step, the text is converted into a sequence of numbers using the text_to_sequence () function and standardized in length using pad_sequences (), adding zeros to shorter texts. Similar to XGBoost, the processed data is divided into training and testing sets with an 80:20 ratio to assess the model's performance.

2.6 Data Mining

XGBoost is an algorithm with a complexity exceeding that of gradient boosting, which combines decision tree-based model learning with linear model splitting. XGBoost has an advantage in speed because it can perform parallel processing on a single processor. Additionally, XGBoost is equipped with cross-validation features and the ability to detect the most influential variables[35]. The formulation of the XGBoost algorithm model can be seen in equation (3) below [16]

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i y_i^{\wedge(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (3)$$

Unlike XGBoost, which is a conventional machine learning approach, LSTM is a derivative architecture adopted from recurrent neural networks or RNNs in the field of deep learning[36]. The reason for choosing the LSTM model is based on its ability to capture and analyze information over long periods, so it is expected to perform tracking of sentiment patterns within review texts [37]. Previous study mention that RNN-derived algorithms like LSTM tend to be designed to learn the sequential context of words within a text [38]. This is quite relevant because it can maximize the contextual understanding of sentences to be classified. LSTM cells store information from the past through three main gates, which are the input, forget, and output. Sutskever, in their research stated that these three gates control the distribution of information within the cell, allowing the cell to control how much past information should be retained, deleted, or passed on to the next process [39]. Mathematically, each gate stage in Long-Short Term Memory namely the forget gate, input gate, output gate, candidate memory cell, representation of the current cell state, and the final

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

output in the form of the current hidden state [40], are sequentially described in equations (4), (5), (6), (7), (8), and (9) as follows.

$$f_t = \delta\left(W_f[h_{t-1}, X_t] + b_f\right) \quad (4)$$

$$i_t = \delta(W_i[h_{t-1}, X_t] + b_i) \quad (5)$$

$$O_t = \delta(W_o[h_{t-1}, X_t] + b_o) \quad (6)$$

$$L_t = tanh(W_o[h_{t-1}, X_t] + b_c) \quad (7)$$

$$C_t = F_t \times C_{t-1} \times i_t \times L_t \quad (8)$$

$$h_t = O_t \times \tanh(C_t) \quad (9)$$

2.7 Data Evaluation

When both models that have been built are trained and tested using training data and testing data, it is vital to assess the model's performance in classifying and predicting data. This evaluation stage can be measured using classification report parameters, including Accuracy, Recall, Precision, and F-1 Score. Calculating these parameters involves using values from the confusion matrix, which displays the model's prediction outcomes on both the test and training data. The value produced by the confusion matrix includes the amount of positive reviews accurately classified by the model (true positive), positive reviews misclassified by the model (false positive), the amount of negative reviews successfully categorized by the model (true negative), and also negative reviews misclassified by the model (false negative). Each of these parameters has its own calculation formula, which are explained as follows [41]

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (11)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (12)$$

$$F1 - Score = \frac{TP}{TP+FP} \times 100\% \quad (13)$$

2.8. Knowledge Presentation

The knowledge presentation stage is a manifestation of the knowledge discovery process aimed at visually presenting the information from dataset analysis results. Knowledge discovery, which in this context is commonly produced through text mining since the past, can be represented in the form of calculating how often a word appears, grouping words, sentiment analysis, and topic extraction from a collection of texts[42]. Several previous studies have successfully applied knowledge discovery across various sectors as a basis for decision-making, such as Pratiwi in 2025 who successfully clustered topics in GoPartner reviews as an application in the transportation sector [43], study by Sofiah in 2024 which reached the environmental sector by comparing the performance of SARIMA and LSTM to extract knowledge related to the Gurugram district quality index [44], and Putri in 2025 which ventured into the fintech sector by comparing topics in BCA Mobile and MyBCA reviews [45]. Through this method, the research will display knowledge from reviews that have been extracted and undergo sentiment identification processes using the VADER technique and Rating to serve as knowledge for application improvement. The information is presented in the form of a Word Cloud Visualization that contains review words with high frequency in a graphical visual form[46]. The visualization results will be analyzed in such a way that they become a useful knowledge representation for the GrokAI application developers to conduct data-driven evaluations, as well as serve as a basis for providing more targeted and evidence-based development recommendations.

3. RESULT AND DISCUSSION

3.1 Data Labelling Results

For the total of 11,108 data entries, labeling and grouping of reviews were carried out by implementing the VADER method. VADER assigns weights to each word in the review so that each line receives positive, negative, neutral, and compound scores. The compound score is calculated based on the sentiment content of the review sentence as the basis for determining the final label. As referenced in the previous chapter's theory, a compound score ≥ 0.05 is grouped as positive, a score ≤ −0.05 is sorted as negative, and scores among these two thresholds are recognized as neutral. As shown in the preview in Table 3, reviews with the phrase "bad app" have a compound score of -0.5423 and are labeled negative, while reviews like "best one filter" with a compound score of 0.6369 are labeled positive. Reviews with a compound score of 0.0000 are categorized as Neutral.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

Table 3. VADER Labelling Score Results

cleaned_content	Positive	Negative	Neutral	Compound	label_vader
bad app	0.000	0.778	0.222	-0.543	Negative
app waste dont download app time want	0.143	0.308	0.549	-0.3612	Negative
prompt response near professional accurate	0.00	0.000	1.000	0.0000	Neutral
direct answer	0.00	0.000	1.000	0.0000	Neutral
…..	…..	…..	……	…..	……
best one filter	0.677	0.000	0.323	0.6369	Positive
amaze	1.000	0.000	0.000	0.5423	Positive

Next, 2,174 data points labeled Neutral by VADER were revalidated using the original rating scores obtained during the scraping stage. Validation was carried out by combining neutral data with rating scores of 1, 2, and 3 into the Negative category, while scores of 4 and 5 were grouped into the Positive category. This step was taken because the study focused only on the two main labels, Positive and Negative, so the Neutral labels assigned by VADER were not included in the further analysis.

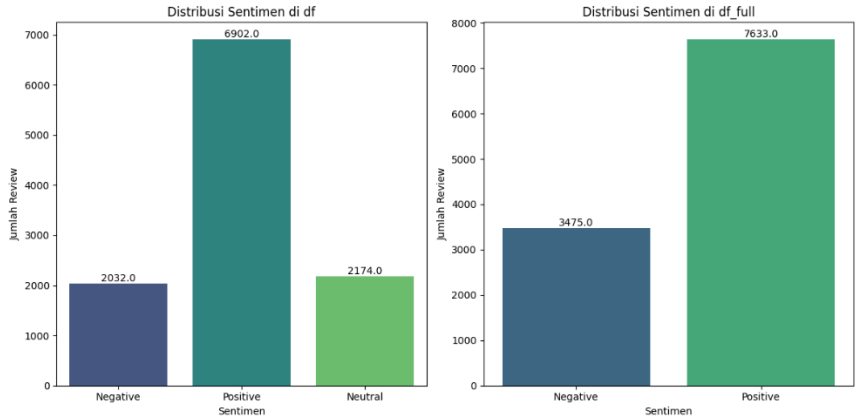

Figure 2. Distribution of Reviews Before and After Neutral Validation

Based on Figure 2, before validation using ratings, the VADER label data showed a distribution of 6.902 positive reviews, 2.032 negative reviews, and 2.174 neutral reviews. After the validation process, which involved combining the original ratings with the neutral data, the number of positive reviews increased to 7.633, while negative reviews also rose to 3.475. The neutral category was no longer found because all data had been adjusted to the two main classifications. This change in distribution indicates that the validation process successfully strengthened the representation of positive and negative sentiments more clearly, while also eliminating potential ambiguity in the data previously labeled as neutral. Therefore, the validated dataset aligns with the research objectives and has a better balance for use in the next sentiment analysis stage.

3.2 Data Mining Results

The data processing for both algorithms starts by transforming sentiment labels into numerical values, with Positive as 1 and Negative as 0. To assess the algorithm's generalization capacity, the dataset is then divide into 80% training data totaling 8.886 samples, and 20% testing data comprising 2.222 samples. In the XGBoost algorithm, the review content functions as the feature, while the label column serves as the target. These text features are transformed by applying the TF-IDF method with "max_features=5000", ensuring only the 5.000 words with the highest TF-IDF weights are employed as features. The transformation results confirm that the feature dimensions for both training and testing data align with these parameters. In contrast, for the LSTM algorithm, the review content column is converted into a list of strings so that it can be processed by the model using the Tokenizer() function from Keras. Subsequently, the text is transformed into sequences of numbers and standardized in length by adding zeros to shorter texts before being used as input to the model.

Entering the core of the classification process, the XGBoost model training stage begins with defining the XGBoost algorithm using predetermined parameters tailored to the research needs. The parameters used include objective = 'binary:logistic' to handle binary classification problems, the number of decision trees built with the n_estimators parameter set to 300, learning_rate = 0.1 to control the model's learning speed, and the random state value set to the default number 42. After initializing the parameters, the training data resulting from TF-IDF transformation is applied along with numerical sentiment labels. This training process produces a model capable of learning the pattern of relationships between text feature representations and the pre-determined sentiment labels, which can then be used to predict sentiment on test data.

After conducting prediction tests based on the training data, the XGBoost algorithm model showed a confusion matrix as displayed in Figure 3. This indicates that the model successfully classified 601 negative test data and 1,319

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

positive test data correctly. There were 94 positive data misclassified as negative and 208 negative data misclassified as positive. Compared to LSTM, the XGBoost model has a higher number of true negatives, indicating better performance in detecting negative classes. However, the larger number of false positives suggests that this model still often misidentifies negative data as positive.

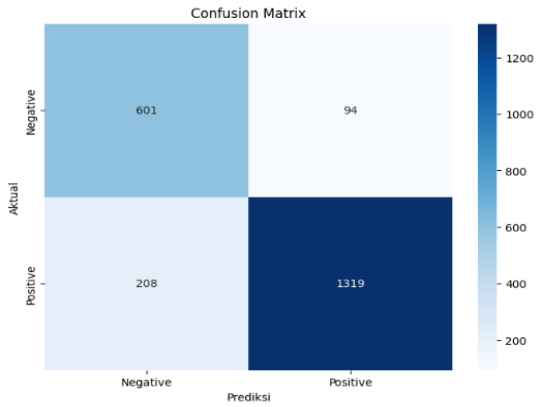

Figure 3. Confusion Matrix XGBoost

On the other hand, the confusion matrix in Figure 4 shows that the LSTM model successfully classified 562 negative data points and 1.362 positive data points correctly. Meanwhile, there are 153 positive data points wrongly classified as negative (false negatives) and 145 negative data points misclassified as positive (false positives). These findings show that the LSTM model has a rather decent capacity to distinguish both classes., although there are still some misclassifications, especially with negative cases being predicted as positive.

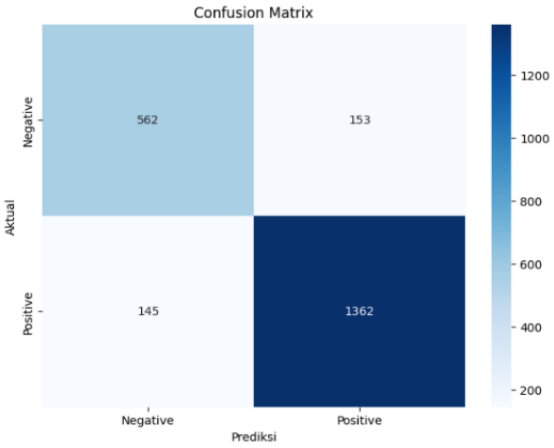

Figure 4. Confusion Matrix LSTM

In the next stage, the four values in the confusion matrix for each model are calculated to produce the classification report percentages. As shown in Table 4, specifically, the performance of both models is similarly lower for the negative class. This condition may be caused by data imbalance, where the number of negatives is fewer than positives, leading to difficulty in prediction. For the negative class, the LSTM model records balanced evaluation values of 79%, which suggests consistent detection ability but still making errors in distinguishing negative reviews from positive ones. Meanwhile, XGBoost shows a higher recall, meaning this model has better sensitivity to negative reviews compared to the LSTM model, although accompanied by a decrease in precision. For the positive class, LSTM demonstrates very stable performance with precision, recall, and F1-score each at 90%, while XGBoost achieves higher precision with slightly lower recall. This indicates that XGBoost is more accurate in identifying positive reviews with fewer prediction errors but has lower sensitivity compared to LSTM in capturing all positive reviews.

Table 4. Comparison of XGBoost and LSTM Classification Report

	LSTM			XGBoost		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Negative	79%	79%	79%	76%	86%	81%
Positive	90%	90%	90%			90%
Accuracy			86%			87%
Macro Avg.	84%	84%	84%	85%	87%	86%
Weighted Avg.	86%	86%	86%	88%	87%	88%

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

Overall, the LSTM model reached an accuracy of 86.58%, showing that the majority of the test data was correctly classified. An average precision of 85% shows that the model is quite effective at correctly identifying positive classes. Meanwhile, a recall of 84% signifies that the model is capable of identifying the majority of actual positive data. The F1-Score result of 85% shows a rather good balance between precision and recall. Overall, the LSTM's performance is relatively stable with a low classification error rate. Meanwhile, the XGBoost model achieved an accuracy of 87.22%, which is slightly higher than the LSTM model. The average precision value of 85% shows the model's correctness value in identifying positive data, while the recall of 87% indicates a better ability to capture all positive data compared to LSTM. With an average F1-Score of 86%, the XGBoost model is considered to have consistent performance in classifying both classes. Therefore, it can be broadly concluded that XGBoost shows slightly superior performance compared to LSTM, especially in terms of recall and accuracy, showing that this algorithm is more effective at finding positive data without sacrificing prediction accuracy in the GrokAI review dataset.

3.3 Knowledge Presentation

After comparing the performance of both models, the most crucial stage of the entire study was the analysis of the words with the highest frequency of occurrence from the review classification results in order to find representative opinions from users regarding the application's performance. The distribution of words was illustrated with word cloud visualization, meaning that the bigger the size of the word that appears, the more dominant its presence in the review dataset. In this context, positive sentiment still outweighs negative sentiment in terms of quantity. As show in Figure 5, the word distribution in the word cloud is dominated by the four largest words, namely "grok," "app," "good," and "ai." The presence of these four words explains that the opinions written by users are related to the positive aspects of GrokAI's performance as Generative AI as a whole. To capture the context further, the words that follow are traced, namely "great," "use," "nice," "free," "amaze," "feature," "like," and "work." This indicates that users appreciate and have a positive emotional view of the features in the GrokAI application and free access, as indicated by the keywords "use," "feature," and "free." This assumption is in line with the reference study by Widaad and Anggraini in 2024 which states that phrases such as "good," "use," and "work" in a positive context describe assessments related to the performance and functionality of the application [18].

Figure 5. Word Cloud Visualization of Postive Sentiment

On the other hand, the word cloud visualization for the negative review group displays several dominant words that are not much different. Referring to Figure 6, the phrases "grok," "ai," and "app" reappear in the word cloud, but are accompanied by the word "bad," which has a high frequency of occurrence, representing expressions of user dissatisfaction directly related to the core performance within the application. The subsequent word forms do not show significant differences in frequency because their sizes are almost the same. These words are "work," "use," "time," "give," and "get," which directly suggest that some users face difficulties when trying to use the main features of the application. Referring to Lobo in 2023, the phrase "work" in a negative context falls under the usable dimension, which tends to contribute to user frustration regarding the application's usability [47]. This frustration can be linked to the context of other words appearing, such as the phrase "use," related to the usage process, "time" representing response time from the application also "give" and "get," related to the input and output aspects provided to and by the AI model within the GrokAI application. Besides the dominant words, there are also words that appear with smaller frequencies. These include "image" and "voice," which more specifically represent features within the GrokAI application. To better understand the complaint patterns from these three phrases, the following paragraphs will show examples of negative reviews containing these words.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

Figure 6. Word Cloud Visualization of Negative Sentiment

Based on an analysis of negative review examples containing those words, several main complaint patterns were identified. The word “image” is often associated with the system’s inability to generate or download images correctly, such as in reviews like “I can’t download images…” and “poor quality image creation.” This indicates issues with GrokAI’s visual output function. Meanwhile, the appearance of the word “voice” suggests problems with the voice interaction feature (voice mode). Some users complain about limited access to voice features on Android devices and functional disruptions like “voice mode is not working.” Additionally, the review “you sometimes can’t interrupt it while it’s answering…” shows that users expect the system to improve the smoothness of two-way interactions.

The knowledge findings extracted based on keywords that appear in both parts of the word cloud visualization classification lead to all points that should be considered by GrokAI for the continuity of performance and user experience. These points include:

a. Positive statements from users can serve as a reference to maintain the application's competitive advantage, especially regarding free access to the application. If possible, developers can expand free access to other essential features as a strategy to retain user loyalty. Additionally, it is important to provide services with added value comparable to users who subscribe to paid access.
b. The complaint regarding the “image” point can be interpreted as a basis for a thorough review of the prompts that generate images and the section for downloading images. If a critical issue with the model is identified after the review, developers can optimize the AI model to produce more accurate and higher-quality images, as well as simplify the download workflow to ensure the process functions stably across various operating systems.
c. Conducting tests related to voice mode performance in various compatible device versions to ensure users can stop or redirect AI responses without having to wait for the answer to be completed.
d. Developers are also advised to focus on simplifying the back end to speed up the application's response time, as latency issues (marked by the word "time") have emerged as one of the complaints. Additionally, it is important to streamline the interaction process by improving the input processing logic. This aims to ensure that the AI model's output is more accurate and relevant to user input, which is currently a concern related to the words "give" and "get" in negative reviews.

4. CONCLUSION

The data collection results that have gone through the preprocessing process show a total of 11,108 review data points, classified as 7,633 positive reviews and 3,475 negative reviews using the VADER Lexicon method. The data was then trained and tested on the XGBoost and LSTM models, which showed a relatively small difference in accuracy. This research highlights that the XGBoost algorithm model is considered to have a more optimal capability compared to the LSTM model, with an accuracy of 87.22% compared to LSTM's 86.58%. Although the overall accuracy results are quite optimal, there is an imbalance in the model's ability in segmentating negative and positive classes. The negative class shows significantly lower precision, recall, and F1-score percentages compared to the positive class, likely due to the substantial difference in the amount of data points between the two classes, making it difficult for the model to predict reviews labeled as negative. In addition to testing the capability of the XGBoost and LSTM models, this study also conducted knowledge mining based on the frequency of phrase appearances in positive and negative reviews to explore users' opinions about the GrokAI application on the Google Play Store platform. Over a six month period, users more frequently provided positive responses regarding the app. The word cloud visualization shows that most users reflect positive appreciation with words of praise for free access and the general functionality of the application. However, the presence of word representations with negative meanings cannot be avoided, even though their frequency is lower compared to positive reviews. Findings indicate that complaints are generally based on performance issues related to response time, output, and some specific features of GrokAI. Therefore, several data-driven recommendations are offered to help developers maximize app improvements in the future. These

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

recommendations include maintaining the advantage of free access and expanding essential features to retain user loyalty. Furthermore, developers can improve the image and voice features to achieve more accurate and stable results across various compatible devices, as well as speed up response times and refine processing logic to maximize the relevance of output and input. In addition to the results and insights gained, this study has several limitations. Therefore, as a contribution to scientific development, future research can test models with more balanced datasets of negative and positive classes, optimize preprocessing to ensure the dataset is free from anomalies, use datasets in other languages, combine machine learning, deep learning algorithms, and other labeling methods to generate more diverse perspectives on GrokAI review classification and prediction models, or implement topic modeling methods to produce more precise knowledge patterns about GrokAI application's evolution.

REFERENCES

[1] V. Corvello, "Generative AI and the future of innovation management: A human centered perspective and an agenda for future research," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 11, no. 1, p. 100456, 2025, doi: https://doi.org/10.1016/j.joitmc.2024.100456.

[2] M. Albashrawi, "Generative AI for decision-making: A multidisciplinary perspective," *Journal of Innovation & Knowledge*, vol. 10, no. 4, p. 100751, 2025, doi: https://doi.org/10.1016/j.jik.2025.100751.

[3] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science (1979)*, vol. 381, no. 6654, pp. 187–192, 2023, doi: 10.1126/science.adh2586.

[4] E. Brynjolfsson, D. Li, and L. Raymond, "Generative AI at Work*," *Q J Econ*, vol. 140, no. 2, pp. 889–942, May 2025, doi: 10.1093/qje/qjae044.

[5] M.-T. Huynh and T. Aichner, "In generative artificial intelligence we trust: unpacking determinants and outcomes for cognitive trust," *AI Soc*, 2025, doi: 10.1007/s00146-025-02378-8.

[6] M. Shukla, I. Goyal, B. Gupta, and J. Sharma, "A Comparative Study of ChatGPT, Gemini, and Perplexity," *International Journal of Innovative Research in Computer Science and Technology*, vol. 12, pp. 10–15, Oct. 2024, doi: 10.55524/ijircst.2024.12.4.2.

[7] K. Wangsa, S. Karim, E. Gide, and M. Elkhodr, "A Systematic Review and Comprehensive Analysis of Pioneering AI Chatbot Models from Education to Healthcare: ChatGPT, Bard, Llama, Ernie and Grok," *Future Internet*, vol. 16, no. 7, 2024, doi: 10.3390/fi16070219.

[8] U. Samet, "The positive influence of large language models on fact-checking practices: A case study of Grok," *World Journal of Advanced Engineering Technology and Sciences*, vol. 15, no. 3, pp. 1727–1738, 2025, doi: https://doi.org/10.30574/wjaets.2025.15.3.1123.

[9] xAI, "Grok." Accessed: Nov. 01, 2025. [Online]. Available: https://play.google.com/store/apps/details?id=ai.x.grok&hl=id

[10] M. Vadla, M. Suresh, and V. Viswanathan, "Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT," *Algorithms*, vol. 17, p. 59, Oct. 2024, doi: 10.3390/a17020059.

[11] V. Novalia, K. D. Tania, A. Meiriza, and A. Wedhasmara, "Knowledge Discovery of Application Review Using Word Embedding's Comparison with CNN-LSTM Model on Sentiment Analysis," in *2024 International Conference on Electrical Engineering and Computer Science (ICECOS)*, IEEE, 2024, pp. 234–238. doi: https://doi.org/10.1109/ICECOS63900.2024.10791113.

[12] C. Singh, T. Imam, S. Wibowo, and S. Grandhi, "A Deep Learning Approach for Sentiment Analysis of COVID-19 Reviews," *Applied Sciences*, vol. 12, no. 8, 2022, doi: 10.3390/app12083709.

[13] R. Kurniawan, H. Oktafia, and R. Aprisusanti, "Sentiment Analysis of Google Play Store User Reviews on Digital Population Identity App Using K- Nearest Neighbors," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, Oct. 2024, doi: 10.32736/sisfokom.v13i2.2071.

[14] U. Kulsum, M. Jajuli, and N. Sulistiyowati, "Analisis Sentimen Aplikasi WETV di Google Play Store Menggunakan Algoritma Support Vector Machine," *Journal of Applied Informatics and Computing*, vol. 6, pp. 205–212, Oct. 2022, doi: 10.30871/jaic.v6i2.4802.

[15] E. Damayanti, A. V. Vitianingsih, S. Kacung, H. Suhartoyo, and A. Lidya Maukar, "Sentiment Analysis of Alfagift Application User Reviews Using Long Short-Term Memory (LSTM) and Support Vector Machine (SVM) Methods," *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 4, no. 2, pp. 509–521, Jun. 2024, doi: 10.51454/decode.v4i2.478.

[16] M. J. Setiawan and V. R. S. Nastiti, "DANA App Sentiment Analysis: Comparison of XGBoost, SVM, and Extra Trees," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, 2024, doi: https://doi.org/10.32736/sisfokom.v13i3.2239.

[17] V. Prasetyo, M. Naufal, and K. Wijaya, "Sentiment Analysis of ChatGPT on Indonesian Text using Hybrid CNN and Bi-LSTM," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, pp. 327–333, Apr. 2025, doi: 10.29207/resti.v9i2.6334.

[18] N. Widaad and D. Anggraini, "SENTIMENT ANALYSIS OF CHATGPT APP USER REVIEWS USING SVM AND CNN METHODS," *Jurnal Teknik Informatika (Jutif)*, 2024, doi: https://doi.org/10.52436/1.jutif.2024.5.6.4010.

[19] C. A. Palacios, J. A. Reyes-Suárez, L. A. Bearzotti, V. Leiva, and C. Marchant, "Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile," *Entropy*, vol. 23, no. 4, 2021, doi: 10.3390/e23040485.

[20] Y. Singgalen, S. Wahyuningtyas, E. Widodo, M. Dasra, and R. Setiawan, "KNOWLEDGE DISCOVERY IN DATABASES FOR HOTEL SERVICE QUALITY IMPROVEMENT THROUGH DATA- MINING APPROACH," *J Theor Appl Inf Technol*, vol. 102, pp. 9004–9020, Dec. 2024.

[21] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Syst Appl*, vol. 166, p. 114060, 2021, doi: https://doi.org/10.1016/j.eswa.2020.114060.

[22] Heri Suroyo and E. J. Pratama, "Comparison of Text Representation Methods for Sentiment Analysis Using Support Vector Machine," *Journal of Advances in Information and Industrial Technology*, vol. 7, no. 1, pp. 21–30, May 2025, doi: 10.52435/jaiit.v7i1.610.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1637−1648
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8651

[23] V. Çetin and O. Yıldız, "A comprehensive review on data preprocessing techniques in data analysis," *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 28, no. 2, pp. 299–312, 2022, doi: doi:10.5505/pajes.2021.62687.
[24] H.-T. Duong and T.-A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Comput Soc Netw*, vol. 8, no. 1, p. 1, 2021, doi: 10.1186/s40649-020-00080-x.
[25] A. Kukkar, R. Mohana, A. Sharma, A. Nayyar, and M. Shah, "Improving Sentiment Analysis in Social Media by Handling Lengthened Words," *IEEE Access*, vol. 11, pp. 9775–9788, Jan. 2023, doi: 10.1109/ACCESS.2023.3238366.
[26] A. Majid, D. Nugraha, and F. Adhinata, "Sentiment Analysis on Tiktok Application Reviews Using Natural Language Processing Approach," *Journal of Embedded Systems, Security and Intelligent Systems*, pp. 32–38, Aug. 2023, doi: 10.59562/jessi.v4i1.471.
[27] J. Fehle, T. Schmidt, and C. Wolff, "Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques," in *Conference on Natural Language Processing*, 2021. doi: 10.5283/epub.50833.
[28] S. Biswas, K. Young, and J. Griffith, *A Comparison of Automatic Labelling Approaches for Sentiment Analysis*. 2022. doi: 10.5220/0011265900003269.
[29] S. Tzimiris, S. Nikiforos, M. N. Nikiforos, D. Mouratidis, and K. L. Kermanidis, "A Comparative Evaluation of Transformer-Based Language Models for Topic-Based Sentiment Analysis," *Electronics (Basel)*, vol. 14, no. 15, 2025, doi: 10.3390/electronics14152957.
[30] V. Nurcahyawati, Z. Mustaffa, and M. Khalaf, "Exceeding Manual Labeling: VADER Lexicon as an Accurate Alternative to Automatic Sentiment Classification," *The International Arab Journal of Information Technology*, vol. 22, Jan. 2025, doi: 10.34028/iajit/22/2/2.
[31] M. Arief and N. A. Samsudin, "Hybrid Approach with VADER and Multinomial Logistic Regression for Multiclass Sentiment Analysis in Online Customer Review," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 12, 2023, doi: 10.14569/IJACSA.2023.0141232.
[32] I. Aggarwal, S. Joseph, N. Jaganathan, A. Patel, V. Kumar, and M. Devarapalli, "Sentiment Analysis in Healthcare: A Comparison of VADER, BERT, and Flair NLP Models on Patient Reviews of Pain Management Physicians," *Cureus*, vol. 17, Jul. 2025, doi: 10.7759/cureus.88902.
[33] R. Wati, S. Ernawati, and H. Rachmi, "Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 13, no. 1, pp. 84–93, Apr. 2023, doi: 10.34010/jamika.v13i1.9424.
[34] A. Erkan and T. Gungor, "Analysis of Deep Learning Model Combinations and Tokenization Approaches in Sentiment Classification," *IEEE Access*, vol. PP, p. 1, Jan. 2023, doi: 10.1109/ACCESS.2023.3337354.
[35] A. Samih, A. Ghadi, and A. Fennan, "Enhanced sentiment analysis based on improved word embeddings and XGboost," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 2, pp. 1827–1836, 2023, doi: http://doi.org/10.11591/ijece.v13i2.pp1827-1836.
[36] G. S. N. Murthy, S. R. Allu, B. Andhavarapu, M. Bagadi, and M. Belusonti, "Text based sentiment analysis using LSTM," *Int. J. Eng. Res. Tech. Res*, vol. 9, no. 05, pp. 32–41, 2020, doi: https://doi.org/10.17577/IJERTV9IS050290.
[37] F. Horasan and B. Bilen, "LSTM Network based Sentiment Analysis for Customer Reviews," *Politeknik Dergisi*, vol. 25, no. 3, pp. 959–966, 2022, doi: 10.2339/politeknik.844019.
[38] W. A. Wily, S. Anggai, and T. Tukiyat, "ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI MEDIA SOSIAL X DI PLAY STORE MENGGUNAKAN ALGORITMA LONG SHORT-TERM MEMORY (LSTM) DAN GATED RECURRENT UNIT (GRU): Studi Kasus pada Ulasan Pengguna di Google Play Store," *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, vol. 9, no. 1, pp. 63–72, 2025, doi: https://doi.org/10.47970/siskom-kb.v9i1.875.
[39] Y. A. Pradana, I. Cholissodin, and D. Kurnianingtyas, "Analisis sentimen pemindahan ibu kota Indonesia pada media sosial Twitter menggunakan metode LSTM dan Word2Vec," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 5, pp. 2389–2397, 2023.
[40] X. Wen and W. Li, "Time series prediction based on LSTM-attention-LSTM model," *IEEE access*, vol. 11, pp. 48322–48331, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3276628.
[41] S. Lonang, A. Yudhana, and M. K. Biddinika, "Analisis Komparatif Kinerja Algoritma Machine Learning untuk Deteksi Stunting," *J. Media Inform. Budidarma*, vol. 7, no. 4, p. 2109, 2023, doi: http://dx.doi.org/10.30865/mib.v7i3.6368.
[42] A. Bagheri, S. Taghvaeian, and D. Delen, "A text analytics model for agricultural knowledge discovery and sustainable food production: A case study from Oklahoma Panhandle," *Decision Analytics Journal*, vol. 9, p. 100350, 2023, doi: https://doi.org/10.1016/j.dajour.2023.100350.
[43] M. Pratiwi and K. Tania, "Knowledge Discovery Through Topic Modeling on GoPartner User Reviews Using BERTopic, LDA, and NMFKnowledge Discovery Melalui Pemodelan Topik pada Ulasan Pengguna Aplikasi GoPartner Menggunakan BERTopic, LDA, dan NMF," *Journal of Applied Informatics and Computing*, vol. 9, pp. 1–7, Jan. 2025, doi: 10.30871/jaic.v9i1.8782.
[44] N. A. Sofiah, K. D. Tania, A. Meiriza, and A. Wedhasmara, "A Comparative Assessment SARIMA and LSTM Models for the Gurugram Air Quality Index's Knowledge Discovery," in *2024 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 2024, pp. 26–31. doi: 10.1109/ICECOS63900.2024.10791243.
[45] S. A. Putri, K. Ditha Tania, N. Kawadha, and P. Gumay, "Knowledge Discovery Through Sentiment Analysis and Topic Modeling of BCA Mobile and MyBCA," *Journal of Mathematics, Computations, and Statistics*, vol. 8, no. 2, pp. 669–682, 2025, doi: 10.35580/jmathcos.v8i2.9782.
[46] J. Zhou, Z. Liang, Y. Fang, and Z. Zhou, "Exploring public response to ChatGPT with sentiment analysis and knowledge mapping," *IEEE Access*, vol. 12, pp. 50504–50516, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3386362.
[47] E. H. Lobo *et al.*, "Detecting user experience issues from mHealth apps that support stroke caregiver needs: an analysis of user reviews," *Front Public Health*, vol. 11, p. 1027667, 2023.