Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 685−695
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7514

Aspect-Based Sentiment Classification of iPhone 15 YouTube Reviews Using VADER-Augmented LSTM

Hasna Rafida Alya*, Yuliant Sibaroni

School of Computing, Informatics, Telkom University, Bandung, Indonesia
Email: [1,*]hasnarafidaalya@student.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id
Email Penulis Korespondensi: hasnarafidaalya@student.telkomuniversity.ac.id
Submitted: **04/06/2025**; Accepted: **30/06/2025**; Published**: 30/06/2025**

Abstract−This research investigates the effectiveness of the Long Short-Term Memory (LSTM) model in performing aspect-based sentiment classification on English-language reviews of the iPhone 15 sourced from the YouTube platform. The study focuses on five key product aspects frequently mentioned by users: charger port, camera, screen, design, and battery. To evaluate the model's performance, two distinct labeling strategies were employed. The first involved manual annotation, where human annotators identified both the relevant aspects and the associated sentiment in each review. The second strategy integrated additional sentiment cues derived from a lexicon-based method, Valence Aware Dictionary and sEntiment Reasoner (VADER). In this approach, the polarity output from VADER was prepended to each review to enrich the input with emotional context. The experimental results demonstrate that supplementing review texts with sentiment polarity information from VADER contributes to a modest but measurable improvement in sentiment classification accuracy. Specifically, using the micro-average accuracy metric, defined as the ratio of correct predictions to the total number of test instances, the model's performance improved from 67% under the manual only annotation to 68% with VADER enhanced input. Additionally, aspect classification remained consistently strong, showing a slight improvement from 90% to 91% after incorporating VADER. Furthermore, based on macro-average accuracy an evaluation metric that calculates the mean performance across all classes regardless of class distribution, accuracy improvements were observed in several aspects, particularly the camera, screen, and design. However, a minor decline in performance was noted for the battery and charger port aspects. These results suggest that enriching review data with sentiment polarity information derived from lexicon-based tools like VADER can enhance the model's ability to comprehend emotional nuance, leading to more accurate identification of user sentiments within aspect-specific reviews.

Keywords: Sentiment Analysis; Aspect-Based; iPhone 15; LSTM; VADER

1. INTRODUCTION

One of the leading companies in the technology industry is Apple, which first introduced the iPhone on January 9, 2007 [1]. Since then, the iPhone has grown to become one of the leading smartphone brands, known for its unique iOS ecosystem. Every year, Apple consistently releases new models, including the iPhone 15, which was launched in September 2023. According to data from Counterpoint [2], the iPhone 15 became the best-selling smartphone in the third quarter of 2024, with a market share of 3.5%, surpassing the iPhone 14, which dominated the previous year. This indicates that improvements in the camera, performance, and other features have successfully attracted consumer interest. However, consumer reviews of various aspects of the iPhone, such as the charging port, camera, display, design, and battery, vary significantly, ranging from positive, negative, to neutral reviews.

Sentiment analysis aims to identify various opinions expressed in a text or comment in order to determine the underlying sentiment trend, whether it is positive, negative, or neutral. Through sentiment analysis, companies can gain deeper insights into consumer perceptions of their products, which can serve as a basis for evaluating and improving product quality to enhance customer satisfaction [3].

Various studies have been conducted on sentiment analysis using different methods. For example, a study conducted by Diny Wahyuni et al. explored the application of deep learning methods, specifically Long Short-Term Memory (LSTM), to classify sentiment in Indonesian-language hotel reviews. The study showed that the LSTM model was effective in understanding and analyzing the aspects and sentiment of hotel review sentences, achieving a relatively high level of accuracy [4].

Furthermore, research by Wahyuni et al. demonstrated that the combination of the Long Short-Term Memory (LSTM) and lexicon-based methods in sentiment analysis of TikTok app user reviews achieved an accuracy of 90.05%, with a precision of 92.14%, recall of 97.35%, and an F1-score of 98.66%. The results of this analysis revealed that the TikTok app tends to receive negative sentiment, with 59.5% of the reviews being negative, compared to 30.0% positive and 10.5% neutral. These findings confirm that the LSTM and lexicon-based methods are effective in accurately classifying and interpreting user opinions, as well as providing an in-depth picture of public sentiment towards the TikTok platform [5].

Meanwhile, Johan Setiawan et al. [6] analyzed sentiments related to tourist destinations in Labuan Bajo using data collected from Instagram via the hashtag "labuanbajo." This study aimed to explore feedback from both local and foreign tourists and to identify the most popular tourist destinations in the area. The analysis results showed that the VADER method achieved an accuracy of 72%.

Marvin Gultom et al. [7] conducted sentiment analysis on English-language text using the VADER sentiment algorithm implemented on a website platform. The study categorized text into four classes: positive, negative, neutral, and mixed, using data from 50 Twitter comments on the YouTube channel "Simplilearn." The final results showed that the system achieved an accuracy of approximately 80%.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 685−695
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7514

Arum Prabowo G. et al. [8] examined the use of the XGBoost algorithm for aspect-based sentiment analysis of iPhone 14 Pro video review comments on YouTube. The aim was to classify reviews into positive and negative sentiments and identify the strengths and weaknesses of the iPhone 14 Pro. The results showed an accuracy ranging from 89% to 97%, with identified strengths including brand image and performance, and weaknesses in design and specifications.

Most previous studies have focused on general sentiment analysis without considering specific aspects. Furthermore, there have been few studies that specifically compare manual labeling approaches with lexicon-based methods in the context of aspect-based sentiment analysis, particularly for iPhone 15 reviews. Based on this gap, this study aims to evaluate the performance of the Long Short-Term Memory (LSTM) model in classifying sentiment toward five main aspects: charger port, camera, screen, design, and battery.

This study compares two data processing scenarios. In the first scenario, aspect and sentiment labels are manually annotated by human annotators through direct reading and understanding of English-language iPhone 15 reviews collected from the YouTube platform. In the second scenario, sentiment labels are generated using the Valence Aware Dictionary and sEntiment Reasoner (VADER) method by applying a threshold value for automatic sentiment classification. The VADER-generated sentiment information is inserted at the beginning of the review text as additional input for the model. However, the target sentiment labels remain based on manual annotation and serve as the ground truth for evaluating model performance. This study aims to compare the two scenarios and determine the extent to which the inclusion of VADER-generated sentiment information at the beginning of the text affects the LSTM model's performance in aspect-based sentiment classification.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This study discusses aspect-based sentiment analysis by evaluating the performance of LSTM models using two labeling methods: manual labeling by annotators and lexicon-based labeling. This section describes the stages of the research process, as illustrated in the system flowchart shown in Figure 1.

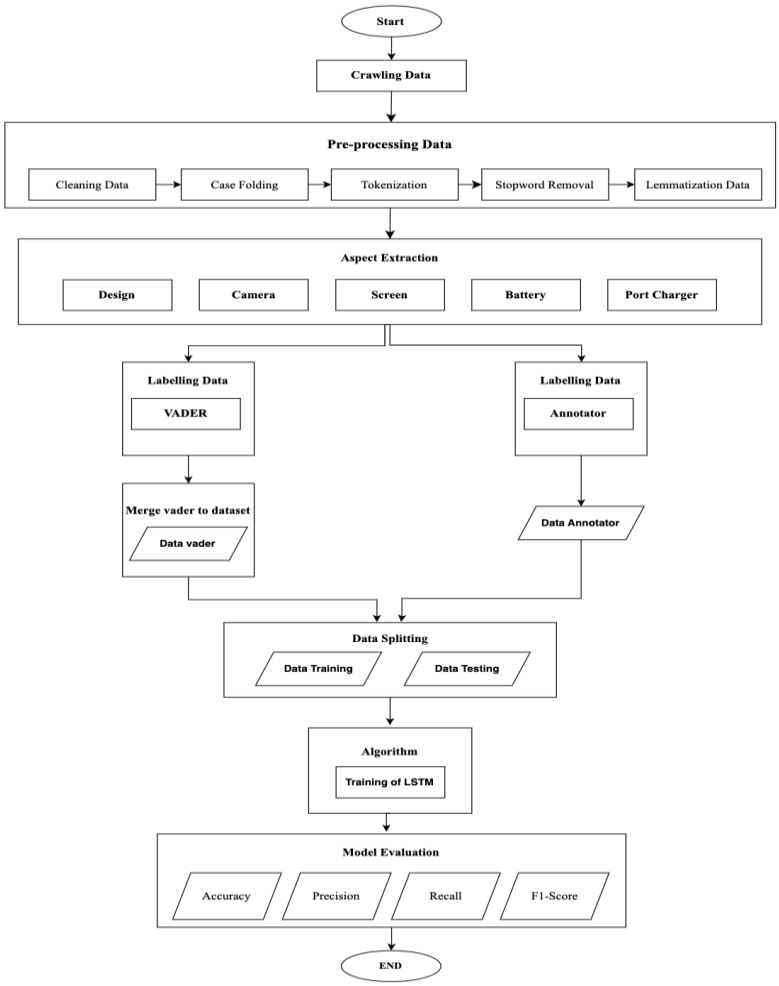

Figure 1. Architecture System

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 685−695
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7514

2.2 Crawling Data

In this study, data were collected from the YouTube platform using an Application Programming Interface (API) to crawl three English-language videos featuring user reviews of the iPhone 15. The videos were selected based on their relevance and level of engagement, such as the number of views and comments. Each video's content was transcribed, and user comments were extracted to obtain responses aligned with the discussion topic. This process resulted in a total of 11,306 comments, reflecting diverse perspectives on the strengths and weaknesses of the iPhone 15, particularly in relation to the charger port, camera, screen, design, and battery.

2.3 Pre-Processing Data

a. Cleaning Data
In the initial stage of data preprocessing, the data cleaning process is conducted by removing irrelevant elements such as symbols, characters, or numbers that do not contribute meaningful information to the analysis [9]. Table 1 presents examples of categories that were removed during the data cleaning process.

Table 1. Examples of Removed Categories in Data Cleaning

Category	Examples
Emojis	😂 😍 😡 😭 💯 👍 👎 🔥 😎 🎉 💔 ❤️ 🙌
Punctuation & Symbols	! ? . , ; : - _ = + * ~ ` ^ " ' ( ) [ ] { } < >
Hashtags	#iPhone15, #apple, #technology
Mentions	@Apple, @user123, @officialstore
URLs	https://bit.ly/abc, www.apple.com

b. Case Folding
Case folding is the process of converting all characters in the text to lowercase [10], aiming to enhance data consistency.

c. Tokenization
Tokenization is carried out to facilitate the identification of individual words in the text, making subsequent processes more efficient. This process involves splitting a sentence or text into the smallest units such as words or phrases [11].

d. Stopwords Removal
In this study, the list of stopwords was sourced from the Natural Language Toolkit (NLTK) library version 3.9.1 using the English language. The list contains 198 common words that generally do not contribute significantly to the overall meaning of a sentence. The stopword list was applied as is, without any additional terms. The stopword removal process aims to filter out less relevant words so that the analysis can focus on terms or words with higher informational value [12]. Table 2 shows examples of stopwords used.

Table 2. Examples of Stopwords Used

Stopwords Removal
a, about, above, after, again, ain, all, am, an, and, any, are, aren, as, at, be, because.

e. Lemmatization Data
In this research, lemmatization was performed using the WordNetLemmatizer component from the Natural Language Toolkit (NLTK). The purpose of this step is to transform words with variations such as different tenses or plural forms into their canonical base forms [13]. This process reduces the complexity of the textual data and effectively filters out tokens that do not contribute to semantic analysis. As a result, each term is normalized to its root form, which improves the precision and efficiency of the subsequent processing. Examples of the lemmatization results are presented in Table 3.

Table 3. Example of Lemmatization Process

Words	Lemmatization
[‘upgrading, ‘better, ‘cameras, ‘using’, ‘charging’, ‘connecting’, ‘designing’]	[‘upgrade’, ‘good’, ‘camera, ‘use’, ‘charge’, ‘connect’, ‘design’]

2.4 Determining Aspects and Sentiment

The aspect determination stage aims to identify the relevant aspects of the iPhone 15 based on user reviews. The analyzed aspects include the charging port, camera, screen, design, and battery. In this process, aspects are identified based on keywords representing different components mentioned in the reviews. The charging port aspect refers to the transition from the Lightning connector to USB-C, which is one of the most significant changes in the iPhone 15. The camera aspect focuses on photo quality, including clarity, megapixel resolution, and image sharpness. The screen

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 685−695
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7514

aspect relates to display quality, especially the refresh rate. The design aspect includes reviews concerning the shape and model of the iPhone, color options, Dynamic Island, and Action Button. Lastly, the battery aspect covers battery life, efficiency, and overheating issues.

Aspect identification was conducted manually, where annotator read each review and determined the relevant aspect based on the content. The number of identified aspects is presented in Table 4.

Table 4. Determining Aspects

Aspect	Negative	Neutral	Positive	Total
Charger port	1931	473	1082	3486
Camera	553	304	668	1525
Screen	903	288	239	1430
Design	1690	377	1075	3142
Battery	1183	184	356	1723
Total	6260	1626	3420	11306

2.5 Labelling Data

The data labeling process is a critical step in classifying review sentiments into positive, negative, or neutral categories, based on the specific aspects of the iPhone 15 that were previously identified. In this study, two different labeling approaches were used. The first approach was manual annotation, where annotators carefully read each review to determine both the sentiment category and the related aspect, based on their contextual understanding. The second approach used an automatic lexicon-based method, namely VADER, which automatically assigned sentiment labels in English and placed the classification result at the beginning of each review text.

The results of this labeling process served as the reference for evaluating the performance of the LSTM model in aspect-based sentiment classification. By comparing the model's predictions with the labels obtained through both manual and automatic methods, this study aims to assess how accurately the model understands user opinions regarding specific aspects of the iPhone 15.

Table 5. Aspect Distribution Based on Manual Sentiment

Labelling	Negative	Neutral	Positive	Total
Manual	6260	1626	3420	11306
VADER	2481	3552	5273	

Based on the distribution shown in Table 5, a noticeable discrepancy exists between manual and automatic sentiment labeling. The manually labeled data is dominated by negative sentiment, totaling 6,260 reviews. In contrast, the VADER-based automatic labeling classified more reviews as positive, with 5,273 positive reviews. Moreover, the automatic approach resulted in a higher proportion of neutral sentiments compared to manual labeling. This distribution difference suggests that the automatic approach tends to yield more optimistic sentiment assessments, positive or neutral, whereas manual labeling is generally more critical, detecting a larger number of negative sentiments.

2.5.1 Valence Aware Dictionary and sEntiment Reasoner (VADER)

One of the commonly used methods in sentiment analysis is the Valence Aware Dictionary and sEntiment Reasoner (VADER). This method was specifically developed to analyze sentiment in short texts on social media, using a lexicon-based approach combined with a set of rules to better understand the emotional context of a statement. The analysis is performed using the polarity_scores() function, which generates sentiment polarity values as numerical scores ranging from -1 to +1. Values closer to -1 indicate negative sentiment, while values closer to +1 indicate positive sentiment [14].

In this research, a specific threshold is applied to improve the accuracy of sentiment classification based on the compound score. The threshold range used is between -0.03 and +0.03. Thus, sentences with a score above +0.03 are categorized as positive sentiment, those below -0.03 as negative sentiment, and scores within the range are classified as neutral. The following is the formula for the compound score, which represents the overall sentiment of a text:

$$Compound_{score} = \frac{\sum_{i=1}^{n} valence\,(w_i)}{\sqrt{\sum valence\,(w_i)^2 + \alpha}} \quad (1)$$

In sentiment analysis, each individual word within the text is denoted by (w_i) referring to the lexical unit under evaluation. Every word carries an associated sentiment score, known as $valence\ (w_i)$, which reflects its emotional polarity, indicating whether the word conveys a positive, negative, or neutral sentiment. The total number of words present in the analyzed text is represented by n, which signifies the overall length of the text. This variable plays a crucial role in determining the aggregate sentiment score, as it directly influences the precision of the analysis outcomes. To ensure that the sentiment values remain proportionate to the length of the text, a normalization factor,

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 685−695
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7514

symbolized by α, is introduced. This parameter functions as an adjustment mechanism to standardize sentiment scores, and is typically set at a default value of 15.

Instead of serving as the primary basis for sentiment labeling, the Valence Aware Dictionary and sEntiment Reasoner (VADER) was utilized as an auxiliary feature to enrich the textual input. After determining sentiment polarity based on a predefined compound score threshold, the VADER-assigned sentiment categorized as positive, negative, or neutral, was appended to the beginning of each review as a tokenized word. For instance, a review classified as positive would be reformatted as "positive battery good." This approach treats the sentiment label from VADER not as ground truth but as additional contextual information embedded directly into the review text. The actual sentiment labels used for model training were obtained through prior manual annotation. It is important to note that these VADER-generated sentiment tags were not encoded into numerical or vectorized forms but were instead processed during tokenization and embedding stages as regular words. Consequently, this method introduces slight variations in input dimensions and vocabulary distribution, potentially influencing model performance and textual representation.

2.6 Data Splitting

In the modeling stage, the dataset is typically divided into two parts: training data and testing data. The training data is used to allow the model to learn from the available data, recognizing patterns, relationships, and underlying characteristics. The objective is for the model to understand the data structure so that it can make accurate predictions in the future. Once the training process is completed, the model is evaluated using the testing data, which consists of data it has not encountered before. This testing phase is conducted to assess how well the model performs in making predictions or classifications, as well as to evaluate the accuracy of the results obtained [15].

In this study, the data was split using an 80:20 ratio, with 80% used for training and 20% for testing. This ratio was chosen because it is considered to provide an appropriate balance between the data available for model learning and the data reserved for performance evaluation. An 80:20 split has been widely used in various previous studies and has proven effective in enhancing model accuracy, particularly in classification tasks, as demonstrated by [16].

The division of data used in this study is shown in Table 6, which shows the amount of data used for training and mode testing.

Table 6. Data Splitting

Data	Total Data
Data Training	9.044
Data Testing	2.262

2.7 Long Short Term-Memory (LSTM)

Long Short-Term Memory (LSTM) is a type of artificial neural network developed from the Recurrent Neural Network (RNN) architecture, specifically designed to address the problem of long-term dependency namely, the difficulty in effectively retaining long-term information in conventional RNNs [17]. The LSTM algorithm was first introduced by Hochreiter and Schmidhuber in 1997 [18], and it has since become a leading method in natural language processing tasks, particularly in sentiment analysis and text classification.

LSTM features a key component known as the cell state, which functions as the network's long-term memory, allowing information to flow consistently through the time series. In addition, LSTM includes three primary gates that regulate the flow of information: the forget gate, input gate, and output gate. Each gate plays a crucial role in deciding which information should be retained, updated, or discarded from the cell state [19]. Figure 2 below illustrates the architecture of the LSTM algorithm.

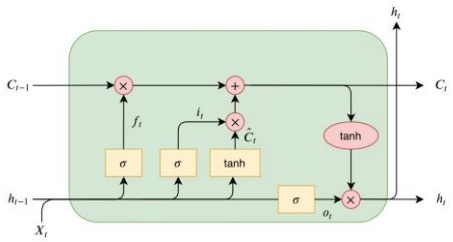

Figure 2. Architecture LSTM

The first step in the LSTM mechanism is the forget gate, which determines which information from the previous cell state should be kept or discarded. The following is the mathematical formulation of the forget gate:

$$f_t = \sigma\left(U_f x_t + W_f h_{t-1} + b_f\right) \quad (2)$$

Where σ is the sigmoid function x_t is the input at time t, and h_{t-1} is the hidden state from the previous time step. The value of f_t lies within the range of 0 to 1, indicating how much of the previous information will be retained in memory.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 685−695
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7514

The second gate is the input gate, which regulates the entry of new information to be stored in the cell state. This stage consists of two components. First, the sigmoid activation function is used to determine the proportion of new information to retain. A value close to 0 means the information will be ignored, whereas a value close to 1 indicates that the information will be preserved. Second, the *tanh* function generates new candidate values to be added to the cell state (main memory) of the LSTM [20].

$$i_t = \sigma\left(U_f x_t + W_i h_{t-1} + b_i\right) \tag{3}$$

$$\tilde{C}_t = tanh\left(U_c x_t + W_c h_{t-1} + b_c\right) \tag{4}$$

These values are then used to update the cell state, which serves as the long-term memory component that distinguishes LSTM from standard RNNs. The memory update is governed by the following equation:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_i) \tag{5}$$

The forget get f_t, ranging form 0 to 1, determines how much of the previous information should be retained. Meanwhile, the input gate i_t generated by a sigmoid function, acts as a weight to determine the extent of new information to be added.

Finally, the output gate decides which parts of the cell state will be output at time *t*. This process is computed using the following equation:

$$O_t = \sigma\left(U_o x_t + W_0 h_{t-1} + b_o\right) \tag{6}$$

$$h_t = O_t \odot \tanh\left(C_t\right) \tag{7}$$

The resulting h_t value is passed to the next time step and used in the prediction process at that time. The gate structure in the LSTM architecture makes it highly effective in selecting and retaining relevant information from the input dataset. Therefore, LSTM is particularly suitable for handling problems that require an understanding of long-term context.

In this research, model training was carried out using the Google Colaboratory platform, utilizing TensorFlow and Keras as the primary libraries. Two separate LSTM-based models were constructed to address the tasks of aspect classification and sentiment classification individually. Rather than applying a multi-task learning strategy, each model was developed independently, with one exclusively dedicated to identifying aspects and the other to analyzing sentiment polarity.

Word representations in the models were generated using Keras built-in embedding layer, without the incorporation of any pre-trained embeddings such as Word2Vec, GloVe, or FastText. Consequently, all embedding vectors were learned entirely from scratch during the training phase. Each word from the review data was transformed into a 128 dimensional vector, which served as input to the neural network. Since the word embeddings were not initialized with external resources, their quality and expressiveness were heavily dependent on the volume and variability of the dataset used.

The key configuration parameters for the LSTM models are summarized in Table 7. This table outlines the core architectural components of the models, including the dimensionality of the embedding vectors, the number of LSTM units, dropout rates, and other relevant hyperparameters that collectively influenced model performance.

Table 7. LSTM Model Parameters

No	Component	Configuration	Description
1	Embedding Layer	input_dim = vocabulary size + 1, output_dim = 128	Converts each word into a 128 dimensional numeric vector so it can be processed by the model.
2	LSTM Layer 1 (Bi-LSTM)	64 unit	Reads the text in both directions to better capture word context.
3	LSTM Layer 2 (Bi-LSTM)	32 unit	Further enhances the model's understanding of the word sequence.
4	Dropout Rate	0.5	Prevents overfitting by randomly deactivating 50% of neurons during training.
5	Dense Hidden Layer	64 units, ReLU activation	A fully connected layer that processes the LSTM output.
6	Output Layer	3 units, Softmax activation	Predict sentiment: positive, neutral, or negative.
7	Optimizer	Adam	Efficiently updates the model's weights during training.
8	Learning Rate	Default (0.001)	Controls how quickly the model adjusts its weights.
9	Loss Function	Sparse Categorical Crossentropy	Calculates prediction error for multi-class classification with numeric labels.
10	Metrics	Accuracy	Measures how many predictions are correct out of all predictions.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 685−695
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7514

No	Component	Configuration	Description
11	Batch Size	32	Number of data samples processed in one training step.
12	Epoch Maximum	20 (with early stopping)	Maximum number of training cycles, may stop earlier if no improvement is seen.
13	Early Stopping	Patience 2	Stops training early if validation performance doesn't improve for 2 consecutive epochs.
14	Sequence Length	100	Limits input text to a maximum of 100 words.
15	Tokenizer	5000 vocabulary size, OOV token "<OOV>"	Converts text into numerical sequences, unknown words are replaced with a designated out-of-vocabulary token.

The LSTM-based model configured in this study demonstrates a strong capability in identifying patterns within textual data, particularly in the context of aspect-based sentiment analysis. By employing a model separation strategy, the architecture is able to distinctly focus on each classification objective, enhancing overall performance. Moreover, the integration of two Bidirectional LSTM layers significantly improves the model's ability to comprehend the intricate and sequential relationships between words, thereby capturing contextual nuances more effectively.

2.8 Model Evaluation

To evaluate the performance of each method in data classification, this study employs several evaluation metrics for sentiment analysis using a confusion matrix. Additionally, adjustments are made to the threshold value in the VADER method to examine how this parameter affects the resulting sentiment classification.

The confusion matrix consists of two main components, the actual values, which represent the true category of the analyzed data, and the predicted values, which are the classification results generated by the model. This matrix is used to assess how accurately the model maps sentiment. Based on the information obtained from the confusion matrix, several evaluation metrics can be used to analyze and measure the model's performance in sentiment classification [21].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (8)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (10)$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall+Precision} \quad (11)$$

3. RESULT AND DISCUSSION

3.1 Performance Evaluation

This section presents the performance evaluation results of the LSTM model in classifying aspects and sentiments of iPhone 15 product reviews from the YouTube platform. The analysis focuses on five main aspects: charger port, camera, screen, design, and battery. The evaluation was carried out using two training data scenarios. The first scenario uses manually labeled aspects and sentiments by annotators, while the second scenario adds sentiment polarity information from VADER to the review text input, while the target labels remain based on manual annotations. The inclusion of VADER polarity aims to provide additional contextual information to help the model better understand sentiment. The following are the results of the LSTM performance evaluation of the two approaches.

3.2 Aspect-Based Sentiment Comparison

The evaluation was performed on the five aspects using precision, recall, F1-score, and accuracy metrics. The results indicate that the model's performance is relatively stable across both scenarios, although variations are observed across individual aspects and sentiment categories. Table 8 shows a comparison of the sentiment classification results between two scenarios based on aspects.

Table 8. Sentiment Classification Results by Aspect

Aspect	Scenario	Sentiment	Precision	Recall	F1- Score	Accuracy
Charger Port	Scenario 1	Negative	0.69	0.90	0.74	0.66
		Neutral	0.25	0.09	0.14	
		Positive	0.67	0.66	0.67	
	Scenario 2	Negative	0.66	0.89	0.75	0.65
		Neutral	0.23	0.12	0.15	
		Positive	0.77	0.50	0.60	

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 685−695
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7514

Aspect	Scenario	Sentiment	Precision	Recall	F1- Score	Accuracy
Camera	Scenario 1	Negative	0.55	0.64	0.59	0.57
		Neutral	0.48	0.23	0.31	
		Positive	0.62	0.67	0.64	
	Scenario 2	Negative	0.59	0.75	0.66	0.61
		Neutral	0.42	0.36	0.39	
		Positive	0.72	0.60	0.65	
Screen	Scenario 1	Negative	0.76	0.82	0.79	0.66
		Neutral	0.31	0.26	0.28	
		Positive	0.60	0.54	0.57	
	Scenario 2	Negative	0.76	0.88	0.81	0.71
		Neutral	0.46	0.38	0.42	
		Positive	0.72	0.46	0.57	
Design	Scenario 1	Negative	0.74	0.83	0.78	0.70
		Neutral	0.76	0.12	0.19	
		Positive	0.66	0.74	0.70	
	Scenario 2	Negative	0.72	0.91	0.89	0.71
		Neutral	0.36	0.16	0.22	
		Positive	0.77	0.64	0.70	
Battery	Scenario 1	Negative	0.78	0.90	0.83	0.73
		Neutral	0.30	0.97	0.31	
		Positive	0.54	0.40	0.75	
	Scenario 2	Negative	0.77	0.90	0.83	0.72
		Neutral	0.14	0.10	0.11	
		Positive	0.64	0.37	0.47	

In general, the LSTM model demonstrates a relatively strong ability to classify sentiment based on specific aspects of iPhone 15 reviews on the YouTube platform. The model performs particularly well in identifying positive and negative sentiments, while its accuracy tends to be lower for neutral sentiments. This indicates that neutral expressions are more difficult to detect, as they typically lack explicit emotional content.

Manual labeling by annotators yields more consistent results, as it takes into account the context and overall meaning of the sentence. In contrast, the addition of sentiment labels using the VADER method at the beginning of the review text provides supplementary information regarding the sentiment direction, which in some cases enhances the model's understanding of the conveyed emotional context.

Overall, the comparison between Scenario 1 and Scenario 2 contributes positively to improving the model's comprehension of review content and meaning. Scenario 2 demonstrates that leveraging additional information in the form of automated sentiment labels can be a promising strategy to enhance model performance in aspect-based sentiment analysis, particularly when dealing with unstructured data such as user comments on social media.

3.3 Overall Model Results Comparison

This section reviews the overall accuracy results for sentiment and aspect classification. The aim is to provide a comprehensive overview of the model's performance.

3.3.1 Accuracy Model

In this research, the performance of sentiment classification models was assessed under two distinct experimental conditions. The first scenario utilized review data that had been manually annotated for both sentiment and aspect categories by human annotators. In contrast, the second scenario employed the same manually labeled dataset, but incorporated additional contextual information in the form of sentiment polarity scores generated through VADER analysis. To ensure a thorough and reliable evaluation, both models were tested using a designated test set consisting of 20% of the entire dataset, which included 2,262 individual YouTube reviews related to the iPhone 15. This approach aimed to provide a comprehensive comparison of model performance across different input configurations.

Based on the evaluation results, the LSTM model trained using manually labeled data by annotators achieved a sentiment classification accuracy of 67% and an aspect classification accuracy of 90%. In comparison, when the model was trained on data labeled with automatically generated sentiment labels using the VADER method, the sentiment accuracy slightly increased to 68%, and the aspect accuracy improved to 91%. The accuracy metrics reported in this study were derived using micro-averaged accuracy, which measures performance by dividing the total number of correct predictions by the total number of instances in the test set. This approach does not consider the distribution or proportion of each individual class. As such, micro-average accuracy provides a holistic evaluation of the model's ability to correctly classify all data points collectively, rather than reflecting per-class performance as in macro-averaging, or a balanced measure of precision and recall as in the F1-score.

This comparison indicates that incorporating polarity information from VADER into the review texts can slightly enhance the model's ability to recognize sentiment, without reducing the accuracy of aspect classification.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 685−695
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7514

Therefore, integrating auto-labeling as additional context in the training data can be an effective strategy to support model performance, particularly in text analysis tasks involving user opinions and perceptions of products.

The detailed performance of the LSTM model on aspect-based sentiment analysis using both annotator-based and lexicon-based labeling methods is presented in Table 9.

Table 9. Overall Model Performance Comparison

Method	Accuracy Sentiment	Accuracy Aspect
LSTM + Annotator	0.67	0.90
LSTM + VADER	0.68	0.91

The strong performance of the VADER-based approach is likely due to its ability to interpret sentence structure, assess emotional intensity, and account for the contextual language typically used on social media platforms. Conversely, the lower accuracy of the manual approach may be attributed to the limited quantity of labeled data and the potential subjectivity of human annotators. These findings underscore the importance of selecting an appropriate lexical approach, as it can significantly impact the accuracy of sentiment and aspect classification in English reviews.

3.3.2 Confusion Matrix

Figure 3 presents the confusion matrix for sentiment and aspect classification in Scenario 1. The model demonstrates strong performance in identifying negative sentiments, with 1,003 samples correctly classified. However, the accuracy for neutral sentiments remains low, with only 90 samples accurately identified. Additionally, 250 positive samples were misclassified as negative, indicating the model's tendency toward a bias in favor of the negative class.

In the aspect classification task under Scenario 1, Figure 3 highlights the best model performance in the charger port and design aspects, with 616 and 578 samples correctly classified, respectively. The battery and camera aspects exhibited lower accuracy, with 318 and 262 correct classifications, respectively. Meanwhile, the screen aspect showed moderate performance, with 262 samples correctly identified.

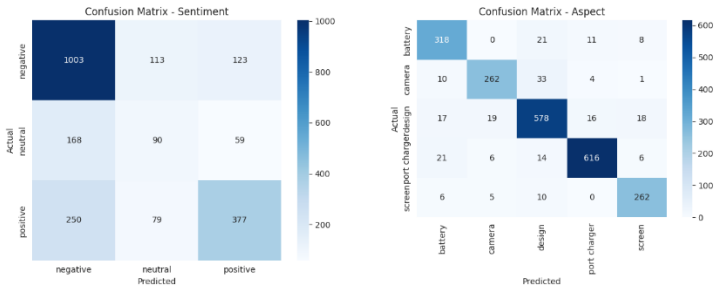

Figure 3. Confusion Matrix in Scenario 1

Overall, the classification results in Scenario 1 indicate that the majority of data points were accurately recognized by the model. Nevertheless, misclassifications still occur across aspects, particularly in categories with overlapping or similar contexts, such as between camera and design, or battery and charger port.

In Scenario 2 of the confusion matrix results presented in Figure 4, the model demonstrated improved overall accuracy, particularly in the positive class, with 458 samples correctly classified. The negative class was also accurately detected in 1,012 samples. However, prediction errors were still observed, with 169 samples incorrectly classified as positive and 58 as neutral. The model's performance in detecting the neutral class remains suboptimal, as only 49 samples were correctly identified.

In aspect classification, the charger port and design aspects yielded the best results, with 614 and 601 correctly classified samples, respectively. Both aspects showed improved accuracy compared to Scenario 1. The battery and camera aspects also achieved good classification performance, although the battery aspect experienced a slight decrease, with 309 correctly identified samples. The screen aspect correctly classified 258 samples, though some misclassifications were still present.

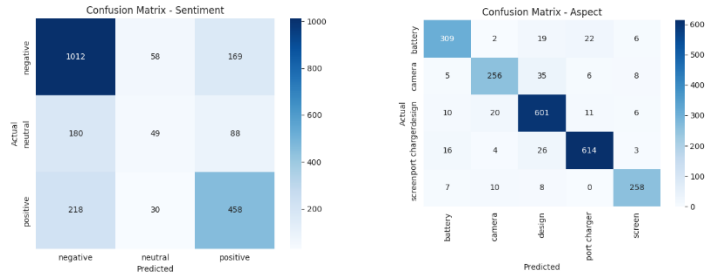

Figure 4. Confusion Matrix in Scenario 2

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 685−695
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7514

Overall, the comparison between Scenario 1 and Scenario 2 indicates that incorporating VADER polarity information into the review data enhanced the model's ability to classify positive sentiments and improved performance in several aspect categories. Nonetheless, challenges remain in accurately identifying neutral sentiments and differentiating between closely related aspect categories.

4. CONCLUSION

When using data that has been manually labeled by annotators, the model demonstrates relatively stable performance, particularly in detecting negative and positive sentiments across most aspects. However, neutral sentiments tend to be more difficult to identify, as evidenced by the low recall and F1-score values in that category. This indicates that the model still struggles to understand opinions that are neutral or lack strong emotional expressions. In the subsequent scenario, sentiment labels generated by VADER were inserted at the beginning of the review text to enrich the input context. As a result, the model's performance improved in several aspects, especially in the *camera*, *screen*, and *design* categories. This improvement is reflected in the increased precision and F1-score values for both positive and negative sentiments. These findings suggest that providing additional polarity information through VADER can assist the model in discerning the emotional direction of review texts, particularly when opinions are explicitly stated. However, not all aspects showed improved performance. In the *battery* and *charger port* aspects, the model's performance remained stagnant or even slightly declined, especially for neutral and positive sentiments. This indicates that the effectiveness of VADER-generated labels is highly dependent on the characteristics of each aspect and the way opinions are expressed in sentences. Overall, it can be concluded that the LSTM model performs well in aspect-based sentiment classification when using manually labeled data. Meanwhile, incorporating sentiment labels from VADER into the input text makes a positive contribution by clarifying the emotional context, particularly in aspects that contain explicit opinions. However, this approach does not fully replace the effectiveness of manual labeling. Therefore, the choice between these methods should be based on the purpose of the analysis and the nature of the data used. For future research, it is recommended to apply data balancing across all aspects, as an imbalanced distribution can affect the model's ability to learn sentiment patterns effectively. In addition, the use of pre-trained word embeddings such as Word2Vec, GloVe, or FastText is encouraged, as these models can capture deeper semantic relationships between words, thereby enhancing the model's understanding of the review context and improving the performance of aspect-based sentiment classification.

REFERENCES

[1] N. Niessen, "Shot on iPhone: Apple's World Picture," *Advertising & Society Quarterly*, vol. 22, no. 2, 2021, doi : https://dx.doi.org/10.1353/asr.2021.0023.
[2] Counterpoint, " iPhone 15 World's Best-selling Smartphone in Q3 2024," *Counterpoint Research,* 2024. [Online]. Available: https://www.counterpointresearch.com/insight/top-10-bestselling-smartphone-q3-2024/.
[3] I. N. Sandri and Y. Sibaroni, "Perbandingan Model LSTM GloVe dengan LSTM Word2Vec dalam Analisis Sentimen Layanan Dompet Digital," *Telkom University Repository,* Sep. 2021.
[4] W. Astriningsih and D. Hatta Fudholi, "Multi Aspect Sentiment Analysis in Hotel Review Using Deep Learning," *Jurnal Teknik Informatika Dan Sistem Informasi (JATISI),* vol. 10, no. 3, 2023, doi : https://doi.org/10.35957/jatisi.v10i3.5321.
[5] D. Wahyuni, N. Fadhillah, and W. W. Ariestya, "Long Short-Term Memory dan Lexicon Based Untuk Analisis Sentimen Ulasan Aplikasi TikTok," *Jurnal Ilmiah Komputasi*, vol. 23, no. 2, Jun. 2024, doi: 10.32409/jikstik.23.2.3579.
[6] J. Setiawan, V. Gousander, and I. Prasetiawan, "Unmasking the Sentiments of Labuan Bajo: An Instagram-based Analysis for Tourism Insights through VADER Sentiment Analysis," *G-Tech: Jurnal Teknologi Terapan*, vol. 7, no. 3, pp. 967–976, Jul. 2023, doi: 10.33379/gtech.v7i3.2615.
[7] M. Gultom, J. Marikros, and W. Rusli, "Penerapan Vader Sentiment untuk Mendeteksi Sentimen Bahasa Inggris berbasis Website," Seminar Nasional Penelitian *(SEMNAS CORISINDO),* pp. 13-18, 2024.
[8] G. Arum Prabowo, B. Rahmat, H. Endah Wahanani, U. Pembangunan Nasional Veteran Jawa Timur, and J. Raya Rungkut Madya Gunung Anyar Surabaya, "Aspect-Based Sentiment Analysis iPhone 14 Pro Menggunakan Algoritma XGBoost," *Jurnal Mahasiswa Teknik Informatika (JATI)*, vol. 7, no. 9, Des. 2023, doi : https://doi.org/10.36040/jati.v7i6.7831.
[9] R. Said and A. Faraby, "Perbandingan Algoritma Machine Learning untuk Analisis Sentimen Berbasis Aspek pada Review Female Daily," *eProceedings of Engineering*, vol. 10, no. 3, pp. 3591–3600, 2023.
[10] M. H. Al-Areef and K. Saputra, "Analisis Sentimen Pengguna Twitter Mengenai Calon Presiden Indonesia Tahun 2024 Menggunakan Algoritma LSTM," *Jurnal Sains Manajemen Informatika dan Komputer (SAINTIKOM)*, vol. 22, no. 2, pp. 270–279, 2023, doi : https://doi.org/10.53513/jis.v22i2.8680.
[11] A. D. Widiantoro, Mustafid and R. Sanjaya, "Pengantar NLP dan Topik Model LDA Sampul Dalam," *Asosiasi Doktor Sistem Informasi Indonesia,* Nov. 2024.
[12] I. H. Kusuma and N. Cahyono, "Analisis Sentimen Masyarakat Terhadap Penggunaan E-Commerce Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Pengembangan IT*, vol. 8, no. 3, 2023, doi 10.30591/jpit.v8i3.5734.
[13] Y. Amalia, N. Jannah, and R. B. Prasetyo, "Analisis Sentimen dan Emosi Publik pada Awal Pandemi COVID-19 Berdasarkan Data Twitter dengan Pendekatan Berbasis Leksikon," *Seminar Nasional Official Statistics*. Vol. 2022. No. 1, pp. 597-608, Nov. 2022, doi : https://doi.org/10.34123/semnasoffstat.v2022i1.1483.
[14] F. Fazrin, O. N. Pratiwi, and R. Andreswari, "Perbandingan Algoritma K-Nearest Neighbor dan Logistic Regression pada Analisis Sentimen terhadap Vaksinasi Covid-19 pada Media Sosial Twitter dengan Pelabelan Vader dan Textblob," *eProceedings of Engineering*, vol. 10, no.2 Apr. 2023, Telkom University.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 685−695
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7514

[15] R. Refianti, A. B. Mutiara, and R. A. Putra, "A Lexicon-Based Long Short-Term Memory (LSTM) Model for Sentiment Analysis to Classify Halodoc Application Reviews on Google Playstore," *Journal of Applied Data Sciences*, vol. 5, no. 1, pp. 146–157, Jan. 2024, doi: 10.47738/jads.v5i1.160.

[16] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 15, no. 2, 2024, doi : 10.14569/IJACSA.2024.0150235.

[17] D. M. Gunarto, S. Sa, and D. Q. Utama, "Predicting Cryptocurrency Price Using RNN and LSTM Method," *Jurnal Sistem Informasi dan Komputer*, vol. 12, pp. 1–8, 2023, doi: 10.32736/sisfokom.v10i3.1554.

[18] K. Sofi, A. S. Sunge, S. R. Riady, and A. Z. Kamalia, "Perbandingan Algoritma Linear Regression, LSTM, dan GRU dalam Memprediksi Harga Saham dengan Model Time Series," *SEMINASTIKA*, vol. 3, no. 1, pp. 39–46, Nov. 2021, doi: 10.47002/seminastika.v3i1.275.

[19] D. R. Alghifari, M. Edi, and L. Firmansyah, "Implementasi Bidirectional LSTM untuk Analisis Sentimen Terhadap Layanan Grab Indonesia," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 12, no. 2, pp. 89–99, Sep. 2022, doi: 10.34010/jamika.v12i2.7764.

[20] N. A. Dirfas and V. R. S. Nastiti, "Perbandingan Kinerja Pre-Trained Word Embedding Terhadap Performa Klasifikasi Sentimen Ulasan Produk Tokopedia Dengan Long Short-Term Memory(LSTM)," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, Sep. 2024, doi: 10.47065/bits.v6i2.5634.

[21] W. Aljedaani, F. Rustam, S. Ludi, A. Ouni, and M. W. Mkaouer, "Learning Sentiment Analysis for Accessibility User Reviews," *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW), 2021*, pp. 239–246. doi: 10.1109/ASEW52652.2021.00053.