Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

Multi-Aspect Sentiment Analysis of Movie Reviews Using BiLSTM on Platform X Data

Astria M P Sinaga*, Yuliant Sibaroni, Sri Suryani Prasetyowati

School of Computing, Informatics, Telkom University, Bandung, Indonesia
Email: [1,*]triameilynp@student.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id, [3]srisuryani@telkomuniversity.ac.id
Correspondence Author Email: triameilynp@student.telkomuniversity.ac.id
Submitted: **04/06/2025**; Accepted: **30/06/2025**; Published**: 30/06/2025**

Abstract−The film industry generates scores of movie reviews annually, reflecting viewer opinion towards various aspects of movies such as story, music, performances, and so on. They are a good source to publicly analyze opinion automatically. Aspect-based and sentiment analysis of movie reviews based on a multitask classification model rooted in the Bidirectional Long Short-Term Memory (BiLSTM) structure is the theme of this study. The objective of this research is to develop and evaluate a multitask BiLSTM-based model capable of simultaneously classifying sentiment polarity and movie review aspects to enhance fine-grained opinion mining. Data was collected from Platform X through web crawling and subjected to various text preprocessing steps before feeding them into the model. Unlike traditional approaches that treat sentiment and aspect classification as independent operations, the method proposed in this work is performing both simultaneously—sentiment prediction (positive, neutral, negative) and aspect categories (plot, music, actors, others). The model was compared between three different sizes of BiLSTM layers—32, 64, and 128 units—to investigate the influence of model capacity on performance. A 10-fold cross-validation scheme also implemented to confirm the reliability and robustness of results. Experiment findings reveal that the 128-unit BiLSTM model outperformed other models across the board, particularly at picking up subtle contextual relationships, to achieve the highest accuracy score in both tasks. Although this model significantly longer to train, its improved generalization—most notably for difficult sentiment- aspect pairs such as neutral or low-resource categories—validated the trade-off. The findings validate the effectiveness of BiLSTM-based multitask learning for comprehensive movie review analysis, demonstrating the importance of model capacity in tackling complex language patterns and fine-grained opinion identification.

Keywords: Sentiment Analysis; Multi-Aspect; BiLSTM; Multitask

1. INTRODUCTION

Film-making is now part of modern living and plays an important role in the personal experience and social setting. The film industry produces millions of films every year, propelled primarily by greed for money. Sadly, quality of the film does not follow the hype. Film trailers do sometimes misrepresent the viewer, and thus spectators end up disenchanted and disappointed when nothing goes as predicted[1].

With internet access, in the digital age, it is now easier than ever before for audiences to access and exchange information. Most individuals check online reviews before they go to see a movie in order to determine the public perception. Social media websites, especially *Platform X*, have become the major platforms on which users give opinions in the shape of reviews, comments, and tweets[2]. Such texts generated by the users are insightful into the general public sentiment and can be tapped using sentiment analysis methods [3].

Sentiment Analysis is a subdiscipline of *natural language processing* (NLP) that is concerned with the identification of emotional sentiments, positive, negative, or neutral, in text content. Sentiment analysis is a necessary tool to record the public sentiment of a product, service, or event. However, observing the overall sentiment of a review is not informative enough, especially when the review contains mixed opinions about different things. For example, a review of a movie can have approbation regarding the storyline but disapproval regarding the music or acting, such as "I liked the plot but hated the music" [4], [5]. Such examples put in perspective the relevance of *aspect-based sentiment analysis* (ABSA), which attempts to classify sentiment at a grain of individual features or elements.

Early efforts in sentiment analysis employed traditional machine learning approaches, such as *Naïve Bayes*. One experiment that used *Naïve Bayes* on tweet datasets did a low accuracy of 65%, an indication of its weakness in processing intricate, context-dependent text [6]. Another experiment had a higher result of 78.96% but was still hindered by subtle expressions of sentiment, which shows that this approach might not be the best for high-complexity datasets [7]. *Random Forest*, another popular classifier, was also superior with accuracy at 86%, though seen to generalize badly to new data [8]. *K-Nearest Neighbor* (KNN), even though it is easy and has a good accuracy of 87%, is prone to overfitting if K is small and is sensitive to the choice of parameter K [9].

Deep learning has revolutionized the performance of sentiment analysis. *Long Short-Term Memory* (LSTM), an extension of *Recurrent Neural Networks* (RNNs), has been shown to extract short- and long-term dependencies in text. For example, LSTM achieved an accuracy of 88.46% on movie review classification and 89.9% on the IMDb data set prepared by Andrew Maas [10], [11]. While LSTM models are effective, they are computationally costly and susceptible to overfitting when trained on small or noisy data. However, these approaches have not always been valuable for yielding significant accuracy improvements over other methods. In addition, their utility remains prejudiced by the lack of multilingual capabilities and dataset English-only dependence, which restricts their generalizability. Future work in this area has proposed combining feature selection methods to improve overall model performance [12].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

Comparative studies have also shown that models like *Bag of Words* with *Random Forest*, *Support Vector Machines* (SVM), or *Logistic Regression* can achieve accuracies as high as 75.59% [13]. These approaches generally fall short in having the context richness required in advanced sentiment tasks, again highlighting the importance of model selection and hyperparameter tuning.

While existing studies have explored various sentiment analysis methods, most either focus solely on overall sentiment or address aspect-based sentiment classification without integrating both tasks in a unified framework. Furthermore, many rely exclusively on English-language datasets and lack adaptability to more contextually nuanced, real-world user reviews. These limitations highlight the need for a model capable of capturing both sentiment polarity and specific aspects within a review simultaneously and in a multilingual setting. To address this gap, we propose the implementation of a *BiLSTM-based multi-aspect sentiment analysis system* for movie reviews collected from *Platform X*.

To counter these limitations, *Bidirectional Long Short-Term Memory* (BiLSTM) has been proposed as a more efficient solution. By processing the data both forward and backward, BiLSTM maintains a richer context, yielding better sentiment classification accuracy. In one study using BiLSTM for the analysis of social media product reviews, a staggering accuracy of 96.58% was achieved, demonstrating its promise for *multi-aspect sentiment analysis* [4]. BiLSTM's ability to access both the previous and next text context enables more comprehensive comprehension of user emotions.

Furthermore, efforts have been made to combine BiLSTM with advanced language models such as BERT (*Bidirectional Encoder Representations from Transformers*). Ma et al. (2021) designed a *Multi-task Learning Double BiLSTM Model* (MLDBM), combining BiLSTM and BERT to conduct *aspect-based sentiment analysis* in clinical reviews with accuracy at 75. 59% [14]. This hybrid approach had great potential for *domain-specific sentiment classification*. Other studies using the combination of BiLSTM and BERT achieved higher accuracy, for instance, 81%, particularly in handling temporal and contextual complexities of text data [15].

The model is designed using a multitask learning framework to classify both sentiment (positive, negative, neutral) and aspects (plot, music, acting, others). By leveraging the strengths of BiLSTM in capturing bidirectional context, this research aims to provide deeper insights into user sentiment and contribute to more granular understanding in the domain of film review analysis.

2. RESEARCH METHODOLOGY

2.1 Research Stages

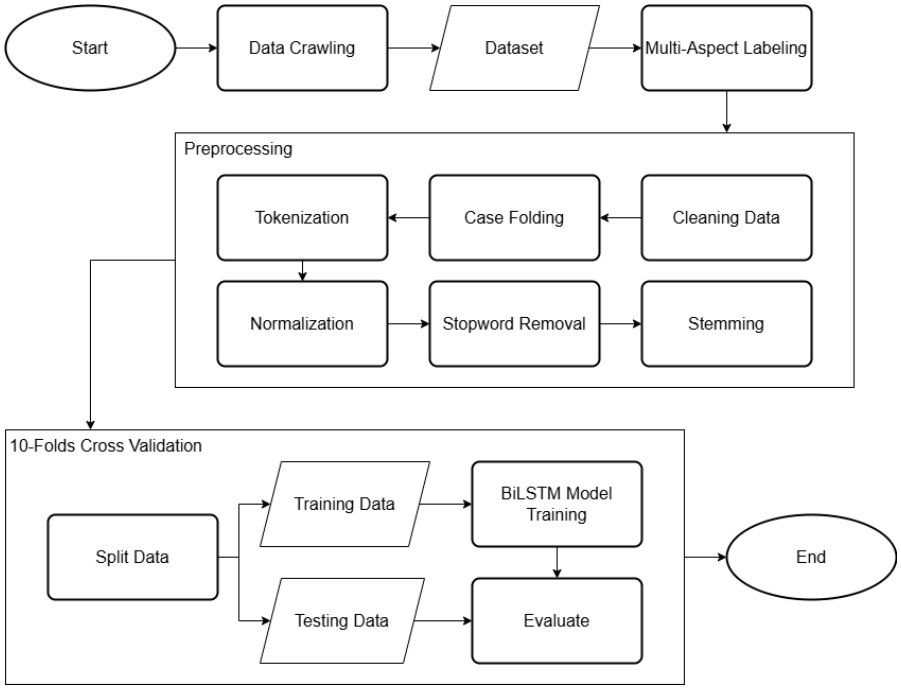


Figure 1. System Flowchart

As shown in Figure 1, the approach used in this sentiment analysis involves several key steps that support effective classification and grading of film reviews, both on the dimension of sentiment polarity as well as aspects. Data gathering is the initial step in the process in which data pertaining to the task is gathered by means of a technique called web crawling from *Platform X*. The platform is a storehouse of user-generated data in the form of movie reviews, comments, and tweets with different opinions that serve as the bricks for sentiment analysis. The data that is

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

obtained from the website is crucial in understanding the view of the general population towards movies, providing a broad range of data regarding the sentiment of the users.

Following data collection, the process moves on to the next stage of the process: data preprocessing. The preprocessing phase plays a vital role in cleaning and prepping the raw text data to render it fit for analysis. Data cleaning is the initial step in preprocessing, where all the irrelevant characters, special characters, or even any HTML tags that have been added while web crawling are removed. This is to ensure only useful textual information is left to be analyzed. Finally, case folding is employed to make the text normalized so that it will then turn into all lowercase so there will be no inconsistency because of capitalization differences, so "Movie" and "movie" will be handled as the same thing. Tokenization, the second process, splits the text into very tiny segments called tokens, typically single words. This enables the model to treat each word as a distinct feature for classification. After tokenization, stemming is carried out, reducing words to their root form. For example, "running" would be reduced to "run, " so the model can examine the root meaning of the words, not their various forms. Stopword removal is then carried out, where common words that don't contribute much to the meaning of the analysis, such as "and, " "the, " or "is, " are removed. This process reduces noise in the data and improves computational efficiency. The final step of preprocessing is normalization, which normalizes text to account for informal language, slang terms, or misspelling and offers consistency in the dataset.

The subsequent step is the application of the BiLSTM algorithm to the training set. BiLSTM is a deep learning model that is capable of processing data in both directions, forward and backward, making it able to see a wider context of the text. This two-way approach is particularly beneficial for sentiment analysis since it allows the model to acquire the interaction between words in both directions, which enhances the accuracy of the sentiment classification. In multi-aspect sentiment analysis, sentiments are classified into three aspects: positive, negative, and neutral, using the BiLSTM model. Furthermore, the model identifies different dimensions of the review, i. e. , plot, music, acting, etc. , which enables it to provide a high-grained breakdown of sentiment for every aspect of the movie.

Finally, the BiLSTM model performance is evaluated based on different critical metrics, i. e. , F1-score, accuracy, precision, and recall. These are used to validate how well the model performs in sentiment classification as well as aspect classification. The F1-score specifically is a balanced score that accounts for both recall and precision in that it is to ensure that the model accurately detects both positive and negative sentiments with minimal false negatives and false positives. The testing process serves the purpose of ensuring the correctness of the model as well as measuring its effectiveness in correctly predicting sentiment and aspects of movie reviews in *Platform X*.

Lastly, the system in Figure 1 attempts to realize an end-to-end sense-integrated solution for multi-aspect sentiment analysis of movie reviews. Through the BiLSTM model, a state-of-the-art deep learning approach, this system aims to achieve richer analysis of user sentiment and identifies the critical factors that influence audience perception of films. The process implemented guarantees that the data is completely prepared, the model is well trained, and the outcomes are compared with set metrics so as to provide worthwhile insights towards understanding the opinion of the public in the movie industry.

2.2 Data Collection

Data used in this research were obtained by extracting film reviews from *Platform X*, which contain users' opinions on several aspects of films, such as plot, actors, and soundtrack. *Platform X* was chosen since it contains a vast and dense amount of user-generated content in its database, which can be utilized to acquire detailed public sentiments in regard to a vast number of films. The vast number of users on the platform allow detailed multi-aspect sentiment analysis, making it a suitable source for the research.

The data was gathered using a web crawling technique, which downloads publicly available text information from the website automatically. After the crawling process was finished, the data collected was filtered such that the reviews contained information that was relevant to the factors being analyzed. Filtering is used to ensure that only reviews that made comments on significant factors, such as plot, acting, and music, were retained for analysis. Processed data was then saved in a structured form, which allows for easy handling of the following preprocessing operations. Preprocessing stages include operations such as text cleaning, tokenizing, and labeling the aspects, prepping the data into a form suitable for the following analysis using the BiLSTM method.

For the purpose of this study, a total of 16705 reviews were successfully extracted. These reviews constituted six distinct movie titles with a different number of reviews for each. The distribution of reviews per movie is presented in the table below:

Table 1. Number of Reviews per Movie Title

No.	Movie Title	Number of Reviews
1	Agak Laen	1926
2	1 Kakak 7 Ponakan	1978
3	Bila Esok Ibu Tiada	1154
4	Home Sweet Loan	1705
5	Ngeri-Ngeri Sedap	647
6	Pengabdi Setan	9344

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

As shown in Table 1, Pengabdi Setan had the highest number of reviews, with 9344 entries, and Ngeri-Ngeri Sedap had the lowest, at 647 reviews. The numbers provide a total dataset on which to train and evaluate the sentiment analysis model.

This process of data collection is crucial for ensuring that the model has enough diverse and representative data to perform accurate multi-aspect sentiment analysis. By leveraging the large pool of user reviews from *Platform X*, the system can effectively analyze various aspects of films, allowing for detailed and context-rich sentiment classification.

2.3 Multi-Aspect Labeling

Multi-Aspect Sentiment Analysis is a more specialized method of sentiment analysis that deals with the assessment of opinions on different aspects or attributes of an entity, e. g. , a product or a film. In the case of films, typical aspects analyzed are plot, music, casting, and visual[5].

This approach differs from general sentiment analysis, which typically identifies document-level sentiment of a text as being positive, negative, or neutral[8]. The distinction is significant since multi-aspect analysis seeks to extract user opinion of individual aspects rather than combining sentiment at the document level.

In this study, manual labeling of the data was conducted. Each review was annotated with respect to two most important categories: sentiment polarity (negative or positive) and aspect (cast, music, or plot). This labeling is done so that the model can be trained to detect and categorize sentiment better in relation to specific aspects of user reviews.

The plot section is where reviews make comments about the plot, such as surprises or the structure of the story. The music section includes things like the OST, background music, or any audio effects (such as crashing and dragging chairs in horror movies) that provide an emotional or atmospheric depth to a film. The casting dimension is comments on actors or the acting out of characters, such as criticism or compliment of performance, expression, or embodiment of character.

The reason for splitting labels into these dimensions is to enable the model to better recognize the sentiment pertaining to certain content in the review[5]. Table 2 provides real-life instances demonstrating how review data were annotated according to sentiment polarity and movie aspect. Each row presents an instance of a movie review sentence and its corresponding sentiment and aspect annotation. These instances are intended to demonstrate how manual annotation is able to capture nuanced opinions targeted towards different aspects of the film.

Table 2. Examples of Labeled Review Data

No.	Review	Sentiment	Aspect
1	Akhirnya gue menyempatkan waktu buat nonton Agak Laen. Gue kagum sama betapa rapi dan detailnya film ini. Tentunya buah dari penulisan yang baik dan kemampuan sutradara "membaca" visual	Positive	Plot
2	Humor yang regresif dalam film, terutama melalui penggunaan karakter dengan disabilitas hanya sebagai alat untuk lelucon	Negative	Casting
3	Musik dalam film Agak Laen mendukung di setiap adegan film. Soundtrack yang digunakan mampu membangun emosi penonton dengan baik	Positive	Music

As shown in Table 2, each review highlights a different component of the movie. The first review reflects appreciation for the storyline and visual interpretation, which is thus labeled as positive under the plot aspect. The second review criticizes how certain characters are portrayed, particularly with regard to sensitivity and representation, hence it is marked as negative under casting. The third review praises the music and emotional impact of the soundtrack, resulting in a positive label under music.

To enable consistent and reproducible annotation, a systematic definition and set of keywords for every aspect were established. The keywords serve as semantic markers to enable manual labeling and future automation.

Table 3 presents explicit definitions of every aspect used in the task of labeling along with corresponding keywords. The keywords were used as reference terms for manual as well as automatic recognition of aspects in the text data.

Table 3. Definitions and Keywords for Each Aspect

No.	Aspect	Definition	Keywords
1	Plot	A review that describes how the story progresses in the film, including plot twists, conflicts, climaxes, and resolutions.	plot twist, conflict, climax, resolution, ending
2	Casting	A review discussing the casting process or comments about actors and the characters they portray.	actor, actress, character, role, performance, chemistry, lead role, expression, embodiment

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

No.	Aspect	Definition	Keywords
3	Music	A review that refers to or supports dramatic scenes, enhances tension or emotion, and focuses on music or sound in the film.	soundtrack, OST, song, music

Table 3 plays a crucial role in defining boundaries between aspects and ensures clarity in the annotation process. These keywords were also later utilized in the cosine similarity matching method to semi-automatically assign aspects to review texts based on their semantic similarity.

To improve the aspect-based labeling precision and consistency, a cosine similarity-based approach was used to make an estimation of the semantic distance between words from the reviews and the pre-determined keywords for each aspect. Cosine similarity is a measurement to determine how similar two text vectors are, with words as high-dimensional spaces.

If the cosine similarity score of a word in the review to an aspect keyword is greater than a threshold, the word is flagged as relevant to the aspect. This allows the model to capture semantic relation even where the very keyword is not there, thereby increasing the robustness and flexibility of the labeling.

This method was adapted from Ghadery et al. (2019), who have used soft cosine similarity in their work to perform unsupervised aspect category detection. Their method not only takes into account exact keyword matching, but also semantic closeness in word vector space of higher dimension [16].

2.4 Bidirectional Long Short-Term Memory (BiLSTM)

Bidirectional Long Short-Term Memory (BiLSTM) is a variation of the Long Short-Term Memory (LSTM) model with the capacity to process sequential data in two directions: forward (left to right) and backward (right to left). The bidirectional approach enables the model to learn the overall context of words in a sentence and, therefore, make sense of past and future dependencies in the text more effectively. Since BiLSTM processes information in both directions, it provides a deeper semantic relationship that is crucial for natural language processing tasks such as NLP, in which context is a necessary ingredient for correct interpretation [17].

BiLSTM is an extension of the LSTM model, which was created especially to address difficulties with sequential data, including the vanishing gradient problem, which is a common issue in traditional neural networks. LSTM units can retain information over long periods, and therefore are highly effective at modeling long-range dependencies of sequences. BiLSTM enhances this capability by processing the sequences in both directions so that it can consider not only the context preceding it but also the context succeeding it in a sentence. This bidirectional flow of information is useful most particularly when one is trying to understand sentence structure or predict a word's meaning based on the context words. [18].

For example, when establishing the meaning of a sentence, a standard LSTM model can struggle to cope with ambiguities due to word order. BiLSTM, however, can cope with such ambiguities more effectively by considering the entire context provided by words that come before and after a specific word. This renders BiLSTM a better approach for applications like language modeling, sentiment analysis, and machine translation, where bidirectional context is essential in making accurate predictions [19].

In addition, BiLSTM architecture contains some very important aspects such as the input layer which is representative of the sequence of words or any other features entering the model. Cell state is another very important parameter that retains long-term memory of the sequence. The output gate decides what information from the cell state should be passed to the output so that only the most important information is passed on. BiLSTM's capacity for sequential time step data processing, where each step has its input and output format, enables it to grasp patterns of change in time-series data or text sequences [20].

This study adopts a multitask learning route for joint classification of both sentiment polarity and aspect categories of movie reviews. This enables the model to learn two types of similar tasks at once, not just conserving computationally expensive time and resources but also generalizing better via shared representation learning between tasks. In this implementation, the Bidirectional Long Short-Term Memory (BiLSTM) network is used as the main feature extractor to process contextual relations in the direction of both text and text reverse. The output of BiLSTM is passed into a shared dense layer that learns general features that can be applied to both tasks. These representations are then passed to two task-specific branches, one for sentiment classification and the other for aspect category classification. Each branch consists of dense and output layers, enabling the model to simultaneously predict sentiment tags (positive, neutral, negative) and aspect categories (actor, plot, music, others) for each review. Such merged architecture enables the model to share common semantic knowledge and optimize performance for each separate classification task.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

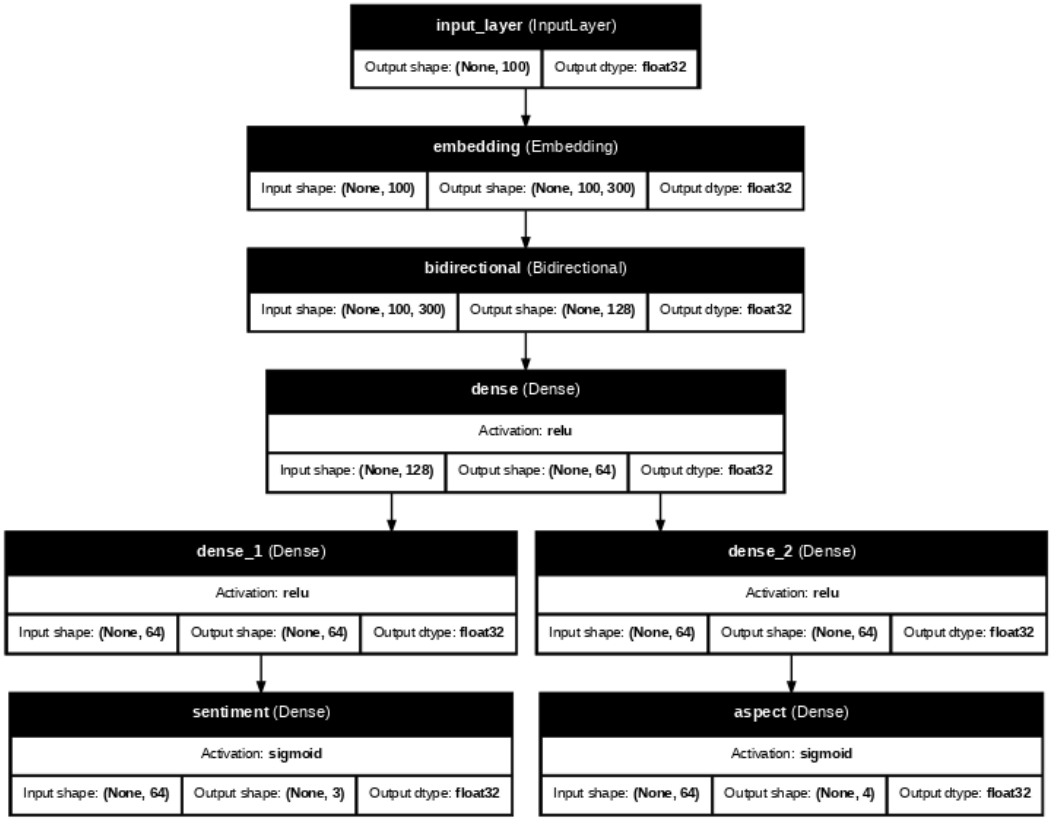


Figure 2. BiLSTM Model Architecture Used in This Study

The architecture, as depicted in Figure 2, is optimized for precision and efficiency in carrying out both the task of classification within an integrated training process. The model can generalize more across tasks by using joint learning, not just in its ability to classify sentiment accurately but also its ability in labeling multiple dimensions contained in a single review. The model is implemented with the TensorFlow Functional API, which gives high flexibility when defining complex network architectures.

The architecture consists of several basic elements, each dedicated to a specific task of feature extraction and classification. The details of the model architecture are provided below:

a. Input Layer
   The input layer accepts preprocessed text data, which is tokenized and padded, and transforms it into fixed-length numeric vectors. Each word sequence in the review is translated into a numeric form that can be handled by the neural network.
   1. Input shape: Sequence length of 100 tokens (100,).
   2. Function: Receives tokenized and padded text sequences of movie reviews.

b. Embedding Layer
   This layer uses a pre-trained matrix of word embedding (with trainable=False), turning every token into a fixed vector that captures the semantic meaning of the word. These embeddings lay the foundation of the meaning for word contexts within the reviews.
   1. Embedding type: Pre-trained FastText embeddings 300 dimensions.
   2. Trainable: False (embedding weights are frozen).
   3. Function: Maps tokens into dense vectors that capture semantic word meanings.

c. Bidirectional LSTM (BiLSTM) Layer
   The key building component of the model is a Bidirectional Long Short-Term Memory (BiLSTM) layer. The network is bidirectionally processed both forward and backward to allow the model to leverage more richer contextual relations between preceding and following words. Bidirectional context is similarly valuable in examining the fine-grained sentiment and aspectual content within user-uploaded reviews.
   1. Tested values : 32, 64, and 128 units.
   2. Units: n units (forward) + n units (backward) = combined output of $2n$ dimensions.
   3. Bidirectionality: Processes input both forward and backward to capture full context.
   4. Function: Learns contextual relationships between words from both directions.

d. Shared Dense Layer
   The output of the BiLSTM is passed through a shared dense layer with ReLU activation. This is a shared feature representation which is fed to both output branches (aspect and sentiment classification). This is the most important

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

component of the multitask learning setup in that it allows the model to learn shared patterns which are beneficial to both tasks.

1. Units: Variable; 32, 64, or 128 units (matching BiLSTM configurations).
2. Activation: ReLU.
3. Function: Extracts shared feature representations used by both tasks.

e. Task-Specific Dense Layers
Following the creation of the shared features, the model diverges into two task-specific branches:
1. x_sentiment : A dense layer with ReLU activation for sentiment classification.
   a) Units: 32, 64, or 128 units (matching BiLSTM configurations).
   b) Activation: ReLU
   c) Purpose: Extracts features specific to sentiment classification.
2. x_aspect : A dense layer with ReLU activation for aspect classification.
   a) Units: 32, 64, or 128 units (matching BiLSTM configurations).
   b) Activation: ReLU
   c) Purpose: Extracts features specific to aspect classification.

f. Output Layers
1. Sentiment Output : This output layer is made up of 3 neurons with sigmoid activation, which is for the three sentiment classes: positive, neutral, and negative. Although sigmoid activation is normally used in multi-label classification, in practice, only one sentiment label is activated for each review.
   a) Units: 3 (positive, neutral, and negative classes)
   b) Activation: Sigmoid
2. Aspect Output : This output layer has 4 neurons, using sigmoid activation, for the aspects: actor, plot, music, and others. A review can belong to more than one aspect simultaneously, so a multi-label format is warranted.
   a) Units: 4 (actor, plot, music, others)
   b) Activation: Sigmoid

With contextualized representations in the form of BiLSTM and task-specific routes of output branches, the model supports intricate interplays among words, emotions, and target spaces, with satisfactory performance on multi-aspect sentiment analysis.

3. RESULT AND DISCUSSION

In this current study, Bidirectional Long Short-Term Memory (BiLSTM) model was introduced to conduct dual-task classification simultaneously, i.e., sentiment classification and aspect classification of movie reviews. The model was trained and tested on a 16705-movie review dataset. Contrary to the traditional train-test splits, in this work, a 10-fold cross-validation method was utilized to obtain a more stable, comprehensive, and unbiased test. The data were divided into ten sets, and nine sets were used for training and one for testing the model in a reiterated process. The process enhances the generalizability and reliability of the findings of the test. The model was designed to generate two primary outputs:

a. Sentiment Classification: The model could classify movie reviews into three sentiment labels: positive, negative, and neutral.
b. Aspect Classification: The model could also identify main aspects discussed in the reviews, such as plot, casting, music, and others.

3.1 WordCloud

Besides quantitative indicators, word clouds were constructed to provide qualitative information on the content of the reviews. Separate word clouds were made for every sentiment class (positive, neutral, and negative) and every found aspect category (plot, casting, music, and others). The word clouds are used to highlight the most frequent words in each category and thereby make it easier to understand what kinds of themes or keywords are normally associated with particular sentiments or aspects.

(a) Positive (b) Neutral (c) Negative

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

Figure 3. Wordcloud of Music Aspect

(a) Positive (b) Neutral (c) Negative

Figure 4. Wordcloud of Plot Aspect

(a) Positive (b) Neutral (c) Negative

Figure 5. Wordcloud of Actor Aspect

(a) Positive (b) Neutral (c) Negative

Figure 6. Wordcloud of Others Aspect

3.2 Experimental Setup

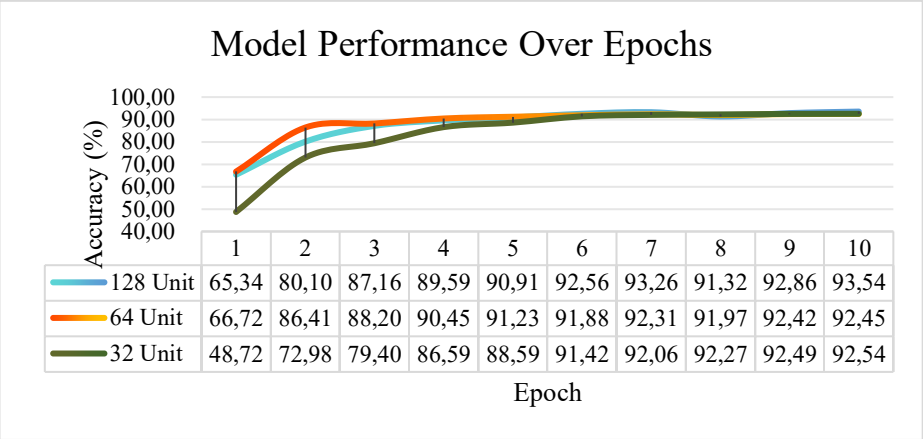
During the experimental phase, the BiLSTM model was executed for aspect and sentiment classification tasks. To examine the effect of model capacity on performance, three experiments were conducted with different LSTM layer sizes: 32, 64, and 128 units. The primary objective of these experiments was to compare model performance across the three LSTM configurations and to determine which layer size yields the best result for each classification task.

All environments were trained on the same train-test data splits, and all experiments were conducted under the same conditions, with the same embedding matrix, optimizer, learning rate, and batch size. This controlled experiment environment ensured that any variation in performance could reasonably be explained by the LSTM layer size.

3.3 Results

Three BiLSTM models were tried out with differences in the number of hidden units (32, 64, and 128). In all three models, training was followed epoch by epoch to search for learning trends and signs of overfitting or convergence, 10-fold cross-validation employed to obtain solid and generalizable estimates of performance. Each configuration was subsequently trained for 10 epochs with each case's performance averaged over folds to mitigate the effects of sampling bias, Figure 7 show accuracy plots by epochs for each model variant. In the initial epoch the 32-unit configuration had the worst accuracy of 48.71%, while the 64-unit and 128-unit configurations performed considerably better at 66.72% and 65.34%, respectively.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509



	1	2	3	4	5	6	7	8	9	10
128 Unit	65,34	80,10	87,16	89,59	90,91	92,56	93,26	91,32	92,86	93,54
64 Unit	66,72	86,41	88,20	90,45	91,23	91,88	92,31	91,97	92,42	92,45
32 Unit	48,72	72,98	79,40	86,59	88,59	91,42	92,06	92,27	92,49	92,54

Figure 7. Model Performance Over Epochs

As training accuracy improved, all configurations consistently show a performance improve. The 64-unit setup had already reached 91.23% at epoch 5, but the 128-unit setup surpassed it in the later epochs, ultimately reaching the best accuracy of 93.54% at epoch 10. Surprisingly, the 32-unit configuration, despite being behind, came close to the final accuracy of the deeper networks and reached 92.54% at epoch 10. These results suggest that while deeper LSTM layers accelerate convergence and provide higher representational capacity, smaller configurations are able to achieve good performance with sufficient training.

Further model prediction analysis breakdown is shown in Table 4, presenting sentiment classification accuracy by aspect (Music, Plot, Casting, Others) and by sentiment category (Negative, Neutral, Positive) for every BiLSTM configuration. The breakdown view gives more grain in perceiving how model capacity affects sentiment detection over content domains.

Table 4. Summary of Sentiment over Aspect Performance across BiLSTM Configurations

Aspect	BiLSTM Units	Sentiment		
		Negative (%)	Neutral (%)	Positive (%)
Music	32 Units	99.95	58.25	99.98
	64 Units	99.9	58.38	99.72
	128 Units	99.95	62.87	100
Plot	32 Units	99.23	26.01	99.96
	64 Units	90.42	44.79	99.86
	128 Units	100	37.95	100
Casting	32 Units	88.38	0	95.53
	64 Units	88.39	0	96.39
	128 Units	87.74	3.08	99.65
Others	32 Units	15.99	3.51	96.85
	64 Units	45.05	13.66	96.07
	128 Units	28.63	18.04	96.36

Figure 8 presents the effect of varying BiLSTM layer units on sentiment classification accuracy within the Music aspect. For negative sentiment, accuracy was consistently high for all unit capacities, with minimal fluctuation between 99.90% and 99.95%, demonstrating model steadiness regardless of capacity. Conversely, for neutral sentiment, there was a trend towards improvement with greater layer units—58.25% when the 32 units are used to 62.87% when 128 units are used—highlighting that more-capacity models better identify subtle emotional expressions in music reviews. Positive sentiment classifying also benefited slightly from additional units, scoring a perfect score (100%) with 128 units.

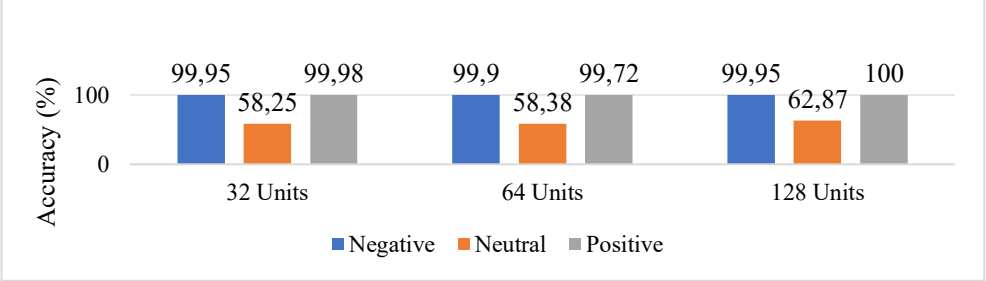


Figure 8. Effect of BiLSTM Layer Units on Sentiment Classification Accuracy within the Music Aspect

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

In Figure 9, how performance changes with BiLSTM layer size in the Plot area is represented. The biggest change was realized in classifying negative sentiment, which changed from 90.42% (64 units) to 100% (128 units), whereas the 32-unit model was already showing a high performance at 99.23%. For neutral sentiment, accuracy showed non-linear performance—beginning low at 26.01%, going as high as 44.79% with 64 units, and decreasing minimally to 37.95% at 128 units—indicating intermediate layer sizes could offer best balance for this class. Positive sentiment classification was always very high, going up to 100% with the greatest number of units.

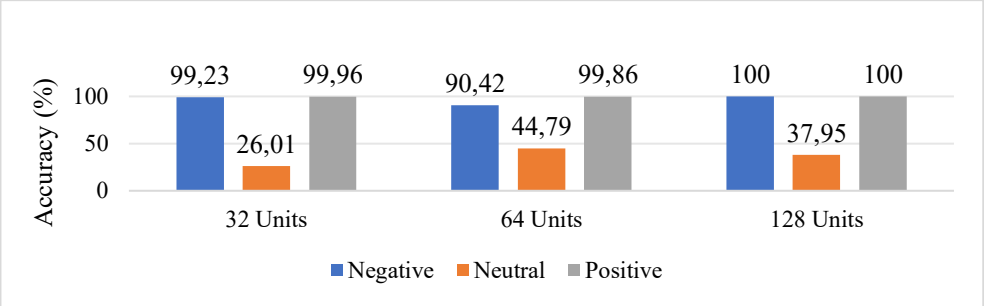


Figure 9. Effect of BiLSTM Layer Units on Sentiment Classification Accuracy within the Plot Aspect

Figure 10 shows the performance variation of the model over different BiLSTM unit sizes for the Casting dimension. Accuracy for the negative sentiment class was relatively uniform across configurations, deviating narrowly between 87.74% and 88.39%, with minimal sensitivity to layer depth. The neutral sentiment class had extremely low figures for smaller models (0% at 32 and 64 units), but marginally improved to 3.08% at 128 units, a small but telling improvement, showing an improvement in representational power. Positive sentiment classification also showed a very steep increase from 95.53% to 99.65% with an increase in the number of units.

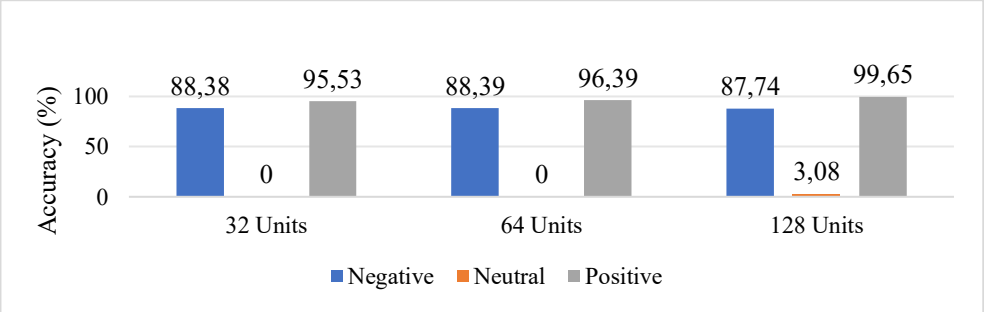


Figure 10. Effect of BiLSTM Layer Units on Sentiment Classification Accuracy within the Casting Aspect

Figure 11 shows how changes in BiLSTM layer size affected performance for the Others aspect. Negative sentiment classification improved from 15.99% at 32 units to 45.05% at 64 units but then declined to 28.63% with 128 units, suggesting diminishing returns or potential overfitting at higher capacities. For neutral sentiment, there was a consistent upward trend, starting at 3.51% and rising to 18.04%, confirming that larger models offered better capability for distinguishing neutral tone. Positive sentiment performance remained strong and relatively stable across all unit sizes, ranging from 96.07% to 96.85%.

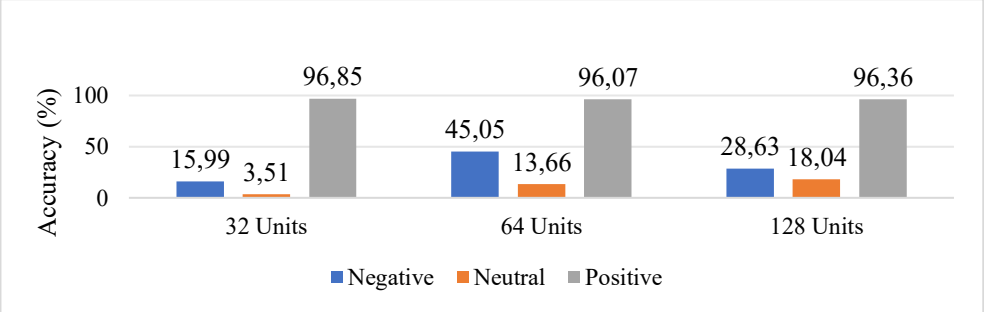


Figure 11. Effect of BiLSTM Layer Units on Sentiment Classification Accuracy within the Others Aspect

The model consistently demonstrated extremely high performance for classifying positive sentiment, particularly across Music and Plot dimensions, with accuracy reaching up to 100% in many configurations. Similarly, classification of negative sentiment for these dimensions was extremely high, well above 99% in most of the instances.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

Such observations indicate that the model performs extremely well for strongly polarized sentiments for rich-content dimensions where emotional expression is more explicit and linguistically differentiated.

Neutral sentiment was much harder to predict across all categories. The most accurately forecast neutral sentiment was in the Music–Neutral category and reached 62.87% with the 128-unit model. The model did very poorly, however, in both the Plot–Neutral and Others–Neutral categories, at best accuracies of 44.79% and 18.04%, respectively. The biggest challenge was, however, in the Casting–Neutral category. For the 32-unit and 64-unit models, the model didn't even manage to correctly classify a single one of the Casting–Neutral samples and scored 0% accuracy. The 128-unit model alone managed to have a little bit of improvement in this class and scored 3.08% accuracy. This class must be mentioned to have had only 12 support samples in the test set and thus was also susceptible to class imbalance and training underrepresentation and quite possibly bore the brunt of the very poor performance. Of the four categories, Music was the best consistently classified category for all types and configurations of sentiments.

This is because the generally categorically sentiment expressions used in music reviews, typically involving emotionally descriptive words that ease classification, are best suited. The Plot category followed, especially for positive and negative sentiments. Casting and Others, however, were more inconsistent and overall did worse. The Casting feature, while with highly high accuracy for positive sentiment to 99.65%, was very limited in classifying into neutral. The Others category, being miscellaneous and possibly confusing content, was also difficult, particularly for identifying negative sentiment where as little as 15.99% with the 32-unit model. Overall, these results point out the strengths and weaknesses of the BiLSTM model for multi-aspect sentiment classification.

The model performs well in polarized sentiment classification—particularly in expressive language applications—although neutral sentiment remains difficult, especially with imbalanced or sparse training data. The more LSTM units employed, generally the better the performance will be, especially for harder-to-classify sentiment classes. Yet, the lower-capacity models (32 and 64 units) can achieve competitive performance, demonstrating the model's robustness and flexibility under different settings.

3.4 Discussion

The experiment results verify that the Bidirectional Long Short-Term Memory (BiLSTM) model is an effective architecture to perform multitask classification, specifically sentiment prediction. and aspect recognition in movie reviews. The bidirectional model was able to capture dependencies both in forward and backward directions in the input sequences, thus it captures the subtlety of sentiment, especially when there are contextual cues before or after significant sentiment-bearing words. This capability was particularly valuable in the highly polarized sentiment class, as seen through the extremely high accuracy for negative and positive sentiment classes for well-established dimensions such as Music and Plot. The model performed extremely high on the classification of Music–Positive and Plot–Positive pairs, to a 100% using the 128-unit configuration, indicating that these dimensions have highly separable rich patterns for the model to learn.

However, as powerful as it was in identifying polarized feelings, the model struggled in identifying neutral feelings, especially in low-resource classes. One of the more inferior points noted was in the Casting–Neutral class, where both 32-unit and 64-unit models failed to accurately predict any example whatsoever, at an accuracy rate of 0%. Even the 128-unit model with greater capacity bested only to 3.08% accuracy. This is largely because of class imbalance since Casting–Neutral only had 12 test instances, and also because neutral sentiment phrases are inherently imprecise, having little emotional information to be captured. Such language is harder for the model to tell apart with only standard word embeddings and sequence modeling. The performance under the Others–Neutral class also remained relatively low across configurations, lending evidence to the challenge of identifying neutrality under weak sentiment cues and sparse training data.

When comparing across model sizes, an interesting trend is evident. The larger models, particularly the 128-unit BiLSTM, generally performed better on more underrepresented or subtle classes. This configuration consistently yielded the best accuracy across most sentiment–aspect pairs, especially where there were subtle distinctions. But that gain was purchased at a great computational cost. The average training time per fold increased from roughly 414 seconds for the 32-unit network to over 1350 seconds for the 128-unit network. Though there was a rise in performance, the increase in final accuracy going from 64 to 128 units was insignificant—at under 1% in some cases—reflecting diminishing returns relative to the cost of training. This is crucial in practical applications where model performance, scalability, and resource constraint must be considered.

The other aspect that must be highlighted is the inconsistency in performance between areas. Areas such as Music and Plot exhibited high consistency in classification, likely due to semantic lucidity and consistent vocabulary being intrinsic with these categories. Nevertheless, the Others area introduced more variability in results, particularly in neutral and negative sentiment classification. This type, being a catch-all, lacks clearly defined thematic boundaries and includes mixed content, contributing semantic noise and hindering learnability. The findings suggest that augmentation of aspect definitions or introduction of aspect-aware components—i.e., dynamically attending attention layers to important words—can enhance classification accuracy in such uncertain domains.

Overall, the results confirm the strength of multitask BiLSTM models for aspect and sentiment classification tasks, particularly with dense data and semantically clear-cut classes. However, the failure of the model to handle neutral sentiment and imbalanced class distributions shows where further strategies are needed. Data augmentation,

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

cost-sensitive learning, or a transition to pre-trained transformer-based models such as BERT may improve the performance, particularly on challenging or underrepresented classes. These results not only demonstrate the potential of BiLSTM-based multitask learning, but also highlight the importance of real-world considerations in model deployment and optimization, as well as in future work.

4 CONCLUSION

This work proposed a Bidirectional Long Short-Term Memory (BiLSTM) based model for multitask learning to perform sentiment and aspect classification on film reviews in parallel. Using a shared classification model, the model was able to transfer learned representations across tasks, improving both efficiency and generalization. The model was validated using a 10-fold cross-validation setup on a corpus of 16705 reviews. Experimental results validated that the BiLSTM model performs extremely well in identifying highly expressed sentiments—positive and negative—particularly when aspect categories are semantically clear and they are highly represented in the dataset. The model achieved accuracy of over 99% for most of the positive sentiment classes, especially Music and Plot aspects. Performance did suffer, though, for the instances of neutral sentiment, especially in the Casting–Neutral class, where smaller configuration predictions were inaccurate. This correctly reflects the fact that it is still difficult to detect neutral sentiments in natural language processing, particularly when there are biased or ambiguous training samples. Moreover, increasing the number of BiLSTM units generally also helped enhance classification performance but at a very high computational cost in the form of a tremendous increase in computational time. The 128-unit model had the best general accuracy but also required more than three times the amount of computational time to train compared to the 32-unit model. This further complicates the process of reaching a balance between model complexity and computational time, especially when running in real-world and resource-constrained environments. In the future, some of the research directions would be very helpful to overcome the challenges encountered in this study. Firstly, methods such as data augmentation, oversampling low-frequency classes, or generating artificially balanced samples would improve the model's capability for handling sentiment classes with sparse data. Secondly, transformer models such as BERT, which offer rich context understanding, may dominate BiLSTM in detecting fine-grained sentiment expressions of ambivalence or neutrality. Thirdly, applying attention mechanisms at word or aspect level might allow the model to selectively concentrate on the most prominent parts of the text and hence further improve aspect detection and overall explainability. Finally, re-structuring or refining the aspect taxonomy, especially for heterogeneous Others category, can reduce semantic overlap and allow for more differentiated classification.

REFERENCES

[1] A. A. Kumar, P. M. Charan Reddy, and N. Gunnam, “Movie Review Based Sentiment Analysis,” *International Journal of Innovative Science and Research Technology (IJISRT)*, vol. 9, no. 8, pp. 706–721, 2024, doi: 10.38124/ijisrt/ijisrt24aug345.

[2] S. Utami, K. M. Lhaksmana, and Y. Sibaroni, “Deep Learning and Imbalance Handling on Movie Review Sentiment Analysis,” *SinkrOn*, vol. 8, no. 3, pp. 1894–1907, 2023, doi: 10.33395/sinkron.v8i3.12770.

[3] J.-H. Wang, T.-W. Liu, and X. Luo, "Combining post sentiments and user participation for extracting public stances from Twitter," Applied Sciences, vol. 10, p. 8035, Nov. 2020, doi: 10.3390/app10228035.

[4] J. Sangeetha and U. Kumaran, “A hybrid optimization algorithm using BiLSTM structure for sentiment analysis,” *Measurement: Sensors*, vol. 25, no. September 2022, p. 100619, 2023, doi: 10.1016/j.measen.2022.100619.

[5] N. Karimah and A. Baita, “Multi-Aspect Sentiment Analysis Pada Review Film Menggunakan Metode Bidirectional Encoder Representations From Transformers (BERT) Multi-Aspect Sentiment Analysis of Film Review Using Bidirectional Encoder Representations from Transformers (BERT),” *Jurnal Sistem Komputer*, vol. 13, no. 1, p. 2020, 2024, doi: 10.34010/komputika.v13i1.11098.

[6] Y. Nurtikasari, Syariful Alam, and Teguh Iman Hermanto, “Analisis Sentimen Opini Masyarakat Terhadap Film Pada Platform Twitter Menggunakan Algoritma Naive Bayes,” *INSOLOGI: Jurnal Sains dan Teknologi*, vol. 1, no. 4, pp. 411–423, 2022, doi: 10.55123/insologi.v1i4.770.

[7] S. Samsir, K. Kusmanto, A. H. Dalimunthe, R. Aditiya, and R. Watrianthos, “Implementation Naïve Bayes Classification for Sentiment Analysis on Internet Movie Database,” *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 1, pp. 1–6, 2022, doi: 10.47065/bits.v4i1.1468.

[8] D. Zheng, “Sentiment Analysis for Film Reviews Based on Random Forest,” *Science and Technology of Engineering, Chemistry and Environmental Protection*, vol. 1, no. 7, pp. 1–5, 2024, doi: 10.61173/5t8epb44.

[9] A. Fadillah, “Sentiment Analysis Towards the Film Dirty Vote on Twitter Social Media Using the K-Nearest Neighbor Algorithm,” vol. 7, no. 2, pp. 541–552, 2024.

[10] J. D. Bodapati, N. Veeranjaneyulu, and S. Shaik, “Sentiment analysis from movie reviews using LSTMs,” *Ingenierie des Systemes d’Information*, vol. 24, no. 1, pp. 125–129, 2019, doi: 10.18280/isi.240119.

[11] S. M. Qaisar, “Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory,” *2020 2nd International Conference on Computer and Information Sciences, ICCIS 2020*, pp. 7–10, 2020, doi: 10.1109/ICCIS49240.2020.9257657.

[12] S. S. Khan and Y. Alharbi, “International Journal of Advanced and Applied Sciences Sentiment analysis of movie review classifications using deep learning approaches,” vol. 11, no. 8, pp. 146–157, 2024.

[13] B. Sangeetha, S. Sangeetha, D. T. Goutham, and N. Vaibhav Ram, “Sentiment Analysis on Movie Reviews: A Comparative Analysis,” *Proceedings of the 2023 International Conference on Intelligent Systems for Communication, IoT and Security, ICISCoIS 2023*, pp. 218–223, 2023, doi: 10.1109/ICISCoIS56541.2023.10100367.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 672−684
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7509

[14] J. Ma, X. Cai, D. Wei, H. Cao, J. Liu, and X. Zhuang, "Aspect-Based Attention LSTM for Aspect-Level Sentiment Analysis," *2021 3rd World Symposium on Artificial Intelligence, WSAI 2021*, pp. 46–50, 2021, doi: 10.1109/WSAI51899.2021.9486323.
[15] Y. Wang, G. Shen, and L. Hu, "Importance evaluation of movie aspects: Aspect-based sentiment analysis," *Proceedings - 2020 5th International Conference on Mechanical, Control and Computer Engineering, ICMCCE 2020*, pp. 2444–2448, 2020, doi: 10.1109/ICMCCE51767.2020.00527.
[16] E. Ghadery, S. Movahedi, H. Faili, and A. Shakery, "An Unsupervised Approach for Aspect Category Detection Using Soft Cosine Similarity Measure," no. September, 2018, doi: 10.48550/arXiv.1812.03361.
[17] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019, doi: 10.1109/ACCESS.2019.2909919.
[18] D. I. Puteri, "Implementasi Long Short Term Memory (LSTM) dan Bidirectional Long Short Term Memory (BiLSTM) Dalam Prediksi Harga Saham Syariah," *Euler : Jurnal Ilmiah Matematika, Sains dan Teknologi*, vol. 11, no. 1, pp. 35–43, 2023, doi: 10.34312/euler.v11i1.19791.
[19] Z. Hameed and B. Garcia-Zapirain, "Sentiment Classification Using a Single-Layered BiLSTM Model," *IEEE Access*, vol. 8, pp. 73992–74001, 2020, doi: 10.1109/ACCESS.2020.2988550.
[20] C. Gupta, G. Chawla, K. Rawlley, K. Bisht, and M. Sharma, *Senti_ALSTM: Sentiment Analysis of Movie Reviews Using Attention-Based-LSTM*, vol. 167. Springer Singapore, 2021. doi: 10.1007/978-981-15-9712-1_18.