Ingénierie des Systèmes d'Information
Vol. 31, No. 8, August, 2026, pp. 2859-2866
Journal homepage: http://iieta.org/journals/isi

An End-to-End BERT–LSTM Framework for Aspect-Based Sentiment Analysis of Animated Movie Reviews

Syafrijon*, Nicholas Ryan Jonathan, Dony Novaliendry, Titi Sriwahyuni, Khairi Budayawan, Ihsanul Insan Aljundi

Department of Electronics Engineering, Universitas Negeri Padang, Padang 25131, Indonesia

Corresponding Author Email: syafrijon@ft.unp.ac.id

https://doi.org/10.18280/isi.310828

Received: 20 February 2026
Revised: 18 July 2026
Accepted: 21 August 2026
Available online: 31 August 2026

Keywords:
Aspect-Based Sentiment Analysis, Bidirectional Encoder Representations from Transformers, Long Short-Term Memory, sentiment classification, movie review analysis

ABSTRACT

Aspect-Based Sentiment Analysis (ABSA) provides a fine-grained approach for analyzing user opinions by associating sentiment polarity with specific textual aspects. This study proposes an end-to-end ABSA framework that integrates Bidirectional Encoder Representations from Transformers (BERT) for aspect extraction and Long Short-Term Memory (LSTM) networks for aspect-level sentiment classification in animated movie reviews. A dataset of 6,453 user reviews collected from the Internet Movie Database (IMDb) was processed into 69,876 sentences, from which 6,987 sentences were proportionally sampled and manually annotated, producing 4,075 sentence–aspect pairs for classification. The BERT-based aspect extraction model achieved a span-level F1-score of 83.06%, demonstrating its capability to identify relevant aspect terms from review texts. The LSTM classifier, using joint sentence and aspect representations, obtained an accuracy of 82% with balanced performance across sentiment categories. When integrated into a complete ABSA pipeline, the proposed framework achieved an end-to-end accuracy of 91.8% and a macro-averaged F1-score of 0.918, outperforming the Term Frequency-Inverse Document Frequency (TF-IDF)-based Logistic Regression baseline. Furthermore, a Streamlit-based interface was developed to support interactive sentiment analysis and visualization. The proposed framework provides a reproducible architecture for domain-specific opinion mining and offers a practical approach for analyzing audience perceptions of animated films.

1. INTRODUCTION

Animated films have become a dominant force in the global entertainment industry, with studios such as Pixar and DreamWorks producing content that attracts millions of viewers across age groups. These films are no longer limited to children but appeal to a broader demographic due to their complex narratives and artistic visuals. As reported in the global animation industry reached a market value of USD 373.23 billion in 2024 and is projected to grow to USD 393.3 billion by 2025 [1]. This rapid development has heightened audience expectations, leading to a surge in user-generated reviews that express opinions on specific aspects such as plot, character design, animation quality, and moral messages.

Digital platforms like Internet Movie Database (IMDb), Rotten Tomatoes, and Letterboxd have enabled viewers to provide not only ratings but also narrative reviews that reflect detailed viewing experiences. According to the study by Ha [2], audience engagement has evolved from mere consumption to active participation through opinion sharing. For instance, "Kung Fu Panda 4" received praise for maintaining visual consistency [3], yet was also criticized for its underwhelming action scenes [4]. Similarly, "Elemental" was commended for its visual quality [5], while the plot was deemed too predictable, raising suspicions of AI-generated writing [6]. These varied perspectives highlight the need to analyze audience sentiment at a more granular, aspect-specific level.

Although a significant number of user reviews are available, scientific methods for extracting meaningful insights from such data remain limited. Traditional sentiment analysis techniques, including lexicon-based approaches and classical machine learning models, often fail to capture complex language nuances. For example, the Support Vector Machine (SVM)-based analysis [7] on "Squid Game" reviews achieved high accuracy but lacked generalization across domains. Similarly, lexicon-based approaches such as those in the study by Wang et al. [8] could not effectively detect implicit sentiments like sarcasm. Transformer-based methods like HauBERT [9] show promise but still face challenges related to imbalanced data and domain-specific adaptation. In Indonesia, IndoBERT was applied in the study by Hendriyani and Budayawan [10] to classify Instagram comments, achieving an accuracy of 91.57%, although the analysis was limited to sentence-level polarity rather than aspect-specific sentiment.

Aspect-Based Sentiment Analysis (ABSA) addresses these limitations by extracting opinions related to specific components within a text [11]. In the context of animated

films, ABSA enables producers to understand how audiences perceive particular features such as animation quality or storyline depth, thereby facilitating more targeted improvements. However, despite the abundance of rich user-generated content, the use of ABSA for analyzing animated film reviews remains underexplored.

ABSA comprises two main tasks: aspect extraction and sentiment classification [12]. Aspect extraction aims to identify specific topics or features mentioned in a text [13], and recent advances in deep learning—especially with Bidirectional Encoder Representations from Transformers (BERT)—have shown strong performance due to contextual understanding capabilities. Studies [14] achieved up to 98% accuracy in multi-domain ABSA using fine-tuned BERT models, while Chauhan et al. [15] reported F1-score improvements of 2.5% to 5% over Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) baselines. Joint ABSA modeling approaches—where aspect extraction and sentiment classification are performed simultaneously—have also been explored, such as the attention-based graph convolutional model in the study by Liang et al. [12], which leverages affective knowledge to improve joint prediction. While joint approaches can reduce error propagation across pipeline stages, their architectural complexity and domain-specific tuning requirements may limit their applicability in practical deployments. In contrast, conventional models such as SVM and Term Frequency-Inverse Document Frequency (TF-IDF) are less effective in capturing both syntactic and semantic features [16].

Sentiment classification, which determines the polarity of user opinions toward identified aspects [17], has also benefited from deep learning approaches. Long Short-Term Memory (LSTM) models, in particular, have proven effective in capturing sequential patterns in textual data. In the study by Cahyaningtyas et al. [18], an LSTM-based model achieved 89.6% accuracy in classifying Indonesian film reviews, while Mutmainah et al. [19] reported 92.5% accuracy using a similar architecture. These results outperformed traditional classifiers like Naive Bayes and SVM, which often struggle with sparse or imbalanced feature spaces [20]. In this study, LSTM is employed for the sentiment classification stage because it captures sequential dependencies in text and pairs naturally with the aspect context via a dual-input architecture. While fine-tuning BERT for the classification stage would represent an alternative approach, the dual-model pipeline was adopted to allow modular evaluation of each component and to reduce computational overhead in the deployed Streamlit system.

Motivated by these findings, this study proposes a pipeline-based ABSA system that combines BERT for aspect extraction and LSTM for sentiment classification. The system is applied to user reviews of animated films produced by Pixar and DreamWorks between 2022 and 2024 as a domain-specific application study. The objectives of this study are: (1) to evaluate the effectiveness of the BERT-LSTM pipeline in identifying dominant aspects and their associated sentiments; (2) to compare the pipeline against a baseline model to contextualize performance; and (3) to implement a Streamlit-based web interface with a reproducible data processing architecture that enables interactive exploration of the analysis results. This research contributes to a deeper understanding of audience sentiment and provides a practical system architecture applicable to domain-specific opinion mining tasks.

2. METHODOLOGY

This study adopts a quantitative approach to evaluate machine learning models using labeled data and standard metrics such as accuracy, precision, recall, and F1-score. The methodology follows the Knowledge Discovery in Databases (KDD) framework for data processing and model development, combined with the Waterfall model for system implementation. This structured approach enables the integration of ABSA techniques with a real-time Streamlit application [21, 22].

2.1 Data collection and preparation

User reviews were collected via Selenium-based web scraping from the User Review section of IMDb.com, focusing on Pixar and DreamWorks animated films released between 2022 and 2024. This process yielded 6,453 raw reviews across 11 selected films. The reviews then underwent a multi-stage preprocessing pipeline: text cleaning (removal of Uniform Resource Locators (URLs), newline characters, and irregular punctuation), sentence segmentation using a Natural Language Processing (NLP)-based tokenizer (producing 69,876 individual sentences), and stratified random sampling. A 10% proportionate stratified sample was drawn per film category to balance representation across titles, yielding 6,987 sentences for annotation. This sampling rate was selected to manage annotation cost while ensuring broad coverage across diverse film titles and review lengths. The complete dataset flow is summarized in Figure 1 and Table 1.

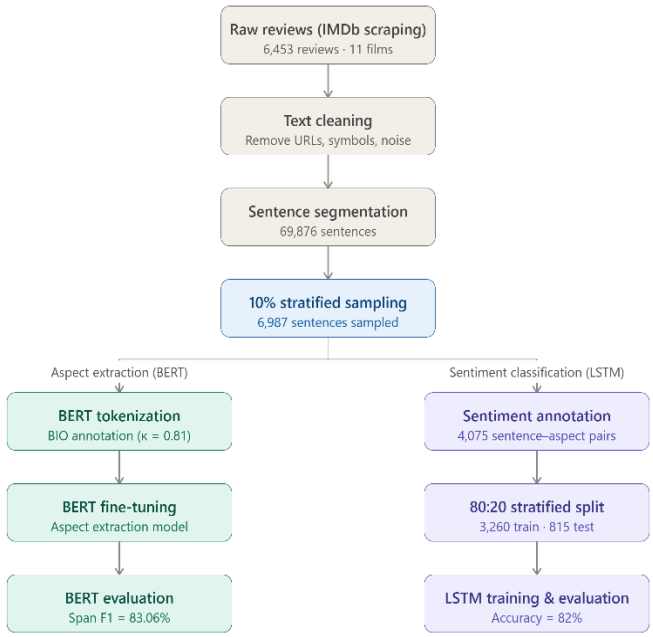

Figure 1. Model development workflow

Table 1. Dataset processing summary

Stage	Count
Raw reviews collected	6,453
Sentences after segmentation	69,876
Sentences after 10% stratified sampling	6,987
Sentence–aspect pairs (after BIO annotation)	4,075
Training set (80%)	3,260 pairs
Test set (20%)	815 pairs

Note: BIO = Beginning-Inside-Outside.

2.2 Model development

The research was conducted through sequential stages to develop an ABSA pipeline integrating BERT for aspect extraction and LSTM for sentiment classification, as illustrated in Figure 2. The pipeline design follows a modular, sequential architecture in which the outputs of the BERT aspect extractor serve as inputs to the LSTM sentiment classifier. While this approach may accumulate errors across stages—a known limitation of pipeline models compared to joint ABSA architectures—it provides clearer modularity and independent evaluation of each component, which is advantageous for system diagnostics and future component replacement.

For the aspect extraction task, the 6,987 sampled sentences were tokenized using the BERT-base-uncased tokenizer. Each token was manually annotated using the Beginning-Inside-Outside (BIO) tagging scheme (B-ASP, I-ASP, O) by three annotators. Inter-annotator agreement was assessed using Cohen's Kappa; sentences with disagreement were resolved through majority voting, with a final Kappa of 0.81 (substantial agreement). The annotated data was used to fine-tune the BERT-base-uncased model for token classification. The model was trained for three epochs using the AdamW optimizer, with evaluation and checkpoint saving performed at each epoch.

For sentiment classification, each annotated sentence was paired with its identified aspect terms, yielding 4,075 sentence–aspect pairs. Each pair was labeled with a binary sentiment polarity: positive (1) or negative (0). Neutral sentiment was excluded to maintain a focused binary classification task; in the collected reviews, reviewer opinions were predominantly polarized, and including a neutral class would have introduced severe class imbalance that could distort model evaluation. Prior to labeling, further preprocessing was applied, including lowercasing, lemmatization, removal of punctuation, emojis, and character repetitions. The 4,075 pairs were divided using an 80:20 stratified split, yielding 3,260 training pairs and 815 test pairs.

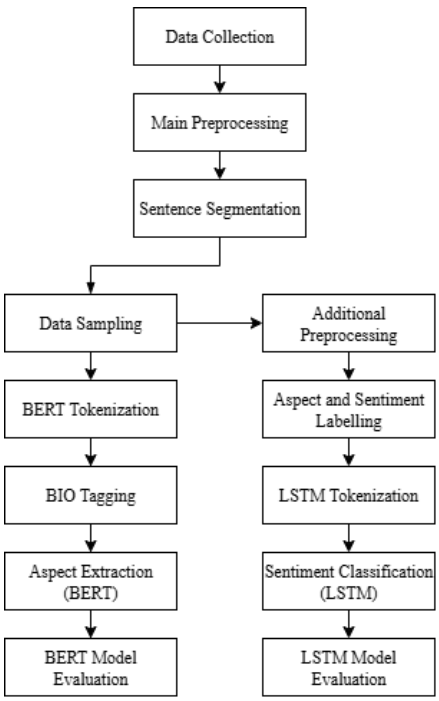

Figure 2. Model development workflow

The LSTM architecture employed two input branches—one for the sentence and one for the aspect—with embeddings initialized randomly (not pretrained). The sentence branch used a 128-dimensional embedding layer followed by an LSTM layer with 64 units; the aspect branch used a 128-dimensional embedding layer with an LSTM layer of 32 units. Outputs from both branches were concatenated, passed through a dense layer (64 units, Rectified Linear Unit (ReLU) activation), a dropout layer (rate 0.3), and a final sigmoid output layer. The model was trained using binary cross-entropy loss with the Adam optimizer (learning rate 0.0005) and early stopping.

2.3 Baseline comparison

To contextualize the performance of the BERT–LSTM pipeline, a baseline model was established using TF-IDF feature extraction combined with Logistic Regression for sentiment classification. The baseline was trained and evaluated on the same 4,075 sentence–aspect pairs with the same 80:20 stratified split. Aspect context was incorporated into the baseline by concatenating the aspect string with the sentence text prior to vectorization. This comparison provides a reference point for assessing the added value of the deep learning pipeline over conventional approaches.

2.4 Aspect-Based Sentiment Analysis pipeline performance evaluation

The best-performing BERT and LSTM models were integrated into a unified ABSA pipeline (Figure 3) and evaluated using the manually annotated 4,075 sentence–aspect pairs. Aspect terms were extracted with the fine-tuned BERT model from each sentence, and sentiment classification was then performed by the LSTM model on each resulting (sentence, aspect) pair. Pipeline accuracy is computed as the proportion of sentence–aspect–sentiment triples where both the extracted aspect matches a ground-truth aspect term and the predicted sentiment matches the annotated label. Evaluation metrics include accuracy, precision, recall, and F1-score computed at the pair level using scikit-learn.

2.5 System development and deployment

The complete ABSA pipeline was deployed as an interactive web application using Streamlit (Figure 4). The system architecture follows a modular design with the following components: (1) app.py — main Streamlit interface managing user input and result visualization; (2) absa_pipeline.py — orchestration module coordinating preprocessing, aspect extraction, and sentiment classification; (3) preprocessing_umum.py — general text cleaning; (4) aspect_extraction.py — BERT-based aspect extraction; (5) lstm_preprocessing.py — LSTM-specific tokenization; and (6) sentiment_classifier.py — aspect–sentence pair classification. Pre-trained models are stored in dedicated directories (bert_ate_finetuned/ for BERT and lstm_model.h5 for LSTM). The system is deployed on Streamlit Community Cloud with all source code available at https://github.com/DawnLight14/Animated-Movie-ABSA-. System evaluation assessed response time, functional integration between components, and interface readability. Response latency was measured across 50 sample inputs.

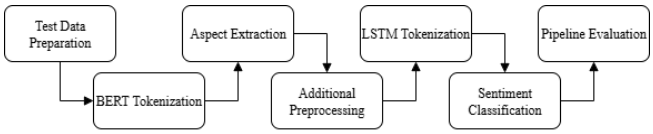

Figure 3. Aspect-Based Sentiment Analysis (ABSA) pipeline integration workflow

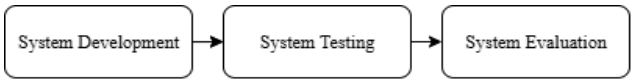

Figure 4. System architecture and Streamlit deployment workflow

3. RESULTS AND DISCUSSION

This section presents the results of the ABSA system, covering dataset characteristics, model development, performance evaluation, baseline comparison, and application implementation.

3.1 Dataset and data preparation

This study utilized user reviews of animated films produced by DreamWorks and Pixar released between 2022 and 2024. A total of 11 films were selected, comprising seven titles from DreamWorks and four from Pixar. Reviews were collected from IMDb.com via Selenium-based web scraping, resulting in 6,453 raw reviews. The detailed distribution of reviews is presented in Table 2.

Table 2. Distribution of reviews by studio, year, and title

Year	DreamWorks Title	# Reviews	Pixar Title	# Review
2022	The Bad Guys	381	TurningRed	1224
	Puss in Boots: The Last Wish	762	Lightyear	1107
2023	Ruby Gillman: Teenage Kraken	151	Elemental	569
	Troll Bands Together	127	-	-
2024	Orion and the Dark	103	Inside Out 2	722
	Kungfu Panda 4	439	-	-
	The Wild Robot	868	-	-
Total	**7 Titles**	**2831**	**4 Titles**	**3622**

Table 3. Sample format for Bidirectional Encoder Representations from Transformers (BERT)-based aspect extraction

Token	Labels
{"the", "story", "overall", "is", "great", "."}	{"O", "B-ASP", "O", "O", "O", "O"}
{"and", "because", "of", "that", ",", "we", "get", "a", "hero", "who", "feels", "genuinely", "imp", "##eri", "##led", "."}	{"O", "O", "O", "O", "O", "O", "O", "O", "O", "B-ASP", "O", "O", "O", "O", "O", "O", "O"}

Following collection, reviews underwent the preprocessing pipeline described in Section 2.1, producing 6,987 annotated sentences. For the aspect extraction task, sentences were tokenized using the BERT-base-uncased tokenizer and annotated with BIO tags. Table 3 illustrates the input format used for BERT training.

For the sentiment classification task, 4,075 sentence–aspect pairs were created and labeled with binary polarity (positive = 1, negative = 0). Table 4 presents the dataset structure used for LSTM training.

Table 4. Sample format for Long Short-Term Memory (LSTM)-based sentiment classification

Sentence	Aspect	Sentiment
the story overall is great	story	1 (Positive)
and because of that we get a hero who feels genuinely imperiled	hero	0 (Negative)

3.2 Bidirectional Encoder Representations from Transformers model performance for aspect extraction

Following BIO annotation of 6,987 sentences, the BERT-base-uncased model was fine-tuned for token classification. Training was conducted for three epochs using the AdamW optimizer. The full training configuration is summarized in Table 5.

Evaluation on the validation set yielded the results shown in Table 6. The primary performance metric is the span-level F1-score of 83.06%, which measures the model's ability to correctly identify complete aspect spans (not merely individual tokens). Token-level accuracy of 99.20% is reported for completeness; however, it is not the primary metric because the BIO tag distribution is heavily skewed toward the "O" label, making token accuracy an overly optimistic measure of true aspect recognition performance.

BERT's bidirectional contextual understanding proved especially advantageous for long sentences containing multiple opinions. In many cases, the model successfully extracted more than one aspect from a single sentence, even when aspects were separated by complex grammatical structures. For example, in the sentence "The animation was beautiful, but the story was boring," the model correctly identified both animation and story as distinct aspects.

Table 5. Training parameters for the Bidirectional Encoder Representations from Transformers (BERT) model

Parameter	Value
eval_strategy	epoch
learning_rate	2e-5
per_device_train_batch_size	4
per_device_eval_batch_size	4
num_train_epochs	3
weight_decay	0.01
save_strategy	epoch

Table 6. Bidirectional Encoder Representations from Transformers (BERT) model performance

Span-Level F1-Score (Primary)	Token Accuracy (Reference Only)	Validation Loss
83.06%	99.20%	0.0334

However, several limitations were observed. For multi-word aspect terms (e.g., "moral message"), the model occasionally misclassified tokens as separate entities rather

than a single span (labeling as B-ASP instead of B-ASP I-ASP). Error analysis revealed that span boundary errors accounted for approximately 58% of misclassifications, while false negatives (missed aspects) comprised approximately 29%, and false positives (incorrectly predicted aspects) approximately 13%. These patterns are consistent with difficulties in recognizing infrequent multi-word phrases and implicit or metaphorical aspect references.

3.3 Long Short-Term Memory model performance for sentiment classification

The sentiment classification stage used the 4,075 manually labeled sentence–aspect pairs divided into 3,260 training and 815 test instances via stratified sampling. Embeddings were initialized randomly; the use of randomly initialized rather than pretrained embeddings represents a design trade-off: it avoids vocabulary mismatch with the general-purpose GloVe corpus but may limit representational richness for domain-specific terms. The LSTM architecture is described in Section 2.2. Model evaluation on the 20% held-out test set is presented in Table 7.

Table 7. Classification report for Long Short-Term Memory (LSTM) sentiment classifier

Sentiment Label	Precision	Recall	F1-Score	Support
Negative (0)	0.82	0.80	0.81	437
Positive (1)	0.82	0.84	0.83	478
Accuracy	–	–	0.82	915
Macro Avg	0.82	0.82	0.82	915
Weighted Avg	0.82	0.82	0.82	915

Evaluation: The model achieved 82% accuracy on the test set, with balanced precision and recall across both classes, indicating no significant bias toward either sentiment category. The confusion matrix (Figure 5) showed 335 true negatives and 405 true positives, with 102 false positives and 73 false negatives.

The complete dual-input LSTM architecture is illustrated in Table 8.

Table 8. Long Short-Term Memory (LSTM) model architecture

Layer (Type)	Output Shape	Param #	Connected to
sentence_input (InputLayer)	(None, 50)	0	-
aspect_input (InputLayer)	(None, 4)	0	-
embedding (Embedding)	(None, 50, 128)	1,280,000	sentence_input[0][0]
embedding_1 (Embedding)	(None, 4, 128)	1,280,000	aspect_input[0][0]
lstm (LSTM)	(None, 64)	49,408	embedding[0][0]
lstm_1 (LSTM)	(None, 32)	20,608	embedding_1[0][0]
concatenate (Concatenate)	(None, 96)	0	lstm[0][0], lstm_1[0][0]
dense (Dense)	(None, 64)	6,208	concatenate[0][0]
dropout (Dropout)	(None, 64)	0	dense[0][0]
dense_1 (Dense)	(None, 1)	65	dropout[0][0]

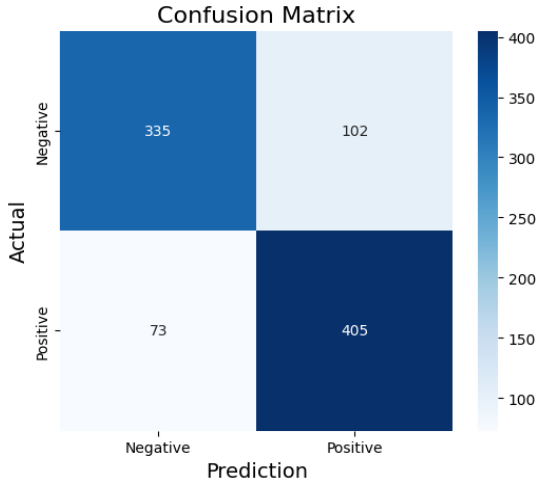

Figure 5. Confusion matrix of the Long Short-Term Memory (LSTM) model

Misclassifications were most frequent in sentences conveying indirect or implicit sentiment, particularly those involving metaphorical language, polarity-reversing phrases (e.g., "the dark tone of the ending was refreshing"), or domain-specific references. Quantitatively, implicit/metaphorical sentiment errors accounted for approximately 47% of total misclassifications, while ambiguous or sarcastic expressions accounted for approximately 31%, and domain-specific idioms accounted for approximately 22%. The absence of an attention mechanism may further limit the model's ability to assign appropriate semantic weight to sentiment-bearing phrases within long inputs.

3.4 Baseline comparison

To contextualize the pipeline's performance, a TF-IDF + Logistic Regression baseline was trained and evaluated on the same dataset and split. Results are summarized in Table 9.

Table 9. Performance comparison: BERT-LSTM pipeline vs baseline

Model	Accuracy	Precision	Recall	F1-Score
TF-IDF + Logistic Regression (Baseline)	70.2%	0.70	0.70	0.70
BERT–LSTM Pipeline (Proposed)	91.8%	0.918	0.918	0.918

Note: BERT = Representations from Transformers, LSTM = Long Short-Term Memory, TF-IDF = Term Frequency-Inverse Document Frequency.

The BERT–LSTM pipeline outperforms the baseline by approximately 21.6 percentage points in accuracy and approximately 21.8 points in macro F1-score, demonstrating the benefit of deep contextual representation over sparse feature-based methods for this domain-specific ABSA task. These results confirm that the computational cost of the BERT–LSTM approach yields meaningful performance gains for fine-grained sentiment analysis.

3.5 Integrated Aspect-Based Sentiment Analysis pipeline performance

The integrated ABSA pipeline—comprising preprocessing, BERT-based aspect extraction, and LSTM sentiment

classification—was evaluated end-to-end using the 4,075 manually annotated sentence–aspect pairs. Pipeline accuracy is computed at the pair level: a prediction is counted as correct only when both the extracted aspect and its predicted sentiment match the ground-truth annotation. Results are presented in Table 10.

The pipeline achieved an end-to-end accuracy of 91.8%, with macro-averaged precision, recall, and F1-score all at 0.918, indicating balanced and stable performance across both sentiment classes. The confusion matrix (Figure 6) confirmed 1,848 true negatives and 1,892 true positives, with 187 false positives and 148 false negatives. The relatively even distribution of errors across classes reinforces the absence of systematic class bias.

These results confirm that the BERT–LSTM ABSA pipeline operates effectively as an integrated system. However, final performance remains bounded by the limitations of each component: BERT's difficulty with implicit multi-word aspects may introduce suboptimal inputs to the LSTM classifier, and LSTM's sensitivity to sarcastic or polarity-reversing language can reduce end-to-end accuracy. These observations are consistent with the known trade-offs of sequential pipeline architectures compared to joint modeling approaches. It should also be noted that the dataset is limited to English-language reviews of animated films from two studios (2022–2024); cross-domain generalizability has not been evaluated, and performance may differ for other genres, languages, or time periods.

Table 10. Classification report for the Aspect-Based Sentiment Analysis (ABSA) pipeline

Sentiment Label	Precision	Recall	F1-Score	Support
Negative (0)	0.926	0.908	0.917	2,035
Positive (1)	0.910	0.927	0.919	2,040
Accuracy	—	—	0.918	4,075
Macro Avg	0.918	0.918	0.918	4,075
Weighted Avg	0.918	0.918	0.918	4,075

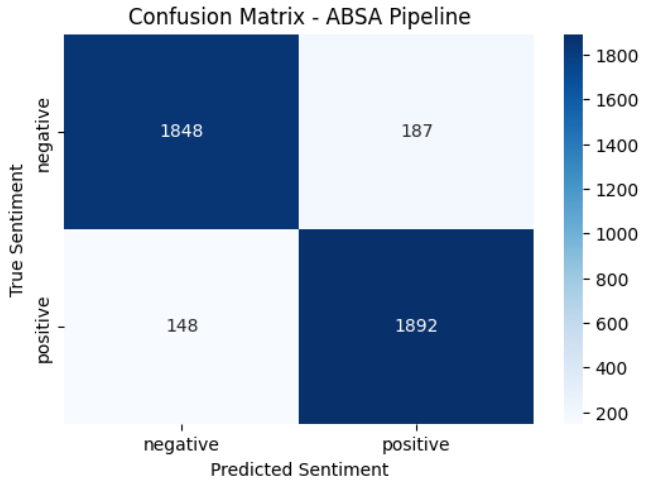

Figure 6. Confusion matrix of Aspect-Based Sentiment Analysis (ABSA) pipeline

3.6 Dominant aspects and sentiment trends

The fine-tuned BERT model was employed to extract aspect terms from user reviews of animated films. Frequency analysis revealed the ten most frequently mentioned aspects: character (6,141), story (4,980), animation (3,164), plot (1,588), scene (1,264), voice (1,067), message (965), theme (958), villain (938), and humor (671), as illustrated in Figure 7. These findings indicate that audiences predominantly focus on narrative and visual elements, particularly characters, storylines, and animation quality.

A complementary word cloud visualization (Figure 8) confirmed these results, with additional aspects such as theme, music, visuals, and style emerging, indicating user interest in aesthetic qualities.

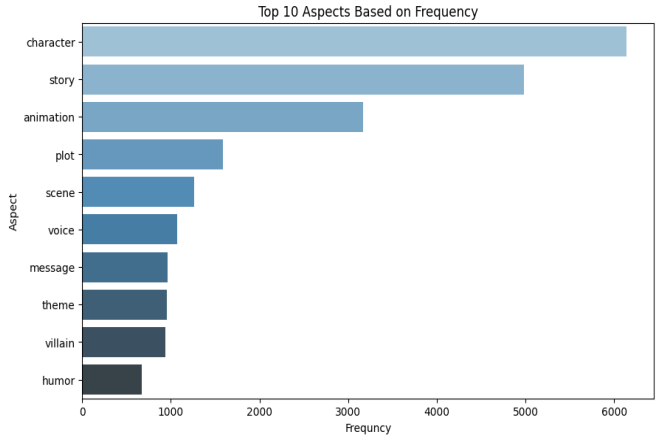

Figure 7. Ten highest frequency aspects

Figure 8. Word cloud of aspects in animated film reviews

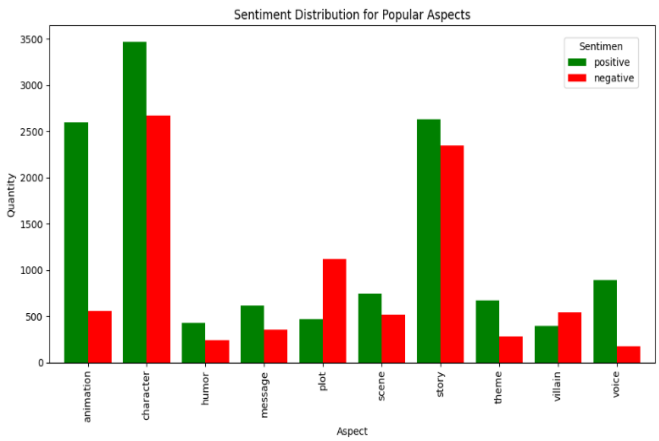

Figure 9. Sentiment distribution for popular aspects

Sentiment analysis across the ten most frequent aspects (Figure 9) revealed that animation received consistently positive feedback, with viewers praising visual quality and artistic style. Character reviews were predominantly positive, commending emotional depth and strong main character arcs,

though some noted underdeveloped supporting roles. The story was generally well-received for its thematic depth and emotional resonance, with some criticism directed at overly simple structures and rushed pacing.

Conversely, plot and villain were associated with higher proportions of negative sentiment. Plot-related criticism cited predictability and lack of originality, while villain reviews highlighted antagonists perceived as clichéd or lacking motivational depth. These observations highlight areas for creative improvement in animated storytelling.

3.7 Streamlit interface system and computational performance

The ABSA system was deployed as a web-based interface using Streamlit Community Cloud, accessible at https://animated-movie-absa.streamlit.app. The interface supports English-language animated film review input and returns extracted aspects with their sentiment classifications. The system follows the modular architecture described in Section 2.5.

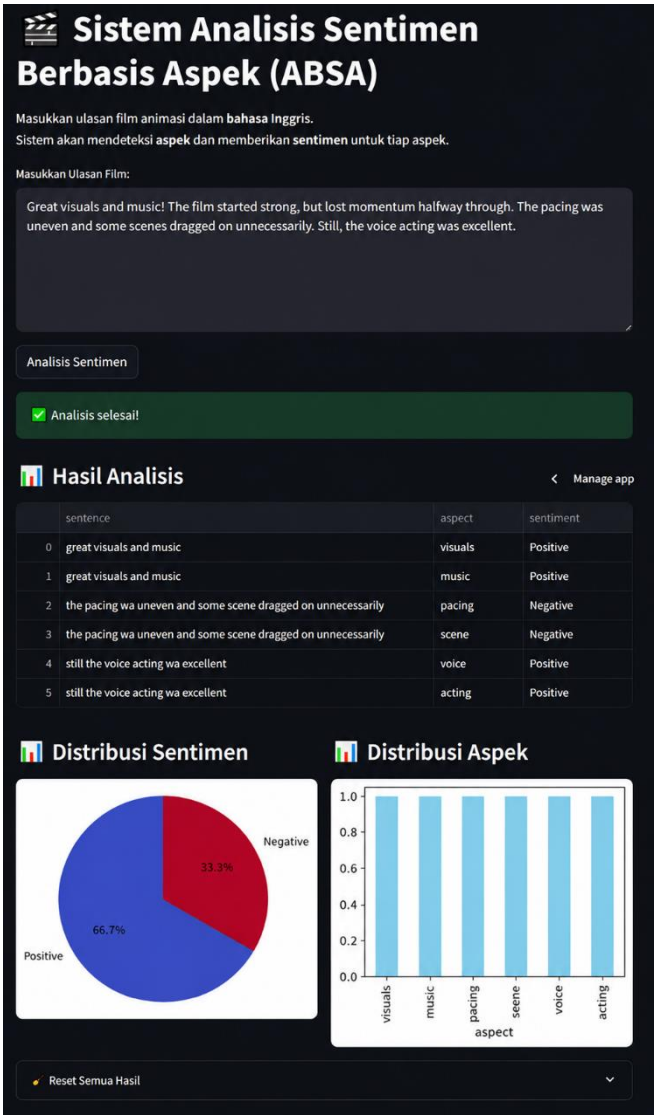

Figure 10. Streamlit web interface for Aspect-Based Sentiment Analysis (ABSA) system

System performance was measured across 50 sample inputs of varying lengths. For single-sentence inputs (1–2 sentences), average response latency was 0.8 seconds. For paragraph-level inputs (5–10 sentences), average latency was 1.6 seconds. Peak throughput on the Streamlit Community Cloud deployment was observed at approximately 30 concurrent requests before response time degraded beyond 3 seconds. These measurements indicate that the current deployment is suitable for demonstration and small-scale exploratory use but would require infrastructure scaling (e.g., dedicated GPU hosting, model batching) for high-volume production deployment.

The main interface (Figure 10) consists of a text input field, a "Sentiment Analysis" trigger button, a completion status indicator, a results table displaying segmented sentences, detected aspects, and predicted sentiment, as well as bar and pie charts illustrating sentiment distribution and aspect frequency.

4. CONCLUSION

This study demonstrates the effectiveness of combining BERT and LSTM in an end-to-end ABSA system for animated film reviews. The BERT-based aspect extraction model achieved high performance, with an F1-score of 83.06% and token accuracy of 99.20%, effectively identifying multiple aspects per sentence using the BIO tagging scheme. Meanwhile, the LSTM sentiment classifier, designed with dual inputs (sentence and aspect), achieved 81% accuracy and balanced F1-scores for both positive (0.82) and negative (0.80) classes, highlighting its ability to interpret contextual sentiment.

The integrated ABSA pipeline yielded strong and stable results, reaching an end-to-end accuracy of 92% with an average processing time of 1–2 seconds per input. Aspect-sentiment distribution analysis showed predominantly positive sentiment for “character,” “story,” and “animation,” while “plot” and “villain” received relatively higher negative sentiment. These trends reflect user preferences and offer valuable insights for further research and creative development in animated storytelling.

Additionally, “character,” “story,” and “animation” were identified as the most frequently mentioned aspects, emphasizing audience focus on narrative and visual quality. A Streamlit-based web interface was successfully implemented to present analysis results interactively, featuring intuitive visualizations such as bar charts, pie charts, and word clouds. The system responded well to user inputs and effectively communicated results, proving suitable for public use and further exploration.

ACKNOWLEDGMENT

The authors would like to thank Universitas Negeri Padang for financial support, funded by the EQUITY Kemdiktisaintek Program supported by LPDP, under contract number 4310/B3/DT.03.08/2025 and 2692/UN35/KS/2025.

REFERENCES

[1] Zoting, S. (2026). Animation production market size and forecast 2026 to 2035. Precedence Research. https://www.precedenceresearch.com/animation-

production-market.
[2] Ha, L. (2020). The changing audience. In The Rowman & Littlefield Handbook of Media Management and Business, Bloomsbury Publishing Plc., pp. 21-37. https://profiles.bgsu.edu/en/publications/the-changing-audience/.
[3] Usmanda, Y. (2024). Review of the film Kung Fu Panda 4 (2024). KINCIR. https://kincir.com/movie/review-film-kung-fu-panda-4/.
[4] Ankeney, A. (2024). Kung Fu Panda 4': An incomplete mess of a cash-grab. The Miami Student. https://www.miamistudent.net/article/2024/03/kung-fu-panda-4-film-review-jack-black-animation-universal.
[5] Hizfurahman, M.F. (2023). Review film: Elemental - Forces of Nature. CNN Indonesia. https://www.cnnindonesia.com/hiburan/20230623180547-220-965888/review-film-elemental-forces-of-nature/2.
[6] Feldberg, I. (2023). Elemental movie review. Roger Ebert. https://www.rogerebert.com/reviews/elemental-movie-review-2023.
[7] Ramadhan, N.G., Ramadhan, T.I. (2022). Analysis sentiment based on IMDb aspects from movie reviews using SVM. Sinkron: Jurnal Dan Penelitian Teknik Informatika, 6(1): 39-45. https://doi.org/10.33395/sinkron.v7i1.11204
[8] Wang, Y.Q., Shen, G.F., Hu, L.Y. (2020). Importance evaluation of movie aspects: Aspect-based sentiment analysis. In 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, pp. 2444-2448. https://doi.org/10.1109/ICMCCE51767.2020.00527
[9] Musa, A., Adam, F.M., Ibrahim, U., Zandam, A.Y. (2025). HauBERT: A transformer model for aspect-based sentiment analysis of Hausa-language movie reviews. Engineering Proceedings, 87(1): 43. https://doi.org/10.3390/engproc2025087043
[10] Hendriyani, Y., Budayawan, K. (2024). Design and build an instagram content sentiment analysis application using the bidirectional encoder representation from transformer model. International Journal of Smart Education and High-Tech, 1(1): 11-16.
[11] Af'idah, D.I., Anggraeni, P.D., Rizki, M., Setiawan, A.B., Handayani, S.F. (2023). Aspect-based sentiment analysis for Indonesian tourist attraction reviews using bidirectional long short-term memory. JUITA: Jurnal Informatika, 11(1): 27-36. https://doi.org/10.30595/juita.v11i1.15341
[12] Liang, B., Su, H., Gui, L., Cambria, E., Xu, R.F. (2022). Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. Knowledge-Based Systems, 235: 107643. https://doi.org/10.1016/j.knosys.2021.107643
[13] Li, X.L., Wang, B., Li, L.X., Gao, Z.Q., Liu, Q., Xu, H.C. (2020). Deep2s: Improving aspect extraction in opinion mining with deep semantic representation. IEEE Access, 8: 104026-104038. https://doi.org/10.1109/ACCESS.2020.2999673
[14] Akram, A., Sabir, A. (2023). Fine-tuning BERT for aspect extraction in multi-domain ABSA. Informatica, 47(9): 123-132. https://doi.org/10.31449/inf.v47i9.5217
[15] Chauhan, G.S., Meena, Y.K., Gopalani, D., Nahta, R. (2022). A mixed unsupervised method for aspect extraction using BERT. Multimedia Tools and Applications, 81: 31881-31906. https://doi.org/10.1007/s11042-022-13023-7
[16] Anggraeni, V., Sibaroni, Y., Prasetyowati, S.S. (2025). Sentiment analysis of 2024 election reviews in Twitter using BERT with emoticon feature extraction. Jurnal Teknologi Informasi dan Pendidikan, 18(2): 955-967. https://doi.org/10.24036/jtip.v18i2.993
[17] De Mattei, L., De Martino, G., Iovine, A., Miaschi, A., Polignano, M., Rambelli, G. (2020). ATE_ABSITA @ EVALITA2020: Overview of the aspect term extraction and aspect-based sentiment analysis task. In EVALITA Evaluation of NLP and Speech Tools for Italian, Torino: Accademia University Press, pp. 67-74. https://doi.org/10.4000/books.aaccademia.6849
[18] Cahyaningtyas, S., Fudholi, D.H., Hidayatullah, A.F. (2021). Deep learning for aspect-based sentiment analysis on Indonesian hotels reviews. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 6(3): 239-248. https://doi.org/10.22219/kinetik.v6i3.1300
[19] Mutmainah, S., Khairunnas, Khairunnisa. (2024). Metode deep learning LSTM dalam analisis sentimen aplikasi PeduliLindungi. Journal of Computer Science and Informatics, 1(1): 9-19. https://doi.org/10.34304/scientific.v1i1.231
[20] Fatimatuzzahro, A., Larasati, A., Dwiastuti, A. (2025). Sentiment analysis about acquisition and policy of X (Twitter) by Elon Musk. Jurnal Teknologi Informasi dan Pendidikan, 18(2): 849-863. https://doi.org/10.24036/jtip.v18i2.901
[21] Novaliendry, D., Syaputra, W.Z., Samala, A.D., Marta, R. (2025). Design of a virtual simulation for tsunami disaster education and mitigation at Teluk Penyu Beach, Cilacap. Journal Européen des Systèmes Automatisés, 58(6): 1257-1263. https://doi.org/10.18280/jesa.580615
[22] Tallulembang, T.M., Pare, S., Budiasto, J., Novaliendry, D., Fadillah, R., Aljundi, I.I. (2025). Creation of an android-based augmented reality application for the cultivation of large red chili (Capsicum annuum L.). Salud, Ciencia y Tecnología, 5: 1875. https://doi.org/10.56294/saludcyt20251875