JASMINE
Journal of Intelligent Systems and Machine Learning

Articles in Press

Aspect-Based Sentiment Analysis on Webtoon Reviews Using Ensemble Learning

Rival Fakhri Amrullah[1,*], Fathoni Mahardika[1], Dani Indra Junaedi[1]
[1]Department of Informatics, Universitas Sebelas April, Sumedang, Indonesia

E-mail: rivalfakhria@gmail.com

Received: November 18, 2025
Accepted for publication: January 6, 2026

ABSTRACT
The rapid growth of online comic platforms such as LINE Webtoon has produced large volumes of user comments that reflect diverse opinions on storytelling elements. Analyzing these comments provides meaningful insights into reader perceptions of aspects such as plot, characters, and visuals. This study proposes an Aspect-Based Sentiment Analysis (ABSA) framework using ensemble learning to classify sentiment in Indonesian Webtoon reviews. The research follows an experimental quantitative methodology consisting of data collection, text preprocessing, manual annotation of aspects and sentiments, feature extraction using Term Frequency–Inverse Document Frequency (TF-IDF), ensemble model training with Random Forest and XGBoost, and performance evaluation. A total of 1,010 annotated comments were used, covering three aspects, plot, character and visual, and three sentiment categories: negative, neutral, and positive. The results demonstrate that incorporating aspect information enhances sentiment classification performance. While the tuned XGBoost model using TF-IDF features achieved an accuracy of 0.6238 in the text-only scenario, the best performance was obtained by the tuned Random Forest ABSA model, which combined TF-IDF and One-Hot Encoded aspect features and achieved an accuracy of 0.6584 with a weighted F1-score of 0.6541. Class-level analysis shows that neutral comments are the easiest to classify, while negative sentiment remains the most challenging due to informal and context-dependent expressions. A 10-fold cross-validation yielded a mean accuracy of 0.6238 with a standard deviation of 0.0572, indicating stable generalization. These findings highlight the effectiveness of aspect-enhanced ensemble learning for sentiment analysis in Indonesian Webtoon reviews.

Keywords: Aspect-Based Sentiment Analysis, Ensemble Learning, Random Forest, XGBoost, Webtoon Reviews

I. Introduction

The rapid development of digital technology has significantly increased the production of web-based content, including online comic platforms such as LINE Webtoon [1], [2]. Beyond entertainment, Webtoon provides an interactive space where readers express opinions through comment sections [3]. These comments often contain subjective evaluations of narrative elements such as plot, characters, and visuals [4], [5], offering valuable data for sentiment analysis, particularly at the aspect level [6], [7]. However, Indonesian Webtoon comments are characterized by informal language, slang, abbreviations, and code-mixing, which pose substantial challenges for sentiment classification [8], [9].

Most existing sentiment analysis studies focus on document-level or sentence-level classification, overlooking aspect-level distinctions [10], [11], [12]. Such approaches are insufficient for multidimensional reviews like Webtoon comments, where users may simultaneously express different sentiments toward different aspects of a story [13]. Aspect-Based Sentiment Analysis (ABSA) addresses this limitation by assigning sentiment polarity to specific aspects within a text [14], [15]. Despite its effectiveness, ABSA research in Indonesia remains limited, particularly in the digital entertainment domain, with most studies

concentrating on e-commerce or application reviews [2], [16]. Moreover, many existing works rely on single classifiers such as Naïve Bayes or Support Vector Machine (SVM), which often struggle with highly variable and informal text [17], [18].

Given the short, sparse, and noisy nature of Indonesian Webtoon comments, selecting an appropriate feature representation and learning strategy is critical. Lexical-based approaches such as Term Frequency–Inverse Document Frequency (TF–IDF) remain effective for short-text sentiment analysis due to their ability to emphasize discriminative terms while maintaining computational efficiency [2], [19]. However, when used with single classifiers, TF–IDF-based models are prone to instability in the presence of linguistic variability and contextual ambiguity.

To address these limitations, this study adopts ensemble learning to improve the robustness and generalization of aspect-based sentiment classification. The proposed approach formulates ABSA as an aspect-conditioned sentiment classification task, where predefined aspect labels provide contextual guidance for sentiment prediction. Two complementary ensemble paradigms are evaluated: Random Forest, representing bagging, and XGBoost, representing boosting. Random Forest enhances stability by aggregating multiple decision trees trained on bootstrapped samples, while XGBoost improves performance through iterative error correction and the modeling of complex feature interactions [20], [21].

This study aims to fill existing research gaps by applying ABSA to Indonesian Webtoon comments and systematically evaluating ensemble learning strategies for sentiment classification across three narrative aspects: plot, character, and visual. By integrating TF–IDF features with both bagging and boosting ensembles, this research provides a comparative analysis supported by class-level evaluation, feature importance analysis, and qualitative error analysis, contributing to a more comprehensive understanding of ensemble-based ABSA in the digital entertainment domain.

II. Related work

Research on Aspect-Based Sentiment Analysis (ABSA) has expanded considerably in recent years, driven by the need for more detailed interpretation of user opinions across multiple aspects within a text. ABSA operates by identifying aspect categories and determining sentiment polarity for each aspect, allowing a more granular analysis compared to traditional document-level sentiment classification. Several comprehensive surveys highlight the increasing importance of ABSA for analyzing complex user-generated content across various domains [7], [12]. These studies emphasize that ABSA yields richer insights when user opinions involve evaluations of multiple components, such as product features or narrative elements in media content.

In the Indonesian context, ABSA studies have been conducted primarily in domains such as e-commerce and application reviews. Research on Indonesian product reviews demonstrates that ABSA is capable of capturing nuanced sentiment across multiple feature categories, although its effectiveness is largely influenced by preprocessing quality and the linguistic characteristics of the text [15]. More recent work introduces generative and multitask-based ABSA architectures for Indonesian text, showing promising improvements but requiring substantial data and computational resources [16]. Despite these advancements, the application of ABSA to digital entertainment platforms especially Webtoon comments remains limited.

Recent research has increasingly applied deep learning approaches to Aspect-Based Sentiment Analysis, including convolutional neural networks (CNN), recurrent neural networks such as LSTM and BiLSTM, and transformer-based models. CNN-based ABSA has been shown to effectively capture local aspect–sentiment patterns, while LSTM and BiLSTM architectures model long-range dependencies between aspect terms and sentiment cues [22], [23]. More recently, transformer-based models such as BERT have achieved state-of-the-art performance in ABSA by leveraging contextualized representations and attention mechanisms to model complex aspect–sentiment interactions [24]. For Indonesian text, IndoBERT has demonstrated superior performance over traditional machine learning approaches in several sentiment analysis and ABSA studies, particularly in handling implicit sentiment and negation [25]. However, deep learning-based ABSA models typically require large annotated datasets, substantial computational resources, and careful fine-tuning, and often exhibit limited interpretability [26]. Given the relatively small dataset size and highly informal nature of Indonesian Webtoon comments, this study adopts ensemble-based classical machine learning as a controlled and interpretable baseline, while transformer-based approaches are reserved for future work.

A major obstacle in Indonesian sentiment analysis is the prevalence of informal, conversational language. Digital user comments often contain slang, contractions, phonetic variations, expressive noises, and Indonesian–English code-mixing. Studies on Indonesian text normalization show that slang and non-standard spelling significantly reduce classifier performance if not properly addressed [8]. Additional

research on code-mixed Indonesian–English data shows similar challenges, where noise and linguistic inconsistency hinder the extraction of meaningful features for classification [9]. These findings underline the need for robust models capable of handling noisy, unstructured Webtoon comments.

Sentiment analysis specifically related to Webtoon platforms is still scarce, particularly in the Indonesian context. One study compared Naïve Bayes and Random Forest on Webtoon application reviews and reported that Random Forest achieved an accuracy of 88%, outperforming Naïve Bayes which obtained 74% accuracy, indicating the advantage of ensemble-based classifiers in this domain [2]. Another study on digital comic interactions highlights how user comments in Webtoon environments often express emotional reactions tied to story events and character dynamics, suggesting that ABSA may be highly suitable for extracting narrative-related insights [3]. However, these studies do not implement aspect-based sentiment classification, leaving a gap in understanding how users evaluate specific story elements such as plot, character design, and visual direction.

In terms of classification methods, many Indonesian sentiment analysis studies continue to rely on single-model approaches such as Naïve Bayes, SVM, or basic deep learning models. While these methods can perform adequately on structured text, they tend to struggle with high textual variability and informal language [17]. Ensemble learning offers an alternative that has demonstrated greater robustness in multiple experiments. Previous research has reported that ensemble-based classifiers, particularly Random Forest, achieved higher classification performance than individual algorithms on Indonesian app review datasets, with accuracy reaching 0.9415 and F1-score 0.9419 on Zoom application reviews [17]. Boosting algorithms such as XGBoost further improves classification performance by iteratively correcting errors made by previous learners. Daniati and Utama demonstrated that ensemble-based sentiment analysis models achieved higher accuracy and more stable performance than single classifiers on Indonesian Twitter data [18]. In addition, Chen and Guestrin showed that XGBoost is specifically designed to handle sparse and high-dimensional data efficiently, achieving state-of-the-art results across various large-scale machine learning benchmarks [20].

These quantitative findings indicate that ensemble-based approaches consistently outperform single classifiers on noisy Indonesian text, motivating their adoption in aspect-based sentiment analysis tasks. From the reviewed literature, two key research gaps can be identified. First, ABSA has not been widely applied to Indonesian digital entertainment platforms, including Webtoon, despite the abundance of aspect-rich user comments. Second, there is a lack of studies combining ABSA with ensemble learning, even though ensemble methods have been shown to outperform single classifiers on Indonesian text with high variability. Addressing these gaps, the present study integrates ABSA with ensemble learning methods, namely Random Forest and XGBoost, to analyze sentiment toward three key aspects (plot, character, visual) in Indonesian Webtoon comments.

III. Material and Methods

Figure 1 presents an overview of the research workflow, which is further described in detail in the subsequent subsections.

A. Dataset and Data Collection

The dataset used in this study consists of user-generated comments collected from the Webtoon platform. These comments reflect readers' opinions and emotional reactions toward various episodes, making them suitable for sentiment and aspect-based analysis. The initial dataset contained 1,034 raw entries. After removing incomplete and invalid records such as empty comments, unreadable text, or mislabeled entries the final dataset used for model training and evaluation consisted of 1,010 valid samples.

Each record in the dataset includes three main attributes:

1. User Comment – the original textual content written by Webtoon readers.
2. Aspect Category – one of three predefined aspects relevant to comic evaluation:
   - Storyline
   - Character
   - Visuals
3. Sentiment Polarity - sentiment assigned to each comment, categorized into:
   - Positive
   - Neutral
   - Negative

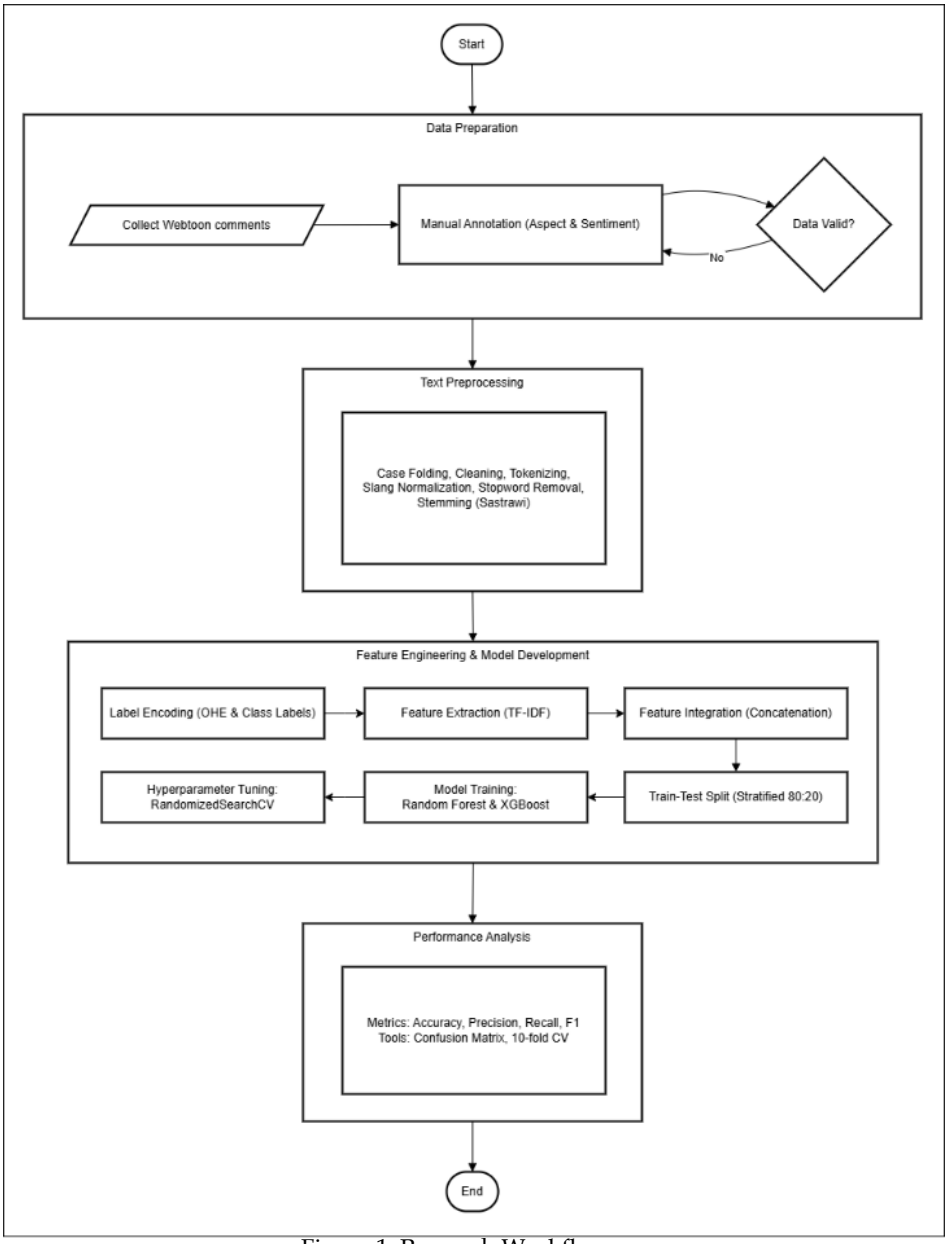

Figure 1. Research Workflow

The dataset was annotated manually. All comments were labeled based on their dominant aspect and sentiment orientation. The labeling process included iterative review to ensure consistency and correctness before the dataset was used as ground truth.

For model evaluation, the dataset was divided into training and testing subsets using an 80:20 stratified split, ensuring that the distribution of sentiment classes remained proportional across both subsets. This approach helps maintain balanced representation and reduces the risk of bias during model evaluation.

B. Aspect and Sentiment Annotation

Aspect and sentiment labels were assigned manually to each Webtoon comment to form the ground truth dataset for Aspect-Based Sentiment Analysis. The annotation process focused on determining (1) the dominant aspect discussed and (2) the sentiment polarity expressed by the reader.

Three aspects were defined for categorization: Storyline, referring to narrative progression and story-related remarks; Character, covering comments about personality traits, behaviors, or development of characters; and Visual, involving artwork quality, drawing style, and overall aesthetic impressions. During

annotation, variations and inconsistencies (e.g., *alur*, *alur cerita*, *visual art*, *karakter utama*) were standardized into these predefined categories to ensure uniform aspect representation.

Sentiment polarity was labeled as Positive, Neutral, or Negative. Positive labels captured expressions of enjoyment, praise, excitement, or admiration. Neutral labels included objective statements or descriptive comments with no emotional orientation. Negative labels represented criticism, dissatisfaction, discomfort, or unfavorable reactions. Comments that were ambiguous, extremely unclear, or irrelevant to the Webtoon content were excluded from the final dataset.

After annotation, all labels underwent a normalization process, including trimming unnecessary whitespace, converting labels to consistent casing, and merging spelling variations. This normalization ensured that every entry aligned with the predefined aspect and sentiment classes, enabling reliable encoding and model training.

To assess the reliability of the manual labeling process, inter-annotator agreement (IAA) was measured using Cohen's Kappa coefficient. A subset of the dataset was independently annotated by two annotators following the same annotation guidelines for both aspect category and sentiment polarity. Cohen's Kappa was selected because it accounts for agreement occurring by chance and is widely adopted in sentiment analysis and natural language processing research.

The agreement analysis resulted in a Cohen's Kappa score of $\kappa = 0.8333$, indicating an almost perfect level of agreement between annotators. This high level of agreement suggests that the annotation guidelines were clearly defined and consistently applied. Any remaining disagreements were resolved through discussion to produce the final ground truth labels used for model training and evaluation.

C. Text Preprocessing

Text preprocessing was performed to standardize raw comments and prepare them for feature extraction and model training. The preprocessing pipeline consisted of two main stages: basic text cleaning and advanced linguistic normalization.

1) Basic Text Cleaning

The initial cleaning process converts raw comments into a normalized textual format. The following operations were applied:

- Case Folding – converting all characters to lowercase to maintain uniformity.
- Punctuation and Symbol Removal – removing punctuation marks, special symbols, and emoji.
- Numeric Removal – eliminating standalone numbers that do not contribute to sentiment.
- Whitespace Normalization – replacing multiple spaces with a single space and trimming leading or trailing whitespace.
- Missing Value Handling – dropping comments that were empty or invalid after cleaning.

This stage ensures that the textual data is free from noise and inconsistencies that could negatively impact feature extraction.

2) Advanced Preprocessing

After basic cleaning, the text underwent further linguistic processing using the Sastrawi library, which is widely used for Indonesian language normalization. Two key operations were performed:

- Stopword Removal – eliminating common Indonesian words (e.g., *yang*, *dan*, *itu*) that carry little discriminative value for sentiment classification.
- Stemming – reducing words to their morphological root (e.g., *membantu* → *bantu*) to unify semantic representations.

The result of this stage is the Komentar_Preprocessed field, which contains the final normalized text used for TF–IDF feature extraction. In addition to stemming and stopword removal, lightweight normalization was applied to address informal slang and limited code-mixing commonly found in Indonesian Webtoon comments. Frequently occurring slang terms and informal variants (e.g., "gak" → "tidak", "banget" → "sangat", "anjg" → "anjing") were normalized using a manually curated slang mapping derived from common Indonesian social media usage.

For code-mixed expressions, English terms that frequently appear in Webtoon discussions (e.g., "plot", "ending", "scene", "visual") were retained without translation to preserve semantic meaning, while non-informative foreign tokens were filtered during preprocessing. This lightweight normalization strategy aims to reduce lexical variability while maintaining the original contextual meaning of user comments.

D. Feature Representation

The preprocessed comments were converted into numerical features to enable machine learning-based sentiment classification. This study employed a combination of lexical-based feature extraction using TF–IDF and categorical aspect encoding to support Aspect-Based Sentiment Analysis. In this study, Aspect-Based Sentiment Analysis is formulated as an aspect-conditioned sentiment classification task. Each comment is associated with a predefined aspect label obtained through manual annotation, and the model is trained to predict sentiment polarity conditioned on the given aspect. The aspect categories are not automatically extracted from the text; instead, they are incorporated as auxiliary features to guide the sentiment classification process.

1) TF–IDF Feature Representation

The textual content of each comment was represented using the Term Frequency–Inverse Document Frequency (TF–IDF) method, which quantifies the importance of a term relative to its frequency within the entire corpus. TF–IDF is effective for short user-generated text because it emphasizes discriminative lexical cues while maintaining computational efficiency. Formally, the TF–IDF weight of a term t in a document d is defined as shown in Equation 1.

$$\mathrm{tfidf(t, d)} = \mathrm{tf(t, d)} \times \log\left(\frac{\mathrm{N}}{\mathrm{df(t)}}\right) \quad (1)$$

2) Aspect Encoding and Feature Integration

To support Aspect-Based Sentiment Analysis, categorical aspect labels were incorporated into the feature space. Each aspect category, Plot, Character, and Visual, was transformed using One-Hot Encoding (OHE), producing three binary indicator features.

These encoded aspect features were horizontally concatenated with the TF–IDF feature matrix through sparse matrix stacking, resulting in a unified representation comprising 2,425 features (2,422 TF–IDF features and 3 OHE aspect features). This integrated structure enables the model to jointly capture sentiment-related lexical information and the contextual aspect referenced in each comment, thereby enhancing the relevance and accuracy of aspect-based sentiment classification.

E. Ensemble Learning Models

Ensemble learning models were employed to address the high variability, sparsity, and linguistic noise present in Indonesian Webtoon user comments. By combining multiple learners, ensemble methods improve robustness and generalization when handling high-dimensional and informal textual features.

Two ensemble approaches were evaluated in this study: Random Forest, representing the bagging paradigm, and XGBoost, representing the boosting paradigm. The inclusion of both methods enables a comparative assessment of variance reduction through bagging and iterative error correction through boosting when applied to TF–IDF-based feature representations.

1) Random Forest (Bagging Approach)

Random Forest is a bagging-based ensemble learning algorithm that constructs multiple decision trees using bootstrap sampling and random feature selection. Each tree is trained on a different bootstrap sample of the training data, while at each split only a random subset of features is considered. This mechanism reduces correlation among trees and improves robustness when handling noisy and high-dimensional textual features such as TF–IDF representations.

Formally, let $\{h_1(x), h_2(x), \ldots., h_B(x)\}$ denote the set of B decision trees in the Random Forest ensemble. For a classification task with class set C, the predicted class label $\hat{y}$ for an input instance x is determined by majority voting, as described In Equation 2.

$$\hat{y} \max_{c \in C} \sum_{b=1}^{B} 1\left(h_1(x) = c\right) \quad (2)$$

where 1(·) is an indicator function that returns 1 if the condition is true and 0 otherwise. By aggregating predictions from multiple trees, Random Forest effectively reduces variance and increases model stability, which is particularly beneficial for informal and noisy user-generated text. In this study, Random Forest was implemented in three configurations: (1) Baseline Random Forest (TF–IDF only); (2) ABSA Random Forest (TF–IDF + OHE); and (3) Tuned Random Forest.

2) XGBoost (Boosting Approach)

XGBoost is an optimized gradient boosting algorithm that builds an ensemble of decision trees sequentially, where each new tree is trained to correct the errors made by the previous ensemble. Unlike bagging-based methods, boosting focuses on reducing bias by emphasizing difficult-to-classify instances during training, allowing the model to capture complex decision boundaries in structured and high-dimensional feature spaces.

In XGBoost, the prediction for an instance x_1 at iteration m is given by an additive model as described In Equation 3:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + fm(x_i), \quad (3)$$

where $\int m\,(\cdot)$ represents the decision tree added at iteration m. The model is trained by minimizing a regularized objective function, as described In Equation 4:

$$L = \sum_{i=1}^{n} \ell\,(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \quad (4)$$

where $\ell(\cdot)$ is a differentiable loss function (e.g., logistic loss for classification) and Ω(fk) is a regularization term that penalizes model complexity to prevent overfitting. This regularization mechanism enables XGBoost to maintain strong generalization performance even when applied to sparse TF–IDF feature representations.

XGBoost was applied in three configurations: (1) Baseline XGBoost (TF–IDF features only); (2) Tuned XGBoost (TF–IDF features with optimized hyperparameters); and (3) ABSA XGBoost (TF–IDF features combined with OHE aspect features). All model configurations were trained and evaluated using identical preprocessing steps and data splits to ensure fair and consistent comparison across experiments.

F. Hyperparameter Tuning

Hyperparameter tuning was conducted to optimize the performance of selected ensemble learning models by identifying parameter configurations that improve generalization while reducing overfitting. This process was applied to models that demonstrated competitive baseline performance to assess the impact of optimization in both text-only and aspect-based classification settings.

1) Randomized Search Strategy

The tuning process used RandomizedSearchCV, a widely adopted approach that samples combinations of hyperparameters from predefined distributions. This method offers an efficient alternative to exhaustive grid search, especially for high-dimensional feature spaces such as TF–IDF and sparse concatenated matrices used in ABSA. The tuning procedure was applied separately to:

a. Random Forest, evaluating parameters such as number of estimators, maximum tree depth, and feature sampling strategy (e.g., *sqrt*). the tuning process assessed constraints on the minimum samples per split and leaf, as well as the application of bootstrap usage.
b. XGBoost, evaluating parameters such as number of boosting rounds, maximum tree depth, learning rate, subsampling ratio, and column sampling ratio.

2) Cross-Validation Procedure

Each parameter configuration was evaluated using 5-fold stratified cross-validation to ensure that class distribution remained consistent across folds. This procedure reduces the risk of overfitting by validating model performance across multiple data partitions.

The tuned models obtained from this process were subsequently used in the comparative evaluation described in the next chapter. All tuning experiments were performed on the training set only, ensuring that the test set remained unseen during optimization and preserved its role as an unbiased evaluation set.

G. Model Evaluation

To assess the performance of the ensemble learning models, several evaluation metrics and validation procedures were applied. These metrics were selected to measure not only overall prediction accuracy but also the model's ability to correctly classify each sentiment class within the dataset. Evaluation was conducted consistently across all model configurations including baseline, tuned, and ABSA variants to ensure fair comparison.

1) Evaluation Metrics

Four standard metrics were used to evaluate the classification performance of the proposed models. These metrics were selected to measure both overall predictive accuracy and class-level performance in multi-class sentiment classification. Accuracy measures the proportion of correctly predicted sentiment labels relative to all predictions and provides an overall view of model performance. Formally, accuracy formula is defined in Equation 5:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. Precision indicates the proportion of correctly predicted positive instances among all instances predicted as positive for a given class, reflecting the model's ability to avoid false positives. Precision formula is defined in Equation 6:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

Recall measures the proportion of correctly predicted positive instances among all actual positive instances for a given class, indicating the model's ability to correctly identify true positives. Recall formula is defined in Equation 7:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{7}$$

F1-score represents the harmonic mean of precision and recall, providing a balanced measure of classification performance, particularly under class imbalance. The F1-score formula is defined in Equation 8:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} \pm \text{Recall}} \tag{8}$$

For multi-class sentiment classification, both macro-averaged and weighted-average F1-scores were reported. Macro-averaged F1 assigns equal weight to each sentiment class, while weighted-average F1 accounts for the relative support of each class, providing a more representative measure of overall performance when class distributions are imbalanced.

2) Confusion Matrix

In addition to aggregate metrics, a confusion matrix was used to visualize misclassification patterns across sentiment classes. This matrix provides insight into the types of errors made by the model, such as confusion between neutral and positive sentiments or between negative and neutral expressions. The confusion matrix was generated for each major model configuration to support qualitative performance interpretation.

3) Train–Test Evaluation Setup

All models were evaluated using the same 80:20 stratified train–test split, ensuring proportional representation of sentiment classes in both the training and testing subsets. This approach maintains consistency across experiments and minimizes sampling bias.

4) Cross-Validation

To further validate model robustness, a 10-fold stratified cross-validation was conducted using the combined TF–IDF and OHE feature representation. Cross-validation provides a more reliable estimate of generalization performance by averaging results across multiple partitions of the dataset. This evaluation helps assess how well the models would perform on unseen data beyond the initial train–test split.

IV. Results and Discussion

A. Dataset Exploration

After data cleaning, a total of 1,010 valid comments were retained for analysis, each annotated with one of three sentiment classes (Negative, Neutral, and Positive) and one of three aspect categories (Storyline, Visuals, and Characters).

Figure 2 shows a relatively balanced sentiment distribution, with Neutral sentiment as the dominant class, a pattern commonly observed in user-generated content where comments tend to be descriptive rather than explicitly evaluative [27], [28]. Figure 3 shows the aspect of distribution of webtoon comments.

Storyline emerges as the most frequently discussed aspect, indicating that narrative progression is a prominent focus of Webtoon readers [3], [29]. Although Visuals and Characters appear less frequently, both remain sufficiently represented for Aspect-Based Sentiment Analysis. To further examine aspect-dependent sentiment patterns, Table 1 summarizes the distribution of sentiment polarity across aspect categories.

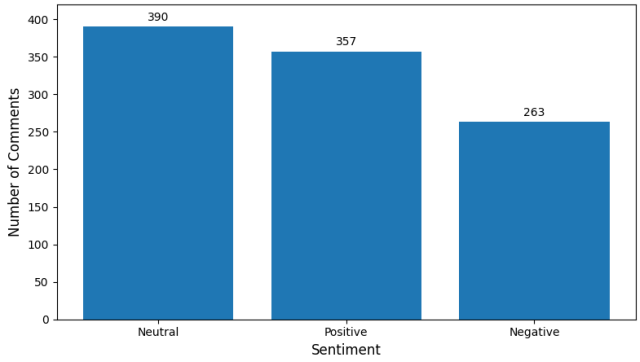

Figure 2. Sentiment Distribution of Webtoon Comments.

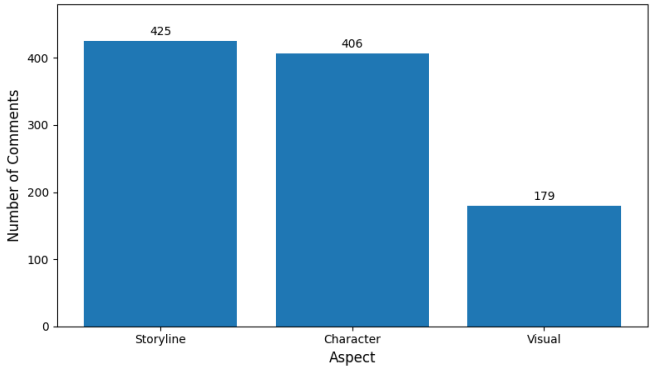

Figure 3. Aspect Distribution of Webtoon Comments

Table 1. Sentiment Distribution per Aspect.

Aspect	Negative	Neutral	Positive	Total
Storyline	48	220	157	425
Characters	111	134	161	406
Visuals	104	36	39	179

The distribution of sentiment polarity across aspects reveals distinct patterns of reader perception. Comments related to the storyline are predominantly neutral, indicating that many users provide descriptive or observational feedback rather than explicit praise or criticism [27], [28]. In contrast, the character aspect shows a higher proportion of positive sentiment, suggesting that character development and portrayal are generally well received by readers [3], [30]. Meanwhile, comments addressing visual aspects are dominated by negative sentiment, reflecting more frequent criticism related to artwork quality, drawing style, or visual consistency. This finding aligns with prior research emphasizing that visual presentation is a salient factor in the Webtoon reading experience, which often elicits strong reader responses, including critical evaluations [29],[30]. These variations across aspects highlight the importance of aspect-based analysis, as aggregating sentiment at the document level would obscure such nuanced differences in reader opinions [31][32].

To gain insight into the dominant vocabulary used across the dataset, a word cloud was generated from the preprocessed text, as shown in Figure 4. Frequently appearing words include expressions of emotional reactions (e.g., *"seru"*, *"keren"*, *"merinding"*), references to story elements, and comments related to

character behavior or visual details. The presence of informal expressions and slang terms highlights the noisy and conversational nature of Indonesian Webtoon reader comments, a characteristic that has been widely reported in studies on Indonesian social media text and user-generated content [8], [33]. This observation further justifies the use of extensive text preprocessing and robust ensemble learning models to effectively handle high linguistic variability and noise in aspect-based sentiment classification [8], [28].

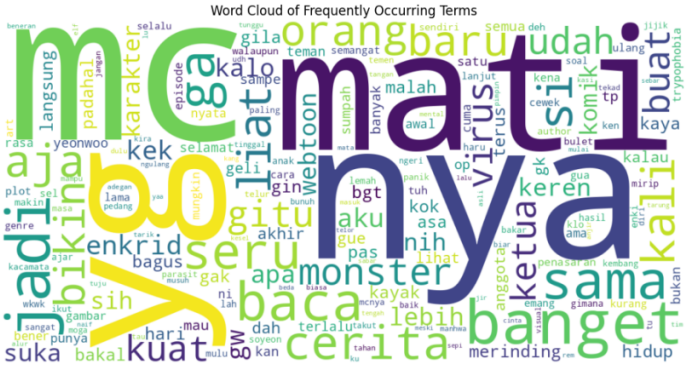

Figure 4. Word Cloud of Frequently Occurring Terms

B. Performance of Ensemble Models and Optimization

In the initial stage of experimentation, ensemble learning models were evaluated using text-only features represented by TF–IDF vectors of the pre-processed comments. Three model configurations were tested on the stratified test set of 202 comments: (1) Random Forest, (2) default XGBoost, and (3) tuned XGBoost, as described in Table 2.

Table 2. Performance Comparison of Text-Only Ensemble Models.

Model	Features	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
Random Forest	TF–IDF only	0.6188	0.6604	0.6188	0.6069
XGBoost (default)	TF–IDF only	0.6089	0.6227	0.6089	0.6064
XGBoost (tuned)	TF–IDF only	0.6238	0.6555	0.6238	0.6161

Overall, the tuned XGBoost model achieved the highest performance among the text-only configurations, indicating that hyperparameter optimization provides a modest improvement over the baseline boosting model. However, the performance gap between the tuned and baseline models remains relatively small, suggesting that text-only representations still impose limitations on sentiment classification in this context.

A class-level examination shows consistent patterns across all text-only models. The Negative class exhibits high precision but lower recall, suggesting that models tend to be conservative when assigning negative sentiment. The Neutral class achieves the highest recall (up to 0.8500 for Random Forest), reflecting its descriptive and less polarized nature. The Positive class maintains balanced precision and recall values with F1-scores ranging from 0.5900 to 0.6300. These observations indicate that while ensemble models can handle informal Indonesian text reasonably well, additional semantic cues such as aspect indicators are needed to improve the classification of subtle negative expressions.

C. ABSA Model Performance

To evaluate the effectiveness of incorporating aspect information into sentiment classification, the aspect labels (storyline, visual, character) were encoded using One-Hot Encoding (OHE) and concatenated with TF–IDF features, as shown in Table 3. Three ensemble models were then evaluated using the combined feature set: (1) XGBoost ABSA, (2) baseline Random Forest ABSA, and (3) tuned Random Forest ABSA.

The results show that all ABSA models outperform the text-only configurations, demonstrating that incorporating aspect indicators enhances sentiment prediction in multi-dimensional review contexts. Among all evaluated models, the tuned Random Forest ABSA model achieved the highest accuracy (0.6584) and weighted F1-score (0.6541), making it the best-performing configuration in this study.

Table 3. Performance Comparison of ABSA Ensemble Models.

Model	Features	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
XGBoost ABSA	TF–IDF + Aspect (OHE)	0.6436	0.6604	0.6436	0.6355
Random Forest ABSA	TF–IDF + Aspect (OHE)	0.6485	0.6745	0.6485	0.6378
Random Forest ABSA (Tuned)	TF–IDF + Aspect (OHE)	0.6584	0.6876	0.6584	0.6541

A detailed class-level performance analysis indicates that the Neutral class remains the easiest to classify, with recall values reaching up to 0.82. This reflects the descriptive and less polarized nature of neutral comments. The Positive class obtains relatively stable F1-scores across all models, ranging from 0.6500 to 0.6600. Meanwhile, the Negative class shows improved recall when aspect features are included, increasing from 0.4300 in the text-only tuned XGBoost model to 0.4900 under the tuned Random Forest ABSA configuration. This improvement suggests that aspect cues help the model better identify complaint-driven or criticism-related comments.

1) Confusion Matrix of the Best Model

To further analyze prediction behavior, a confusion matrix was generated for the tuned Random Forest ABSA model, as shown in Figure 5. The matrix shows clearer cluster separation between the three sentiment classes compared to text-only models, particularly in distinguishing negative versus neutral expressions. Some misclassifications remain between the Neutral and Positive categories, which is expected due to the informal tone and mixed emotional content of many Webtoon comments.

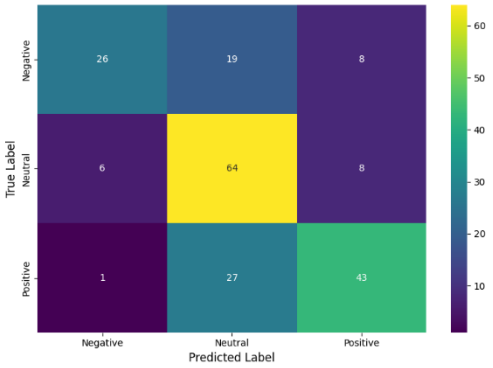

Figure 5. Confusion Matrix of the Tuned Random Forest ABSA Model.

2) Feature Importance Analysis

To understand which features, contribute most to the model's decision-making, feature importance values were extracted from the tuned Random Forest ABSA model. Interestingly, two of the top five most important features originate from aspect indicators, specifically Visuals_Aspect and Storyline_Aspect.

This emphasizes that aspect information is not only additive but significantly influences model predictions. The remaining top-ranked features are high-frequency sentiment-bearing words such as *"seru"*, *"keren"*, *"meringding"*, and *"bagus"*. Figure 6 presents the top 20 most influential features contributing to sentiment prediction.

3) Error Analysis

To better understand the limitations of the tuned Random Forest ABSA model, a qualitative error analysis was conducted on a subset of misclassified Webtoon comments. Table 4 presents five representative examples that illustrate common linguistic challenges present in Indonesian Webtoon reader comments.

Based on Table 4, the observed misclassifications can be grouped into three dominant error categories. First, the model struggles to capture negation and intensifiers simultaneously. In Example 3, the presence of intensifiers such as "sumpah" and "banget" increases positive lexical weights, overshadowing the negative meaning expressed through negation ("gak jelas").

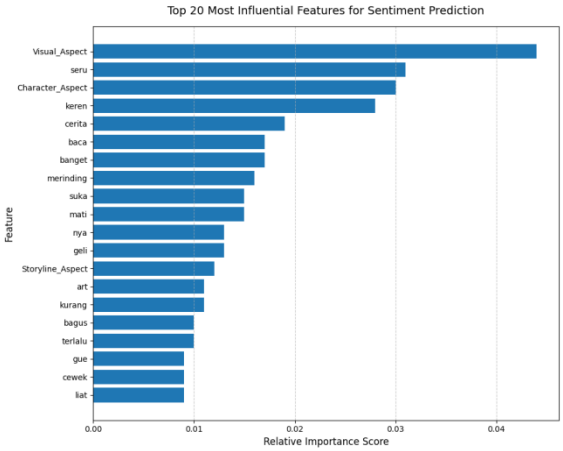

Figure 6. Top 20 Most Influential Features Contributing to Sentiment Prediction

Second, idiomatic and rhetorical expressions pose a challenge for TF–IDF-based representations. Comments such as "ini baru temen??" (Example 4) convey strong appreciation but lack standard sentiment adjectives, leading the classifier to predict Neutral sentiment.

Third, negative sentiment in Webtoon comments is often expressed implicitly through narrative critique rather than explicit evaluative language. In Example 5, dissatisfaction with story logic is communicated using lexically neutral terms, which reduces the model's ability to identify the underlying negative sentiment.

These findings indicate that although the proposed ABSA framework improves overall performance, misclassifications are primarily driven by sarcasm, slang, negation, and implicit narrative criticism. Incorporating contextualized embeddings and improved handling of negation and slang normalization may further reduce such errors in future work.

Table 4. Qualitative Error Analysis of the Proposed Model.

No	Comment (Raw Text)	Ground Truth	Predicted	Potential Cause of Error
1	*"lucu bangett pada peduli semuaa sama ketuaa"*	Positive	Neutral	Lexical ambiguity: Words such as lucu and peduli are interpreted as descriptive rather than strongly sentiment-bearing by TF–IDF features.
2	*"temannya yg lain kek anjg gitu cape² ditolong..."*	Negative	Neutral	Slang & profanity: Informal slang (e.g., anjg [sic]) and colloquial expressions dilute sentiment signals and bias the model toward the Neutral class.
3	*"sumpah si ini komik gak jelas banget…"*	Negative	Positive	Negation failure: Intensifiers (sumpah, banget) receive high positive weights, while negation (gak jelas) is not effectively captured.
4	*"ini baru temen??"*	Positive	Neutral	Idiomatic expression: Rhetorical appreciation without explicit positive adjectives is difficult for bag-of-words representations to capture.
5	*"butuh setahun untuk menang lawan dia tapi lupa…"*	Negative	Neutral	Implicit criticism: Critique of storyline logic is conveyed implicitly without explicit negative adjectives.

D. Cross-Validation

To evaluate the stability and robustness of the proposed model, a 10-fold cross-validation procedure was conducted using the combined TF–IDF and aspect feature set. The accuracy scores obtained across the ten folds show a moderate level of variation, with values ranging from 0.5247 to 0.7029. Overall, the model

achieved a mean cross-validation accuracy of 0.6238 with a standard deviation of 0.0572, indicating a fairly consistent performance across different data partitions. Although some folds produced lower accuracy due to the presence of noisier or more imbalanced subsets, the majority of folds remained within a relatively stable range. These results demonstrate that the ensemble-based ABSA model particularly the tuned Random Forest configuration exhibits reliable generalization and is not overly sensitive to the specific composition of the training data.

E. Discussion

The experimental results demonstrate that incorporating aspect information substantially enhances sentiment classification performance. Compared to text-only configurations, all ABSA models achieve higher accuracy and F1-scores, with the tuned Random Forest ABSA model achieving the best performance (accuracy 0.6584, weighted F1-score 0.6541), confirming that aspect cues provide additional contextual signals for distinguishing subtle sentiment expressions in Webtoon comments.

The Neutral class consistently achieves the highest recall, reflecting the descriptive and less emotional nature of neutral comments, whereas negative comments often contain informal slang, sarcasm, or figurative expressions that make them harder to distinguish from neutral statements. Nevertheless, the inclusion of aspect features improves Negative class recall, indicating that aspect cues help the model better capture criticism related to specific narrative or visual elements.

Feature importance analysis further strengthens the role of aspect information in the ABSA framework, with Visuals and Storyline appearing among the top contributors in the tuned Random Forest model. This indicates that Webtoon readers frequently associate sentiment with specific narrative and visual elements, aligning with common reader behavior observed in online comic communities.

While boosting-based methods often outperform bagging approaches in structured datasets, the tuned Random Forest model surpasses both default and tuned XGBoost configurations in this ABSA setting, suggesting that bagging offers greater robustness for noisy and highly variable Indonesian user comments.

Although only lightweight slang normalization was applied, the error analysis indicates that more comprehensive slang dictionaries and improved code-mixing handling may further enhance sentiment detection, particularly for implicit negative expressions.

Overall, the results confirm that ABSA is more effective than document-level sentiment analysis for understanding user opinions on digital entertainment content, as aspect features improve classification accuracy and reveal narrative and visual factors driving audience sentiment. Beyond performance improvement, this study demonstrates that aspect-conditioned ensemble learning, combined with qualitative error analysis, extends existing ABSA studies in the Indonesian language context.

Although transformer-based models such as IndoBERT have shown strong performance in Indonesian NLP tasks, this study intentionally focuses on ensemble-based classical machine learning under a controlled TF–IDF representation, with direct comparison to deep contextual models left for future work.

V. Conclusion

This study presents an Aspect-Based Sentiment Analysis (ABSA) framework for classifying Indonesian Webtoon comments using ensemble learning methods. A dataset of 1,010 manually annotated comments was processed through multiple stages, including text cleaning, normalization, stemming, TF–IDF vectorization, and aspect encoding using One-Hot Encoding. Several ensemble configurations were evaluated to measure the impact of incorporating aspect information into the sentiment classification process.

The results show that ABSA demonstrates improved performance compared to text-only sentiment analysis. Among all tested models, the tuned Random Forest ABSA configuration achieved the best performance, with an accuracy of 0.6584 and a weighted F1-score of 0.6541 on the test set. The inclusion of aspect features enhanced the model's ability to distinguish subtle sentiment expressions, particularly in identifying negative comments that are often expressed using informal or ambiguous language. Feature importance analysis also indicates that aspect indicators especially Visuals and Storyline play a significant role in guiding the model's decisions.

Cross-validation results further confirm that the proposed approach demonstrates stable and reliable generalization, with a mean 10-fold accuracy of 0.6238 and a standard deviation of 0.0572. These findings indicate that ensemble-based ABSA enables more fine-grained sentiment analysis across predefined aspects, as reflected in improved classification performance and error reduction compared to document-level sentiment analysis.

For future work, the performance of ABSA can be enhanced by incorporating deep contextual embeddings such as IndoBERT or multilingual transformer models, expanding the dataset to cover more

genres and linguistic variations, or employing semi-supervised learning to reduce annotation costs. Aspect-level insights derived from the proposed ABSA framework may potentially support content creators in understanding audience reactions and improving narrative or visual elements in digital comic production.

Conflicts of Interest
The authors declare no conflicts of interest.

Author Contributions Statement
All authors contributed significantly to the completion of this study. Rival Fakhri Amrullah conducted the data collection, annotation, preprocessing, model development, analysis, and manuscript preparation. Fathoni provided supervision, methodological guidance, and refinement of the research framework. Dani Indra contributed to validation, manuscript review, and structural improvements. All authors have read and approved the final published version of the manuscript.

Acknowledgment
The authors would like to thank Universitas Sebelas April (UNSAP) for providing academic support throughout the research process. The authors also extend their appreciation to the annotators and peers who assisted in validating the dataset, as well as those who provided technical input during the experimental phase. No external funding was received for this study.

References
[1] A. F. Lestari and I. Irwansyah, "Line Webtoon Sebagai Industri Komik Digital," SOURCE J. Ilmu Komun., vol. 6, no. 2, p. 134, 2020, doi: 10.35308/source.v6i2.1609.
[2] F. T. Admojo, S. Risnanto, A. W. Windiawati, M. Innuddin, and D. Mualfah, "Comparison of Naïve Bayes and Random Forest Algorithm in Webtoon Application Sentiment Analysis," Innov. Res. Informatics, vol. 6, no. 1, pp. 23–28, 2024, doi: 10.37058/innovatics.v6i1.10636.
[3] B. Yecies, D. Wang, M. Amirghasemi, M. Lu, and K. Kariippanon, "Communicating Through Chaos in the Webtoon Parasocial Intimacy Chamber," Int. J. Commun., vol. 18, no. 006021001, pp. 2919–2947, 2024.
[4] A. Chamekh, M. Mahfoudh, and G. Forestier, "Sentiment Analysis Based on Deep Learning in E-Commerce BT - Knowledge Science, Engineering and Management," G. Memmi, B. Yang, L. Kong, T. Zhang, and M. Qiu, Eds., Cham: Springer International Publishing, 2022, pp. 498–507.
[5] N. Amalia Putri, A. Srirahayu, and N. Arif Sudibyo, "Sentiment Analysis Towards the KitaLulus Application Using the Naive Bayes Method from Google Play Store Reviews," J. Indones. Sos. Teknol., vol. 5, no. 10, pp. 4593–4603, 2024, doi: 10.59141/jist.v5i10.1244.
[6] A. Muzaki, V. Febriana, and W. N. Cholifah, "Analisis Sentimen Pada Ulasan Produk di E-Commerce dengan Metode Naive Bayes," J. Ris. dan Apl. Mhs. Inform., vol. 5, no. 4, pp. 758–765, 2024, doi: 10.30998/jrami.v5i4.9647.
[7] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey," IEEE Trans. Affect. Comput., vol. 13, no. 2, pp. 845–863, 2022, doi: 10.1109/TAFFC.2020.2970399.
[8] A. Bustamin, A. A. Prayogi, D. Siswanto, M. Rafrin, and A. Nurdin, "Text Normalization for Indonesian Slang Words in Sentiment Analysis Development," ICIC Express Lett. Part B Appl., vol. 16, no. 2, pp. 121–129, 2025, doi: 10.24507/icicelb.16.02.121.
[9] A. M. Barik, R. Mahendra, and M. Adriani, "Normalization of indonesian-english code-mixed twitter data," W-NUT@EMNLP 2019 - 5th Work. Noisy User-Generated Text, Proc., pp. 417–424, 2019, doi: 10.18653/v1/d19-5554.
[10] F. Greco, "Sentiment analysis and opinion mining," Elgar Encycl. Technol. Polit., no. May, pp. 105–108, 2022, doi: 10.4337/9781800374263.sentiment.analysis.
[11] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artif. Intell. Rev., vol. 55, no. 7, pp. 5731–5780, 2022, doi: 10.1007/s10462-022-10144-1.
[12] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," IEEE Trans. Knowl. Data Eng., vol. 35, no. 11, pp. 11019–11038, 2023, doi: 10.1109/TKDE.2022.3230975.
[13] H. Hayat, C. Ventura, and A. Lapedriza, "Modeling Subjective Affect Annotations with Multi-Task Learning," Sensors, vol. 22, no. 14, 2022, doi: 10.3390/s22145245.
[14] M. Água, N. António, P. Carrasco, and C. Rassal, "Large Language Models Powered Aspect-Based

Sentiment Analysis for Enhanced Customer Insights," Tour. Manag. Stud., vol. 21, no. 1, pp. 1–19, 2025, doi: 10.18089/tms.20250101.

[15] H. G. Sulistio and A. Handojo, "Aspect-Based Sentiment Analysis pada Ulasan ECommerce dengan Metode Support Vector Machine untuk Mendapatkan Informasi Sentimen dari Beberapa Aspek," J. Infra, vol. 10, no. 2, pp. 1–5, 2022.

[16] R. Z. Suchrady and A. Purwarianti, "Indo LEGO-ABSA: A Multitask Generative Aspect Based Sentiment Analysis for Indonesian Language," Proc. Int. Conf. Electr. Eng. Informatics, 2023, doi: 10.1109/ICEEI59426.2023.10346852.

[17] Y. A. Mustofa and I. S. K. Idris, "Ensemble Approach to Sentiment Analysis of Google Play Store App Reviews," Jambura J. Electr. Electron. Eng., vol. 6, no. 2, pp. 181–188, 2024, doi: 10.37905/jjeee.v6i2.25184.

[18] E. Daniati and H. Utama, "Analisis Sentimen Dengan Pendekatan Ensemble Learning Dan Word Embedding Pada Twitter," J. Inf. Syst. Manag., vol. 4, no. 2, pp. 125–131, 2023, doi: 10.24076/joism.2023v4i2.973.

[19] K. P. Harmandini and K. M. L, "Analysis of TF-IDF and TF-RF Feature Extraction on Product Review Sentiment," Sinkron, vol. 8, no. 2, pp. 929–937, 2024, doi: 10.33395/sinkron.v8i2.13376.

[20] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[21] Z. An, T. Xiong, Z. Zou, and H. Wan, "Aspect-based Sentiment Analysis with an Ensemble Learning Framework for Requirements Elicitation from App Reviews," J. Internet Technol., vol. 25, no. 7, pp. 1083–1090, 2024, doi: 10.70003/160792642024122507012.

[22] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," WIREs Data Min. Knowl. Discov., vol. 8, no. 4, p. e1253, Jul. 2018, doi: https://doi.org/10.1002/widm.1253.

[23] W. Meng, Y. Wei, P. Liu, Z. Zhu, and H. Yin, "Aspect Based Sentiment Analysis with Feature Enhanced Attention CNN-BiLSTM," IEEE Access, vol. 7, pp. 167240–167249, 2019, doi: 10.1109/ACCESS.2019.2952888.

[24] L. Xu and W. Wang, "Aspect-based sentiment classification with BERT and AI feedback," Nat. Lang. Process. J., vol. 10, no. September 2024, p. 100136, 2025, doi: 10.1016/j.nlp.2025.100136.

[25] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," COLING 2020 - 28th Int. Conf. Comput. Linguist. Proc. Conf., pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess., pp. 97–101, 2016, doi: 10.18653/v1/n16-3020.

[27] J. Cui, Z. Wang, S. B. Ho, and E. Cambria, Survey on sentiment analysis: evolution of research methods and topics, vol. 56, no. 8. Springer Netherlands, 2023. doi: 10.1007/s10462-022-10386-z.

[28] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," Decis. Anal. J., vol. 3, no. June, p. 100073, 2022, doi: 10.1016/j.dajour.2022.100073.

[29] S. P. Nurohmah, G. F. Putri, K. Imawan, and A. D. Lestari, "Content Analysis of Comics on Line Webtoon," Int. J. Business, Econ. Soc. Dev., vol. 5, no. 4, pp. 483–492, 2024.

[30] J. L. Budianto and D. Ratri, "Influence of Character ' s Visual Style o n Reader ' s Empathy on Sad Emotional Story (Case Study : Webtoon ' Bingkai Titik ')," pp. 36–49, 2022.

[31] Y. C. Hua, P. Denny, J. Wicker, and K. Taskova, A systematic review of aspect-based sentiment analysis: domains, methods, and trends, vol. 57, no. 11. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10906-z.

[32] G. Brauwers and F. Frasincar, "A Survey on Aspect-Based Sentiment Classification," ACM Comput. Surv., vol. 55, no. 4, 2023, doi: 10.1145/3503044.

[33] A. N. A. Saputra, R. E. Saputro, and D. I. S. Saputra, "Enhancing Sentiment Analysis Accuracy Using SVM and Slang Word Normalization on YouTube Comments," Sinkron, vol. 9, no. 2, pp. 687–699, 2025, doi: 10.33395/sinkron.v9i2.14613.