Journal of Information Systems and Informatics
Vol. 7, No. 2, June 2025 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v7i2.1150 Published By DRPM-UBD

Sentiment Analysis of Consumer Acceptance of Honda's Digital Marketing Strategy Using Lexicon-Based Algorithm

Bartolomius Dias[1], Asma' Khoirunnisa'[2], Yosef Budiman[3], Setiyawami[4]

[1,4]Department of Business and Finance, Universitas Negeri Yogyakarta, Indonesia
[2]Statistics Study Program, Universitas Negeri Yogyakarta, Indonesia
[3]School of Manufacturing System and Mechanical Engineering, Thammasat University, Thailand
Email: [1]bartolomius0064fe.2022@student.uny.ac.id, [2]asmakhoirunnisa.2021@student.uny.ac.id, [3]yoshimabudiman@gmail.com, [4]setiyawami@uny.ac.id

Abstract

This study analyzes customer sentiment toward Honda's digital marketing strategy via the Wahana Honda application. A total of 2,000 customer reviews were collected from the Google Play Store using web-scraping techniques. Text data underwent preprocessing (e.g. cleansing, tokenization, stop-word removal, stemming, and translation into English). Sentiment classification using a lexicon-based approach revealed that 56.7% of reviews were positive, 20.8% neutral, and 22.5% negative. The model demonstrated high precision in identifying negative sentiment, though it showed limitations in classifying neutral opinions due to linguistic ambiguity. These findings highlight the need for more adaptive sentiment models and offer strategic insights for Honda's digital marketing. Specifically, the analysis can help prioritize improvements in app functionality, excellence service priority, enhance personalized customer engagement, and shape targeted digital marketing strategies based on real user feedback. Leveraging these insights enables Honda to optimize user experience, increase retention, and align digital campaigns with customer expectations.

Keywords: Customer Review, Digital Marketing, Lexicon-Based Algorithm, Sentiment Analysis, Wahana Honda

1. INTRODUCTION

HONDA is one of the few leading companies in the world that consistently produces high-quality automotive products. In line with this commitment, HONDA continues to prioritize innovation, especially in digital marketing strategies aimed at enhancing user satisfaction. One such innovation is the launch of a virtual assistant platform known as Wahana Honda (WANDA). This platform was designed to address customer needs, offering free services such as emergency assistance, vehicle maintenance support, and insurance-related help. WANDA's creation reflects HONDA's proactive approach to integrating technology into

customer service. However, despite its intended advantages, the platform has received substantial criticism from users.

Behind the strengths of WANDA, various limitations in the software have triggered widespread complaints from users of the Honda brand. These complaints are not limited to app reviews but are also frequently voiced across HONDA's digital media platforms like Instagram, TikTok, and Twitter. Many users express dissatisfaction related to slow response, technical glitches, or unaddressed inquiries. Unfortunately, such feedback often goes unnoticed or unresolved by the company, leading to growing frustration. As a result, public trust in the HONDA brand has been affected, giving rise to widespread negative sentiment and reducing the overall brand image. This problem reflects a critical issue in the feedback-response loop between users and platform developers.

This phenomenon signals a relevant research opportunity within the domain of sentiment analysis, particularly through the application of machine learning-based methods like lexicon algorithms. Lexicon is a frequently used algorithm for computing the sentiment orientation of an entire document or reference set [1], which relies on the power of sentiment-expressing words to predict the subjective nature of textual content [2]. These methods perform consistently across multiple domains, categorizing sentiment as positive, negative, or neutral. However, their effectiveness is limited when dealing with sarcasm, as sarcasm can appear even in text that outwardly seems positive or neutral [3]. This shortcoming reveals a key challenge in using lexicon-based methods for complex sentiment interpretation.

Moreover, sentiment analysis or opinion mining is defined as a computational method used to detect, extract, and interpret subjective information from text to assess the overall sentiment or attitude expressed about a particular entity, product, or service [4]. Existing research, such as that by Barik and Misra (2024), has explored sentiment analysis using lexicon methods; however, their study only touches on general cases and fails to offer platform-specific insights [5]. This lack of focus on particular services or applications presents a significant research gap. Therefore, this study intends to explore sentiment analysis specifically within the context of the WANDA platform a novel direction that has not yet been addressed in prior academic literature.

To bridge this gap, the main aim of this research is to analyze public sentiment toward HONDA's WANDA platform based on user opinions and comments posted on digital platforms, particularly Google Play. This study applies a lexicon-based sentiment analysis algorithm to understand the polarity and tone of the comments, identifying patterns that could influence customer perception. Ultimately, the results will be correlated with strategic digital marketing insights to reduce negative sentiment and improve user experience. By focusing on a specific,

under-researched platform and applying an established yet evolving methodology, this study provides actionable knowledge that can help refine HONDA's digital service performance and strengthen its brand image in the long run.

2. METHODS

A systematic workflow involving data collection, preprocessing, and sentiment classification, which aims to extract meaningful insights from customer reviews through lexicon-based sentiment analysis, including the following sections below.

2.1. Data Collection

The data collection process in this study focused on analyzing customer service performance on Wahana Honda application by utilizing transaction and service records from the company's internal database. Data was extracted and processed using Python, with the dataset comprising various service attributes such as customer username and review. These records span a defined operational period in 2019 – 2025, enabling the identification of customer sentiment. Data cleaning and preprocessing were carried out to handle missing values and standardize time formats, ensuring the dataset's reliability for subsequent analysis. The structured and automated collection process allowed for objective insights into service performance without relying on subjective assessments.

2.2. Text Preprocessing

Text preprocessing was a critical step in preparing the dataset for analysis, particularly for handling textual attributes related to customer feedback about general review of application. The preprocessing workflow involved several standard natural language processing (NLP) techniques implemented using Python that is represented on flowchart in Figure 1. Initially, all text data were converted to lowercase to ensure uniformity. This was followed by the removal of punctuation, numbers, and special characters that did not contribute to semantic meaning. Tokenization was applied to split text into individual words, and stop words were removed to reduce noise. Furthermore, stemming techniques were employed to normalize words to their root forms, aiding in dimensionality reduction and improving the consistency of the data. These preprocessing steps ensured that the textual data were clean, structured, and suitable for further qualitative and quantitative analysis.

The workflow depicted in Figure 1 illustrates a comprehensive text-preprocessing pipeline commonly employed in sentiment and lexicon-based classification studies. After importing essential libraries (e.g., pandas and NumPy for data handling, google-play-scraper for review extraction, Sastrawi for Indonesian stemming,

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

VADER for sentiment scoring, and visualization tools such as WordCloud and matplotlib), raw text is first cleansed by removing noise (HTML tags, punctuation, stop words). Zhu et al., (2024) comprehensively reviews these AI-based methods for error detection, data imputation, and repair, comparing them with traditional techniques across key dimensions to accelerate into cleaned data [6], then cleaned data normalized via case folding. Case folding is a fundamental text-normalization step in many journal research methodologies for information retrieval and natural language processing [7].

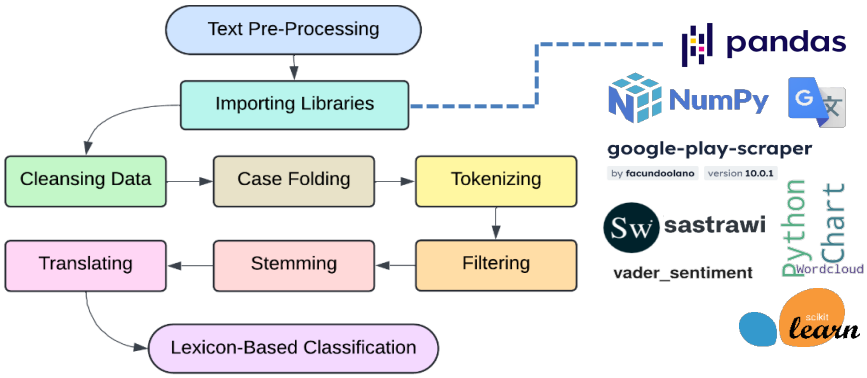

Figure 1. Flowchart of text preprocessing

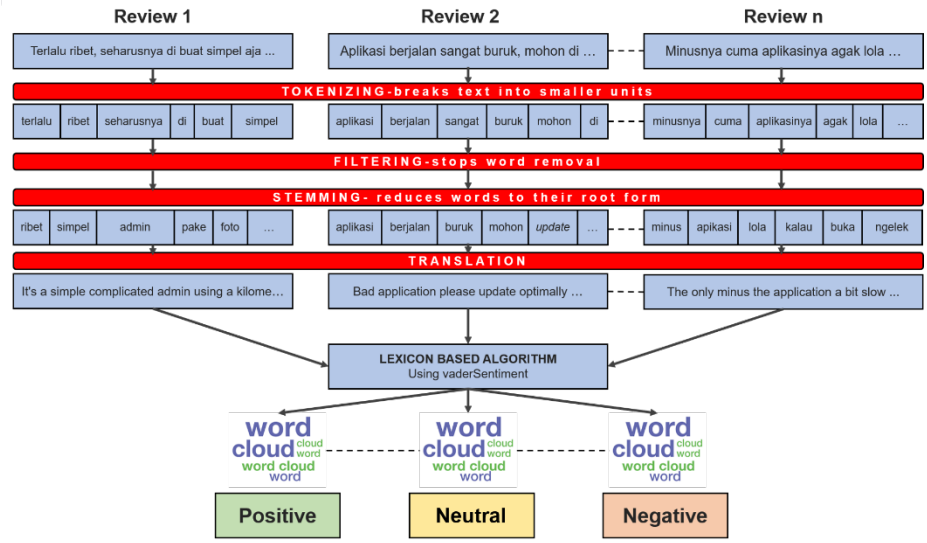

Figure 2. Text preprocessing steps

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Furthermore, the text split into discrete units via tokenization, which are subsequently filtered to eliminate irrelevant tokens as illustrated in Figure 2. The filtered tokens undergo stemming to reduce them to their root forms, before an optional translation step aligns multilingual inputs into a single target language. Finally, the processed tokens feed into a lexicon-based classifier to assign sentiment or thematic labels.

2.3. Lexicon-Based Classification

The lexicon-based classification approach was employed to analyze sentiment orientations within textual data [8]. This method relies on predefined sentiment lexicons, where each term is assigned a polarity score reflecting its emotional orientation (positive, negative, or neutral). In this study, we constructed a domain-adapted lexicon by integrating publicly available sentiment dictionaries with domain-specific modifications to capture contextual nuances effectively.

The classification process involved multiple preprocessing stages, including tokenization, lowercasing, lemmatization, and stop-word removal. Following preprocessing, the lexicon was applied by matching tokens to the sentiment dictionary, and an aggregated sentiment score was calculated for each text instance. The final sentiment class is determined based on cumulative score thresholding technique, relevant to the research conducted by [9]. Although lexicon-based approaches have been noted for their interpretability and transparency, they also face challenges when handling sarcasm, context-dependent expressions, and domain-specific jargon. Nevertheless, recent advancements in hybrid models have demonstrated that enhancing lexicon-based methods with contextual weighting can significantly improve their robustness[10].

2.4. Visualization

Visualization methods were employed to clarify the distribution of sentiment and highlight dominant lexical patterns in the data. A bar chart was used to display the proportions of texts classified as positive, negative, and neutral, providing a concise quantitative overview of sentiment prevalence[11]. Word clouds were then generated for each sentiment class to visualize the most frequent and sentiment-revealing terms [12]. These visual representations aid in qualitatively identifying prominent topics and emotional vocabulary associated with each sentiment category [13]. All visualizations were implemented in Python using Matplotlib for bar charts and the WordCloud library for word clouds, ensuring both clarity and reproducibility. The combined use of these visualization approaches allows for both quantitative assessment of sentiment distribution and qualitative insight into underlying lexical drivers, consistent with contemporary best practices in sentiment analysis research.

2.5. Model Performance

The performance of the lexicon-based sentiment classification will be evaluated using standard metrics: accuracy, precision, recall, and F1-score. The model achieved satisfactory results but showed limitations in handling complex linguistic constructs such as sarcasm, negations, and domain-specific expressions. Recent studies demonstrated that improved lexicon-based models, such as IVADER, reached high accuracy levels (98.64% accuracy, 97% precision) in multiple domains[14]. Meanwhile, hybrid approaches that combine lexicon-based methods with deep learning architectures achieved approximately 93.5% accuracy and demonstrated better robustness against contextual and nuanced language [15]. While lexicon-based methods offer high interpretability and computational efficiency, hybrid models significantly enhance performance, especially in complex textual environments.

3. RESULTS AND DISCUSSION

3.1. Data Analysis and Findings

A total of 2000 distinct customer-feedback entries were subjected to sentiment classification using the preprocessed text corpus. Customer feedback was gathered via an automated web-scraping procedure. A Python-based bot utilizing the google-play-scraper library, a robust open-source framework for google play review, was developed for this task. The script is performed in identifying review sections on each page. The harvested data were then cleaned, organized, and saved in a CSV format. A summary of the scraped products alongside their associated parameters is provided in Table 1. Those parameters were used to identify the quantitative analysis of sentiment from application users.

Table 1. Dataset parameters

Features	Type of Data	Description
Review ID	String / Alphanumeric	Unique identifier for each review
Username	String	Name or alias of the customer who posted the review
User Image	URL / String	Link to the profile picture of the reviewer
Review Content	Text	Main body of the customer's review
Score	Integer / Float	Rating value given by the customer (e.g., 1–5 stars)
Thumbs Up	Integer	Number of likes or helpful votes the review received
Review Created Version	String	Version of the app or platform when the review was submitted
Time Stamp	DateTime	Date and time the review was posted

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Features	Type of Data	Description
Reply Content	Text	Response message from the business or system
Reply Time	DateTime	Date and time when the reply was made

The compiled dataset comprises unique customer reviews, each indexed by a distinct alphanumeric Review ID and associated with a Username. These identifier attributes ensured traceability throughout the analysis pipeline, enabling us to link individual feedback entries back to their source while preserving anonymity. By maintaining a one-to-one mapping between Review ID and Username, we eliminated the risk of duplicating or mismatched records, thereby guaranteeing dataset integrity. This rigorous indexing also facilitated time series analyses of reviewer behavior, as each Review ID could be chronologically ordered via its timestamp metadata.

Focusing on the Username and Review Content, this study used reviewer identities to observe engagement patterns and relate them to customer sentiment. After text preprocessing, positive reviews mostly mentioned staff professionalism, fast service, and cleanliness, while negative reviews often pointed out issues like long queues and unclear billing. Repeat reviewers tended to give more detailed and useful feedback, helping to highlight service problems. These findings show that linking reviewer identity with review content helps us better understand customer experience and guide service improvements.

3.2. Text Preprocessing

3.2.1. Cleansing Data

The total user reviews were extracted from the Google Play Store using the google-play-scraper library to construct the dataset for analysis. Prior to modeling, comprehensive data cleansing was conducted to ensure data quality and consistency. The cleansing process involved removing duplicates, handling missing values, filtering non-English reviews, and eliminating noise such as special characters, emojis, and URLs. Furthermore, text normalization steps such as lowercasing, stop-word removal, and lemmatization were applied to standardize the review content. These preprocessing techniques are essential for reducing inconsistencies and improving the accuracy of subsequent sentiment analysis [16].

3.2.2. Case Folding

By converting all alphabetic characters in a corpus to a single case (usually lower case), case folding reduces the dimensionality of the term space and unifies tokens

that would otherwise be treated as distinct due solely to casing differences. This process improves term matching during indexing and retrieval, helping to boost recall without substantially harming precision in most applications. Recent systematic studies in case folding across user-generated and formal texts, highlighting its role in enhancing downstream NLP tasks [17]. Moreover, hybrid frameworks that integrate context-aware normalization (case folding plus neural sequence models) have demonstrated significant improvements in noisy social-media text analysis, reinforcing case folding as a simple yet powerful preprocessing step. The case folding results are written in Table 2.

Table 2. Case folding results

Column	Username	Review Content
0	ElmyLove	dan sepertinya aplikasi ini ada dan dibuat dgn...
1	Ozie	terlalu ribet seharusnya di buat simpel aja ad...
2	Cwo Cprcrn	bertahun tahun service di wahana gunung sahari...
...	...	...
1997	Iday Cha	pelayanannya bagus dan cakapmenggunakan kpb bi...
1998	Pengguna Google	logonya lebih baik yg lama dari pada yg sekara...
1999	firrean A1_09_A2_04	mekanik jual barang second fuelpump 300rb baru...

3.2.3. Tokenizing

Tokenization served as a crucial first step in processing the customer review dataset, where each review was broken down into individual word-level tokens by using standard NLP. This segmentation enabled the conversion of raw text into analyzable units, facilitating further stages such as filtering and sentiment labeling. As reflected in the output, each row in the dataset contains a user identifier and a corresponding list of tokenized words, clearly demonstrating the successful isolation of key textual components. The tokenization process effectively handled common linguistic features such as punctuation and stopwords while preserving meaningful expressions that are essential for subsequent sentiment analysis. Prior studies have emphasized that proper tokenization significantly affects the accuracy of text classification models, and subword or language-specific methods may further enhance these outcomes [18], [19].

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3.2.4. Filtering

After tokenization and stop word removal, a multi-stage filtering process was conducted to ensure the quality and relevance of the review comments. Then, very short comments, those containing fewer than 10 valid tokens after stop-word removal and stemming, were retained, in line with standard practices for retaining meaningful text in sentiment analysis [20]. Reviews lacking actual comment text were also discarded to maintain consistency in the dataset. Lastly, reviews were limited to those submitted between 2019 – 2025, ensuring alignment with the study period and reducing variation from outdated service experiences [21]. With a total of 2,000 review records, Table 3 illustrates the effectiveness of preprocessing in transforming informal, user-generated content into a clean and consistent format suitable for natural language processing tasks, comparing review content with clear content by removing unnecessary words after tokenizing.

Table 3. Tokenized vs filtering customer feedback

Column	Tokenized Review Content	Filtered Review Content
0	[dan, sepertinya, aplikasi, ini, ada, dan, dib...]	[aplikasi, dgn, fitur, sebatas, fitur, simboli...]
1	[terlalu, ribet, seharusnya, di, buat, simpel,...]	[ribet, simpel, aja, admin, pake, foto, kilome...]
2	[bertahun, tahun, service, di, wahana, gunung,...]	[bertahun, service, wahana, gunung, sahariserv...]
...	...	...
1997	[pelayanannya, bagus, dan, cakapmenggunakan, k...]	[layan, bagus, cakap, menggunakan, kpb...]
1998	[logonya, lebih, baik, yg, lama, dari, pada, y...]	[logonya, yg, yg, yg, logo, aplikasi, pencari,...]
1999	[mekanik, jual, barang, second, fuelpump, 300r...]	[mekanik, jual, barang, second, fuelpump, 300r...]

3.2.5. Stemming

To further standardize the textual data and reduce dimensionality, a stemming process was applied to the review content. Stemming involves converting inflected or derived words into their base or root form, allowing similar terms (e.g., *waiting, waited, waits*) to be treated as a single lexical item (*wait*) [22]. In this study, the Porter stemming algorithm was implemented mainly using Python's Sastrawi, Wordcloud and Vader Sentiment library, due to its simplicity and proven efficiency in sentiment and topic modeling tasks. This process not only reduced the vocabulary size but also improved the accuracy of subsequent feature extraction by minimizing semantic fragmentation caused by morphological variations. While stemming may occasionally result in non-dictionary root forms, its computational benefits for

large-scale review analysis outweigh such limitations [23]. The normalized text resulting from stemming provided a consistent basis for sentiment classification and keyword frequency analysis in later stages of the research.

3.2.6. Translating

Subsequent to the stemming phase, a translation process was implemented to convert all textual data from Bahasa Indonesia into English. This step was essential to enable the application of standardized English-language natural language processing (NLP) models and lexicons, thereby ensuring methodological consistency and enhancing cross-linguistic comparability. Translation was executed using the Google Translate API, which has demonstrated acceptable semantic preservation in previous studies involving user-generated content [24]. Prior to translation, linguistic normalization was performed to address region-specific abbreviations, idiomatic expressions, and informal constructions that could impair translation fidelity. Although automatic translation may introduce minimal semantic drift, its efficacy for large-scale textual analysis in multilingual corpora has been substantiated in recent computational linguistics literature [25]. The resulting English corpus facilitated accurate sentiment annotation, keyword extraction, and topic modeling, thereby supporting scalable and reproducible NLP workflows in multilingual research contexts.

3.3. Classification Performance

After completing the preprocessing steps, including cleansing, tokenization, stop-word removal, and stemming, the cleaned text data was subjected to lexicon-based sentiment classification, where each review was assigned a compound score to reflect its overall emotional polarity. The compound score is a normalized metric ranging from –1 to +1, calculated by aggregating individual word sentiments and accounting for the intensity of emotional expressions. Reviews with a compound score less than 0 were labeled as negative, indicating an overall unfavorable sentiment. These typically contained words or phrases associated with dissatisfaction, such as *slow, confusing,* or *error.* Furthermore, Reviews that scored exactly 0 were labeled as neutral, indicating comments that were purely informational or vague such as *OK* or *sed the app* without expressing praise or criticism. In contrast, reviews with scores above 0 were deemed positive, reflecting satisfaction and approval, often using words like *helpful, easy, fast,* or *recommended.*

The use of the compound score as a continuous measure supports effective visualization and trend analysis across time periods or user groups. The classification thresholds negative (<0), neutral (=0), and positive (>0) were adopted in alignment with previous lexicon-based sentiment research to ensure consistency and interpretability in downstream analysis. Hutto et al., (2014)

calculate the compound score in VADER sentiment analysis which is computed by first summing the valence scores of each token in each text and then normalizing this sum to fall between –1 and +1, that is given by Equation 1 [26].

$$C = \frac{\sum_{i=1}^{n} s_i}{\sqrt{(\sum_{i=1}^{n} s_i)^2 + a}} \quad (1)$$

Where:
C = compound score
s_i = denotation of the valence score of the i-th token
α = normalization constant
n = total number of tokens in the text

Building on the per review compound score C, Figure 3 visualizes these proportions, typically via a segmented bar making it immediately clear what fraction of customer feedback is positive ($p+$), neutral ($p0$), or negative ($p-$). This graphical representation enhances interpretability by showing briefly the balance of sentiment across the entire dataset.

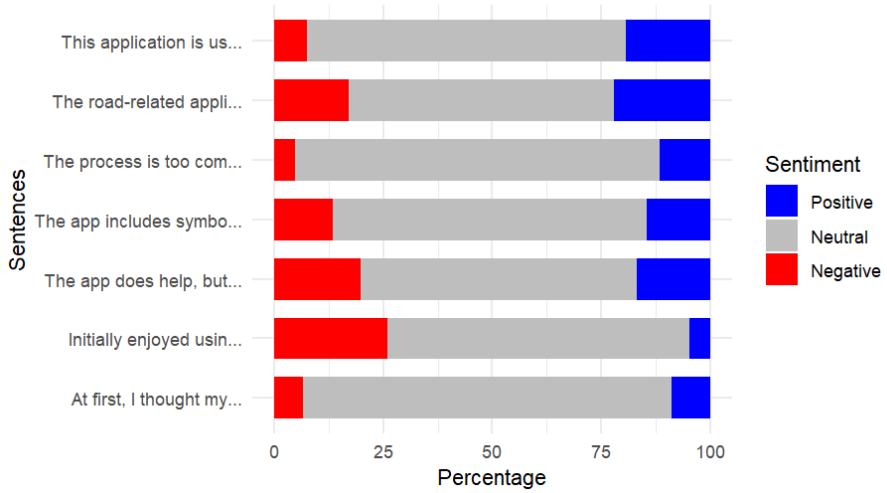

Figure 3. Basic sentiment analysis using VADER

The sentiment analysis illustrates the proportion of positive, neutral, and negative sentiment identified across a selection of user-generated sentences related to application usage. The analysis reveals that neutral sentiment predominates across all sentences, indicating that user expressions tend to be perceived as informational rather than emotionally charged. Sentences such as "*This application is useful...*" and "*The road-related application...*" exhibit a higher proportion of positive sentiment

approximately 6 – 23%, suggesting favorable user perception. In contrast, statements introducing critical elements or complexity e.g., *"The process is too complicated..."* and *"The app includes symbols..."* still maintain a largely neutral classification, with only minor shifts toward negative sentiment. Notably, compound or transitional expressions such as *"The app does help, but..."* and *"Initially enjoyed using..."* display increased negative sentiment, accounting for 6 – 26%, reflecting how contrastive or ambiguous phrasing affects sentiment polarity. These findings underscore the influence of linguistic structure and phrasing on automated sentiment analysis outcomes, with transitional cues and mixed sentiments prompting more negative interpretations despite generally neutral language.

3.4. Sentiment Analysis

The distribution of sentiment proportions shown in Figure 3 highlights two key points that motivate our deeper examination of automated sentiment analysis. First, although most user comments are categorized as neutral, subtle shifts in phrasing, particularly the introduction of contrastive conjunctions (e.g., *but* and *however*) or evaluative adjectives (e.g., *complicated* and *useful*), can meaningfully alter the balance of positive and negative labels. Second, even short fragments of feedback yield a non-trivial negative component (up to 30 %), underscoring the classifier's sensitivity to mixed or hedged expressions. These observations point to the necessity of a robust sentiment analysis pipeline capable of (a) accurately disambiguating neutral informational content from genuinely positive or negative effects, and (b) handling the linguistic cues that drive polarity shifts.

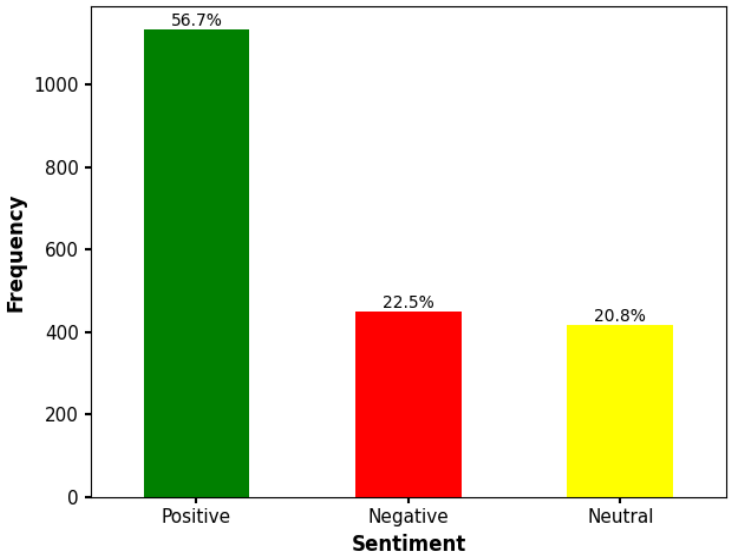

Figure 4. Distribution of sentiment counts

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Although the Figure 3 examines the proportion of sentiments on a few representative user statements, an extension of the analysis is required to the entire corpus of 2,000 review comments (see Figure 4) which results in a clear aggregate picture: approximately 56.7% of the comments are classified as positive-praising staff courtesy, prompt service delivery, and professionalism-while 20.8% remain neutral, providing factual observations without clear judgment. The remaining 22.5% fell into the negative category, mostly mentioning long waiting times, miscommunication about service status, or billing discrepancies. This broader breakdown not only confirms the largely favorable customer perception but also points out specific trouble spots that require targeted improvements.

To further elucidate the overall sentiment distribution across the full corpus of 2,000 customer review comments in Figure 4, we generated a word-cloud representation (see Figure 5) in general analysis including positive, neutral, and negative which each term's font size is proportional to its frequency in the dataset and sentiment is encoded by color: dark-shaded words denote positive valence (e.g., *admin, dealer, clear, service,* and *selling*), yellow indicates neutral descriptors (e.g., *KPB* and *fuel pump*), and green highlights negative expressions (e.g., *update, complicated,* and *features*). This visualization contributes into sharp relief the most recurrent themes: the prominence of dark words attests to generally strong perceptions of staff courtesy and operational professionalism, whereas the visibility of green terms such as *update* and *features* underscores specific service-delivery bottlenecks that warrant targeted improvement efforts. Moreover, dissect these patterns, Figure 6 presents three distinct word-clouds that arranged left to right for positive, neutral, and negative comments respectively, allowing a direct comparison of the most salient terms within each sentiment category.

A sentiment analysis-based classified word cloud visualization illustrates user feedback for a service application, categorized into positive, negative, and neutral sentiments in parts a, b, and c respectively, as shown in Figure 6. Within the positive sentiment section, words like *service*, *application*, *complimentary*, *free*, *friendly*, *good*, *easy*, and *thank* suggest that users appreciate the interface with the application and its overall utility. Terms such as *workshop*, *motorbike*, and *Honda* also indicate satisfaction with the motorbike services, particularly regarding booking and maintenance.

On the other hand, the focus on negative sentiment captures user frustrations with technical problems, ordering mistakes, and complicated procedures associated with the application through *difficult, complicated, disappointed, failed, problem*, etc. The users already suggested that they felt they had taken many actions that had not resolved problems. These points to low application reliability and poor user experience design. The neutral sentiment word cloud captures the terms Booking, process, system, using, and click. These comments tend to be neutral or objective.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Waiting, voucher, and choose articulate objective observations of the app's functions or workflow. In general, this visualization indicates neutral user comments about the application's strengths and weaknesses related to its technical performance.

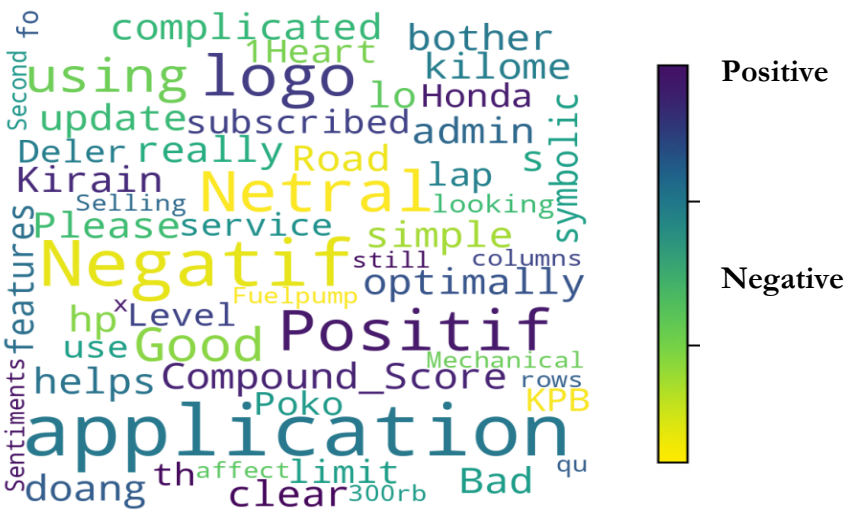

Figure 5. General wordcloud visualization

Figure 6. Classified wordcloud visualization; a) positive, b) negative, c) neutral

To assess the performance metrics for sentiment classification further, the confusion matrix as illustrated in Figure 7 shows the true versus predicted distributions for the model across three sentiments: negative, neutral, and positive. The diagonal components (537 for negative, 113 for neutral, and 871 for positive) are the true positives for each class while the rest of the elements are the false positives.

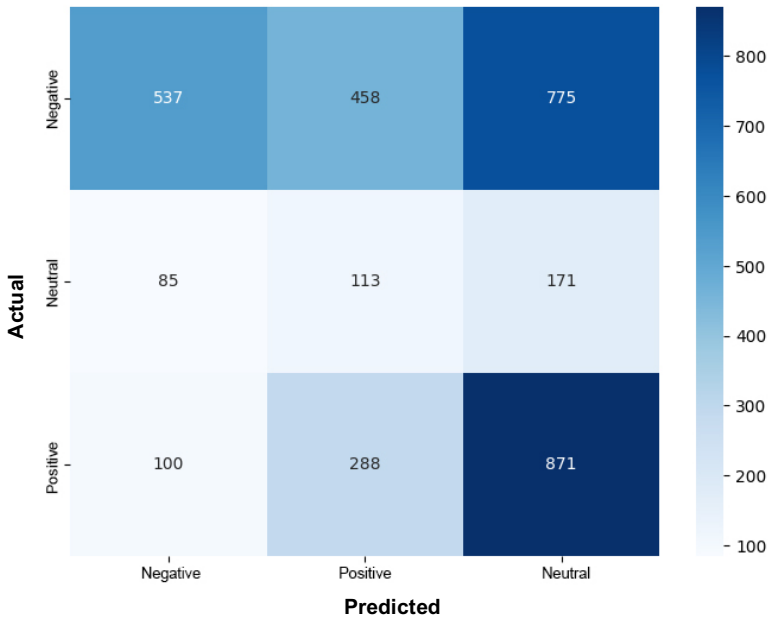

Figure 7. Confusion matrix of sentiment classification

The results on the confusion matrix also demonstrate that the model's strongest capability is in discerning positive sentiments, where 871 instances were correctly classified. This suggests that positive feedback features are easier to classify with the model and features chosen. Nevertheless, a large proportion of negative (775 instances) and neutral (171 instances) instances were wrongly classified as neutral. This suggests that the model is biased towards the neutral class. This indicates there may be shared attributes between neutral sentiment and the other two sentiments that complicate accurate classification.

Furthermore, 458 negative cases were mistakenly identified as positive, and 288 positive cases were incorrectly regarded as positive, illustrating difficulties in differentiating between minimal positive feedback and mild negative comments. This insight reveals the essence of user sentiment as a subjective phenomenon, as some terms may involve context-dependent meanings that are difficult to decipher precisely. To conclude, the model appears to efficiently capture positive feedback; however, it requires more work to improve accuracy in classifying negative and neutral sentiments, especially regarding misclassification into the neutral category.

In addition to the confusion matrix analysis, the performance metrics further reveal the model's classification capability across sentiments. Although positive sentiments demonstrate acceptable precision, the model's overall low accuracy

reflects persistent challenges in differentiating between neutral and negative sentiments, as illustrated in the confusion matrix in Figure 7. These challenges often arise from linguistic ambiguity, domain-specific terminology, and cultural variability in sentiment expression. The model achieved its highest precision for the negative class (0.74), but with a low recall (0.30), suggesting that while it confidently predicts negative labels, it often fails to capture many actual negative instances. Conversely, the neutral class showed both low precision (0.13) and recall (0.32), confirming that the model struggles considerably with neutral sentiment detection, likely due to overlapping linguistic features with the other classes. For positive sentiment, the model yielded a precision of 0.48 and a recall of 0.69, showing better performance in identifying positive feedback. The overall macro-average F1-score of 0.39 and weighted-average F1-score of 0.45 reflect the imbalanced nature of the model's performance, with stronger classification for positive sentiments while facing notable challenges in accurately distinguishing negative and neutral sentiments. This highlights the need for further improvements, such as incorporating domain-specific lexicons or more advanced machine learning approaches, to enhance sentiment classification accuracy. However, the considerable rate of misclassification for negative and neutral sentiments exposes limitations in identifying dissatisfied or ambivalent customer feedback. Such deficiencies can lead to an underestimation of emerging consumer concerns, thereby constraining organizational responsiveness in areas such as crisis management, customer retention, and service improvement [27]. To address these limitations, it is advisable to adopt hybrid sentiment analysis frameworks that integrate automated algorithms with human-in-the-loop verification processes, enhancing the validity and contextual accuracy of sentiment classifications.

The performance metrics, including precision, recall, and F1-score, further highlight the model's imbalanced capability across sentiment categories. While the model exhibits reasonable precision for positive sentiments, it continues to encounter substantial challenges in accurately distinguishing between neutral and negative sentiments. These limitations are often attributed to linguistic ambiguity, the complexity of domain-specific vocabulary, and cultural nuances in sentiment expression [28]. Addressing these challenges requires the application of domain-adaptive natural language processing models, incorporating pre-trained architectures such as BERT, BiLSTM, graph attention networks, and hybrid multi-task learning frameworks[29]. Furthermore, recent studies suggest that combining sentiment analysis with topic modeling provides deeper insights into consumer psychological dynamics and emerging market behaviors, particularly within technologically driven sectors [30]. By adopting these advanced analytical techniques, marketing practitioners can extract more reliable, nuanced, and actionable insights, ultimately enhancing decision-making in areas such as customer relationship management, product innovation, targeted marketing campaigns, and strategic brand management.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3.5. Relevance with Marketing Field

The capacity to accurately classify customer sentiment has become increasingly pivotal in contemporary marketing analytics, enabling organizations to extract actionable insights from large-scale textual data and systematically inform decisions related to branding, consumer engagement, and service development. Improving the precision of marketing constructs requires continuous refinement of measurement tools, ensuring that sentiment analysis models are sensitive to evolving customer expectations and contextual shifts. One effective approach involves embedding sentiment analysis outputs into personalized recommendation engines, where user behavior patterns, semantic knowledge graphs, and contextual analysis jointly enhance customer targeting and message personalization. Furthermore, the incorporation of trend analysis methodologies allows marketers to monitor sentiment fluctuations over time, enabling proactive adjustments to campaign messaging and product positioning based on emerging consumer sentiment patterns [31]. These strategic enhancements position sentiment analysis not merely as a diagnostic tool but as an adaptive mechanism for sustaining competitive advantage in dynamic market environments.

Building on Herlambang & Hartono (2020) findings, which demonstrate that domain-specific adversarial-augmented BERT models significantly improve sentiment classification accuracy in automotive reviews via whole-word masking and adversarial training [32]. Wahana Honda can adopt a similar methodology to refine its sentiment analysis pipeline. By training models on Indonesian-language motorcycle reviews and social media discussions around Wahana's products, the marketing team can capture nuanced sentiments such as rider comfort, engine performance, or after-sales service that generic models miss. These precise insights enable real-time campaign optimization: campaigns emphasizing newly recognized positive aspects (e.g., "*smooth gear shifts*") can be rapidly promoted, while initiatives tied to emerging negative sentiment (e.g., "*maintenance cost concerns*") can be adjusted or rephrased proactively.

According to Xue (2024) study on automotive social media analytics, there is a bidirectional influence between marketing strategies and user sentiment companies' messaging both shapes and is shaped by online sentiment, influencing competitive positioning and brand image [33]. Applying this framework, Wahana Honda should integrate continuous sentiment monitoring into its digital marketing feedback loop: sentiment trends extracted from platforms like Instagram, Facebook, and Twitter inform next-phase content, while subsequent marketing outputs (e.g. launch videos, promo deals) are immediately evaluated through the same sentiment lens. Over time, this cyclical model fosters data-driven content creation, reinforced brand trust, and more effective product positioning ensuring that Wahana Honda's digital strategy evolves in concert with rider perceptions and

market dynamics. Over successive iterations, this closedloop system will refine Wahana Honda's content strategy, elevate the relevance of its promotional activities, and reinforce a customercentric brand image, thereby driving both shortterm engagement spikes and longterm loyalty.

3.6. Discussion

To evaluate the effectiveness of the sentiment classification, the model was assessed using precision, recall, and F1-score for each sentiment category. The positive sentiment class yielded a precision of 0.48, recall of 0.69, and F1-score of 0.56. The negative sentiment class produced a higher precision (0.74) but lower recall (0.30), resulting in an F1-score of 0.43. The neutral class, meanwhile, had the lowest performance, with precision 0.13, recall 0.32, and F1-score 0.18. These metrics highlight the model's stronger performance in identifying positive feedback while also revealing challenges in accurately detecting neutral or weakly expressed sentiments. The overall macro-average F1-score was 0.39, and the weighted-average F1-score was 0.45, indicating an imbalanced classification performance.

Despite offering computational simplicity and high interpretability, the lexicon-based model has notable limitations in handling complex linguistic constructs such as sarcasm, ambiguous phrasing, or mixed sentiments. For instance, a review like 'The app was helpful... eventually' may contain a positive keyword but is framed in a sarcastic or critical context, which lexicon scores often fail to detect. This oversimplification leads to misclassification, particularly when emotionally nuanced content contradicts word-level sentiment cues. To mitigate this, future work should consider incorporating context-aware sentiment classification techniques such as BERT or BiLSTM, which better capture sentiment shifts within a sentence.

A major source of misclassification was the neutral sentiment category, which showed the lowest precision and recall. Neutral feedback is often descriptive or factual, such as 'used the app yesterday' or 'system running as expected', lacking clear emotional cues. The challenge lies in distinguishing between emotionally neutral and mildly positive or negative expressions, particularly in short or technical reviews. These findings suggest that lexicon-based methods struggle with semantic subtlety and that syntactic features, discourse cues, or domain-specific embeddings could enhance detection of neutral sentiment in future studies.

This study analyzed a cross-sectional snapshot of user sentiment; however, incorporating a temporal dimension could uncover valuable insights into how sentiment shifts over time. By leveraging metadata such as timestamps and application versions, future research could track sentiment fluctuations in response

to feature updates, bug fixes, or promotional campaigns. Such analysis would allow marketers and developers to correlate sentiment trends with product lifecycle events, leading to more data-driven decision-making in customer experience management, release strategy planning, and brand communication effectiveness.

In conclusion, while lexicon-based sentiment classification remains a powerful tool for large-scale textual analysis due to its efficiency and transparency, it is essential to recognize its limitations in detecting nuanced or context-dependent expressions. Therefore, future sentiment analysis frameworks should explore hybrid models that combine rule-based methods with deep learning architectures to enable more accurate, context-aware, and dynamic sentiment classification aligned with real-world complexities in consumer feedback.

4. CONCLUSION

This study classifies that the features available in the Wahana Honda application have provided convenience for users, this is illustrated by the following sentence "This application is useful..." , one of the features that provides this convenience is the tutorial feature, so it is necessary for Honda to be able to add a direct briefing feature like this in order to accommodate the barrier to entry of customers who use the Wahana Honda application. The research applied sentiment analysis to 2,000 Wahana Honda customer reviews, finding 56.7 % positive, 20.8 % neutral, and 22.5 % negative feedback. The evidence of generally positive user perceptions still dominates. The classification model excelled at identifying negative comments (precision 0.74) and positive comments (recall 0.69) but struggled with neutral sentiment (precision 0.13, recall 0.32), resulting in an overall macro-F1 of 0.39. To address these gaps, future work should leverage domain-adaptive NLP techniques (e.g., BERT, BiLSTM, graph attention, multi-task learning) and embed real-time sentiment monitoring. Furthermore, by integrating continuous segments into digital marketing feedback loop, this approach will reinforce a customer-centric brand image that can increase engagement and build lasting customer loyalty. Wahana Honda application can innovate continuous features to improve the user experience of application users. Such as the daily check in feature which is one of the excellent features representing good customer relationship management to ensure that customers continue to stay in touch with the application so that this feature is personalized to be more exclusive to each user.

REFERENCES

[1] N. Gupta and R. Agrawal, "Application and techniques of opinion mining," in *Hybrid Computational Intelligence*, Singapore: Springer, 2020, pp. 1–23.

[2] E. M. Mercha and H. Benbrahim, "Machine learning and deep learning for sentiment analysis across languages: A survey," *Neurocomputing*, vol. 531, pp. 195–216, 2023.
[3] J. H. Balanke and V. Haripriya, "Extension of the lexicon algorithm for sarcasm detection," in *Proc. 3rd Int. Conf. Computing Methodologies and Communication (ICCMC)*, Erode, India: IEEE, 2019, pp. 1063–1068.
[4] N. A. Sharma, A. B. M. S. Ali, and M. A. Kabir, "A review of sentiment analysis: tasks, applications, and deep learning techniques," *Int. J. Data Sci. Anal.*, pp. 1–38, 2024.
[5] K. Barik and S. Misra, "Analysis of customer reviews with an improved VADER lexicon classifier," *J. Big Data*, vol. 11, no. 1, p. 10, 2024.
[6] J. Zhu, X. Zhao, Y. Sun, S. Song, and X. Yuan, "Relational data cleaning meets artificial intelligence: A survey," *Data Sci. Eng.*, pp. 1–28, 2024.
[7] J. Khan and S. Lee, "Enhancement of text analysis using context-aware normalization of social media informal text," *Appl. Sci.*, vol. 11, no. 17, p. 8172, 2021.
[8] 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.
[9] N. Darraz, I. Karabila, A. El-Ansari, N. Alami, and M. El Mallahi, "Advancing recommendation systems with DeepMF and hybrid sentiment analysis: Deep learning and Lexicon-based integration," *Expert Syst. Appl.*, vol. 279, p. 127432, 2025.
[10] V. Bonta, N. Kumaresh, and N. Janardhan, "A comprehensive study on lexicon based approaches for sentiment analysis," *Asian J. Comput. Sci. Technol.*, vol. 8, no. S2, pp. 1–6, 2019.
[11] S. Y. Shih, F. K. Sun, and H. Y. Lee, "Temporal pattern attention for multivariate time series forecasting," *Mach. Learn.*, vol. 108, no. 8–9, pp. 1421–1441, Sep. 2019, doi: 10.1007/s10994-019-05815-0.
[12] G. Anese, M. Corazza, M. Costola, and L. Pelizzon, "Impact of public news sentiment on stock market index return and volatility," *Comput. Manag. Sci.*, vol. 20, no. 1, p. 20, 2023.
[13] H. Khalilia et al., "Crowdsourcing lexical diversity," *arXiv preprint arXiv:2410.23133*, 2024.
[14] K. Barik and S. Misra, "Analysis of customer reviews with an improved VADER lexicon classifier," *J. Big Data*, vol. 11, no. 1, p. 10, 2024.
[15] L. G. Atlas et al., "A modernized approach to sentiment analysis of product reviews using BiGRU and RNN based LSTM deep learning models," *Sci. Rep.*, vol. 15, no. 1, pp. 1–24, 2025.
[16] C. Zong, R. Xia, and J. Zhang, *Text Data Mining*, vol. 711–712, Singapore: Springer, 2021, pp. 978–981.

[17] A. A. Aliero et al., "Systematic review on text normalization techniques and its approach to non-standard words," unpublished, 2023.
[18] J. Camacho-Collados and M. T. Pilehvar, "On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis," *arXiv preprint arXiv:1707.01780*, 2017.
[19] S. Choo and W. Kim, "A study on the evaluation of tokenizer performance in natural language processing," *Appl. Artif. Intell.*, vol. 37, no. 1, p. 2175112, 2023.
[20] L. Ashbaugh and Y. Zhang, "A comparative study of sentiment analysis on customer reviews using machine learning and deep learning," *Computers*, vol. 13, no. 12, Dec. 2024, doi: 10.3390/computers13120340.
[21] J. Kim et al., "A comprehensive survey of deep learning for time series forecasting: architectural diversity and open challenges," *Artif. Intell. Rev.*, vol. 58, no. 7, Jul. 2025, doi: 10.1007/s10462-025-11223-9.
[22] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in *Proc. SDAIR-94, 3rd Annu. Symp. Document Analysis and Information Retrieval*, Ann Arbor, MI, USA, 1994, p. 14.
[23] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM J. Res. Dev.*, vol. 1, no. 4, pp. 309–317, 1957.
[24] S. V. Vadivu, P. Nagaraj, and B. S. Murugan, "Opinion mining on social media text using optimized deep belief networks," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 2024.
[25] J. Tiedemann and S. Thottingal, "OPUS-MT--Building open translation services for the World," in *Proc. Conf. Eur. Assoc. Mach. Transl.*, Lisboa, Portugal: EAMT, 2020, pp. 479–480.
[26] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, 2014, pp. 216–225.
[27] C. Graham and R. Stough, "Consumer perceptions of AI chatbots on Twitter (X) and Reddit: an analysis of social media sentiment and interactive marketing strategies," *J. Res. Interact. Mark.*, 2025.
[28] G. Xu, Z. Chen, and Z. Zhang, "Aspect category sentiment analysis based on pre-trained BiLSTM and syntax-aware graph attention network," *Sci. Rep.*, vol. 15, no. 1, p. 3333, 2025.
[29] M. M. Hossain et al., "Multi task opinion enhanced hybrid BERT model for mental health analysis," *Sci. Rep.*, vol. 15, no. 1, p. 3332, 2025.
[30] F. J. Costello and C. Kim, "Leveraging sentiment–topic analysis for understanding the psychological role of hype in emerging technologies—A case study of electric vehicles," *Behav. Sci.*, vol. 15, no. 2, p. 137, 2025.
[31] K. Y. Youssef, "Evaluating the performance of non-profit organizations using trend analysis: The future impacts of the present performance," *Arab J. Admin.*, vol. 45, no. 2, pp. 405–416, 2025, doi: 10.21608/aja.2022.131632.1229.

[32] R. T. Herlambang and D. S. H. MM, "Analysis price, perception of quality, and promotion with intervening brand trust toward purchase intention Honda Vario 150cc (case study at PT Wahana Makmur Sejati)," *Int. J. Innov. Sci. Res. Technol.*, vol. 5, no. 8, pp. 1276–1284, 2020.

[33] S. Xue, "Social media data analytics in the automotive industry: A study of the interactive impact of marketing strategies and user ratings," *Adv. Econ. Manag. Political Sci.*, vol. 78, no. 1, pp. 142–147, Apr. 2024, doi: 10.54254/2754-1169/78/20241663.