JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 325 – 332 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4476
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

SENTIMENT ANALYSIS USING MACHINE LEARNING FOR DIGITAL SERVICE DEVELOPMENT

Rugaiyah Balqis [1], Jahda Rusti Putri [1], Mira Afrina [1], Ali Ibrahim [1*], Fathoni [1]
[1]Computer Science, Sriwijaya University
email: *aliibrahim@unsri.ac.id

Abstract: The rapid growth of e-commerce mobile applications has generated large volumes of user reviews, making manual sentiment analysis increasingly impractical. This study aims to compare the effectiveness of three machine learning algorithms Support Vector Machine (SVM), Random Forest, and Naive Bayes for automated sentiment classification of Indonesian-language mobile application reviews. A dataset of 3,000 user reviews from the RupaRupa application on the Google Play Store was collected and preprocessed through normalization, tokenization, stopword removal, and stemming. TF-IDF vectorization was applied for feature extraction, while the Synthetic Minority Over-sampling Technique (SMOTE) was used to address class imbalance across three sentiment categories: positive, negative, and neutral. The results show that SVM achieved the highest accuracy of 90.02%, while Random Forest obtained the best F1-score of 88.08% when sufficient training data were available. Naive Bayes demonstrated relatively stable performance across varying training data sizes. Furthermore, TF-IDF keyword analysis revealed that negative reviews were primarily associated with delivery issues, technical problems, and pricing concerns. These findings demonstrate the effectiveness of machine learning approaches for sentiment classification and provide practical insights for improving mobile application services.

Keywords: sentiment analysis; machine learning; SMOTE; TF-IDF; text classification

Abstrak: Pertumbuhan pesat aplikasi mobile e-commerce telah menghasilkan volume ulasan pengguna yang sangat besar, sehingga analisis sentimen secara manual menjadi semakin tidak praktis. Penelitian ini bertujuan untuk membandingkan efektivitas tiga algoritma machine learning Support Vector Machine (SVM), Random Forest, dan Naive Bayes dalam melakukan klasifikasi sentimen otomatis terhadap ulasan aplikasi mobile berbahasa Indonesia. Dataset yang digunakan terdiri dari 3.000 ulasan pengguna aplikasi RupaRupa yang dikumpulkan dari Google Play Store. Data kemudian diproses melalui tahapan preprocessing yang meliputi normalisasi, tokenisasi, penghapusan stopword, dan stemming. Ekstraksi fitur dilakukan menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF), sedangkan ketidakseimbangan kelas ditangani menggunakan Synthetic Minority Over-sampling Technique (SMOTE) pada tiga kategori sentimen, yaitu positif, negatif, dan netral. Hasil penelitian menunjukkan bahwa SVM mencapai tingkat akurasi tertinggi sebesar 90,02%, sementara Random Forest memperoleh nilai F1-score terbaik sebesar 88,08% ketika tersedia data pelatihan yang memadai. Naive Bayes menunjukkan performa yang relatif stabil pada berbagai ukuran data pelatihan. Selain itu, analisis kata kunci berbasis TF-IDF mengungkapkan bahwa ulasan negatif terutama berkaitan dengan masalah pengiriman, kendala teknis aplikasi, dan isu harga. Temuan ini menunjukkan bahwa pendekatan machine learning efektif untuk klasifikasi sentimen serta memberikan wawasan yang bermanfaat dalam meningkatkan kualitas layanan aplikasi mobile.

Kata Kunci: analisis sentimen; pembelajaran mesin; SMOTE; TF-IDF; klasifikasi teks.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 325 – 332 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4476
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

The rapid growth of e-commerce mobile applications has significantly transformed consumer behavior [1], [2], generating vast quantities of user-generated reviews on platforms such as the Google Play Store. These reviews play a critical role in providing developers with actionable insights into service quality [3], technical issues [4], and overall user satisfaction [5]. As daily review volumes continue to increase, manual analysis becomes increasingly impractical, creating a pressing demand for automated sentiment classification systems [6].

Sentiment analysis has emerged as a key computational approach to address this challenge, enabling the automatic categorization of textual data into positive, negative, or neutral classes [7]. Several prior studies have explored sentiment classification and related techniques in various contexts. Madyatmadja et al. [8] applied senti ment analysis on user reviews of mutual fund applications using machine learni ng approaches, demonstrating the feasi bility of automated review classification in Indonesian financial app contexts. Tan et al. [9] conducted a compre hensive survey of sentiment analysis approaches and datasets, establishing that SVM, Naïve Bayes, and ensemble methods such as Random Forest remain strong baselines across diverse NLP tasks. Xiao et al. [10] demonstrated that TF-IDF combined with word embedd ings significantly improves text classifi cation performance on domain-specific datasets. To address class imbalance, Raveendhran and Krishnan [11] propo sed a hybrid SMOTE oversampling app roach for social media text classifi cation, showing substantial impro vement in minority class detection. Furthermore, Zhang [12] and Birann avar [13] evaluated machine learning methods including Naïve Bayes and SVM for sentiment analysis on user-generated content, with results indica ting that algorithm performance varies considerably depending on dataset characteristics and preprocessing strate gies [14], [15].

Despite this growing body of research, Indonesian-language reviews present unique and underexplored challe nges, including agglutinative morpholo gy, extensive affixation, and frequent code-switching with English in digital communication contexts [16]. Prior work on improving Indonesian text classification has highlighted that informal language and non-standard vocabulary substantially reduce classifi er accuracy when adequate preproce ssing is not applied [17]. Critically, studies that systematically compare SVM, Random Forest, and Naïve Bayes under conditions of class imbalance and varying data availability specifically within the Indonesian e-commerce review domain remain scarce. While SMOTE has been shown to be effective for imbalanced datasets in general [18], [19], its combined application with TF-IDF for Indonesian e-commerce sentiment classification has not been thoroughly investigated.

Therefore, this study aims to compare the performance of SVM, Random Forest, and Naïve Bayes for automated sentiment classification of Indonesian e-commerce application reviews from the Google Play Store, employing TF-IDF for feature extraction and SMOTE to address class imbalance, thereby contributing empirical evidence to guide algorithm

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 325 – 332 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4476
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

selection for Indonesian-language sentiment analysis tasks.

METHOD

Data Collection

Data were collected through web scraping of 3,000 customer reviews for the RupaRupa application from the Google Play Store, covering the period from 2019 to 2025. RupaRupa is a prominent Indonesian e-commerce application specializing in home furnishing, lifestyle products, and home improvement items.

This data encompassing three primary attributes: review text, numerical ratings on a scale of 1–5 stars, and submission date. This data collection approach ensured a representative sample of user feedback across different time periods, providing a robust foundation for sentiment analysis.

Table 1. Data Scrapping

Content	Score	Day	Month	Year
product quality is good	5	22	04	2019
the application is very useful for buying	1	22	04	2019
cannot be opened. it keeps saying update…	1	24	04	2019
cannot be opened yet	1	24	04	2019
easy to use with many discounts	5	24	04	2019

Data Preprocessing

Following best practices in Indonesian NLP [17], the preprocessing pipeline consisted of five sequential stages, including nrmalization, tokenization, stopword removal, stemming, and length-based filtering [8]. Text was lowercased and cleaned from non-informative elements, tokenized using NLTK, filtered using Indonesian stopwords with domain-specific additions, and stemmed with Sastrawi [20]. Tokens shorter than three characters were removed, and the cleaned text was used for TF-IDF feature extraction[10] .

Feature Extraction

Term Frequency-Inverse Docu ment Frequency (TF-IDF) was emplo yed to transform preprocessed text into numerical vectors [10]. TF-IDF assigns weights to terms based on their frequency within individual documents relative to their prevalence across the entire corpus [10].

Smote

Preliminary analysis revealed substantial class imbalance in the dataset, with positive sentiment reviews significantly outnumbering negative and neutral categories a pattern commonly observed in e-commerce review datasets. To address this imbalance and prevent model bias toward majority classes, SMOTE was applied to the TF-IDF feature matrix, resulting in balanced classes of approximately 2,050 samples each [11].

Machine Learning Algorithms

This study evaluates three machine learning algorithms to determine the most effective approach for Indonesian e-commerce sentiment classification. Support Vector Machine (SVM) with a linear kernel was selected for its efficiency in handling high-dimensional sparse text data [12]. Random Forest, an ensemble method utilizing 100 decision trees, was implemented to capture complex nonlinear feature interactions and reduce overfitting [21]. Multinomial Naive Bayes was employed as a probabilistic baseline specifically optimized for discrete word frequency features [13].

Experimental Design

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 325 – 332 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4476
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

To comprehensively evaluate algorithm performance across varying data availability scenarios, experiments were conducted using nine train-test split ratios: 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80, and 10:90. This systematic variation enabled assessment of how each algorithm scales with data availability and identification of minimum data requirements for acceptable performance, resulting in 27 distinct experimental conditions (3 algorithms × 9 split ratios). All experiments were implemented in Python using the scikit-learn library for machine learning and imbalanced-learn for SMOTE application [19].

Performance Metrics

Classification quality was quantified using Accuracy, Precision, Recall, and F1-Score. Given the multi-class nature of the sentiment data (positive, negative, and neutral), macro-averaged metrics were calculated.

RESULT AND DISCUSSION

Data Review

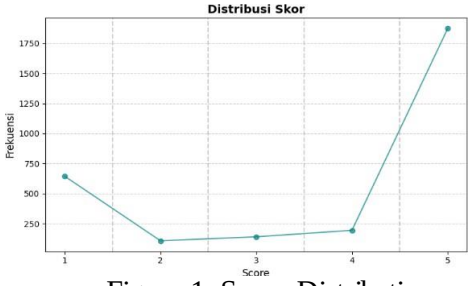

Figure 1. Score Distribution

The distribution of user scores (Figure 1) shows a polarized pattern. Score 5 dominates with more than 1,800 reviews, indicating high user satisfaction, while score 1 ranks second with around 650 reviews, reflecting substantial dissatisfaction. Intermediate scores (2–4) occur far less frequently, each below 250 reviews. This bimodal pattern suggests users tend to express extreme opinions rather than moderate evaluations, indicating strong positive or negative reactions toward the application.

Following score analysis, reviews were classified into three sentiment categories (Figure 2). Positive sentiment (score > 3) accounts for approximately 60% of reviews, negative sentiment (score < 3) represents 35.4%, and neutral sentiment (score = 3) comprises 4.6%. The results indicate that most users are satisfied, while a notable minority experience dissatisfaction. The small neutral proportion suggests users generally hold strong opinions about the application.

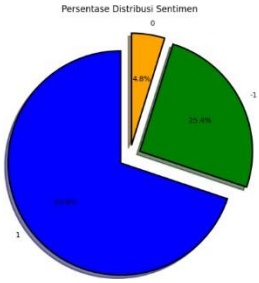

Figure 2. Sentiment Percentage

Temporal analysis (Figure 3) illustrates sentiment trends from 2019 to 2025. Positive sentiment shows steady growth, increasing from approximately 75 reviews in 2019 to over 520 in 2025, indicating rising user satisfaction and engagement. Negative sentiment fluctuates, peaking in 2024 before slightly declining in 2025. Neutral sentiment remains consistently low throughout the observed period. Overall, the increasing review volume over time suggests expanding user engagement, while persistent negative sentiment highlights the need for continuous service improvement.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 325 – 332 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4476
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

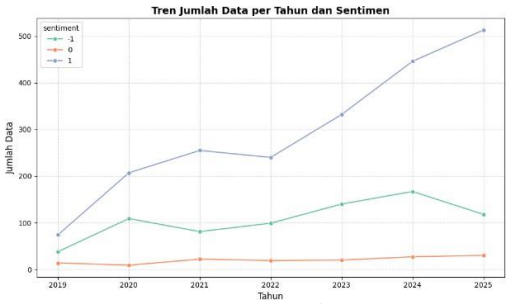

Figure 3. The evolution of user reviews

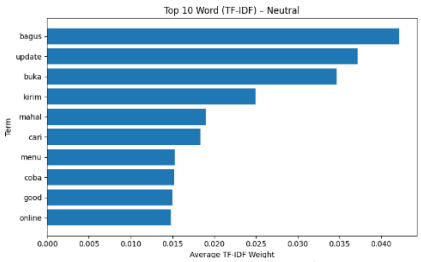

Figure 4. Top 10 Bar Chart of Neutral Sentiment Word

Data Pre-Processing

Table 2 presents a sample of the preprocessing output, demonstrating how raw reviews were transformed into cleaned tokens ready for feature extraction.

Table 2. Data Preprocessing Result

Normalization	Tokenization	Stopword	Stemming
product quality is good	[quality, product, good]	[quality, product, good]	[quality, product, good]
the application is very useful for buying	[application, useful, buy, buying]	[application, useful, buy, buying]	[application, useful, buy]
cannot be opened. it keeps saying update…	[not, opened, the message, update, ...]	[not, opened, the message, update, ...]	[not, open, message, update, ...]
cannot be opened yet	[not yet, opened]	[not yet, opened]	[not yet, open]
easy to use with many discounts	[easy, discounts]	[easy, discounts]	[easy, discount]

Feature Extraction

Analysis of the top-weighted TF-IDF terms across sentiment categories, presented in Figures 4, 5, and 6, reveals distinct lexical patterns characteristic of each sentiment class.

Neutral sentiment reviews (Figure 4) predominantly feature terms such as "bagus" (good), "update", "kirim" (send), and "mahal" (expen sive), suggesting that neutral reviewers tend to discuss operational aspects and service quality in a balanced manner without strong emotional valence.

Positive sentiment reviews (Figure 5) exhibit a distinctly different lexical profile, with dominant terms including "bagus" (good), "mantap" (excellent), "mudah" (easy), "bantu" (help), and "cepat" (fast). This vocabulary emphasizes satisfaction, usability, and service quality. The bilingual inclusion of both "bagus" and "good" reflects code-switching practices common in Indonesian digital communication.

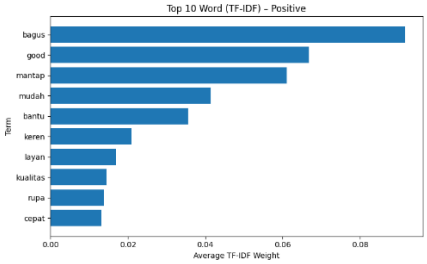

Figure 5. Top 10 Bar of Positive Sentiment

In contrast, negative sentiment reviews (Figure 6) are characterized by terms such as "kirim" (delivery), "mahal" (expensive), "loading", "kecewa" (disappointed), and "parah" (terrible). This lexicon reveals that user dissatisfaction predominantly stems from delivery issues, technical problems, and pricing concerns providing actionable insights for application improvement.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 325 – 332 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4476
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

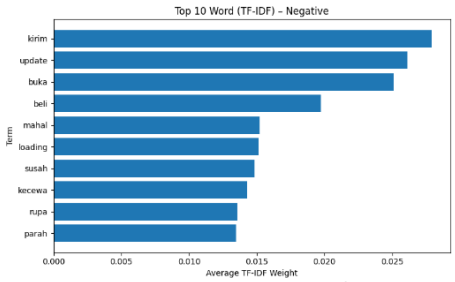

Figure 6. Top 10 Bar Chart of Negative Sentiment Word

Smote

Prior to model training, SMOTE was applied to address the substantial class imbalance identified in the dataset. As illustrated in Figure 7, the application of SMOTE successfully balanced all three sentiment classes to approximately 2,050 samples each, creating equal representation across positive, negative, and neutral categories. This balancing intervention ensures that machine learning models receive equal exposure to examples from each sentiment class during training, preventing the development of biased classifiers that might otherwise favor the majority class.

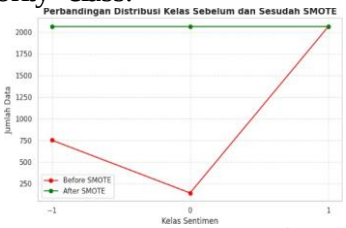

Figure 7. Class Distribution Before and After Using SMOTE

Performance Analysis Across Train-Test Split Rations

The performance evaluation phase examined three machine learning algorithms SVM, Random Forest, and Naive Bayes across nine different train-test split configurations to assess how each model scales with varying data availability.

Figure 8 presents a comparative line chart depicting accuracy trajectories across the nine split ratios for all three models. SVM achieved the highest accuracy of 90.02% at the 90:10 split, maintaining above 85% for training proportions ≥50% before declining to 75.76% at the 10:90 split. Random Forest peaked at 87.64% at the 80:20 split but degraded sharply below the 50% threshold due to insufficient data for robust tree construction. Naive Bayes achieved 83.67% at the 90:10 split with the most stable curve across all ratios (77–83%), reflecting its resilience under varying data conditions.

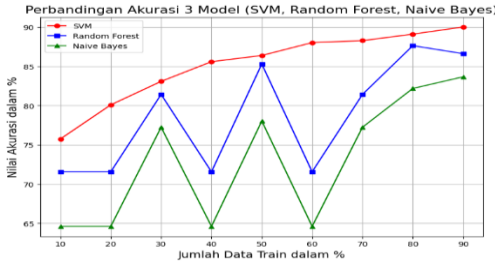

Figure 8. Comparative Line Chart of 3 Model

The accuracy heatmap (Figure 9) confirms that the 70–90% training data range is optimal across all models. Below 50%, all models degrade significantly, with Random Forest being the most affected. At ≤30% training data, all three algorithms converge to similar accuracy levels (65–75%), indicating that data acquisition should be prioritized over algorithm selection under severely data-constrained conditions.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 325 – 332 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4476
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

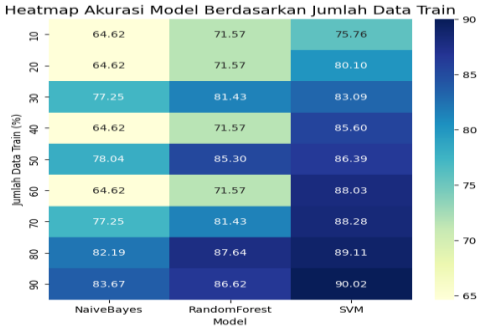

Figure 9. The Heatmap of Model Accuracy

F1-Score Comparison

Figure 10 presents the F1-score heatmap across all models and train-test split ratios. SVM maintained consistent F1-scores ranging from 82.73% at the 90:10 split to 71.16% at the 10:90 split, demonstrating reliable performance across varying data conditions. Random Forest recorded the highest F1-score of 88.08% at the 90:10 split, exceeding SVM by over 5 percentage points, but dropped sharply to 69.87% at 10% training data. Naive Bayes showed the lowest yet most stable F1-scores, ranging from 65.40% to 77.02%, making it a suitable option for data-constrained scenarios.

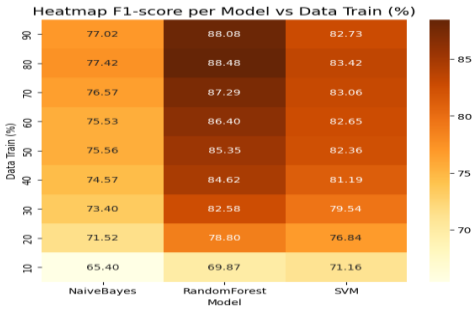

Figure 10. The Heatmap of F1-Score

Overall, these findings suggest that Random Forest is preferable when sufficient training data (80–90%) is available, while SVM offers greater reliability under suboptimal data conditions. Naive Bayes remains a viable baseline for real-time monitoring with limited resources.

CONCLUSION

This study compared SVM, Random Forest, and Naive Bayes on 3,000 Indonesian RupaRupa application reviews across nine train-test split ratios, with SMOTE for class balancing and TF-IDF for feature extraction. The dataset revealed a polarized sentiment distribution: 60% positive, 35.4% negative, and 4.6% neutral.

SVM achieved the highest accuracy of 90.02% with robust performance down to 50% training data. Random Forest recorded the highest F1-score of 88.08% but degraded sharply below 50% training data. Naive Bayes provided the most stable baseline suitable for data-constrained scenarios. The optimal training range for all models is 70–90%. Temporal analysis (2019–2025) revealed nearly sevenfold growth in positive sentiment, while TF-IDF keyword profiling identified delivery issues, technical problems, and pricing as primary drivers of negative sentiment.

From a practical standpoint, SVM is recommended when computational resources and training data are abundant, Random Forest when model interpretability is required, and Naive Bayes for real-time monitoring with limited resources. SMOTE application is strongly recommended for e-commerce platforms given their naturally skewed sentiment distributions, as it substantially improves detection of negative sentiments the most actionable category for service quality improvement.

Future work should explore BERT and LSTM architectures, aspect-based sentiment analysis, cross-domain validation, and active learning strategies to reduce annotation costs.

BIBLIOGRAPHY

[1] S. Sharma and A. Sharma, "Insights into customer engagement in a mobile app context: review and research agenda," 2024, *Cogent OA*. doi: 10.1080/233 11975.2024.2382922.

[2] J. S. Nayyar, T. Khosla, and V. K. Saini, "Trend Analysis of E Comme rce," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 5, pp. 6455–6463, May 2023, doi: 10.22214/ijraset .2023.53203.

[3] R. Moosa, "Service Quality Preferences Among Customers at Islamic Banks in South Africa," *International Journal of Professional Business Review*, vol. 8, no. 10, p. e03281, Oct. 2023, doi: 10.2 6668/businessreview/2023.v8i10.3281.

[4] S. Nauhaus, J. Luger, and S. Raisch, "Strategic Decision Making in the Digital Age: Expert Sentiment and Corporate Capital Allocation," *Journal of Management Studies*, vol. 58, no. 7, pp. 1933–1961, Nov. 2021, doi: 10.1111/joms.12742.

[5] F. B. Harlan, Y. Tarigan, S. Riadi, and A. M. Sitompul, "Analysis of E-Commerce Logistic Service Quality on Customer Satisfaction, Loyalty, and Brand Image in Indonesia," *International Review of Management and Marketing* , vol. 15, no. 1, pp. 118–127, 2025, doi: 10.32479/irmm.17503.

[6] Z. Jiang, V. Liu, and M. Erne, "Examining the Usefulness of Customer Reviews for Mobile Applications: The Role of Developer Responsiveness," *Journal of Database Management*, vol. 35, no. 1, 2024, doi: 10.4018/JDM .343543.

[7] E. S. Yusifov, "An Intelligent System for Assessing the Emotional Conno tation of Textual Statements," *Wave Electronics and its Application in Infor mation and Telecommunication Syste ms, WECONF - Conference Proceedi ngs*, 2022, doi: 10.1109/ WECONF55058.2022.9803516.

[8] E. Madyatmadja, Shinta, D. Susanti, F. Anggreani, and D. Sembiring, "Sentiment Analysis on User Reviews of Mutual Fund Applications," *Journal of Computer Science*, vol. 18, pp. 885–895, Feb. 2022, doi: 10.3844/jcs sp.2022.885.895.

[9] K. L. Tan, C. P. Lee, and K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," Applied Sciences, vol. 13, no. 7, 2023, doi: 10.3390/app13074550.

[10] L. Xiao, Q. Li, Q. Ma, J. Shen, Y. Yang, and D. Li, "Text classification algorithm of tourist attractions subcategories with modified TF-IDF and Word2Vec," *PLoS One*, vol. 19, Feb. 2024, doi: 10.1371/journal.pone.0305095.

[11] N. Raveendhran and N. Krishnan, "A novel hybrid SMOTE oversampling approach for balancing class distribu tion on social media text," Bulletin of Electrical Engineering and Informatics, vol. 14, no. 1, pp. 638–646, Feb. 2025, doi: 10.11591/eei.v14i1.8380.

[12] X. Zhang, "Performance Evaluation of Reddit Comments Using Machine Learning and Natural Language Processing Methods in Sentiment Analysis," *Mechanisms and Machine Science*, vol. 173, pp. 14–24, 2025, doi: 10.1007/978-3-031-77489-8_2.

[13] N. Birannavar, "Performance Evaluation of Sentiment Analysis on Reddit Comments: Insights and Improvement Opportunities for Naive Bayes, SVM, and BERT Models," *ICCECE 2025 - International Conference on Computer, Electrical and Communication Engineering*, 2025, doi: 10.1109/ICCECE61355.2025.1 0940395.

[14] S. Ramakrishnan, "Improving Multi-Label Emotion Classification on Imbalanced Social Media Data With BERT and Clipped Asymmetric Loss," IEEE Access, vol. 13, pp. 60589–60601, 2025, doi: 10.1109/ACCESS.2025.3557091.

[15] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," Multimed. Tools Appl., vol. 82, pp. 3713–3744, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusI D:7678100