Journal of Data Science, Informatics and Computer Engineering (JDSICE),
Vol. 1, Issue 1
February 2026
ISSN: XXXX-YYYY

Sentiment Classification and Category Mapping of Public Perception on Tax Service Quality at KPP Pratama Surabaya Gubeng Using Naive Bayes and Support Vector Machine

Puput Ayu Setiawati[1], Fitrah Maharani Humaira[2], Diah Kusumadewi Sugiarto[3]
[1]*Electronic Engineering Polytechnic Institute of Surabaya, Indonesia*
[2]*Electronic Engineering Polytechnic Institute of Surabaya, Indonesia*
[3]*KPP Pratama Surabaya Gubeng, Indonesia*
**Corresponding author: Third Author; email: puputayusetiawati2@gmail.com*

Manuscript received April 9, 2026, revised September 17, 2026, accepted September 21, 2026
doi: zz.yyyyy/xxxxx.pp134-1399

Abstract – *Public feedback on tax services provides valuable information for evaluating service quality, but the large volume of textual opinions requires systematic analysis to identify sentiment and service-related concerns. This study aims to classify public sentiment toward the services of KPP Pratama Surabaya Gubeng, map reviews into service-related categories, and compare the performance of Naive Bayes and Support Vector Machine (SVM). Data were collected from Google Maps reviews and Google Forms, resulting in 465 records. After preprocessing, 365 valid reviews remained and were mapped into three categories is service, place, and overall. Sentiment labeling produced 293 positive and 72 negative reviews. Undersampling was then applied to obtain a balanced dataset of 120 reviews, consisting of 60 positive and 60 negative samples. The balanced dataset was evaluated using 70:30, 80:20, and 90:10 training-testing ratios, corresponding to 84/36, 96/24, and 108/12 training/testing samples, respectively. The best reported performance was achieved by Naive Bayes, with 75.00% accuracy, 77.14% precision, 75.00% recall, and 75.56% F1-score, while SVM achieved 72.22% across the reported metrics. These findings indicate that Naive Bayes provides better classification performance for the characteristics of the collected review data and that category mapping provides additional information about the service aspects discussed by the public.*

CC BY SA

Keywords: *Tax, Sentiment, Naive Bayes, Support Vector Machine, Public opinion*

I. Introduction

The Tax Service Office (KPP) Pratama Surabaya Gubeng is a government institution responsible for providing taxation services to the public. The quality of tax services is important because public experiences and perceptions can provide information for evaluating service delivery. In the digital era, public feedback can be collected through online platforms, including Google Maps and online questionnaires such as Google Forms [1]. Such textual feedback can be analyzed using text mining techniques to identify information and patterns within large collections of data [2].

The increasing availability of textual feedback requires a systematic approach to process and interpret public opinions. Reviews may express different sentiments and address various aspects of service delivery. Sentiment analysis can be used to classify opinions according to their polarity [3], while category mapping can provide additional information about the aspects most frequently discussed. In tax services, this approach can distinguish opinions related to service quality, physical facilities, and overall evaluations.

Machine learning methods are widely used for text classification, including Naive Bayes and Support Vector Machine (SVM) [4]. Naive Bayes uses a probabilistic approach that is relatively simple and efficient for text classification [5], whereas SVM constructs a decision boundary and is suitable for high-dimensional feature spaces commonly found in text data [6]. Previous studies have applied Naive Bayes to public service sentiment analysis [7] and compared Naive Bayes and SVM in review classification, showing that performance may vary according to dataset characteristics [8]. Other research has highlighted the importance of preprocessing and data

balancing in sentiment classification [9]. These findings demonstrate the applicability of machine learning for digital opinion analysis, while research on taxation services in Indonesia using multi-source public feedback remains limited.

Based on this research gap, this study analyzes public perceptions of KPP Pratama Surabaya Gubeng by integrating Google Maps reviews and Google Forms responses. The data are processed through preprocessing, topic mapping, sentiment labeling, data balancing, and classification using Naive Bayes and SVM. The reviews are mapped into three categories, namely Service to "pelayanan", Place to "tempat", and Overall, to identify the aspects most frequently discussed by the public. Sentiment labeling combines numerical ratings with textual polarity analysis [10]. The performance of Naive Bayes and SVM is compared using accuracy, precision, recall, and F1-score. The contribution of this study is the integration of multi-source public opinion data with sentiment classification and category mapping in the context of tax service quality at KPP Pratama Surabaya Gubeng. The study also provides an empirical comparison of Naive Bayes and SVM on the collected review data. The findings are expected to provide a data-driven view of public perceptions and identify service-related issues that can support the evaluation and improvement of tax service quality.

II. Methodology

The research methodology consists of a series of systematic stages for collecting, processing, classifying, and evaluating public opinion regarding the tax services provided by KPP Pratama Surabaya Gubeng. The research process consists of data collection, data exploration, text preprocessing, topic mapping, sentiment labeling, data balancing, classification modeling, and model evaluation. The overall research procedure is presented in Fig. 1.

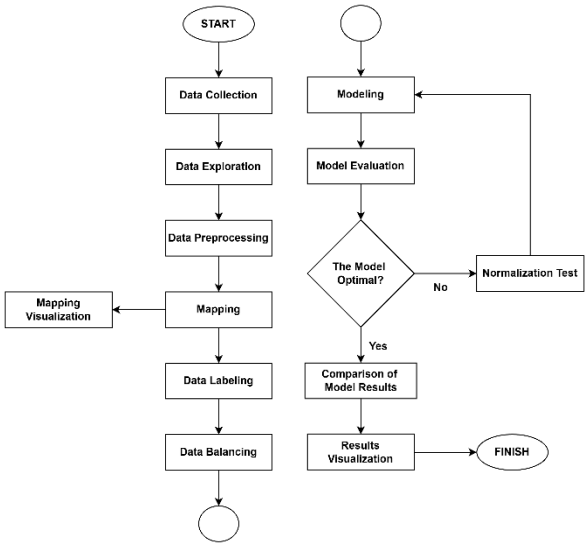


Fig. 1. Research Procedure Flowchart

A. Data Collection

The data collection stage aims to obtain public opinions regarding the services of KPP Pratama Surabaya Gubeng. Data were collected from two primary sources: user reviews on Google Maps and responses to an online questionnaire distributed through Google Forms [1], [11]. Google Maps reviews were collected using a web scraping technique [12], [13], while Google Forms was used to obtain direct responses from participants.

The Google Maps data cover reviews from 2016 to 2025, while the Google Forms questionnaire was distributed voluntarily through platform X from January 20, 2025, to June 1, 2025. A total of 465 records were collected, consisting of 311 Google Maps reviews and 154 Google Forms responses. Each record contains several attributes, including username, date, rating, and review text. These attributes were used as the primary data for subsequent analysis and classification.

TABLE I
DATA INFO

No.	Variabel	Description
1	*Username*	The name of the user who provided the review
2	*Date*	The date when the review was submitted
3	*Rating*	Numerical rating on a scale of 1 to 5 stars
4	*Review*	Textual content of the user's opinion

B. Data Exploration

The data exploration stage begins by integrating two primary data sources, namely Google Maps reviews and responses from Google Forms questionnaires, into a unified dataset. This process involves schema harmonization, including renaming columns such as 'Tanggapan' to 'Review', standardizing 'Timestamp' into 'Date', and consolidating identity-related fields into 'Username'. Irrelevant columns that do not contribute to sentiment analysis are removed to maintain data consistency and quality.

Following the integration process, an initial analysis is conducted to examine the distribution of ratings. In addition, temporal analysis is performed to observe the trend of review counts per year from 2016 to 2025, providing insights into fluctuations in data volume over time. Visualization techniques are employed to present an overview of the dataset characteristics before proceeding to further data processing stages [2].

The data exploration indicate that the dataset is highly imbalanced. The majority of reviews are dominated by 5-star ratings (68.6%), followed by 4-star ratings (15.5%). In contrast, lower ratings (1 to 3 stars) account for less than 10% of the dataset. This distribution suggests that users tend to provide feedback primarily when they are either highly satisfied or dissatisfied. The distribution of reviews

by rating and year is presented in Fig. 2 and Fig. 3, respectively.

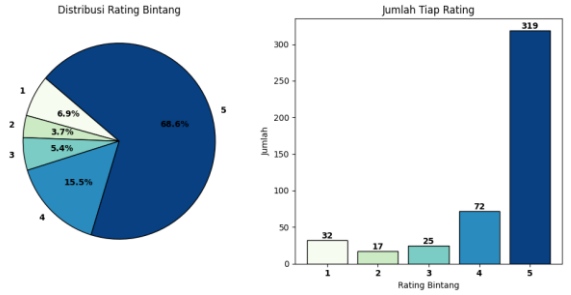


Fig. 2. Visualization of Star Rating Distribution

In addition, temporal analysis shows fluctuations in the number of reviews over time, with significant increases observed in 2022 and 2025. These trends may indicate the influence of external factors such as service improvements or policy changes.

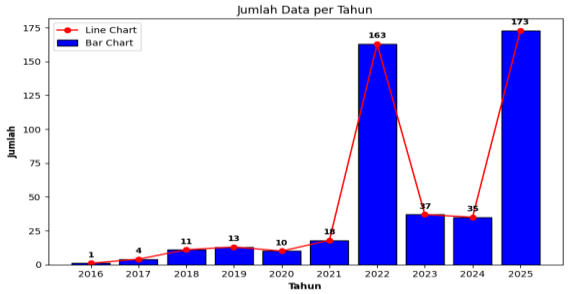


Fig. 3. Number of Review Data per Year

C. Data Preprocessing

The data preprocessing stage is conducted to clean and prepare textual data before the classification process. This step is essential to ensure that the input data are consistent, meaningful, and free from noise. Proper preprocessing can significantly improve the performance of text classification models [14], [15], especially for public opinion data, which often contain variations in writing styles. In this study, preprocessing is performed through six main steps as summarized in Table II.

TABLE II
DATA PREPROCESSING STAGES

No.	Step	Description
1	*Text Cleaning*	Removing numbers, punctuation, and irrelevant symbols.
2	*Case Folding*	Converting all text to lowercase.
3	*Normalization*	Standardizing informal or non-standard words.
4	*Tokenization*	Splitting sentences into individual tokens
5	*Stopword Removal*	Removing common words with low semantic value
6	*Stemming*	Converting words into their root forms.

This process refers to common practices applied in various text-based sentiment analysis studies in Indonesia. Proper application of preprocessing steps can significantly improve model performance in Indonesian language sentiment classification [16].

After preprocessing, 365 valid reviews remained. The reduction from 465 to 365 records was caused by the removal of empty, duplicate, or irrelevant records. The dataset was subsequently used for topic mapping and sentiment labeling. The preprocessing stage improves data quality by ensuring consistency and reducing noise, which is essential for accurate sentiment classification. An example of the preprocessing result is presented in Fig. 4.

Review	Year	Normalization	Tokenizing	Stopword	Stemming
pelayanan di kpp pratama surabaya gubeng sanga...	2025	pelayanan di kpp pratama surabaya gubeng sanga...	pelayanan, di, kpp, pratama, surabaya, gubeng,...	pelayanan, kpp, pratama, surabaya, gubeng, mem...	layan, kpp, pratama, surabaya, gubeng, muas, t...
pelayanan di kpp pratama surabaya gubeng sanga...	2025	pelayanan di kpp pratama surabaya gubeng sanga...	pelayanan, di, kpp, pratama, surabaya, gubeng,...	pelayanan, kpp, pratama, surabaya, gubeng, ram...	layan, kpp, pratama, surabaya, gubeng, ramah, ...
saya sangat mengapresiasi pelayanan yang diber...	2025	saya sangat mengapresiasi pelayanan yang diber...	saya, sangat, mengapresiasi, pelayanan, yang, ...	mengapresiasi, pelayanan, kpp, pratama, suraba...	apresiasi, layan, kpp, pratama, surabaya, gube...
saya merasa puas dengan pelayanan di kpp prata...	2025	saya merasa puas dengan pelayanan di kpp prata...	saya, merasa, puas, dengan, pelayanan, di, kpp...	puas, pelayanan, kpp, pratama, surabaya, guben...	puas, layan, kpp, pratama, surabaya, gubeng, a...

Fig. 4. Data Preprocessing Result

D. Mapping

After completing the preprocessing stage, the next step is data mapping to categorize each review based on the discussed topic. This process aims to identify the aspects most frequently highlighted in public opinions. The mapping is performed using a rule-based approach by detecting specific keywords within each preprocessed and stemmed review. This approach is suitable for relatively small datasets and enables fast and interpretable initial classification [17]. In this study, three main categories are used for topic mapping, as summarized in Table III.

TABLE III
TOPIC MAPPING CATEGORIES

No.	Variabel	Description
1	*Service*	Reviews containing keywords related to service quality or staff interaction (e.g., ‘layan’, ‘cepat’, 'ramah').
2	*Place*	Reviews referring to physical conditions or facilities (e.g., ‘nyaman’, ‘parkir’, ‘bersih’)
3	*Overall*	Reviews that do not contain keywords from the above categories and represent general opinions

E. Data Labeling

Sentiment labeling was performed to classify each review into two sentiment classes Positive and Negative. A hybrid labeling approach was applied by combining numerical rating information with textual polarity analysis using TextBlob [10].

TextBlob provides a polarity score ranging from -1 to +1, where a higher score indicates a more positive textual orientation and a lower score indicates a more negative orientation. This approach was applied particularly to reviews with neutral ratings to determine their sentiment based on the textual content.

The labeling rules applied in this study are as follows:

1. Reviews with a rating of 4 or 5 were labeled Positive.
2. Reviews with a rating of 1 or 2 were labeled Negative.
3. Reviews with a rating of 3 were analyzed using TextBlob to obtain the polarity of the review text.

F. *Data Balancing*

Class imbalance in a dataset can cause classification models to become biased toward the majority class, leading to reduced performance, especially in identifying patterns from the minority class. To address this issue, a data balancing process is applied to ensure a more balanced class distribution. In this study, an undersampling technique is used by reducing the number of samples from the majority class (Positive) to match the number of samples in the minority class (Negative). Undersampling is chosen because the initial dataset is relatively small, making it more efficient while minimizing the risk of introducing noise compared to oversampling techniques [18], [19].

G. *Modelling*

The modeling stage in this study employs two widely used classification algorithms in sentiment analysis, namely Naive Bayes and Support Vector Machine (SVM). Before modeling, the dataset is split into training and testing sets using three different ratios: 70:30, 80:20, and 90:10. Both algorithms are selected due to their effectiveness in handling text data, particularly short and direct reviews commonly found in digital opinion datasets [4].

1. Naive Bayes

   Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with the assumption of feature independence [5]. In this study, Naive Bayes is used to classify textual reviews from Google Maps and Google Forms into sentiment categories such as positive and negative. The model calculates the probability of a text belonging to a particular class based on the frequency of words appearing in the text.
2. Support Vector Machine (SVM)

   Support Vector Machine (SVM) is a machine learning algorithm widely used for classification and regression tasks. SVM works by identifying an optimal hyperplane in a high-dimensional space that separates data into different classes with maximum margin. The margin represents the distance between the hyperplane and the closest data points from each class. A larger margin indicates better generalization performance when applied to unseen data [8].

H. *Model Evaluation*

Model evaluation is a crucial stage in the development of machine learning systems, particularly in sentiment analysis. The main objective is to assess how well the model can classify data into the correct sentiment categories, such as positive and negative. In this study, model performance is evaluated using standard metrics, including accuracy, precision, recall, and F1-score. These evaluation metrics were used to compare the performance of Naive Bayes and SVM across the three training-testing ratios. The comparison provides a basis for determining the classification model that performs better on the balanced public opinion dataset [20], [21].

III. Results and Discussion

A. *Topic Mapping Results*

The topic mapping results show that the "Pelayanan" category is the most dominant, with a total of 292 reviews. This finding indicates that service quality is the primary concern for taxpayers when providing feedback. Other categories such as "Tempat" and "Overall" appear less frequently, suggesting that physical infrastructure is not the main focus of public evaluation. The dominance of the Service category indicates that service delivery is the primary aspect discussed in the collected public feedback. This result provides additional information beyond sentiment polarity because it identifies the specific aspect of the tax service that receives the most attention from the public.

The distribution of reviews across the three categories is visualized in Fig. 5.

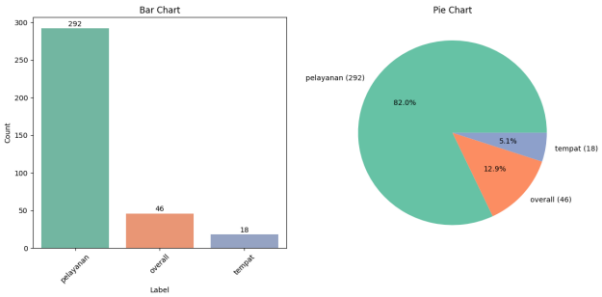


Fig. 5. Topic Mapping Chart

B. *Sentiment Labeling Results*

The sentiment labeling process reveals an imbalanced class distribution. Out of 365 reviews, 293 are classified

as positive and 72 as negative. This imbalance reflects a strong tendency toward positive sentiment in public perception of the services. The sentiment distribution is presented in Fig. 6.

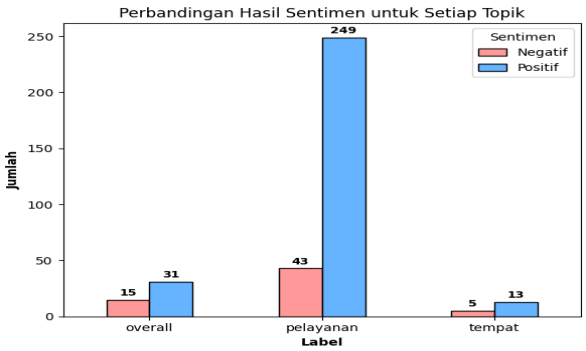


Fig. 6. Sentiment Distribution Chart

The Examples of sentiments from the collected reviews are presented in Table IV.

TABLE IV
EXAMPLE OF SENTIMENT

No.	Review	Rating	Sentiment
1	saya mengurus npwp baru dan semua prosesnya berjalan lancar petugasnya responsif dan ramah	4	Positive
2	saya agak kecewa karena harus bolak balik cuma untuk satu urusan	3	Negative

C. Data Balancing Results

To address class imbalance, an undersampling technique was applied. The majority class (Positive) is reduced to match the minority class (Negative), resulting in a balanced dataset of 120 reviews, consisting of 60 Positive and 60 Negative reviews. This balanced dataset was then used for model training and testing.

D. Model Classification Results

The performance of Naive Bayes and SVM was evaluated using accuracy, precision, recall, and F1-score. The results indicate that Naive Bayes slightly outperformed SVM across all evaluation metrics using the specified training and testing sample ratios, in Table V.

TABLE V
TRAINING AND TESTING DATA DISTRIBUTION

Ratio	Train	Test
70:30	84	36
80:20	96	24
90:10	108	12

1. Classification Using the Naive Bayes (NB) Model
The Naive Bayes model was applied to a balanced dataset and evaluated using an 80:20 train-test split. The model achieved 63.64% accuracy for the negative class and 61.54% for the positive class. To assess performance consistency, additional experiments were conducted using 70%, 80%, and 90% training splits. Among these, the 90:10 ratio yielded the best overall performance and was selected as the optimal configuration.

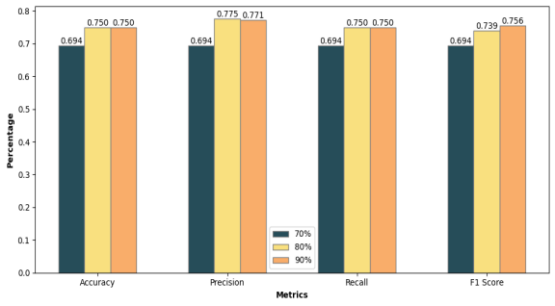


Fig. 7. Naive Bayes Model Performance Evaluation

2. Classification Using Support Vector Machine (SVM)
The second model, Support Vector Machine (SVM), followed the same processing pipeline as Naive Bayes and was evaluated on a balanced dataset using an 80:20 train-test split. The model achieved similar accuracy to Naive Bayes, with 63.64% for the negative class and 61.54% for the positive class. Further experiments with 70:30, 80:20, and 90:10 splits showed that the best performance was obtained at 70:30 with 72.22% accuracy, while performance declined at higher training proportions, indicating potential overfitting and sensitivity to data distribution.

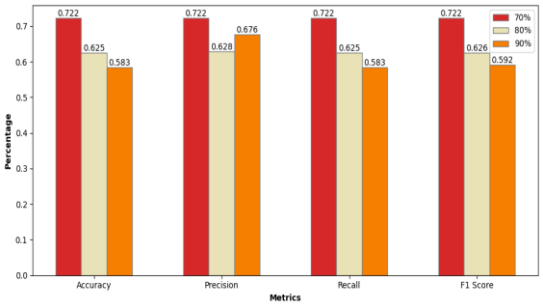


Fig. 8. SVM Model Performance Evaluation

E. Model Performance Evaluation

The performance of the two classification models, Naive Bayes and Support Vector Machine (SVM), was evaluated to compare their effectiveness in classifying sentiment. The evaluation was conducted using standard metrics, including accuracy, precision, recall, and F1-score. In addition, further analysis was performed to assess the models' ability to predict sentiment across different topics. The overall performance comparison is presented in Table VI.

TABLE VI
PERFORMANCE METRICS COMPARISON

Metric	Naive Bayes	SVM
Accuracy	0.7500	0.7222
Precision	0.7714	0.7222
Recall	0.7500	0.7222
F1-Score	0.7556	0.7222

The results indicate that Naive Bayes slightly outperforms SVM across all evaluation metrics. Naive Bayes achieves an accuracy of 75% and a precision of 77.14%, indicating that most of its positive predictions are correct. Meanwhile, SVM shows consistent performance with values of 72.22% across all metrics, demonstrating stable but slightly lower performance compared to Naive Bayes. Overall, these findings suggest that Naive Bayes is more effective for this dataset.

F. Visualization

Visualization is used to identify the main focus of public opinion regarding the services of KPP Pratama Surabaya Gubeng. One of the methods applied is a word cloud, which represents word frequency based on size. The results show that terms such as “tax”, “service”, “friendly”, and “fast” are the most dominant.

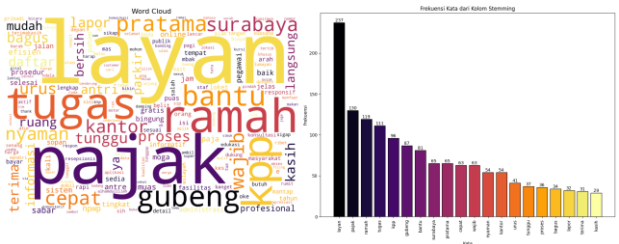


Fig. 9. Word Cloud and Word Frequency

For a more quantitative analysis, word frequency is also calculated and visualized using a bar chart. The term “service” appears most frequently (237 occurrences), followed by “tax” (130), “friendly” (119), and “staff” (111). The presence of the term “staff” indicates references to service personnel in public reviews. Overall, these visualizations confirm that service quality and interaction are the primary concerns of users. The dominance of positively connoted words such as “friendly”, “helpful”, “fast”, and “comfortable” also reflects a tendency toward positive sentiment in the dataset. The wordcloud provides an additional visualization of frequently occurring terms in the processed reviews.



Fig. 10. Word Cloud for Each Topic

IV. Conclusion

This study classified public sentiment toward the services of KPP Pratama Surabaya Gubeng using Naive Bayes and Support Vector Machine (SVM). The analysis of public feedback collected from Google Maps and Google Forms resulted in 365 valid reviews after preprocessing, consisting of 293 positive and 72 negative reviews. After undersampling, a balanced dataset of 120 reviews was obtained, consisting of 60 positive and 60 negative samples. In addition to sentiment classification, the reviews were mapped into three categories: Service, Place, and Overall. The Service category was the most frequently discussed aspect, indicating that service quality and staff interaction were prominent aspects in public feedback. This category mapping complements sentiment classification by providing information about the service aspects discussed by the public.

The classification results show that Naive Bayes outperformed SVM based on the reported evaluation metrics. Naive Bayes achieved 75.00% accuracy, 77.14% precision, 75.00% recall, and 75.56% F1-score, while SVM achieved 72.22% for the reported metrics. These results indicate that Naive Bayes was more suitable for the characteristics of the collected short-text review data.

Overall, the integration of multi-source public feedback, sentiment classification, and category mapping provides a useful approach for analyzing public perceptions of tax service quality. The findings can support the identification of service-related concerns and provide information for the evaluation and improvement of tax services at KPP Pratama Surabaya Gubeng.

Appendix

Program Repository:
https://github.com/puputayus/KP_KPP-Pratama-Surabaya-Gubeng

Acknowledgements

The authors would like to express their sincere gratitude to all parties who contributed to this research, including the respondents to the data collection and examiner for their continuous support, guidance, and encouragement throughout the research process.

References

[1] A. Ubarok, N. T. Aprilia, and S. Susanti, “Analisis kepuasan pengguna layanan Google Forms sebagai media survei online menggunakan model DeLone & McLean,” Jurnal Informatika, vol. 7, no. 2, 2020.

[2] V. Gupta and G. S. Lehal, “A survey of text mining techniques and applications,” IEEE Access, vol. 7, pp. 21687–21712, 2019.

[3] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, “Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research,” arXiv preprint arXiv:2005.00357, 2020.

[4] H. S. Bhatti, B. Raza, S. Khalid, and J. Iqbal, “A comparative analysis of machine learning algorithms for sentiment analysis using Python,” IEEE Access, vol. 9, pp. 123612–123625, 2021.

[5] “Naive Bayes classifier,” Wikipedia, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[6] S. Tripathy, A. Samal, and D. P. Dogra, “Sentiment analysis using rule-based classification and Min–Max normalization for social media text,” IEEE Trans. Comput. Social Systems, vol. 8, no. 2, pp. 315–326, 2021.

[7] A. Syafira, E. Prasetyo, and D. Lestari, “Sentiment analysis on public services using Naive Bayes classifier,” Journal of Data Science Applications, vol. 4, no. 2, pp. 101–108, 2022.

[8] R. Oktaviani, H. Widodo, and S. Anwar, “Comparative study of sentiment classification using SVM and Naive Bayes on e-commerce reviews,” Procedia Computer Science, vol. 179, pp. 876–882, 2021.

[9] N. Rachmawati and A. Santoso, “Optimizing sentiment classification with balanced datasets in public feedback analysis,” Indonesian Journal of Artificial Intelligence, vol. 6, no. 1, pp. 33–40, 2023.

[10] F. Lestari, R. Aditya, and A. Ramadhan, “Hybrid sentiment labeling based on rating and textual polarity for review classification,” Jurnal Teknologi dan Sistem Informasi, vol. 11, no. 1, pp. 45–53, 2023.

[11] “Google Maps,” Wikipedia, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Google_Maps

[12] “Web scraping,” Wikipedia, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Web_scraping

[13] A. Brenning and S. Henn, “Web scraping: A promising tool for geographic data acquisition,” arXiv preprint arXiv:2305.19893, 2023.

[14] A. F. Aufar et al., “Mengoptimalkan preprocessing teks untuk analisis sentimen yang akurat pada ulasan e-wallet,” Journal of Information and Computer Technology Education (JICTE), vol. 7, no. 2, 2023.

[15] M. Arief and M. B. M. Deris, “Text preprocessing impact for sentiment classification in product review,” in Proc. Sixth Int. Conf. Informatics and Computing (ICIC), 2021, pp. 1–6.

[16] D. P. Putra and A. Wibowo, “Text preprocessing optimization for Indonesian sentiment analysis,” Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 9, no. 1, pp. 35–42, 2022.

[17] A. Kusumawardani, D. Maulidina, and D. Fitria, “Topic classification in customer reviews using rule-based and keyword matching approach,” Jurnal Teknologi Informasi dan Komputer, vol. 7, no. 2, pp. 112–119, 2022.

[18] D. Maulidya, T. Rahman, and M. Hidayatullah, “Handling class imbalance in Indonesian sentiment analysis: A comparative study,” Jurnal Informatika dan Data Sains, vol. 7, no. 1, pp. 55–64, 2023.

[19] B. Yousefimehr et al., “Data balancing strategies: A survey of resampling and augmentation methods,” arXiv preprint arXiv:2505.13518, 2025.

[20] F. Wulandari, U. Hasanah, and B. Pramudyo, “Evaluation metrics for sentiment classification in short texts,” Jurnal Ilmu Komputer dan Aplikasi, vol. 11, no. 2, pp. 65–72, 2023.

[21] D. M. W. Powers, “Evaluation: From precision, recall and F-score to ROC, informedness, markedness & correlation,” arXiv preprint arXiv:2010.16061, 2020.