IJASST International Journal of Applied Sciences and Smart Technologies

Volume 06
Issue 01

p-ISSN 2655-8564, e-ISSN 2685-943, pp. 183-196, 2024

Sentiment Analysis on Tweets about Waste Problem in Yogyakarta using SVM

Robertus Adi Nugroho[1,*], Sri Hartati Wijono[1],
Kartono Pinaryanto[1], Ridowati Gunawan[1], F.X. Sinungharjo[2]

[1]*Faculty of Science and Technology, Sanata Dharma University, Yogyakarta, Indonesia*
[2] *Faculty of Literature, Sanata Dharma University, Yogyakarta, Indonesia*
[*]*Corresponding Author: robertus.adi@usd.ac.id*

(Received 26-10-2023; Revised 26-05-2024; Accepted 30-05-2024)

Abstract

Yogyakarta Province is facing a waste management problem. The closure of the only Integrated Waste Treatment Plant in Piyungan, Yogyakarta, has a huge impact in society life. Much waste generated from industries and houses cannot be handled appropriately until final disposal. This problem can be solved through government policies. Its effectiveness can be seen from the public response on social media. Sentiment analysis on social media, especially X, can be efficiently conducted using Support Vector Machines (SVM). Data is directly obtained from X, and text processing is performed on it. The accuracy rate of sentiment analysis using SVM on the topic of garbage in Yogyakarta is quite good at 87%. This accuracy is obtained using paramater C = 1 in SVM and k = 10 in K-Fold Cross Validation. By using these C and k values, 40% of the data is identified as positive sentiment and 60% as negative sentiment.

Keywords: Garbage, Sentiment Analysis, Support Vector Machine, Yogyakarta

1 Introduction

The world's waste problem requires serious attention. This is also the case in Yogyakarta, a province in Indonesia. Yogyakarta faces waste management problems that are quite troubling to the community. The closure of the only Integrated Waste Treatment Plant in Piyungan, Yogyakarta, has hampered the waste management process in the community. The closure of the Integrated Waste Treatment Plant in Piyungan was urged by the local community, who had been disturbed by the waste pollution caused by the Integrated Waste Treatment Plant [1]. The waste management problem in Yogyakarta can

be solved through local government policies. However, the steps taken by the government have not been effective in solving the waste problem.

Currently, social media such as Facebook, X, and Instagram are places often used by the public to give opinions or discuss a problem happening in society. The more social media users who talk about an issue causes the emergence of a trending topic that can attract the attention of parties with an interest in the issue. This can be an early warning system for parties with an interest in the issue that becomes the trending topic. If the trending topic is the waste management problem in Yogyakarta, then public opinion about it can be a warning for the local government in Yogyakarta. Hopefully, the government can immediately take steps to solve the problem.

The development of science in the field of Machine Learning, especially on Natural Language Processing, can help analyze the sentiment of opinions that develop on social media. Some research about it has been done succesfully. Diekson et al. [2] conducted a sentiment analysis of Traveloka user satisfaction with the services provided by Traveloka to its users. The study used three classification methods including: Support Vector Model (SVM), Logistic Regression, and Naïve Bayes. User satisfaction data used is taken from Twitter. Pavitha et al. [3] conducted research on a movie recommendation system which involved a sentiment analysis process on movie reviews. Sentiment analysis uses two machine learning algorithms, namely Naïve Bayes and Support Vector Machine (SVM). Wang and Zhao [4] used Support Vector Machine to predict investor sentiment towards news about stock prices. The prediction results are used by users to decide the right investment. Borg and Boldt [5] investigated sentiment analysis of Customer Support at the Swedish Telecom company. The data used in this study are emails sent to the company's Customer Support. The sentiment analysis process uses Vader for the labeling process and SVM for the classification process. Jaya Hidayat et al. [6] conducted research on public sentiment towards development on Rinca Island, Indonesia. Data is taken from Twitter. The classification algorithms used in this study are SVM and Logistic Regression. Isnan et al. [7] tried to analyze user sentiment towards the TikTok application. The criticism and response data analyzed were taken from the Google Play Store. The classification algorithm used to analyze the sentiment is SVM. Styawati et al. [8] have an interest in the phenomenon of using online transportation such as Gojek and

Grab. Some users give positive opinions and some give negative opinions. Styawati et al. tried to do sentiment analysis on the responses of these online transportation users on the Google Play Store using SVM. Chen and Zhang [9] conducted research on text sentiment analysis using CNN (Convolutional Neural Networks) and SVM. Chen and Zhang tried to improve the accuracy of sentiment analysis by combining CNN and SVM. Taufik et al. [10] used the Support Vector Machine (SVM) approach to analyze the sentiment towards public figures on Twitter. Saputri et al. [11] researched the classification of Borobudur Temple tourist sentiment on the TripAdvisor site using SVM and K-Nearest Neighbor (KNN).

From these studies, it can be seen that the use of machine learning approach was appropriate to be applied in sentiment analysis, especially in classification process. The reliability of machine learning in performing classification is not only in sentiment analysis but also in other cases. Yadav et al. [12] tried to classify traffic sign images that appear in a video using SVM . The classification results would be used to generate sounds that match the traffic sign. Then the sound would be used to remind the driver. Kumalasanti [13] also used SVM to classify a person's handwriting with his personality. Therefore, this research tried to conduct sentiment analysis for the topic of waste management in Yogyakarta using machine learning approach so that the polarization of public opinion on waste management can be known. The algorithm that was used in this research was Support Vector Machine (SVM).

2 Material and Methods

The purpose of sentiment analysis of a document or sentence is to determine the polarity of the document or sentence. The three types of polarity are positive, negative, or neutral [14]. According to [15], sentiment analysis methods are categorized into three types, namely:

a) Lexicon Based Method
b) Machine learning Based Method
c) Hybrid Method

The lexicon-based method uses a word sentiment dictionary to determine the sentiment of a document or sentence. Machine learning methods use datasets with sentiment to determine the sentiment of new sentences or documents. The hybrid method will combine the two methods above.

Support Vector Machine (SVM) is one of the popular algorithms in machine learning. Based on the research conducted by Diekson et al. [2] , Pavitha et al. [3], Wang and Zhao [4], Borg and Boldt [5], Jaya Hidayat et al. [6], Isnan et al. [7], Styawati et al. [8], Chen and Zhang [9], Taufik et al. [10], and Saputri et al. [11], the use of SVM algorithm in performing sentiment analysis gave good results. In fact, when compared to several other algorithms, the use of SVM gave better results. From the research conducted by Diekson et al. [2], based on F1 score evaluation, SVM gave better results than Logistic Regression and Naïve Bayes in classifying user satisfaction about Traveloka services. Pavitha et al. [3] concluded the accuracy score of SVM is better than Naïve Bayes in classifying movie reviews. Then, Wang and Zhao [4] successfully predict the investor sentiment from the news. The accuracy reached 59%. Borg and Boldt [5] showed us that SVM could predict the sentiment of customer email succesfully with the mean F1 score of 0.688. Jaya Hidayat et al. [6] used SVM and Logistic Regression to analyse the sentiment of citizen's opinion about Rinca Island development. The result was good, the accuracy rate of SVM was about 86% and Logistic Regression was about 75%. Isnan et al. [7] and Styawati [8] gave the same conclussion that the accuracy of the sentiment analysis by using SVM was above 80%. Chen and Zhang [9] got the result that using CNN combined with SVM could improve the accuracy of the sentiment classification. Taufik et al. [10] tried to use SVM in classifying the sentiment to the public figure. They got the accuracy was about 80%. Saputri et al. [11] compared SVM and KNN in classifying the sentiment of tourist opinion abot Borobudur Temple in TripAdvisor. They got the result that SVM gave a better accuracy than KNN. From these researches, the use of SVM to analyze sentiment is very appropriate. Therefore, in this research, SVM was used to analyze sentiment of public opinion in X about waste problem in Yogayakarta.

Next, we will explain the steps for sentiment analysis regarding the waste problem in Yogyakarta. In the first stage, this research collected data from X and saved it into a CSV file. The second stage is preprocessing the raw data. The process is carried out to

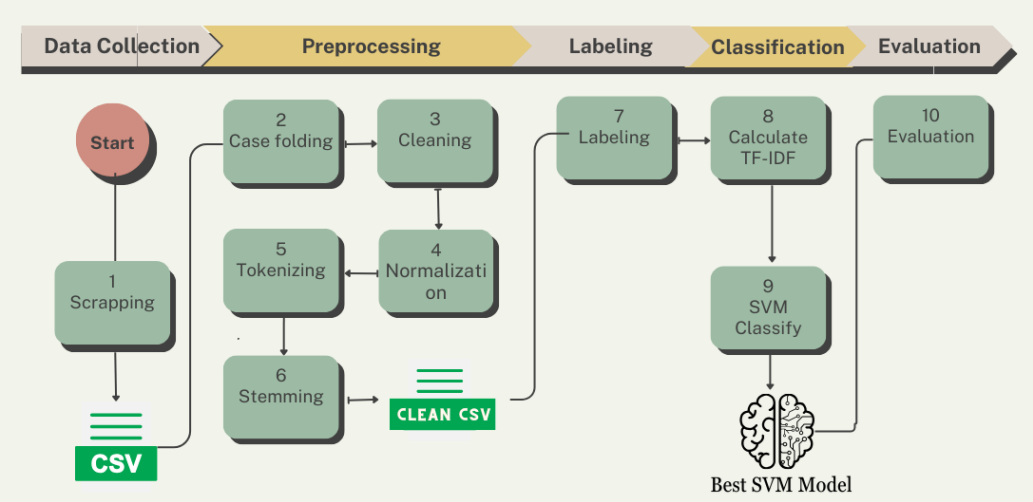


Figure 1. Process diagram of this research

clean the data of unnecessary characters. After the data is clean, the next step is labeling positive or negative sentiment. The fourth stage is the classification process using the Support Vector Machine (SVM) method. After the classification process, the best SVM model will be obtained. The best SVM model is used for the evaluation process. This step is shown in the Fig 1.

2.1. Data Collection

The data used is a collection of Indonesian language reviews on X social media that discuss the topic of waste. The hashtags used are sampah, sampahyogya, sampahpiyungan, bakarsampahyogya. Data was gathered by web scraping with the *tweepy*[1] library from Python. The data collection period spanned from September 2023 to October 2023. The web scraping yielded 1211 tweets. The tweet results were selected by removing duplicate and irrelevant tweets. This selection process produces 1000 tweets, which are used for the next step. The web scrapping results contain tweet data, user data, URL address, retrieval time, and other data. The data used is only tweet data for further processing.

[1] tweepy.org

2.2. Data Preprocessing

This step aims to remove unnecessary characters that would affect the classification process. The first step is case folding, namely making all the letters lowercase. Next is to remove unnecessary characters such as @, #, $, ?, !, and others. Apart from that, unnecessary tagging was also removed. The only characters used are letters, so other than that, they will not be used. This process is carried out using the *re* library from Python.

The next step is the normalization, which is conducted to correct incorrect spellings of words. Another goal is to restore original words from abbreviations or slang words commonly used in tweets. This process requires checking Kamus Besar Bahasa Indonesia[2] (KBBI) and the Alay dictionary[3]. After the tweet does not contain abbreviations and slang words, the following process is tokenizing the tweet using the *nltk* library from Python. The results of this stage are in the form of a list of tokens.

The final preprocessing step is the lemmatization or stemming process, namely removing affixes to get the stem word. The stemming process is conducted using the Sastrawi library[4].

We do not carry out a stopword removal process because it can lose meaning. For example, " Dia tidak suka membuang sampah" (He does not like throwing rubbish). The sentence's sentiment will change if the word "tidak" (not) becomes a stopword.

2.3. Labeling

Labeling was done manually by two annotators. One of the annotators is a linguist. Linguists train other annotators to determine the polarity of tweets. Every tweet annotated by one annotator will be double-checked by a linguist so that there is no bias in determining the polarity of the tweet.

2.4. Classification

Before classification, term frequency (TF) and index document frequency (IDF) are calculated. TF-IDF weight evaluates how important a word is in a sentence. Term

[2] https://kbbi.kemdikbud.go.id/Beranda
[3] https://github.com/fendiirfan/Kamus-Alay
[4] https://github.com/sastrawi/sastrawi

Frequency (TF) is the opposite of IDF (eq. 1); the higher the frequency of occurrence of a term in a document, the higher the weight of the term itself. IDF is the opposite of TF; the higher the frequency of occurrence of a term, the lower the weight of the term itself. TF-IDF calculation is done using Python program code.

$$IDF(t) = \log \frac{N}{df(t)} \quad (1)$$

IDF(t) = IDF of the word t

N = total sentence

df = number of documents containing the word t

The method used for the sentiment-based X classification process uses SVM. SVM attempts to identify the best hyperplane to separate large data into two optimal classes. For example, if we want to divide two-dimensional space, we need a one-dimensional hyperplane, namely a line. SVM will classify data using a hyperplane that maximizes the boundaries between classes in the training data. First of all, SVM tries to divide the data into two dimensions. If the data classes formed cannot divide the data linearly, the algorithm will change the margin until the hyperplane can separate the data into its classes linearly. Margin is the distance between the hyperplane and each class's closest member (support vector). The following (eq. 2) is the hyperplane calculation formula:

$$f(x) = w.x + b \quad (2)$$

w = the chosen hyperplane parameter

x = data input x

b= the hyperplane parameter that was selected

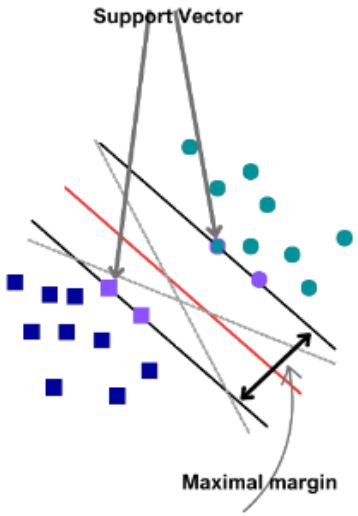


Figure 2. Support Vector Machine

SVM tries to maximize the margin in order to find the right hyperplane to divide classes and minimize classification errors (see Fig 2). So, the two interests contradict each other. If we increase the margin, we will obtain a higher misclassification rate; if we decrease the margin, we will get a lower misclassification rate. Parameter C is the answer to control between maximizing margin and minimizing error. In this research, experiments were conducted to find the C value that best suited our data [16].

SVM increased in size when used for classification with many classes, and SVM was theoretically developed to solve classification problems with two classes. This research uses linear kernels for SVM classification. Program implementation uses the Sklearn library from Python.

2.5. Evaluation

This research evaluates a sentiment analysis system using recall, precision, F1 and accuracy. Recall, precision, F1, and accuracy are used to assess the accuracy of the results. Where recall (True Positive Rate) (eq.3) is the proportion of true positive data identified versus all true positive data. Precision (Positive Predictive Value) (eq.4) is the proportion of correct identifications of all identified positive findings. F1 score (eq. 6) is the sum of precision and recall.Accuracy (eq.5) is the correct identifications (positive and negative)

ratio to the total data. Evaluation was conducted on separated data using k fold cross-validation. Testing with different *k* values will produce an SVM model. This research will evaluate this SVM model.

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

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

$$F1 = 2 * \frac{precision + recall}{precision * recall} \tag{6}$$

3 Results and Discussions

Several experiments have been conducted on the sentiment analysis approach proposed in this research. Experiments were conducted for several types of C to determine the most optimal C parameter values in SVM. Experiments were conducted on data that was divided using k-fold cross-validation for each type of C.

From several experiments with a combination of parameters C and k, the best combination will be obtained in carrying out sentiment analysis from the prepared tweet data set. Some of the C values used include C=1, C=10, and C=100. At the same time, the value of k in k-Fold Cross Validation includes k = 2, 3, 4, 5, 6, 7, 8, 9, and 10.

In the first experiment, parameter C is set to 1. Table 1 shows that the best accuracy is 87%. It is got when the k-value equals 10.

Table 1. Experiment Result with C=1

k	Accuracy	Precision	Recall	F1 Score
2	77.60%	77%	78%	77%
3	81.98%	82%	82%	82%
4	80.40%	80%	80%	80%
5	84.50%	85%	84%	84%

6	82.23%	84%	83%	83%
7	83.21%	83%	83%	83%
8	86.40%	87%	86%	86%
9	84.68%	85%	85%	85%
10	87.00%	88%	87%	87%

Table 2. Experiment Result with C=10

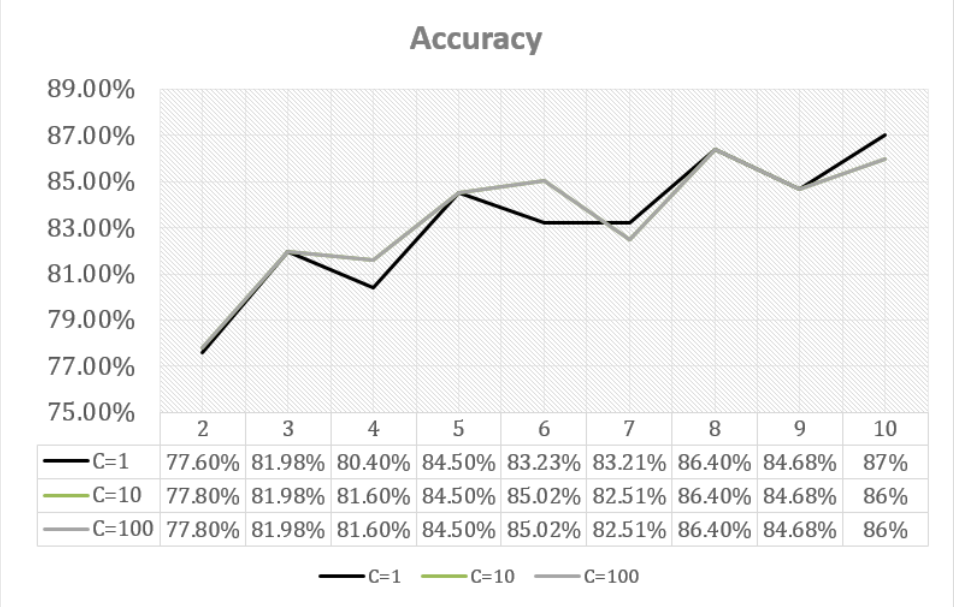
K	Accuracy	Precision	Recall	F1 Score
2	77.80%	78%	78%	78%
3	81.98%	82%	82%	82%
4	81.60%	83%	82%	81%
5	84.50%	85%	84%	84%
6	85.02%	85%	85%	85%
7	82.51%	83%	83%	82%
8	86.40%	86%	86%	86%
9	84.68%	85%	85%	85%
10	86.00%	86%	86%	86%

In the second experiment, parameter C is set to 10. Table 2 show that the best accuracy occurs when k-value is 8, which is 86.40%. In the third experiment, parameter C is set to 100. As shown in Table 3, the best accuracy is 86.40%. It occurs when the k-value equals to 8.

Table 3. Experiment Result with C=100

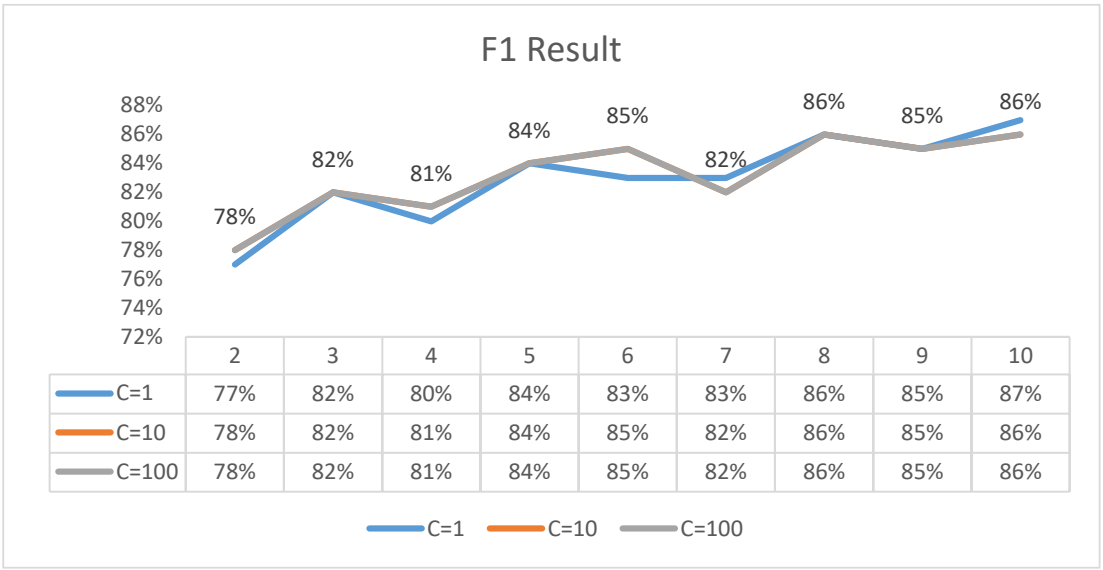
K	Accuracy	Precision	Recall	F1 Score
2	77.80%	78%	78%	78%
3	81.98%	82%	82%	82%
4	81.60%	83%	82%	81%
5	84.50%	85%	84%	84%
6	85.02%	85%	85%	85%
7	82.51%	83%	82%	82%
8	86.40%	86%	86%	86%
9	84.68%	85%	85%	85%
10	86%	86%	86%	86%

From those experiments, it can be seen that C parameters influence the accuracy of the sentiment analysis result, although it is not significant. Moreover, in C=10 and C=100, the experiment result shows no differences for all evaluation aspects, such as accuracy, precision, recall, and F1 score. It means that C parameters could not improve the prediction results anymore. (see Figs 3 and 4).



	2	3	4	5	6	7	8	9	10
C=1	77.60%	81.98%	80.40%	84.50%	83.23%	83.21%	86.40%	84.68%	87%
C=10	77.80%	81.98%	81.60%	84.50%	85.02%	82.51%	86.40%	84.68%	86%
C=100	77.80%	81.98%	81.60%	84.50%	85.02%	82.51%	86.40%	84.68%	86%

Figure 3. Accuracy comparison



	2	3	4	5	6	7	8	9	10
C=1	77%	82%	80%	84%	83%	83%	86%	85%	87%
C=10	78%	82%	81%	84%	85%	82%	86%	85%	86%
C=100	78%	82%	81%	84%	85%	82%	86%	85%	86%

Figure 4. F1 comparison

As seen in figure 3 and figure 4, the best sentiment analysis result is obtained using C=1 and k=10. By this configuration, the proposed method could predict 397 tweets as positive and 603 tweets as negative tweets. It means that the sentiment of the X data about waste management in Yogyakarta is 40% positive and 60% negative.

From these results it can be seen that as many as 60% of tweets from data taken from X regarding waste management in Yogyakarta are categorized as tweets with negative sentiment. This information can illustrate that the handling of waste in Yogyakarta has not fully received positive appreciation from the public. The majority of people still give negative sentiments. The government can utilize this information to follow up by evaluating policies on waste management in Yogyakarta.

4 Conclusions

From the experiments that have been carried out, it can be concluded that the sentiment analysis approach using SVM can be applied to analyze sentiment regarding waste management in Yogyakarta province. Using parameters C = 1 and k = 10 is the best combination for sentiment analysis, which can produce an accuracy of 87%. From the configuration values of C and k, it is found that X data about waste management in Yogyakarta province has 40% positive sentiment and 60% negative sentiment.

Acknowledgements

This research was funded by LPPM Sanata Dharma University in the framework of a Research Assignment with the theme Universal Apostolic Preferences (UAP).

References

[1] M. Fakhruddin, “Yogyakarta Darurat Sampah,” *Republika*, May 13, 2022. [Online]. Available: https://news.republika.co.id/berita/rbq3q8318/yogyakarta-darurat-sampah

[2] Z. A. Diekson, M. R. B. Prakoso, M. S. Q. Putra, M. S. A. F. Syaputra, S. Achmad, and R. Sutoyo, “Sentiment Analysis for Customer Review: Case Study of

Traveloka," *Procedia Computer Science*, vol. 216, pp. 682–690, 2023, doi: 10.1016/j.procs.2022.12.184.

[3] N. Pavitha *et al.*, "Movie Recommendation and Sentiment Analysis Using Machine Learning," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 279–284, Jun. 2022, doi: 10.1016/j.gltp.2022.03.012.

[4] D. Wang and Y. Zhao, "Using News to Predict Investor Sentiment: Based on SVM Model," *Procedia Computer Science*, vol. 174, pp. 191–199, 2020, doi: 10.1016/j.procs.2020.06.074.

[5] A. Borg and M. Boldt, "Using VADER Sentiment and SVM for Predicting Customer Response Sentiment," *Expert Systems with Applications*, vol. 162, p. 113746, Dec. 2020, doi: 10.1016/j.eswa.2020.113746.

[6] T. H. Jaya Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, "Sentiment Analysis of Twitter Data Related to Rinca Island Development Using Doc2Vec and SVM and Logistic Regression as Classifier," *Procedia Computer Science*, vol. 197, pp. 660–667, 2022, doi: 10.1016/j.procs.2021.12.187.

[7] M. Isnan, G. N. Elwirehardja, and B. Pardamean, "Sentiment Analysis for TikTok Review Using VADER Sentiment and SVM Model," *Procedia Computer Science*, vol. 227, pp. 168–175, 2023, doi: 10.1016/j.procs.2023.10.514.

[8] S. Styawati, A. Nurkholis, A. A. Aldino, S. Samsugi, E. Suryati, and R. P. Cahyono, "Sentiment Analysis on Online Transportation Reviews Using Word2Vec Text Embedding Model Feature Extraction and Support Vector Machine (SVM) Algorithm," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, Jakarta, Indonesia: IEEE, Jan. 2022, pp. 163–167. doi: 10.1109/ISMODE53584.2022.9742906.

[9] Y. Chen and Z. Zhang, "Research on Text Sentiment Analysis Based on CNNs and SVM," in *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, Wuhan: IEEE, May 2018, pp. 2731–2734. doi: 10.1109/ICIEA.2018.8398173.

[10] I. Taufik and S. A. Pamungkas, "Analisis Sentimen Terhadap Tokoh Publik Menggunakan Algoritma Support Vector Machine (SVM)," *Jurnal Logika*, vol. 8, no. 1, 2018.

[11] R. P. Saputri, W. S. Winahju, and K. Fithriasari, "Klasifikasi Sentimen Wisatawan Candi Borobudur pada Situs TripAdvisor Menggunakan Support Vector Machine dan K-Nearest Neighbor," *Jurnal Sains dan Seni Institut Teknologi Sepuluh November*, vol. 8, no. 2, 2019, doi: 10.12962/j23373520.v8i2.44391.

[12] S. Yadav, A. Patwa, S. Rane, and C. Narvekar, "Indian Traffic Signboard Recognition and Driver Alert System Using Machine Learning," *Int.J.Appl.Sci.Smart Technol.*, vol. 1, no. 1, pp. 1–10, Jun. 2019, doi: 10.24071/ijasst.v1i1.1843.

[13] R. A. Kumalasanti, "Design of Someone's Character Identification Based on Handwriting Patterns Using Support Vector Machine," *Int.J.Appl.Sci.Smart Technol.*, vol. 4, no. 2, pp. 233–240, Dec. 2022, doi: 10.24071/ijasst.v4i2.5417.

[14] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1–2, pp. 1–135, Jan. 2008, doi: 10.1561/1500000011.

[15] M. Rathan, V. R. Hulipalled, P. Murugeshwari, and H. M. Sushmitha, "Every Post Matters: A Survey on Applications of Sentiment Analysis in Social Media," *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pp. 709–714, 2017.

[16] C. Campbell and Y. Ying, *Learning with Support Vector Machines*, vol. 5. 2011. doi: 10.2200/S00324ED1V01Y201102AIM010.