Journal of Information System Exploration and Research Vol. 2, No. 1, January 2024, pp. 49-60

SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Information System Exploration and Research

https://shmpublisher.com/index.php/joiser

p-ISSN 2964-1160 | e-ISSN 2963-6361

Classification of Hate Comments on Twitter Using a Combination of Logistic Regression and Support Vector Machine Algorithm

Nabila Putri Damayanti[1*], Della Egyta Prameswari[2], Putri Susi Sundari[3], Wiyanda Puspita[4]

[1,2,3,4] *Department of Computer Science, Universitas Negeri Semarang, Indonesia*

DOI: https://doi.org/10.52465/joiser.v2i1.229

Received 22 October 2023; Accepted 27 January 2024; Available online 29 January 2024

Article Info

Keywords:
Hate speech;
Classification;
Logistic regression;
Support vector machine

Abstract

This research was conducted to increase accuracy in classifying sentences containing hate speech and non-hate speech on Twitter. This is important to do because, as technology develops, it also comes with negative impacts, one of which is hate speech. This classification is carried out using a combination of Logistic Regression (LR) and Support Vector Machine (SVM) methods. This combination is based on the ease of implementation and speed of LR as well as SVM's ability to handle more complex and non-linear data. In this context, LR is used to model the probability that a comment on Twitter contains hate elements or not. The model can then provide probability predictions for each class, and a threshold can be set to determine the final class. This research shows that combining these methods can build a good classification model with an accuracy of 96%.

1. Introduction

Technological developments nowadays are very rapid. Everything humans do can be easily assisted by technology [1]‑[4]. Apart from that, the dissemination of information also occurs in just a matter of minutes. Technology itself can have positive and negative impacts, depending on how each individual uses it [5]‑[7]. The positive influence is that we can quickly find out what is trending at the moment as a means of business introduction or cultural preservation. However, behind these positive influences, there are, of course, negative influences that we cannot prevent. When someone uploads or expresses themselves on social media, not everyone sees it as a positive thing. There may be negative comments from other people. These negative comments can trigger pros and cons. There is significant global concern regarding the phenomenon of hate speech. The international community is increasingly concerned about types of communication that are considered expressions of hatred towards

* ***Corresponding Author:***

Nabila Putri Damayanti,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Central Java 50229, Indonesia.
Email: nabilaptrd@student.unnes.ac.id

individuals or groups based on social characteristics such as race, ethnicity, gender, religion, sexual orientation, and other attributes that are an integral part of being human [8]–[10].

According to a PEW Research Institute survey in 2014, 73% of adult internet users had seen someone being harassed online, 40% of internet users had personally experienced online harassment, and 45% of them had experienced severe harassment [11]. Sometimes extreme cases of cyberbullying even lead the victim to commit suicide. Social media platforms are particularly interested in determining online abuse by users who report such abuse [11], [12]. The spread of the phenomenon of hate speech and disinformation online has the potential to disrupt democratic debate and practice and has the potential to facilitate serious human rights violations. In addition, this can strengthen the exclusion of minority groups. Even in Ethiopia, despite having one of the lowest internet connectivity penetration rates on the continent, it cannot be denied that this phenomenon is still felt [10], [13]–[15]. The Nigeria Stability and Reconciliation Program describes hate speech as a catalyst for violence, considering it a degrading form of communication The program notes that the content of hate speech "has the potential to create a destructive cycle as viewers surround it while acting as an alternative source of information that dilutes positive information [16].

Based on these problems, it is necessary to classify positive and negative comments. This classification aims to ensure that words that contain rude, hurtful, or detrimental elements can be detected using machine learning. Machine learning has penetrated various fields today, be it science, education, finance, or business [17]–[20]. Machine learning has found its application everywhere. However, machine learning is only limited to "machine learning" and has not even realized its full potential. Machine learning is the ability of a computer to learn the relationship between input and output without being explicitly programmed [21]. That way, when using machine-learning-based machines, algorithms need to be assisted. Algorithms that are often used for classification are Logistic Regression (LR) and Support Vector Machine (SVM) [22]–[25].

Logistic Regression (LR) is an algorithm in machine learning that is used for binary classification purposes [26]. Binary classification refers to the predictions of two different classes, for example, positive and negative. In the framework of detecting hate speech, LR can be used to form a model that can identify whether a text is included in the hate speech category or not. One of the advantages of LR is the ease and speed of implementation, especially in the case of binary classification [22]. The Support Vector Machine (SVM) is also an algorithm in machine learning that is useful for classification purposes. SVM is an algorithm that aims to find the separator with the maximum margin, which means finding the line (hyperplane) that separates two classes with the largest margin. The specialty of SVM lies in its capacity to handle data that has complex and non-linear features [23].

In previous studies, some studies discussed the classification of hate speech. [11] conducted research using collaboration between the BiLSTM and CNN algorithms, CNN-LSTM, GRU, and LSTM. The collaboration of the BiLSTM and CNN algorithms obtained 87% accuracy, CNN-LSTM obtained 85% accuracy, GRU produced 86% accuracy, and LSTM produced 82% accuracy. This research shows that the collaboration of BiLSTM and CNN produces the highest accuracy. However, the results are still below 90%. Therefore, efforts need to be made to increase accuracy to the maximum. In this research, efforts were made to increase accuracy by combining the LR and SVM algorithms.

Research on the same topic was conducted by [27] using the multinomial logistic regression method. The weakness of this research is the lack of exploration of the characteristics of the method used. This research obtained an accuracy of 84%. These accuracy results can still be improved by exploring the features further or by trying to combine them with other methods. [20] conducted research using the semantic fuzzy logic model, which produced an F1-score of 91.5%. There is a lack of feature exploration, and the time used is still relatively long for computing. There is a need to increase the semantic fuzzy logic variables to accommodate more sentiment classes for better accuracy. Machine learning algorithms can also be applied to detect hate speech. Research by [28] produced an accuracy of 95% using the rule-based clustering method. Then research by [29] using (BERT + ANN) produced an accuracy of 93.55%, and (BERT + MLP) produced an accuracy of 93.67%. [30] also conducted research, which resulted in an accuracy of 77.6% and a balanced accuracy of 83%.

Based on the description above, in this research, a classification was carried out between sentences that are hate speech and those that are not hate speech using the LR and SVM algorithms. The combination of Logistic Regression (LR) and Support Vector Machine (SVM) is expected to provide a more potential solution by combining the advantages of Logistic Regression in ease of implementation and speed, as well as SVM's ability to handle more complex and non-linear data. It is hoped that the combination of these two algorithms can increase accuracy so that it can maximize effectiveness in classifying hate speech words or sentences.

2. Literature Review

References from several previous studies are important before researching to determine the relationship between the research that will be carried out and previous research to avoid duplication or similarities in the research that will be carried out. It is used as a reference in research based on the references taken, and the focus of research can be defined and classified.

There are previous studies that conducted research with the aim of classification using the LR and SVM methods. Research conducted by [22] discusses the LR method for classification. This research resulted in the conclusion that traditional LR classification requires many iterations and a long time, so it is necessary to adjust the logistic mathematical model, error function, and regression coefficient using the gradient descent method, and improve the sigmoid function. Meanwhile, [23] conducted research discussing the implementation of SVM in classifying images. This research shows that in classifying images using SVM, the things that influence model performance are high-quality datasets, ancillary data, feature extraction, and designing models that are adapted to existing data.

In research [25] used Logistic Regression, Random Forest, and KNN algorithms for text classification. This research shows that of the three algorithms used, Logistic Regression provides performance as expected in all parameters. This makes Logistic Regression the algorithm that produces the highest accuracy among the three algorithms compared. In addition, research conducted by [24] describes a comparison between Random Forest and SVM algorithms. This research aims to statistically measure the characteristics of methods in terms of frequency and accuracy. The results obtained show that the Random Forest algorithm is superior in the case of article databases and low spatial resolution image cases. Meanwhile, in classifying data that contains more features, the SVM method is superior in comparison to the average accuracy of the method compared to Random Forest.

In research [11], Afaan Oromo hate speech was detected using a model with five different deep-learning methods. These models are CNN, LSTM, GRU, BiLSTM, and CNN-LSTM. Based on the research conducted, the best performance was achieved by BiLSTM with an F1-Score of 91%. Apart from that, the research also shows that training embedded representations with models and combining samples can improve the performance of the models built. However, the research conducted has not yet collaborated on methods that might provide better performance.

3. Method

This research uses an experimental approach to classifying hate speech in cyberspace. This type of research focuses on collecting and analyzing numerical data to produce findings that can be expressed quantitatively. The methods used in this research are LR and SVM with a combination voting classifier technique. The framework for this research consists of several stages, which are described below.

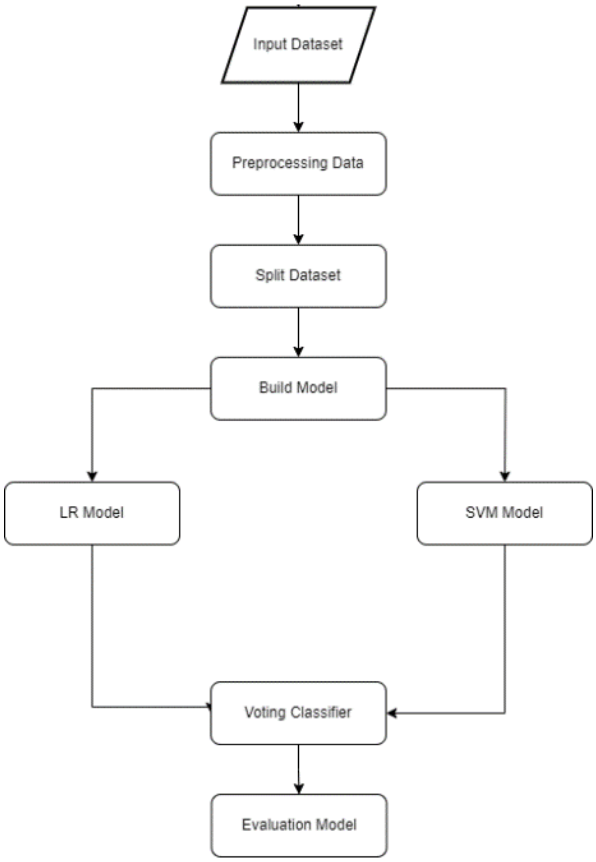

Figure 1. Hate speech classification flowchart using a combination of LR and SVM

Based on the framework above, the explanation of each stage is as follows:

Data collection

The first step in this research was to carry out a data collection process related to hate speech in cyberspace. This research uses the Twitter Sentiment Analysis dataset [21], which comes from the Kaggle Repository. The dataset contains 29,530 sentences that have a label column. The label column with the number '1' is the racist sentence category, and the number '0' is the non-racist sentence category.

Preprocessing Data

Dataset processing is carried out by cleaning the tweets in stages including case folding, noise removal, tokenizing, stopword removal, and stemming.

1. Case Folding: The process of changing all capital or uppercase letters in the dataset to lowercase or lowercase.
2. Noise Removal: The process of removing noise in a dataset, such as hashtags, numbers, punctuation, URLs, and other attributes that contain missing values.
3. Tokenizing: The process of breaking down a sequence of characters or sentences into several parts of words called tokens.
4. Stopwords Removal: The process of removing several unimportant words from the dataset.
5. Stemming: The process of changing each word in the dataset into a basic form.

Split Dataset

The dataset that has gone through the preprocessing stage will be divided into two parts, which are 75% as training data and 25% as test data. Training data will be used when training the model, while test data will be used when testing the model. This is done to ensure that the model being built can work well on new data.

Logistic Regression (LR)

LR is a method for solving regression and classification problems. LR predicts the classification of categorical data using probability. LR models can be formed with a logistic function using numerical values to predict the outcome. Maximum probability is calculated using a logistic function to predict the maximum data, and determining a probability between 0 and 1 determines whether an event will occur or not [31]. LR analysis is used to explain the relationship between response variables in the form of dichotomous or binary data and independent variables in the form of interval and/or categorical scale data [31]. The LR method has the advantage of being able to provide an easy analysis process because it does not require previous assumptions, making the classification process easier and being able to analyze the relationships between variables. Method illustration logistic regression [32] can be seen in Figure 2 below:

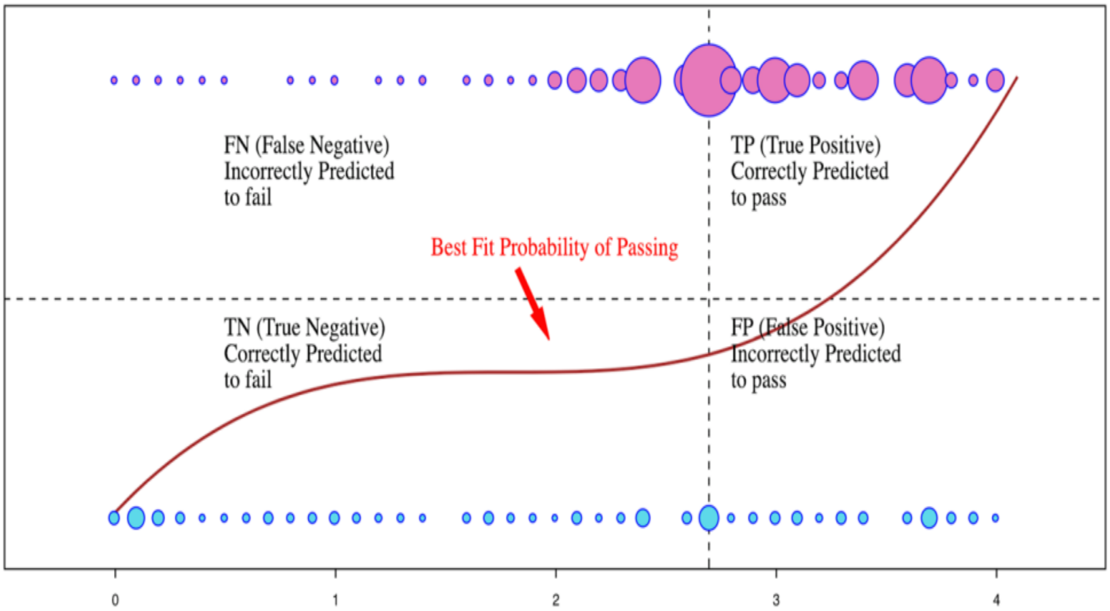

Figure 2. A visual illustration of logistic regression
Source: [32]

The mathematical expression of the logistic regression [33] method is as follows:

$$\hat{p} = \pi(x) = \frac{exp(g(x\,))}{1+exp(g(x))} \tag{1}$$

where:

$\hat{P}$ = Logistic probability.
exp = 2.71828183 or 2.72
B_0, B_1X = Logit function of logistic regression models.

Support Vector Machine (SVM)

SVM is a classification method that predicts classes using a model from the training process [34]. Grading is done by looking for a hyperplane or decision boundary that separates one class from another, which in this case plays a role in separating tweets with racist sentences, or number 1, and tweets with non-racist sentences, or number 0. The use of SVM in this research is to classify tweets that are utterances of hatred or not. The advantage of SVM is that it focuses on finding the best hyperplane that separates data classes, and this can function well even in high dimensions. SVM has a good ability to generalize training data to previously unseen data. SVM tends to avoid overfitting training data. Assume that the input data consists of n data vectors, where each data vector is represented by $x_i \in Rn$, where $i = 1, 2, 3, \ldots, n$. Let the class label that needs to be assigned to the data vectors to implement supervised classification be denoted by y_i, which is $+1$ for one category of data vectors and 0 for the other category of data vectors. The data set can be geometrically separated by a hyperplane. Since the hyperplane is represented by a line, it can also be mathematically represented by [35]:

$$wx_i + b = 0 \; as \; y_i \; = 0 \; (hyperplane) \tag{2}$$
$$w^T x_i + b \geq +1 \; as \; y_i \; = +1 \; (positive \; class) \tag{3}$$
$$w^T x_i + b \; \leq -1 \; as \; y_i \; = -1 \; (negative \; class) \tag{4}$$

$$sign\,(x) = \begin{cases} 1 \; if \; x > 0 \\ 0 \; if \; x = 0 \\ -1 \; if \; x < 0 \end{cases} \tag{5}$$

The distance *D* of a data point x from the hyperplane is represented mathematically by the equation:

$$D = \frac{|w^T x + b|}{|w|} \tag{6}$$

where:

w = weight vector
x = data (input)
b = bias
w^T = transpose weight

In sentiment analysis, if an equation is positive, then the algorithm will output positive, and vice versa if it is negative. SVM will generate an optimal hyperplane when separating the two categories [36]. The method illustration from the support vector machine [37] can be seen in Figure 2 below:

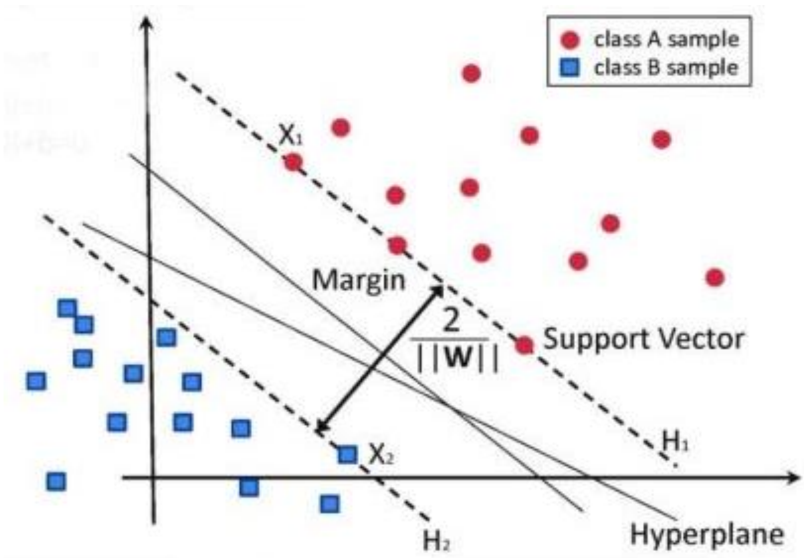

Figure 3. An illustration of a support vector machine
Source: [37]

Voting Classifier

At this stage, the models that have been created previously will be combined, namely logistic regression and SVM, using a voting classifier. In this research, we will use the 'hard' voting type, which makes decisions based on the dominant vote from the combined model. Thus, the voting classifier will make predictions based on the dominant results from the two combined models [38]. This voting classifier is a powerful approach to improving prediction accuracy, especially when the models have different advantages that can complement each other.

Evaluation Method

This research uses the confusion matrix evaluation method. The confusion matrix table consists of 4 parts, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [39]. The confusion matrix table can be seen in Table 1 below:

Table 1. Confusion matrix

	Predicted Class	
True Class	TP	FN
	FP	TN

The model performance evaluation measurement parameters used are as follows [40]:

a. Accuracy

Accuracy is the ratio of valid predictions (positive and negative) using holistic data. The following is the equation:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100\% \tag{7}$$

b. F1 Score

The F1 score is a parameter used in evaluation to describe the comparison of weighted average precision and recall. The following is the equation:

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

c. Recall

This is the ratio of true positive predictions compared to the total true positive data. The following is the equation:

$$Recall = \frac{TP}{(FN+TP)} \times 100\% \tag{9}$$

d. Precision

This is the ratio of true positive predictions compared to the overall positive predicted results. The formula is as follows:

$$Precision = \frac{TP}{(FP+TP)} \times 100\% \tag{10}$$

4. Results and Discussion

This article utilizes an approach that combines LR and SVM models to carry out the classification process. The LR and SVM methods used implement a voting classifier to combine different models and are used to make decisions based on labels that appear frequently or the majority of votes.

Train Test Split Dataset

In evaluating model performance, the dataset is divided into two parts: training data used to train the model and test data used to test the model regarding its ability to generalize data that has never been seen before. The data distribution can be observed in Figure 4.

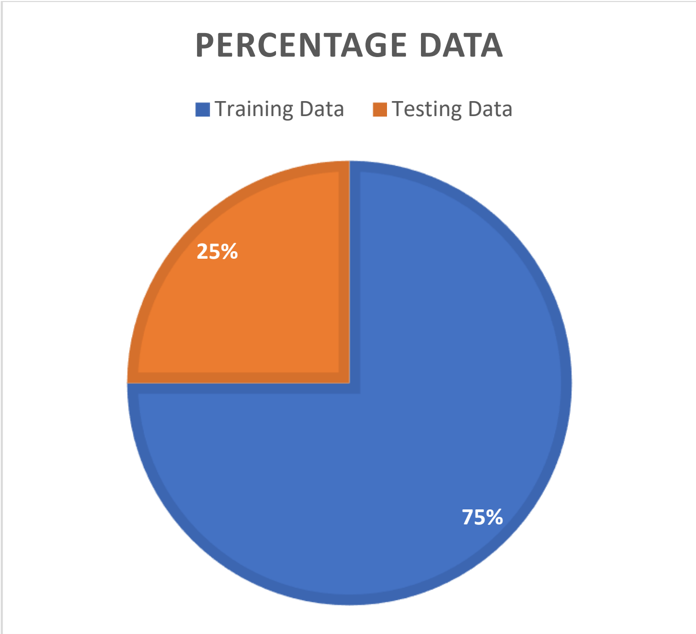

Figure 4. Train test split data

Dataset Labeling

Figure 5 shows the results of labeling the dataset into two categories, namely racist and non-racist categories.

Figure 5. Dataset labeling

Result of Model Evaluation

The results of research and experiments on the combination model of the LR algorithm and SVM show quite satisfactory performance in classifying hate comments. SVM has excellent performance in processing fairly high-dimensional data, as well as the ability to generalize better [28]. The LR algorithm is used to train a multi-level classification to separate speech that contains hate, offensive sentences, and those that do not contain hate elements [41]. These performance results can be observed in Table 2.

Table 2. Classification report

	precision	recall	f1-score	support
0	0.96	0.99	0.98	6,880
1	0.86	0.49	0.62	503
Accuracy			0.96	7,383
Macro avg	0.91	0.74	0.80	7,383
Weight avg	0.96	0.96	0.95	7,383

Table 2 provides quite satisfactory accuracy results, namely 96%. With a precision of 0 of 96% and a recall of 99%. This certainly produces a positive indication of the effectiveness of the LR and SVM models. A comparison of the accuracy results in this study with previous research can be seen in Table 3.

Table 3. Comparison result of the proposed method with other research

Method	**Accuracy**
BiLSTM and CNN [11]	87%
CNN and LSTM [11]	85%
GRU [11]	86%
LSTM [11]	82%
Multinomial Logistic Regression [27]	84%
Rule-Based Clustering [28]	95%
BERTand ANN [29]	93.55%
BERT and MLP [29]	93.67%
Proposed Method	96%

Based on Table 3, it can be concluded that combining the LR and SVM methods for analyzing hate comments is superior when compared to using accuracy parameters from previous research. These accuracy results were obtained because LR uses one-versus-rest performance, which means that the classification will be trained on each class using the highest probability predictions compared to all classification methods used in sentiment labels as well as the ability to predict the probability of binary target variables [42]. While SVM can work in binary classification, modeling nonlinear decisions, as well as the ability to overcome overfitting [43].

5. Conclusion

This research was conducted using a combination of LR and SVM methods to classify hate comments. The integration of these two methods provides a more potential solution by combining the advantages of LR in ease of implementation as well as the speed and ability of SVM in handling more complex and non-linear data. Research shows that combining these methods can produce high accuracy, with an average accuracy of 96%. It is hoped that the research results can help overcome the increasingly widespread problem of hate comments and can serve as a reference for further research. In future research, it is hoped to expand the dataset to obtain a greater variety of data and explore advanced text processing techniques such as word embeddings to improve the feature representation of comment text.

References

[1] B. Mathew *et al.*, "Thou shalt not hate: Countering online hate speech," in *Proceedings of the international AAAI conference on web and social media*, 2019, vol. 13, pp. 369–380.

[2] N. Alkiviadou, "Hate speech on social media networks: towards a regulatory framework?," *Inf. Commun. Technol. Law*, vol. 28, no. 1, pp. 19–35, 2019.

[3] S. Ullmann and M. Tomalin, "Quarantining online hate speech: technical and ethical perspectives," *Ethics Inf. Technol.*, vol. 22, pp. 69–80, 2020.

[4] K. Florio, V. Basile, M. Polignano, P. Basile, and V. Patti, "Time of your hate: The challenge of time in hate speech detection on social media," *Appl. Sci.*, vol. 10, no. 12, p. 4180, 2020.

[5] M. R. S. P. Pamungkas, M. N. Huda, D. A. Fauzan, A. H. Itsna, and F. M. Al-Hijri, "Sistem Klasifikasi Otomatis Dengan Konsep Machine Learning As A Service (MLaaS) Pada Kasus Pesan Berindikasi Cyberbullying," *Ilk. J. Comput. Sci. Appl. Informatics*, vol. 4, no. 3, pp. 252–261, 2022.

[6] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PLoS One*, vol. 14, no. 8, p. e0221152, 2019.

[7] F. Alkomah and X. Ma, "A literature review of textual hate speech detection methods and datasets," *Information*, vol. 13, no. 6, p. 273, 2022.

[8] K. R. Sylwander, "Affective atmospheres of sexualized hate among youth online: A contribution to bullying and cyberbullying research on social atmosphere," *Int. J. bullying Prev.*, vol. 1, no. 4, pp. 269–284, 2019.

[9] A. Tontodimamma, E. Nissi, A. Sarra, and L. Fontanella, "Thirty years of research into hate speech: topics of interest and their evolution," *Scientometrics*, vol. 126, pp. 157–179, 2021.

[10] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, "Detecting and monitoring hate speech in Twitter," *Sensors*, vol. 19, no. 21, p. 4654, 2019.

[11] G. O. Ganfure, "Comparative analysis of deep learning based Afaan Oromo hate speech detection," *J. Big Data*, vol. 9, no. 1, pp. 1–13, 2022.

[12] E. Barendt, "What is the harm of hate speech?," *Ethical Theory Moral Pract.*, vol. 22, pp. 539–553, 2019.

[13] G. Nguyen *et al.*, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 77–124, 2019, doi: 10.1007/s10462-018-09679-z.

[14] L. Vrysis *et al.*, "A web interface for analyzing hate speech," *Futur. Internet*, vol. 13, no. 3, p. 80, 2021.

[15] M. S. Jahan and M. Oussalah, "A systematic review of Hate Speech automatic detection using Natural Language Processing.," *Neurocomputing*, p. 126232, 2023.

[16] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," *IEEE Access*, vol. 7, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.

[17] I. Temitayo, S. Solomon, and S. Oyelere, *A systematic review of teaching and learning machine learning in K - 12 education*, vol. 28, no. 5. Springer US, 2023. doi: 10.1007/s10639-022-11416-

7.
[18] D. Ni, "Machine learning in recycling business : an investigation of its practicality , benefits and future trends," vol. 7, pp. 7907–7927, 2021, doi: 10.1007/s00500-021-05579-7.
[19] H. Ghoddusi, G. G. Creamer, and N. Rafizadeh, "Machine learning in energy economics and finance : A review," *Energy Econ.*, vol. 81, pp. 709–727, 2019, doi: 10.1016/j.eneco.2019.05.006.
[20] F. Emmanuel, O. Folorunso, F. Thomas, I. Ademola, and A. Abayomi-alli, "A probabilistic clustering model for hate speech classification in twitter," *Expert Syst. Appl.*, vol. 173, no. February 2020, p. 114762, 2021, doi: 10.1016/j.eswa.2021.114762.
[21] A. Toosi, "Twitter Sentiment Analysis," 2019. https://www.kaggle.com/
[22] X. Zou, "Logistic Regression Model Optimization and Case Analysis," pp. 135–139, 2019.
[23] M. Arya and C. S. S. Bedi, "Survey on SVM and their application in image classification," *Int. J. Inf. Technol.*, vol. 13, no. 5, pp. 1867–1877, 2021, doi: 10.1007/s41870-017-0080-1.
[24] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Member, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification : A Meta-Analysis and Systematic Review," vol. 13, pp. 6308–6325, 2020.
[25] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and KNN models for the text classification," *Augment. Hum. Res.*, vol. 5, pp. 1–16, 2020.
[26] G. Ambrish, B. Ganesh, A. Ganesh, C. Srinivas, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," vol. 3, no. April, pp. 127–130, 2022, doi: 10.1016/j.gltp.2022.04.008.
[27] P. S. B. Ginting, B. Irawan, and C. Setianingsih, "Hate speech detection on Twitter using multinomial logistic regression classification method," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 2019, pp. 105–111.
[28] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions," *Comput. Sci. Rev.*, vol. 38, p. 100311, 2020.
[29] H. Mehta and K. Passi, "Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI)," *Algorithms*, vol. 15, no. 8, p. 291, 2022.
[30] B. Vidgen and T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media," *J. Inf. Technol. Polit.*, vol. 17, no. 1, pp. 66–78, 2020.
[31] F. R. Irawan, A. Jazuli, and T. Khotimah, "Analisis Sentimen Terhadap Pengguna Gojek Menggunakan Metode K-Nearset Neighbors," *JIKO (Jurnal Inform. dan Komputer)*, vol. 5, no. 1, pp. 62–68, 2022.
[32] D. Gupta, A. Choudhury, U. Gupta, and P. Singh, "Computational approach to clinical diagnosis of diabetes disease : a comparative study," pp. 30091–30116, 2021.
[33] R. H. Situngkir, "Analisis Regresi Logistik untuk Menentukan Faktor-Faktor yang Mempengaruhi Kesejahteraan Masyarakat Kabupaten/Kota di Pulau Nias." Universitas Sumatera Utara, 2022.
[34] H. Sulastomo, R. Ramadiansyah, K. Gibran, E. Maryansyah, and A. Tegar, "Analisis Sentimen Pada Twitter@ Ovo_Id dengan Metode Support Vectore Machine (SVM)," *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 6, no. 2, pp. 1050–1056, 2022.
[35] M. R. A. Nasution and M. Hayaty, "Comparison of Accuracy and Processing Time of K-NN and SVM Algorithms in Twitter Sentiment Analysis," *J. Inf.*, vol. 6, no. 2, 2019.
[36] M. A. R. Reynaldhi and Y. Sibaroni, "Analisis Sentimen Review Film Pada Twitter Menggunakan Metode Klasifikasi Hybrid Svm, Naïve Bayes, Dan Decision Tree," *eProceedings Eng.*, vol. 8, no. 5, 2021.
[37] Y. S. Triyantono, S. Al Faraby, and M. Dwifebri, "Analisis Sentimen Terhadap Ulasan Film Menggunakan Word2Vec dan SVM," *eProceedings Eng.*, vol. 8, no. 4, 2021.
[38] M. Manzo and S. Pellino, "Voting in transfer learning system for ground-based cloud classification," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 3, pp. 542–553, 2021.
[39] K. Harimoorthy and M. Thangavelu, "Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 3, pp. 3715–3723, 2021, doi: 10.1007/s12652-019-01652-0.
[40] V. Chang, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16157–16173, 2023, doi: 10.1007/s00521-022-07049-z.

[41] R. T. Mutanga, N. Naicker, and O. O. Olugbara, “Detecting Hate Speech on Twitter Network using Ensemble Machine Learning,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 3, 2022.

[42] D. C. Asogwa, C. I. Chukwuneke, C. C. Ngene, and G. N. Anigbogu, “Hate speech classification using SVM and naive BAYES,” *arXiv Prepr. arXiv2204.07057*, 2022.

[43] E. Omran, E. Al Tararwah, and J. Al Qundus, “A comparative analysis of machine learning algorithms for hate speech detection in social media,” *Online J. Commun. Media Technol.*, vol. 13, no. 4, p. e202348, 2023.