Utilizing Support Vector Machines to Detect Hate Speech on Social Media

Muhammad Athoillah[1], Rani Kurnia Putri[2]

[1]*Statistics Department, Universitas PGRI Adi Buana Surabaya, 60234, Indonesia.*
[2]*Mathematics Education Department, Universitas PGRI Adi Buana Surabaya, 60234, Indonesia.*

Abstract

This paper presents a novel application of Support Vector Machines (SVM) in developing automated systems to detect hate speech on social media platforms, addressing a critical need for scalable solutions to enhance online safety and societal cohesion. By leveraging SVM's proven efficacy in managing high-dimensional data and optimizing the balance between precision and recall, the study offers a comprehensive methodology that includes data collection, preprocessing, model training, deployment, and evaluation. The results demonstrate robust average performance across key metrics, affirming the model's reliability in accurately identifying hate speech while minimizing false positives. This research advances the field by showcasing the practical and theoretical contributions of SVM in automated hate speech detection, highlighting its potential to significantly improve content moderation practices. The findings underscore the necessity for ongoing refinement of detection systems and collaborative efforts among researchers, technology firms, and policymakers to create more inclusive online environments that promote respectful discourse and community well-being.

Keywords: *Automated Systems, Classification, Hate Speech Detection, Sentiment Analysis, Text Mining.*

1. Introduction

Hate speech encompasses any form of communication—verbal, written, or otherwise conveyed—that attacks or incites hatred, violence, discrimination, or prejudice against individuals or groups based on characteristics such as race, ethnicity, nationality, religion, sexual orientation, gender identity, disability, or other identifiable traits. The targeted characteristics can be diverse, and the purpose behind hate speech can vary from expressing bigotry and prejudice to inciting violence or discrimination against particular groups [1]. Hate speech frequently aims to belittle, intimidate, or dehumanize individuals or entire communities, fostering an atmosphere of fear, hostility, and division. It can appear in many forms, such as verbal insults, slurs, derogatory remarks, threats, harassment, and the spread of hateful propaganda through various media, including social media, public speeches, publications, and online forums [2], [3].

Hate speech represents a complex threat that goes well beyond mere words. Fundamentally, it serves as a tool for division and dehumanization, targeting individuals or groups based on inherent traits like race, religion, ethnicity, gender, sexual orientation, or disability. Its dangerous nature lies in its capacity to incite prejudice, discrimination, and violence, fostering a toxic environment where intolerance can grow and persist [4]. Hate speech not only undermines the dignity and rights of its targets but also erodes social cohesion, fostering mistrust and animosity within communities. Moreover, it can have tangible, devastating consequences, ranging from psychological harm and social exclusion to physical

Corresponding author: *Muhammad Athoillah (athoillah@unipasby.ac.id)*
Received: *20 February 2024*; Revised: *28 August 2024*; Accepted: *3 September 2024*; Published: *6 September 2024*

violence and even genocide. By normalizing bigotry and promoting hostility towards marginalized groups, hate speech corrodes the foundations of democracy, equality, and human rights, posing a grave threat to the fabric of society. It perpetuates cycles of oppression and conflict, hindering progress towards a more just, inclusive world where all individuals are valued and respected. Thus, the danger of hate speech lies not only in its words but in its power to sow seeds of hatred and division that reverberate throughout society, leaving lasting scars on individuals and communities alike[1], [5].

Given these dangers, there is an urgent need to develop automated systems to detect hate speech. Human moderation alone cannot keep pace with the vast volume of content generated online, making it essential to leverage technology for scalable solutions. Automated detection systems can swiftly identify and remove hateful content, mitigating its harmful effects and promoting online safety. Additionally, these systems can help platforms comply with legal regulations aimed at combating hate speech, ensuring that online spaces adhere to standards of decency and respect [6].

The development of automated systems for detecting hate speech on social media has become increasingly sophisticated, leveraging advances in machine learning and natural language processing (NLP). Recent trends indicate a focus on creating models that can understand context and nuance in language, which are crucial for accurately identifying hate speech. These models utilize deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to analyze vast datasets and learn to recognize patterns associated with hate speech[7]. Additionally, there is a growing emphasis on cross-linguistic and cross-cultural adaptability, ensuring that detection systems can function effectively across different languages and cultural contexts. Another significant trend is the collaboration between tech companies, academic researchers, and non-profit organizations to continuously improve these systems, sharing data and best practices to enhance the accuracy and reliability of hate speech detection technologies. These efforts aim to create safer online environments by effectively identifying and mitigating the spread of hateful content[8].

The challenge of automatically identifying hate speech can be addressed by employing different classification algorithms, among which Support Vector Machines (SVM) stand out as particularly potent. SVM is a good choice for building automated systems to detect hate speech because of its effectiveness in handling high-dimensional data, robustness to overfitting, ability to capture non-linear relationships, interpretability, versatility, efficiency, and extensive research support. SVM's capability to effectively process textual data, its lesser susceptibility to overfitting, and its ability to handle non-linear relationships make it suitable for hate speech detection tasks. Additionally, SVM's interpretability allows for a better understanding of hate speech characteristics, while its efficiency makes it practical for real-time applications. Moreover, SVM is a well-studied and widely used algorithm, providing a solid foundation for implementation and optimization. These factors collectively position SVM as a compelling choice for developing accurate and efficient automated systems to combat hate speech on online platforms [9], [10], [11].

The scientific contribution of this study lies in its novel application of SVM to the problem of hate speech detection on social media platforms. While SVM has been widely used in various text mining tasks, its specific application to hate speech detection can advance the current state of automated content moderation tools. This research not only demonstrates the practical effectiveness of SVM in identifying hate speech but also provides a framework for further refinement and development of automated moderation systems. By showcasing the algorithm's strengths in this specific context, this study contributes to both theoretical and practical advancements in the field of machine learning and its applications in social media analysis.

There exists compelling evidence supporting the efficacy of this algorithm in effectively resolving classification tasks, notably within the domain of text mining. Numerous studies have contributed to substantiating its success in this regard. For example, Andraini et al. analyzed public sentiment towards the Russia-Ukraine conflict on Twitter[12], finding that SVM could identify key topics discussed by Twitter users related to the conflict, such as support for Ukraine and condemnation of Russia's actions. Another notable example is the research conducted by Athoillah and Rani, who used this algorithm to identify hoax news related to Covid-19 [13]. In this study, the dataset consisted of Covid-19-related news collected from online sources, and

the results showed that SVM could classify news with good accuracy and effectively assist in detecting Covid-19-related hoaxes.

Given this context, the research team decided to create an automated system for detecting hate speech on social media. The goal of this study is to design and implement a system that identifies hate speech using the Support Vector Machine (SVM) algorithm.

2. Method

The steps involved in constructing this system are outlined below:

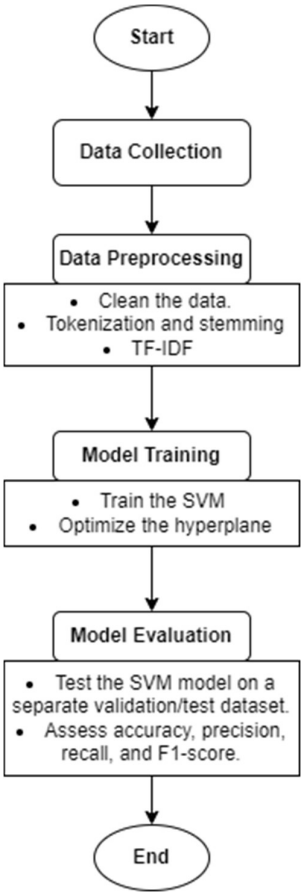

Figure 1. Flowchart of the process.

a. Data Collection

Collect a substantial dataset of social media posts or comments, labeled as hate speech or non-hate speech. The dataset consists of Twitter data from 2017, focusing on discussions around the Jakarta gubernatorial election, particularly the biases and controversies associated with the candidates. The dataset has been collected during a period of heightened social media activity in Indonesia, reflecting the public's sentiments and opinions. The data primarily comprises textual content, including tweets, user identifiers, and possibly timestamps and hashtags. The data originated from Twitter's API or scraping methods, and while there is no explicit mention of data completeness or privacy measures, it's assumed that the dataset might have some gaps and should adhere to privacy norms such as anonymizing user details. The purpose of the dataset is to analyze the political discourse of the time, and valuable for studies examining social media's role in shaping public opinion during elections. This dataset should be diverse, balanced, and representative to ensure the model's accuracy and generalization capability.

b. Data Preprocessing

Preprocess the collected data by eliminating irrelevant elements such as special characters, URLs, and punctuation, and by performing tasks like tokenization and stemming to standardize the text format. Additionally, methods such as TF-IDF (Term Frequency-Inverse Document Frequency) can be applied to transform text into numerical features. TF-IDF is a statistical approach used to assess the significance of a term in a document relative to a collection of documents, often called a corpus. This method is crucial in natural language processing tasks like text mining and information retrieval. The Term Frequency (TF) component measures how often a term appears in a document in relation to the total number of terms, using the formula:

$$TF(t,d) = \frac{n_{t,d}}{\sum_k n_{k,d}} \quad (1)$$

Where $n_{t,d}$ denotes the number of times term t appears in document d, and the denominator represents the total number of terms in document d. On the other hand, the Inverse Document Frequency (IDF) component measures how uncommon a term is across the entire corpus of documents. It is calculated using the formula:

$$IDF(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (2)$$

Where N is the total number of documents in the corpus, and $|\{d \in D : t \in d\}|$ represents the number of documents containing term t. The TF-IDF score for a term in a document is obtained by multiplying its TF by its IDF:

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D) \quad (3)$$

The TF-IDF score reflects the significance of a term in a document by considering both its local importance (TF) and global rarity (IDF). Terms with higher TF-IDF scores are considered more relevant or discriminative for characterizing the content of a document. TF-IDF scores are often normalized to prevent bias towards longer documents, typically using L2 normalization, ensuring that the TF-IDF vector for each document has a unit length. In hate speech detection and various other text analysis tasks, TF-IDF is instrumental in converting textual data into numerical feature vectors, enabling machine learning algorithms to effectively process and analyze text[14], [15].

c. Model Training

Train the SVM model using the preprocessed dataset. During the model training phase in the development of an automatic hate speech detection system using Support Vector Machine (SVM), the processed dataset serves as the foundation for instructing the SVM algorithm on how to discern between instances of hate speech and non-hate speech. Throughout training, the SVM algorithm endeavors to identify patterns inherent in the data and establishes an optimal decision boundary, often termed a hyperplane, within the feature space. This decision boundary is engineered to maximize the margin of separation between various classes of data points, effectively segmenting the feature space into discrete regions corresponding to hate speech and non-hate speech. The main goal of SVM training is to find the hyperplane that best separates the two classes while minimizing classification errors. This is accomplished through an iterative process where the SVM algorithm adjusts the hyperplane based on the training examples, aiming to maximize the margin between support vectors, which are the data points closest to the decision boundary. By refining its understanding of the data through this process, the SVM algorithm learns to generalize from the training dataset, enabling it to accurately classify new, unseen instances of hate speech and non-hate speech. The effectiveness of the trained SVM model is evaluated using validation or test data, and adjustments may be made to model parameters, such as the kernel function or regularization parameter, to enhance performance [11], [16], [17].

The mathematical representation of the decision boundary in SVM can be expressed using the equation of the hyperplane:

$$w.x + b = 0 \quad (4)$$

where

w: represents the weight vector, which defines the orientation of the hyperplane.

x: denotes the input feature vector.

b: the bias term, representing the offset of the hyperplane from the origin.

This equation delineates the decision boundary that separates the feature space into different classes, facilitating the classification of instances as either hate speech or non-hate speech.

d. Deployment

Deploy SVM as part of an automated system for hate speech detection on social media platforms. The system should take user-generated content as input, preprocess it, extract features using the trained SVM model, and classify the content as hate speech or non-hate speech.

e. Model Evaluation

During the model evaluation phase of developing an automated hate speech detection system using the Support Vector Machine (SVM) algorithm, the trained SVM model undergoes comprehensive assessment to ascertain its accuracy in identifying instances of hate speech. This evaluation entails testing the SVM model on a distinct dataset, separate from the training dataset, known as the validation or test dataset. Various performance metrics, including accuracy, precision, recall, and the F1-score, are utilized to gauge the model's effectiveness.

The accuracy metric assesses the overall correctness of the model's classifications by determining the proportion

of correctly classified instances out of the total evaluated. Precision measures the model's capability to accurately identify true positive instances of hate speech, representing the ratio of accurately identified hate speech instances to the total instances classified as hate speech. Recall, or sensitivity, evaluates the model's capacity to capture all instances of hate speech, indicating the ratio of correctly identified hate speech instances to the total actual hate speech instances in the dataset. The F1-score integrates precision and recall into a single metric, offering a balanced assessment of the model's performance by calculating the harmonic mean of precision and recall. The equations for calculating these metrics are as follows:

$$\boldsymbol{Acc} = \frac{Number\ of\ Correctly\ Classified\ Instances}{Total\ Number\ of\ Instances}$$

$$\boldsymbol{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (5)$$

$$\boldsymbol{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$\boldsymbol{F1 - Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

These evaluation metrics provide insights into different aspects of the SVM model's performance, enabling developers to assess its strengths and weaknesses. By comprehensively evaluating the model's performance on the validation or test dataset, developers can make informed decisions regarding model optimization and fine-tuning to enhance its effectiveness in accurately detecting hate speech[18], [19].

f. Model Validation

Validate the model using Holdout validation, a straightforward technique for assessing machine learning model performance. This method involves splitting the dataset into two separate subsets: a training set and a validation set. The model is trained solely on the training set, learning the data's patterns and relationships. After training, the model's performance is tested on the validation set, which consists of data the model has not seen before. This evaluation assesses predictive accuracy and other performance metrics. While Holdout validation is simple and efficient, its reliability can be influenced by the random partitioning of the data. When selecting an appropriate validation method, it is crucial to take into account the specific characteristics of the dataset and the objectives of the machine learning task [20], [21].

3. Result and Discussion

3.1. Test result

Assessing the performance of an application requires examining several key metrics like Accuracy, Precision, Recall, and F1-Score, offering valuable insights into the application's ability to deliver accurate and pertinent results. In this research, the validation was carried out using Holdout Validation, meticulously performed across ten experimental runs. The ensuing table (Table 1) showcases the test outcomes of the developed system.

Table 1. Performance of system.

Trial	Perform			
	Accuracy	**Precision**	**Recall**	**F1-Score**
1	88,03%	85,00%	97,70%	90,91%
2	85,92%	84,62%	92,77%	88,51%
3	86,62%	85,57%	94,32%	89,73%
4	88,03%	85,58%	97,80%	91,28%
5	83,80%	84,16%	92,39%	88,08%
6	87,32%	88,00%	93,62%	90,72%
7	83,80%	81,55%	95,45%	87,96%
8	87,32%	87,10%	93,10%	90,00%
9	88,03%	85,44%	97,78%	91,19%
10	83,80%	83,17%	93,33%	87,96%
AVG	**86,27%**	**85,02%**	**94,83%**	**89,63%**

The dataset presents the performance metrics of a classification model across ten trials, revealing notable insights into its efficacy. On average, the model achieves strong results, with an accuracy of 86.27%, precision of 85.02%, recall of 94.83%, and F1-score of 89.63%. These metrics collectively indicate the model's proficiency in correctly classifying instances and capturing positive cases. Despite minor variability across trials, the model demonstrates consistent performance levels, highlighting its reliability. Notably, the model excels in recall, indicating its ability to accurately identify positive

instances, while maintaining a commendable level of precision, effectively minimizing false positive predictions. Overall, these findings underscore the model's effectiveness and reliability in classification tasks, particularly in scenarios where both precision and recall are crucial metrics for evaluation

However, Further elaboration on the key aspects emphasized concerning this system included:

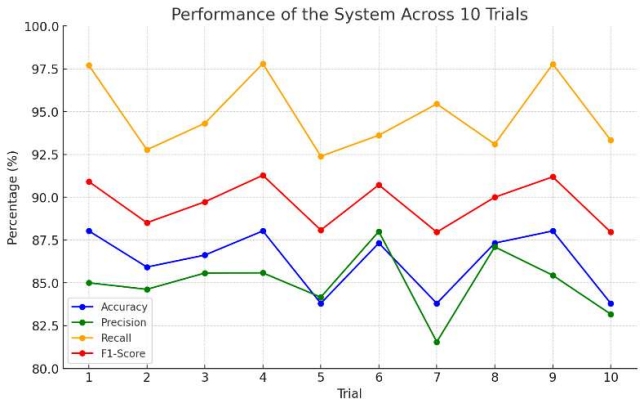

Figure 2. Comparative performance metrics of the proposed system across various trials.

The chart illustrates the performance of a system across 10 different trials, focusing on four key metrics: Accuracy, Precision, Recall, and F1-Score. The dashed lines represent the average values for each metric.

Consistency and average performance:

a. The mean accuracy of 86.27% signifies that, typically, the model accurately predicts the class of instances in the dataset approximately 86.27% of the time. This suggests a consistent level of performance across multiple trials.

b. Precision, averaging at 85.02%, demonstrates the model's capacity to correctly identify true positive instances among all instances predicted as positive. This indicates that, on average, the model's positive predictions are accurate around 85.02% of the time.

c. With an average recall of 94.83%, the model shows its ability to accurately identify a large majority of positive instances from the entire pool of positive instances. This implies that, on average, the model captures nearly 94.83% of all actual positive instances.

d. The F1-score, averaging at 89.63%, signifies the balance between precision and recall. This score serves as a valuable measure to evaluate the overall performance of the model, particularly in situations where both precision and recall hold significance.

Variability across trials:

a. Although the overall performance is strong, there could be some variance in performance among individual trials. This variability is evident in slight fluctuations observed in metrics like precision, recall, and F1-score across various trials.

b. Nonetheless, despite this variability, the general consistency in performance metrics implies that the model sustains a relatively steady level of performance across trials.

Strength in recall:

The highest performance is observed in recall, with a maximum of 97.80% and an average of 94.83%. This indicates that the model excels in correctly identifying positive instances, which is crucial in applications where identifying all positive cases is essential, such as disease diagnosis or fraud detection.

Reliable precision:

While precision exhibits slightly more variability compared to other metrics, with a range from 81.55% to 88.00%, the overall performance remains strong. This indicates that the model effectively minimizes false positive predictions, which is crucial in scenarios where false positives can have significant consequences.

In summary, while there may be minor variability in performance across trials, the overall performance of the model across various metrics highlights its reliability and effectiveness in classification tasks, particularly in terms of accurately identifying positive instances while maintaining a strong level of precision.

3.2. Result comparison

This section compares the performance of our Support Vector Machine (SVM) model for hate speech detection with results from other relevant studies. We focus on key performance metrics: Accuracy, Precision, Recall, and F1-Score. The studies selected for comparison include Davidson et al. [22], Zhang et al. [23], and Badjatiya et al. [24], which utilize various machine learning approaches for hate speech detection.

Table 2. Comparison result.

Study	Algorithm	Precision	Recall	F1-score
This Study	SVM	85,02%	94,83%	89,63%
[22]	Logistic Regression	91,00%	90,00%	90,00%
[23]	Cov-GRU DNN	-	-	93,00%
[24]	SVM	81,60%	81,60%	81,60%

Our SVM model for hate speech detection exhibits robust performance with an average precision of 85.02%, recall of 94.83%, and F1-Score of 89.63%, demonstrating its effectiveness in identifying hate speech and maintaining a good balance between precision and recall. Compared to Davidson et al.'s logistic regression model, which has a higher precision of 91.00% but lower recall of 90.00%, our model excels in recall, crucial for comprehensive hate speech detection. While Zhang et al.'s Conv-GRU DNN reports a higher F1-Score of 93.00%, the lack of specific precision and recall values makes a detailed comparison challenging. Our model outperforms Badjatiya's SVM model, which has a precision, recall, and F1-Score of 81.60%, highlighting its superior balance of precision and recall. This study underscores the viability of SVM in hate speech detection, offering a reliable solution for safer online environments and providing a foundation for future research to enhance precision and explore advanced methodologies.

4. Conclusion

The utilization of Support Vector Machines (SVM) for automated hate speech detection on social media platforms represents a significant advance in addressing online hostility. The system demonstrates strong performance, with an average accuracy of 86.27%, precision of 85.02%, recall of 94.83%, and F1-score of 89.63%. These results highlight the model's proficiency in correctly classifying hate speech instances and capturing most positive cases, aligning with current literature on SVM's effectiveness in text classification. Despite minor variability, the system's consistent performance underscores its reliability. However, limitations include potential performance variation with different datasets and the focus on textual data. Future research should explore multimodal data integration and adaptive models to improve accuracy and coverage. This study underscores the importance of developing sophisticated automated systems to foster safer and more inclusive online environments, necessitating ongoing collaboration among researchers, technology firms, and policymakers.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

Data Availability Statement

Supplementary materials and data used in this research are accessible upon request. For access, please contact the corresponding author via athoillah@unipasby.ac.id

Acknowledgments

The research team would like to thank the Institute for Research and Community Service at PGRI Adi Buana University, Surabaya, for their acknowledgment and appreciation. They are grateful for the financial assistance provided by the Adi Buana Grant program during the 2023/2024 period, which has been instrumental in facilitating the smooth continuation of their research efforts.

Reference

[1] A. Sellars, "Defining hate speech," *Berkman Klein Center Research Publication*, no. 2016–20, pp. 16–48, 2016.

[2] M. A. Paz, J. Montero-Díaz, and A. Moreno-Delgado, "Hate speech: A systematized review," *Sage Open*, vol. 10, no. 4, p. 2158244020973022, 2020.

[3] T. T. A. Putri, S. Sriadhi, R. D. Sari, R. Rahmadani, and H. D. Hutahaean, "A comparison of classification algorithms for hate speech detection," in *Iop conference series: Materials science and engineering*, IOP Publishing, 2020, p. 32006.

[4] S. Malmasi and M. Zampieri, "Detecting hate speech in social media," *arXiv preprint arXiv:1712.06427*, 2017.

[5] K. Amin, M. D. A. Alfarauqi, and K. Khatimah, "Social media, cyber hate, and racism," *Komuniti: Jurnal Komunikasi dan Teknologi Informasi*, vol. 10, no. 1, pp. 3–10, 2018.

[6] S. Datta and S. Sarkar, "Automation, security and surveillance for a smart city: Smart, digital city," in *2017 IEEE Calcutta Conference (CALCON)*, IEEE, 2017, pp. 26–30.

[7] N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: a review," *IEEE Access*, vol. 9, pp. 88364–88376, 2021.

[8] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–30, 2018.

[9] D. Alita, "Multiclass Svm Algorithm For Sarcasm Text In Twitter," *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, vol. 8, no. 1, pp. 118–128, 2021.

[10] Z. Peng, Q. Hu, and J. Dang, "Multi-kernel SVM based depression recognition using social media data," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 1, pp. 43–57, 2019.

[11] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine learning*, Elsevier, 2020, pp. 101–121.

[12] F. Andraini and E. Mahdiyah, "Analisis Sentimen Twitter Terhadap Peperangan Rusia Dan Ukraina Menggunakan Algoritma Support Vector Machine," *Jurnal Aplikasi Komputer*, vol. 2, no. 1, pp. 46–58, 2022.

[13] R. K. Putri and M. Athoillah, "Support Vector Machine Untuk Identifikasi Berita Hoax Terkait Virus Corona (Covid-19)," *Jurnal Informatika Politeknik Harapan Bersama*, vol. 6, no. 3, pp. 162–167, 2021, doi: 10.30591/jpit.v6i3.2489.

[14] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification," in *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, IEEE, 2016, pp. 112–116.

[15] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, IEEE, 2016, pp. 61–66.

[16] C. Campbell and Y. Ying, *Learning with support vector machines*. Springer Nature, 2022.

[17] L. H. Hamel, *Knowledge discovery with support vector machines*. John Wiley & Sons, 2011.

[18] B. Juba and H. S. Le, "Precision-recall versus accuracy and the role of large data sets," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 4039–4048.

[19] M. Athoillah and R. K. Putri, "Handwritten arabic numeral character recognition using multi kernel support vector machine," *KINETIK: Game technology, information system, computer network, computing, electronics, and control*, pp. 99–106, 2019.

[20] R. M. Oxford and L. G. Daniel, "Basic Cross-Validation: Using the" Holdout" Method To Assess the Generalizability of Results.," *Research in the Schools*, vol. 8, no. 1, pp. 83–89, 2001.

[21] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *J Econom*, vol. 187, no. 1, pp. 95–112, 2015.

[22] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *arxiv:1703.04009[cs.CL]*, Jul. 2024.

[23] Z. Zhang, D. Robinson, and J. Tepper, "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," in *The Semantic Web*, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds., Cham: Springer International Publishing, 2018, pp. 745–760.

[24] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," *arxiv:1706.00188[cs.CL,cs.IR]*, Jul. 2024, doi: 10.1145/3041021.3054223.