RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received September 17, 2024; revised October 1, 2024; accepted October 2, 2024; date of publication October 30, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i4.580

How to cite: Siti Ernawati, Frieyadie, and Eka Rini Yulia, “Cybersentinel: The Cyberbullying Detection Application Based on Machine Learning and VADER Lexicon with GridSearchCV Optimization”, Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 4, pp. 533-542, October 2024.

Cybersentinel: The Cyberbullying Detection Application Based on Machine Learning and VADER Lexicon with GridSearchCV Optimization

Siti Ernawati, Frieyadie, and Eka Rini Yulia

Information and Technology Faculty, Universitas Nusa Mandiri, Jakarta, Indonesia

Corresponding author: Siti Ernawati (e-mail: siti.ste@nusamandiri.ac.id)

ABSTRACT Cyberbullying is becoming an increasingly troubling issue in today's digital age, with serious impacts on the well-being of individuals and society as a whole. With the number of social media users continuously rising, there is an urgent need to develop effective solutions for detecting cyberbullying. This urgency negatively affects the well-being of individuals, especially children and adolescents. The Big Data era also brings many new challenges, including the ability of organizations to manage, process, and extract value from available data to generate useful information. The aim of this research is to develop Cybersentinel, a cyberbullying detection application that combines Machine Learning and VADER Lexicon approaches to improve classification accuracy. It involves comparing several Machine Learning algorithms optimized using the GridSearchCV technique to find the best combination of parameters. The dataset used consists of social media comments labeled as bullying and non-bullying. The successfully developed model uses the Support Vector Machine algorithm, achieving a best accuracy of 98.83%. The system is developed using Python with the Streamlit framework. This application development follows the Design Science Research (DSR) approach, which integrates principles, practices, and procedures to facilitate problem-solving and support the design and creation of applications. Testing is conducted using a blackbox testing. The results show that parameter optimization using GridSearchCV can significantly enhance model performance, and applying the DSR method allows for the development of Cybersentinel tailored to specific needs. Thus, Cybersentinel provides an effective solution for detecting cyberbullying and contributes to improving the safety of social media users.

INDEX TERMS Cyberbullying, Design Science Research, GridSearchCV, Machine Learning, Sentiment Analysis, Vader Lexicon.

I. INTRODUCTION

As the number of social media users increases, cyberbullying has become an increasingly serious problem. The negative impacts of cyberbullying include anxiety, depression, self-harm, suicide attempts, as well as mental health, emotional, and economic difficulties for the victims. [1]. Cyberbullying is an attempt to bully by using digital technology, which can be found on social media, chatting platforms, gaming platforms, and mobile phones [2], and is one of the negative aspects of social media. This includes behaviors such as addictive users, trolling, online witch hunts, spreading fake news, and privacy violations. [3]. The early detection of cyberbullying is crucial to prevent and mitigate negative impacts. By detecting and addressing cyberbullying cases as quickly as possible, it can help protect vulnerable individuals and reduce the risk of more serious consequences. Machine learning algorithms offer the potential to overcome the challenges of detecting cyberbullying on social media [4] It also plays an important role in processing large and complex data [5]. The design of a science research framework for cyberbullying detection applications utilizing Machine

Learning (ML) which requires a comprehensive understanding of existing methodologies, challenges, and advancements in this field. Cyberbullying, an increasingly widespread problem due to anonymity in online interactions, has prompted researchers to explore various machine learning techniques for effective detection and prevention [6], [7]. One of the crucial aspects in the development of cyberbullying detection applications is the selection of appropriate machine learning algorithms. Also, the large volume of data generated every day makes it difficult to identify cyberbullying manually. It can be interpreted that big data is a large amount of data derived from the surge of individual social interactions in the digital realm [8]. The big data is then, derived from social media that has an important significance in research because of the online activities that generate large volumes of data [9]. With the abundance of data, it comes the new challenge of how to effectively analyze data which requires the ability to sift through relevant information, identify meaningful patterns, and make accurate predictions. Dataset considerations are critical in the development of robust cyberbullying detection applications. The challenge of class imbalance, where non-bullying cases far outnumber bullying cases, complicates the training of machine learning models [10]. Techniques such as data rebalancing and the use of labelled datasets, such as those curated from social media platforms, are critical to improving model performance [11]. In addition, the quality of the dataset significantly affects the effectiveness of the detection algorithm. For example, [11] provide a labelled dataset specific to Instagram, which can be a valuable source for training and testing cyberbullying detection models.

By using machine learning algorithms, such as data clustering, sentiment classification, anomaly detection or others can be extracted from social media data to support decision making. This research implements a text mining model by performing sentiment classification [12], [13]. One of the labeling techniques can use VADER to classify sentiment into bullying and non-bullying labels. VADER is also known as a lexicon dictionary-based sentiment analysis method that has proven successful in examining natural language text. VADER serves to evaluate sentiment in the form of text such as reviews or documents and determine whether the sentiment falls into the positive, negative, or neutral category. To achieve optimal performance, the parameters of the machine learning model need to be carefully set.

Some previous and relevant research on machine learning-based cyberbullying detection applications has been carried out including, The research focuses on automating the detection of cyberbullying using four machine learning classifiers (NB, LR, DT, SVM), and the proposed model delivers the best results in content classification, with the Logistic Regression classifier achieving an accuracy of 96% [14]. Detecting cyberbullying using machine learning algorithms (RF, NB, SVM, LR, Ensemble) by incorporating the element of sarcasm, the research results show that the SVM classifier performed better than the other classifiers, achieving an average accuracy of 79% [15]. Detecting cyberbullying content using machine learning (SVC, LR, NB, RF, SGD), this study aims to present ideas related to cyberbullying detection on the Twitter social media platform, and it can be concluded that the logistic regression classifier is the most accurate among all other classifiers, with an accuracy of 93% [16]. Designing and developing an effective technique to detect abusive and bullying messages online by combining natural language processing (NLP) and machine learning (DT, NB, SVM, RF), the results show that the SVM algorithm outperforms the others [17].

Based on previous research and aligned with its objectives, this study will develop the Cybersentinel application, a cyberbullying detection tool that combines Machine Learning approaches (NB, SVM, K-NN, DT, LR, RF) with the VADER Lexicon to enhance classification accuracy. This research is headed to compare various Machine Learning algorithms, optimized through the GridSearchCV technique, to find the best parameter combination. The application development follows the Design Science Research (DSR) approach, integrating principles, practices, and procedures, facilitating a focus on problem-solving and supporting the design and creation of the Cybersentinel application.

II. MATERIAL AND METHODS

This research was conducted using the Design Science Research (DSR) model approach [18]–[21] which includes the stages of problem identification, object identification, design and development, evaluation, and communication. Design Science Research (DSR) has emerged as an important methodology within the Information Systems field, characterized by its focus on the creation and evaluation of artifacts designed to address real-world problems. Moreover, the application of DSR is not limited to the context of Information Systems, it has been successfully adapted in various fields such as engineering, healthcare, and management. The iterative nature of DSR enables continuous refinement of artifacts based on empirical feedback, which is essential for achieving practical relevance. DSR methodology facilitates the development of systems that are responsive to user needs through a cycle of design, implementation, and evaluation. This adaptability is particularly important in rapidly evolving fields such as information technology, where user needs and technological capabilities are constantly changing. An overview of the DSR research design is presented in Figure 1, The figure explains that the stages in DSR include Problem

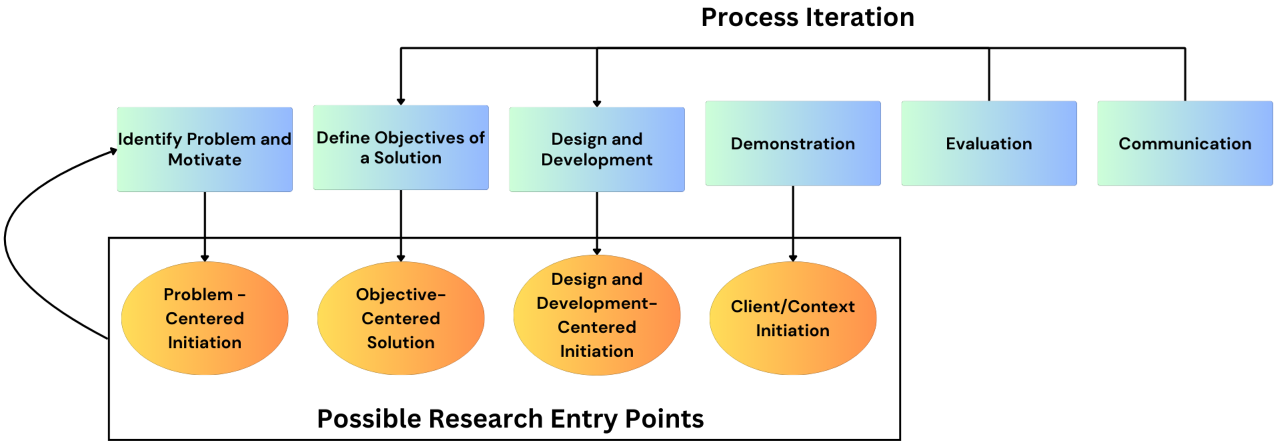

Process Iteration
Identify Problem and Motivate
Define Objectives of a Solution
Design and Development
Demonstration
Evaluation
Communication
Problem - Centered Initiation
Objective-Centered Solution
Design and Development-Centered Initiation
Client/Context Initiation
Possible Research Entry Points

FIGURE 1. Design Science Research (DSR) Model

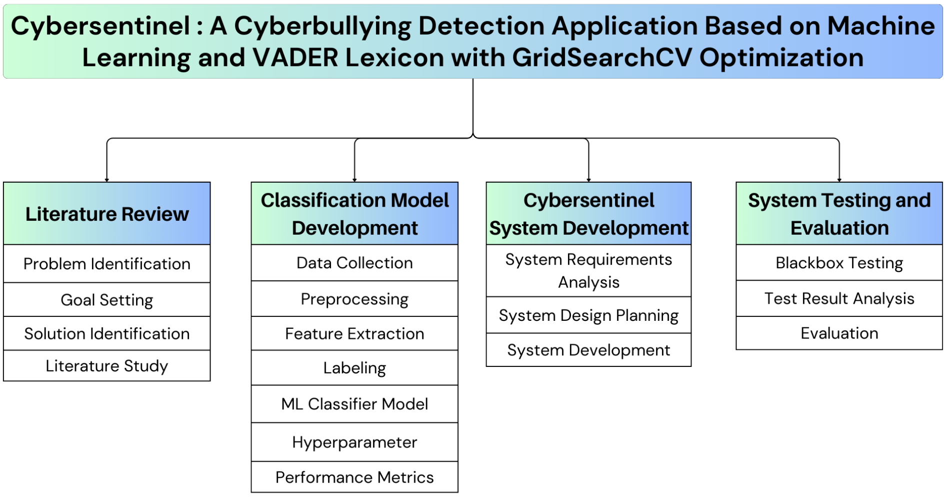


FIGURE 2. Research Stages

Identification and Motivation, Define the Objectives for a Solution, Design and Development, Demonstration, Evaluation, and Communication. This process is carried out iteratively, with each stage being repeated to refine or improve the artifact based on the results of the evaluation and feedback obtained, allowing for the enhancement or development of artifacts that are more relevant and aligned with the needs.

The stages of this research are then organized into four main parts, namely literature review, classification model development, system development, and system testing and evaluation, as shown in Figure 2. All these stages still follow and are based on the Design Science Research (DSR) model proposed by Peffersv [18].

At the literature study stage, defining and identifying the problem to be addressed, namely cyberbullying on social media which is a serious problem that requires an effective solution to detect it early. The use of machine learning algorithms in the big data era requires the right techniques to process, analyze, and extract information. Therefore, a technological solution is needed that can work quickly and efficiently in detecting cyberbullying behavior, so that prevention and intervention measures can be taken early. Formulating the clear and specific research objectives, it is proposing a text mining-based model to detect cyberbullying using machine learning algorithms, classifying cyberbullying behavior, and developing a system that can automatically detect and classify cyberbullying behavior.

At the Classification Model Development stage, collecting data in the form of user comments on social media regarding trending news that contains cyberbullying from three social media; there are: X (Twitter), Instagram, and YouTube. The data was obtained through a specific process on the three social media platforms. A total of 19,377 data points were collected. The available dataset will be pre-processed including case folding, cleaning text, normalization, tokenization, stop removal / stop words and stemming. In the step of preprocessing, especially text processing, the reason for using the aforementioned techniques is to improve data quality, to ensure data consistency, to reduce dimensionality, and to facilitate the analysis [22], [23]. Then the data will be feature extraction with TF-IDF technique, in this stage the data in the form of text will be converted into numbers [24]. The hyperparameter process using gridsearchCV is used to optimize parameters [25], [26]. Then, it divides the data into training data and testing data. Implement it into machine learning algorithms namely NB, SVM, K-NN, DT, LR and RF. Evaluate by testing the performance of the model. These six algorithms will be categorizing into the category of supervised learning, which is an efficient and accurate approach for text classification due to its ability to learn patterns from labeled datasets and apply them to unseen data [27]. The limitations in data collection and analysis include the use of different languages, slang, or informal expressions, which can make text analysis challenging and may result in errors in detecting the context within the comments. However, these issues can be addressed by applying stricter preprocessing to enhance data quality and consistency. At the System Development stage, analyzing system requirements, designing system designs followed by system construction. The System Testing and Evaluation stage is the last stage, namely testing the system by designing test scenarios, viewing and analyzing test results, then evaluating the test results.

A. CLASSIFICATION MODEL DEVELOPMENT

This research uses a dataset of 19,377. Data is taken randomly on social media, in the form of comments from social media, with details as follows: 1,554 from Twitter, 4,845 from Instagram, and 12,978 from YouTube, the amount of data can be seen in TABLE 1. The time span for data collection is from January 2024 to June 2024. Then the collection of comments is integrated into a csv dataset. Furthermore, the data is processed using Python. The collected dataset will be labelled bullying or nonbullying using the VADER Lexicon technique. VADER is used to determine negative, neutral, positive, and combined polarity scores for each sentiment. A combined score less than or equal to -0.05 is considered as negative polarity, while a score greater than or equal to 0.05 is considered as positive polarity. Values between 0.05 and -0.05 are considered neutral [28]. The sample data that has been processed using the VADER Lexicon technique can be seen in TABLE 2. The first data represents a bullying sentiment, while the second data represents a non-bullying sentiment.

TABLE 1
Sample Dataset Used in the Research

Source	Data Total
Twitter (x)	1554
Instagram	4845
YouTube	12978

TABLE 2
Sample dataset processed using VADER Lexicon technique

Comment	Sentiment
The father who died was punished in the grave because he had a son who was a sinner.	Bullying
You judge Depe, then why on Earth the other site keeps wiggling and wearing sexy clothes also; your brain is just dirty; you know it is not Depe's fault either.	Non-Bullying

Before applying VADER, a preprocessing process is performed. Preprocessing is an important stage in data analysis that aims to prepare raw data to be used by machine learning algorithms. The steps taken in preprocessing include data cleaning, such as removal of punctuation marks, stop words, symbols, as well as text normalization by converting all letters to lowercase. Tokenization, stemming, and normalization are also often used to break text into smaller units and unify words. This process ensures that the data being processed is clean, consistent, and in the right format to improve the accuracy and efficiency of the machine learning model to be applied.

It will then perform feature extraction on the dataset with the TF-IDF technique, which is one of the important methods in Natural Language Processing (NLP) to extract features from text and measure the importance of words in a document relative to a collection of other documents. The concept combines two metrics: Term Frequency (TF), which counts how often a word appears in a document, and Inverse Document Frequency (IDF), which counts how rarely the word appears in the entire document. A high TF indicates that the word appears frequently in a particular document, while a high IDF indicates that the word is rarely found in many documents, so the word is considered more informative. The combination of these two metrics results in a TF-IDF value that emphasizes words that have specific meaning in the context of the document. TF-IDF is often used in various applications such as information retrieval, text classification, and sentiment analysis, where an understanding of the importance of words in a particular context is necessary.

Term Frequency (TF) is to measure how often a word appears in a document. The Eq. (1) [24] for TF-IDF is as follows:

$$F\ (t,d) = \frac{\text{Number of times terms } t \text{ appears in document } d}{\text{Total number of terms in document } d} \tag{1}$$

Inverse Document Frequency (IDF) is to measure how rarely a word appears throughout the document (Eq. (2)).

$$IDF\ (t,d) = \log\left(\frac{N}{1+DF(t)}\right) \tag{2}$$

After calculating TF and IDF, we can combine them to get a value of TF-IDF (Eq. (3)).

$$TF - IDF(t,d,D) = TF(t,d)\ x\ IDF(t,D) \tag{3}$$

where t represents the term, d represents the document, D represents the entire corpus of documents, N represents Total number of documents in the corpus.

The dataset is pre-processed first before being applied to the machine learning algorithm. Sentiments with neutral values will be removed, only sentiments that have bullying and nonbullying labels will be processed and used for machine learning algorithms [29]. The data will be split into two sets: training and testing data. The training data is used to train the model, allowing the algorithm to learn from the patterns and characteristics of the data. The testing data, which remains invisible to the model during training, is used to test how well the model can generalize on new data that it has never encountered before. The data split was done with a ratio of 80:20, which means 80% for the training process and 20% for the testing process. This split process is done to optimize model parameters and prevent overfitting. Using 10-fold cross validation and six machine learning algorithms including NB, SVM, K-NN, DT, LR, RF. The performance of each model will be evaluated using confusion matrix, such as accuracy, precision, recall, and F1-score [30].

GridSearchCV helps in automating the process of finding the best parameters to maximize the performance of machine learning models. GridSearchCV works with the basic concepts of cross-validation and parameter grid exploration. GridSearchCV works by defining a grid, which is a range of possible values for each hyperparameter to be adjusted. The method then evaluates all combinations of the defined values using cross-validation techniques, such as k-fold cross-validation, to determine the combination that gives the best performance on the training data. The result is an optimized model with the most suitable hyperparameters for a particular dataset. GridSearchCV helps in automating the process of finding the best parameters to maximize the performance of machine learning models. GridSearchCV works with the basic concepts of cross-validation and parameter grid exploration. Some mathematical elements underlie the way GridSearchCV works. Suppose there are n parameters to be tuned, and each parameter has multiple value options (Eq. (4)).

Where p_i is the number of possible values for the ith parameter.

$$Total\ Combinations = P_1\ x\ P_2\ x\ P_{...}\ x\ P_n \tag{4}$$

The data is divided into k folds. For each parameter combination, GridSearchCV performs k iterations where the model is trained on k-1 folds and tested on the remaining folds. Then, the equation [27] for GridSearchCV is as follows (Eq. (5)):

$$Average\ Score = \frac{1}{k}\sum_{i=1}^{k} Score_i \tag{5}$$

Average Score is the average value of all evaluation scores obtained from each fold during the training and testing process. This average is used to get an overall picture of how well the model performs with a particular parameter combination. Where $\frac{1}{k}$ is a scaling factor that divides the total score by the number of folds k to obtain the average. $\sum_{i=1}^{k} \mathrm{Score}_i$ is the sum of all evaluation scores (Score_i) obtained from each fold from 1 to k. Each Score_i represents the evaluation result of the model on the ith fold.

B. SYSTEM DEVELOPMENT

Models that have been trained with relevant datasets are hosted on servers or cloud services that support machine learning frameworks. One of them, streamlet, is a Python framework that allows the creation of interactive web applications quickly and easily, especially for data science and machine learning purposes. Users can input data through this interface, which is then processed and fed to the model to generate predictions. Streamlet enables live visualization of prediction results, and can integrate additional features such as graphs or tables to help users understand model performance. This deployment offers a fast and effective solution to visualize and utilize models in a real-time context.

C. SYSTEM TESTING

The testing method performed is black-box or functionality type. Testing must be done from the client side or front-end. This process will check whether the system functions according to predetermined specifications or needs. This test aims to ensure that all features work correctly, detect errors in functionality, and verify that the system behaves in accordance with user expectations. Test cases are carried out for the case of checking comments in the form of text and documents, will be checked whether they fall into bullying or non-bullying. The black-box method is used in the final stage of development to ensure product readiness before release.

III. RESULT

A. THE ADVANCED PROCESSING AND DATASET ADJUSTMENT

TABLE 3 shows some words with the highest TF-IDF values in each document. It can be seen that the larger the value produced by TF-IDF, the more important those words become in the document based on their frequency in the document and the whole corpus.

TABLE 3
Sample Calculation Results using TF-IDF

No.	Document	Word	TF-IDF
1	Doc1	good	0.707
1	Doc1	new	0.707
2	Doc2	good	0.353
2	Doc2	more	0.353
2	Doc2	neat	0.353
3	Doc3	good	0.236
3	Doc3	body	0.236
4	Doc4	beautiful	0.5
4	Doc4	correct	0.5

No.	Document	Word	TF-IDF
5	Doc5	beautiful	0.5
5	Doc5	style	0.5
...	...	...	...

The explanation of TABLE 3 is column of 'No' to indicate the sequence number of the document in the document list. Document column, is a snippet or summary of the document. Word column, presents the important words or terms calculated in the TF-IDF process. TF-IDF Shows the TF-IDF value calculated for each word in each document.

B. CLASSIFICATION MODEL

The data is processed using the Python programming language. Data obtained from social media, namely Twitter, Instagram and YouTube, with a total of 19,377 data. Data is grouped using VADER Lexicon into bullying and non-bullying sentiments. Classification results using VADER with a total of 19,377 data, obtained as much as 98% with bullying and non-bullying labels as much as 2%, can be seen in FIGURE 3. Next, data preprocessing is carried out. Data is split into training data and testing data with a data ratio of 80:20. Then, word weighting is done using TF-IDF. Next, the application of machine learning algorithms including SVM, K-NN, NB, DT, LR, RF on data that has been completed through the preprocessing and weighting process. The training data is validated using a 10-fold cross-validation technique. The evaluation process uses Confusion Matrix which serves to visualize the performance results of an algorithm.

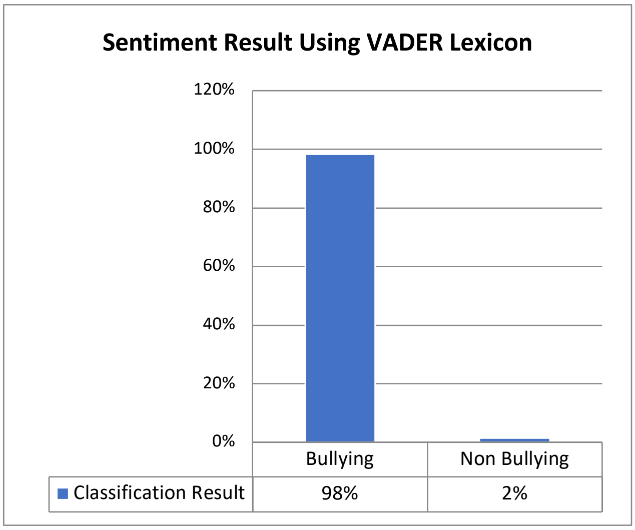


FIGURE 2. The Percentage of Sentiment Results Using VADER Lexicon

Based on the results of the experiments, it is known that the model with the SVM algorithm has the highest accuracy compared to the other five machine learning algorithms. The best model produced was named HyVADSVM (Hybrid Valence Aware Dictionary and Sentiment Reasoner with SVM). The model is a combination of the use of the Vader Lexicon, the SVM algorithm, and the GridSearchCV technique.

The parameters adjusted in the SVM algorithm are Kernel, C, and Gamma. For the Kernel parameter, we adjusted four types of kernels: linear, rbf, poly, and sigmoid. For the C parameter, we adjusted the range of values from [0.1 – 20], and for the Gamma parameter, we adjusted two types: Scale and Auto. The best values obtained from the parameter adjustments are Kernel = linear, C = 1, and Gamma = scale, as shown in TABLE 4.

The results of the model evaluation before and after parameter tuning are presented in TABLE 5, where it can be seen that the highest value was found in the SVM algorithm, followed by RF in second place, RF again in third place, DT in fourth place, KNN in fifth place, and NB in sixth place. A comparison of the model evaluations of the six algorithms before parameter tuning is presented in FIGURE 4, and a comparison of the model evaluations of the six algorithms after parameter tuning is presented in FIGURE 5. Thus, based on this figure, parameter tuning using the GridSearchCV technique has proven to be an effective way to improve classification performance.

TABLE 4
Tunning Parameters of GridSearchCV On SVM

Parameter	Range/ Type	Best Parameter
Kernel	[linear, rbf, poly, sigmoid]	linear
C	[0.1 – 20]	1
Gamma	[scale, auto]	scale

TABLE 5
Evaluation Results of Pre and Post Tunning Parameters

ML	Accuracy Score (%)		Precision Score (%)		Recall Score (%)		F1-Score (%)	
	Before	After	Before	After	Before	After	Before	After
SVM	**98.80**	**98.83**	**98.60**	**98.78**	**98.40**	**98.83**	**98.60**	**98.62**
KNN	98.23	98.33	97.82	98.18	98.23	98.33	97.81	97.81
NB	98.15	98.23	97.96	98.00	98.15	98.23	97.41	97.61
LR	98.15	98.65	98.19	98.58	98.15	98.65	97.37	98.34
DT	98.46	98.62	98.25	98.49	98.46	98.62	98.30	98.34
RF	98.78	98.80	98.70	98.75	98.78	98.80	98.56	98.58

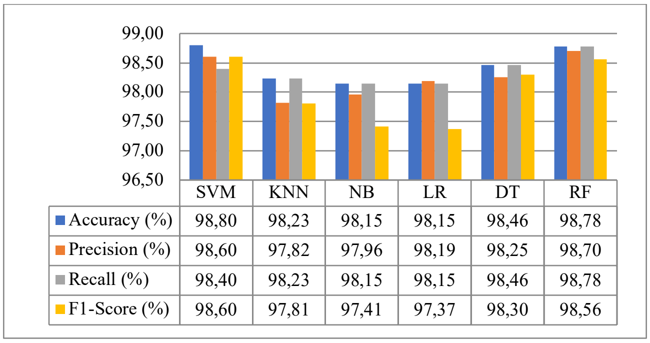
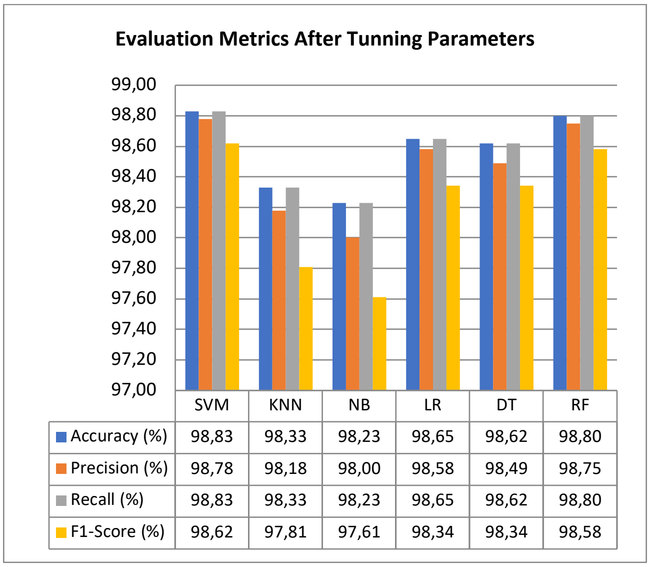


FIGURE 3. Evaluation of the six algorithms before tuning parameter



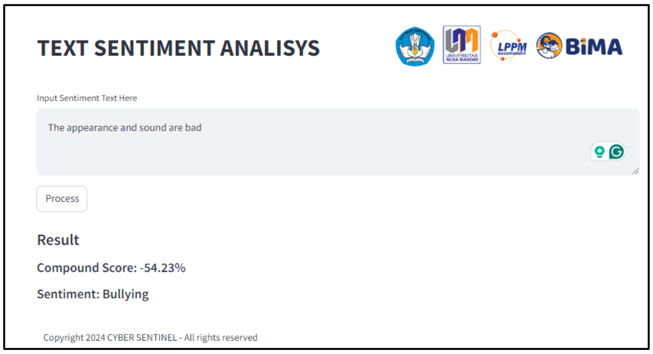
	SVM	KNN	NB	LR	DT	RF
🟦 Accuracy (%)	98,83	98,33	98,23	98,65	98,62	98,80
🟧 Precision (%)	98,78	98,18	98,00	98,58	98,49	98,75
⬜ Recall (%)	98,83	98,33	98,23	98,65	98,62	98,80
🟨 F1-Score (%)	98,62	97,81	97,61	98,34	98,34	98,58

FIGURE 5. Evaluation Of The Six Algorithms After Tuning Parameter

IV. DISCUSSION

A. IMPLEMENTATION OF MODELS IN APPLICATIONS

Based on the experimental results conducted using six machine learning algorithms, it was found that SVM achieved the highest scores with an accuracy of 98.83%, precision of 98.78%, recall of 98.83%, and an F1-Score of 98.62%. The name of this best-performing model was HyVADSVM (Hybrid Valence Aware Dictionary and Sentiment Reasoner with SVM). The use of GridSearchCV proved to improve the performance of the SVM model. The model was then integrated into a web-based cyberbullying detection system called the Cybersentinel application. This application is designed to detect and classify sentiments containing cyberbullying on social media. The application interface can be seen in FIGURE 6, which shows a page for analysing text where users are able to input sentences or upload documents (PDF or DOC). The system then displays the classification results, indicating whether the text or document contains bullying or non-bullying content.


TEXT SENTIMENT ANALISYS

Input Sentiment Text Here

The appearance and sound are bad

Process

Result

Compound Score: -54.23%

Sentiment: Bullying

Copyright 2024 CYBER SENTINEL - All rights reserved

FIGURE 6. Cybersentinel Application Interface Page View

TABLE 6, which compares the experimental model with models from other researchers, concludes that the machine learning, TF-IDF, and GridSearchCV approach delivers the best results.

TABLE 6
Comparison with other existing system

	Best ML Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Zhao et.al [31]	SVM	77.80	77.80	76.60	75.60
Van Hee et.al [32]	SVM	78.50	73.32	57.19	64.26
R.Shah et.al [16]	LR	93.00	91.00	96.00	93.00
K. Alam et.al [14]	LR	94.00	93.00	94.00	93.00
A.Ali et.al [15]	SVM	79.30	79.00	78.10	79.30
Our Algorithm	**SVM**	**98.83**	**98.78**	**98.83**	**98.62**

The limitations of this study include changes in user behaviour on social media, which may affect the ability of models trained on a single dataset to adapt to emerging trends and new languages. Additionally, while the use of techniques like GridSearchCV can provide advantages in finding optimal parameters, it is important to note that this process may require more computation time and resources, especially when working with very large datasets. Furthermore, the primary focus on text analysis can present challenges, considering that cyberbullying is able to occur in various formats, including text, images, and videos. The potential implications of this study's findings for stakeholders are significant. For educators, this application can serve as a valuable tool to identify and address cases of cyberbullying in school environments, fostering a safer learning atmosphere. For parents, Cybersentinel provides insights into their children's behavior on social media, enabling them to take appropriate action if necessary. Additionally, for policymakers, the findings offer a basis for developing better policies to prevent and address cyberbullying, as well as supporting initiatives aimed at raising public awareness about this issue. In this way, the application not only provides a practical solution to the problem of cyberbullying but also contributes to broader efforts to create a safer and more supportive digital environment for all users.

B. FUNCTIONAL TESTING

Functionality, it tests in serving to detect software failures so that defects can be repaired and corrected in the early phase, and to ensure that the produced product functions according to specifications and meets customer needs [33]. In general, the system workflow begins with the receipt of a response in the form of a message. The messages are then processed and analysed to determine whether or not they contain elements of cyberbullying. This analysis was performed by applying the HyVADSVM model. This process includes using appropriate inputs, generating expected outputs, and comparing actual results with predictions to ensure all features in the detection system are functioning properly. The system's functionality flow can be seen in FIGURE 7. In addition, FIGURE 8 shows the block diagram of the Cybersentinel system, which serves to provide a clear and

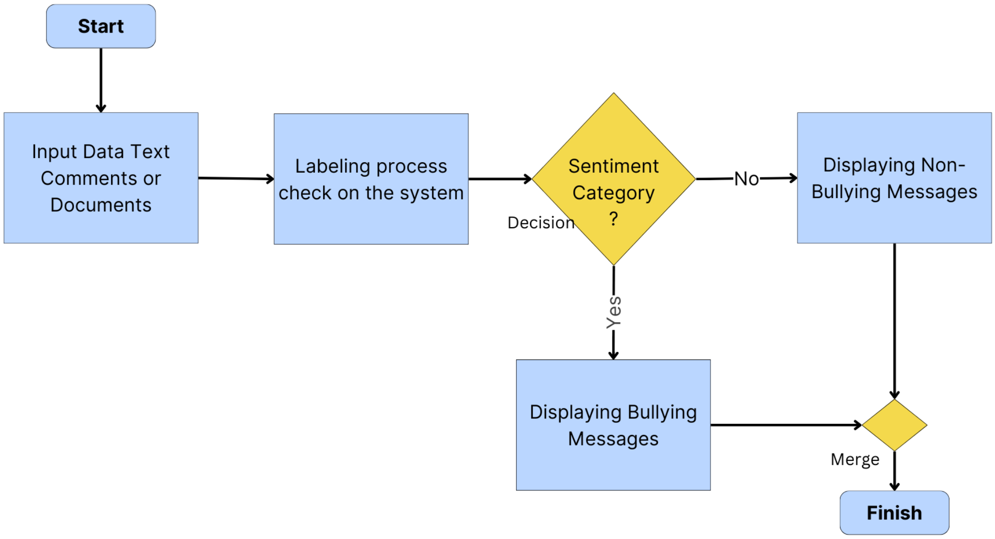

Start
Input Data Text Comments or Documents
Labeling process check on the system
Sentiment Category ?
Decision
No
Displaying Non-Bullying Messages
Yes
Displaying Bullying Messages
Merge
Finish

FIGURE 7. System Functionality Flow

easy-to-understand overview of the workflow or structure of the system

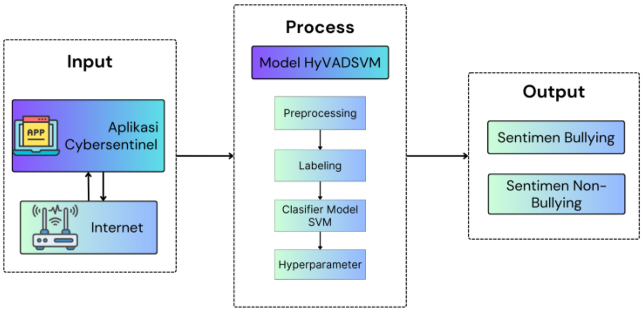


FIGURE 8. Block Diagram of The Cybersentinel System

Based on these conditions, the test is carried out according to the scenario. The results showed that the system can correctly detect the type of messages that contain elements of cyberbullying. A recapitulation of the results of the functionality tests obtained can be seen in TABLE 7.

TABLE 7
System Testing Results using Black-Box on the Text Sentiment menu

Testing Scenarios	Test Case	Expected Results	Test Results	Conclusion
Input messages that contain elements of cyberbullying (such as expletives and reproaches)	Input: [The appearance and sound are bad.]	The system will display the classification results with the message "Bullying"	Conforms to Expectations	Valid
Sending messages that do not contain elements of cyberbullying (such as non-reprehensible words)	Input: [She is beautiful and amazes everyone.]	The system will display the classification results with the message "Non-Bullying"	Conforms to Expectations	Valid

IV. CONCLUSION

This study aims to develop Cybersentinel, an application for detecting cyberbullying that combines Machine Learning and sentiment analysis using the VADER Lexicon to improve classification accuracy. The application has shown significant progress in identifying cyberbullying on social media with high accuracy. Experiments conducted on six algorithms—SVM, KNN, NB, LR, DT, and RF—revealed that the SVM algorithm achieved the highest accuracy compared to the others. Before parameter tuning, the SVM model showed an accuracy of 98.80%, and after tuning, it increased to 98.83%, reflecting a 0.03% improvement. Precision increased from 98.60% to 98.78% (a 0.18% improvement), recall from 98.40% to 98.83% (a 0.43% improvement), and the F1-Score from 98.60% to 98.62% (a 0.02% improvement). These enhancements led to more accurate and efficient detection of cyberbullying behavior.

The functional testing, using the black-box method, demonstrated precise category predictions, confirming that Cybersentinel can effectively identify language patterns indicative of cyberbullying and provide practical solutions for the ever-evolving issue on social media platforms. This application not only contributes theoretically to cyberbullying detection but also offers a practical tool that can enhance the security and well-being of social media users.

The success of this application opens avenues for further research in this field, with potential to continually refine and adapt the technology to meet new challenges in the digital environment. Future research recommendations include increasing the dataset size to further improve model accuracy. Additionally, adding features that allow users to provide feedback on prediction accuracy can help grow the dataset, leading to more accurate and adaptive models in response to evolving language dynamics and cyberbullying patterns on social media.

ACKNOWLEDGMENT

The author expresses deep gratitude to the Direktorat Riset, Teknologi, dan Pengabdian kepada Masyarakat (DRTPM), Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, for the funding support that enabled this research to be carried out. Saying thanks are also addressed to the research colleagues, and to the entire team who have provided direction and support throughout the research process. Also say thanks to the family and friends who have provided moral support, as well as all parties who contributed to the completion of this study.

REFERENCES

[1] A. Muneer, "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter," *Electronics Journal*, vol. 10, no. 22, pp. 1–20, 2020.

[2] UNICEF, "Cyberbullying: What is it and how to stop it," *UNICEF.Org*, 2022. .

[3] C. V Baccarella, T. F. Wagner, J. H. Kietzmann, and I. P. Mccarthy, "Social media ? It ' s serious ! Understanding the dark side of social media," *European Management Journal journal*, vol. 36, pp. 2017–2019, 2018.

[4] T. K. Balaji, C. Sekhara, R. Annavarapu, and A. Bablani, "Machine Learning Algorithms for Social Media Analysis : A Survey," *Computer Science Review*, vol. 40, no. 100395, pp. 1–32, 2021.

[5] S. Ray, "A Quick Review of Machine Learning Algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 35–39.

[6] D. Sultan *et al.*, "A Review of Machine Learning Techniques in Cyberbullying Detection," *Tech Science Press*, vol. 74, no. 3, pp. 5625–5640, 2022.

[7] S. Salawu, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying : A Survey," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, vol. 3045, no. c, pp. 1–20, 2017.

[8] E. Olshannikova, T. Olsson, J. Huhtamäki, and H. Kärkkäinen, "Conceptualizing Big Social Data," *Journal of Big Data*, 2017.

[9] M. Dreier, M. E. Beutel, E. Duven, and S. Giralt, "A hidden type of internet addiction ? Intense and addictive use of social networking sites in adolescents," *Computers in Human Behavior*, vol. 55, pp. 172–177, 2016.

[10] M. Khairy, T. M. Mahmoud, and T. Abd-El-Hafeez, "The effect of rebalancing techniques on the classification performance in cyberbullying datasets," *Neural Computing and Applications*, vol. 36, no. 3, pp. 1049–1065, 2024.

[11] M. Hamlett, G. Powell, Y. N. Silva, and D. Hall, "A Labeled Dataset for Investigating Cyberbullying Content Patterns in Instagram," in *Proceedings of the Sixteenth International AAAI Conference onWeb and Social Media (ICWSM 2022)*, 2022, pp. 1251–1258.

[12] S. Ernawati, R. Wati, N. Nuris, and L. S. Marita, "Comparison of Naïve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application Comparison of Na ̈ ıve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application."

[13] S. Ernawati, "Implementation of The Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies," in *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, 2018, pp. 1–5.

[14] K. S. Alam, S. Bhowmik, and P. R. K. Prosun, "Cyberbullying Detection : An Ensemble Based Machine Learning Approach," in *Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021)*, 2021, pp. 710–715.

[15] A. Ali and A. M. Syed, "Cyberbullying Detection Using Machine Learning," *Pakistan Journal of Engineering and Technology, PakJET*, vol. SI, no. 01, pp. 45–50, 2020.

[16] R. Shah, S. Aparajit, R. Chopdekar, and R. Patil, "Machine Learning based Approach for Detection of Cyberbullying Tweets," vol. 175, no. 37, pp. 52–57, 2020.

[17] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, "Cyberbullying Detection on Social Networks Using Machine Learning Approaches," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020, pp. 1–6.

[18] K. Peffers *et al.*, "The Design Science Research Process: A Model For Producing And Presenting Information Systems Research," in *First International Conference on Design Science Research in Information Systems and Technology*, 2020, pp. 1–24.

[19] J. R. Venable, J. Pries-heje, and R. L. Baskerville, "Choosing a Design Science Research Methodology," in *Australasian Conference on Information Systems 2017*, 2017, pp. 1–11.

[20] C. Lawrence, T. Tuunanen, and M. D. Myers, "Extending Design Science Research Methodology for a Multicultural World," in *IFIP Advances in Information and Communication Technology*, 2020, no. March, pp. 112–126.

[21] J. Q. Azasoo, "A Retrofit Design Science Methodology for Smart Metering Design in Developing Countries," in *15th International Conference on Computational Science and Its Applications (ICCSA)*, 2015, no. June.

[22] A. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 16, no. 6, pp. 22–32, 2018.

[23] C. P. Chai, "Comparison of Text Preprocessing Methods," *Natural Language Engineering*, vol. 29, no. 3, 2023.

[24] M. Chiny, M. Chihab, and Y. Chihab, "LSTM , VADER and TF-IDF based Hybrid Sentiment Analysis Model," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, pp. 265–275, 2021.

[25] T. Yan, S.-L. Shen, A. Zhou, and X. Chen, "Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm," *Journal of Rock Mechanics and Geotechnical Engineering*, vol. 14, no. 4, pp. 1292–1303, 2022.

[26] R. G. S. K, A. K. Verma, and S. Radhika, "K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries," in *2019 5th International Conference for Convergence in Technology (I2CT)*, 2019, pp. 9–13.

[27] A. I. Kadhim, "Survey on Supervised Machine Learning Techniques," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 273–292, 2019.

[28] T. Pano and R. Kashef, "A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19," *MDPI-Big Data and Cognitive Computing*, vol. 4, no. 33, pp. 2–17, 2020.

[29] V. D. Chaithra, "Hybrid Approach : Naive Bayes and Sentiment VADER for Analyzing Sentiment of Mobile Unboxing Video Comments," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, pp. 4452–4459, 2019.

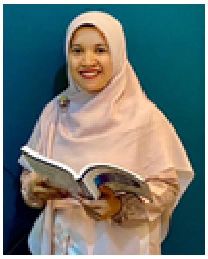
[30] E. P. Costa, C. Postal, A. C. Lorena, R. S. Ad, C. Postal, and A. A. Freitas, "A Review of Performance Evaluation Measures for Hierarchical Classifiers," *Association for the Advancement of Artificial Intellegence*, pp. 1–6, 2007.

[31] A. Zhou, "Automatic Detection of Cyberbullying on Social Networks based on Bullying Features," in *Proceedings of the 17th international conference on distributed computing and networking*, 2016, pp. 1–6.

[32] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, and B. Desmet, "Automatic Detection and Prevention of Cyberbullying," in *International Conference on Human and Social Analytics (HUSO 2015)*, 2015, pp. 1–6.

[33] M. Kumar, S. K. Singh, and D. R. K. Dwivedi, "A Comparative Study of Black Box Testing and White Box Testing Techniques," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 10, no. 10, pp. 32–44, 2015.

AUTHOR BIOGRAPHY



Siti Ernawati is a lecturer at Universitas Nusa Mandiri, where she has been teaching since 2015. She completed her graduate studies at Universitas Nusa Mandiri, deepening her knowledge and expertise in data science and information systems. She has a strong background in research and development in the field of information technology. Her research focuses on data science, involving the analysis of big data and the use of statistical analysis techniques to generate insights useful for decision-making. Additionally, she is also involved in information systems, including the development and implementation of technology-based systems to enhance organizational efficiency and effectiveness.
ORCID : https://orcid.org/0000-0002-0086-7320

Frieyadie is a lecturer at Universitas Nusa Mandiri, where he has been teaching since 2008. He completed his graduate studies at Universitas Nusa Mandiri and has taught various courses such as Object-Oriented System Modeling (PSBO) and Database Modeling at Universitas Nusa Mandiri, Jakarta. Holding a Master’s degree in Computer Science and currently pursuing doctoral studies at Universiti Kuala Lumpur, he is also actively involved in writing programming books and participating in various research and international conferences. Additionally, Frieyadie has extensive teaching experience across different areas of information technology. His research focuses on data science and information systems.
ORCID : https://orcid.org/0000-0002-8282-0672

Eka Rini Yulia is a lecturer at Universitas Nusa Mandiri, where she has been teaching since 2015. Her research interests lie in the fields of information systems and decision support systems (DSS). With a strong focus on these areas, she is dedicated to exploring how technology can be leveraged to improve organizational decision-making and system efficiency. Her work involves both theoretical research and practical applications, contributing to advancements in how information systems are designed and utilized in various organizational contexts.
ORCID : https://orcid.org/0000-0002-3851-2066