[1]Mr. Mritunjaykumar Ojha

[2]Dr.Nilesh M. Patil

[3]Dr.Manuj Joshi

Assessment of Classification Models for Identifying Cyberbullying Detection

Abstract: - The most critical challenge in cybersecurity is dealing with cyberbullying. The increase in the complicated dynamics of social media, which are marked by their complexity, variety, subjectivity, and multimodal nature, provide obstacles to the identification of cyberbullying. This has led to the need for automated mechanisms that can identify these harmful behaviors. This study aims to assess how well various categorization methods detect cyberbullying. For training and testing, our study uses cybersecurity-related data. The models that have been selected include the Linear SVC, Random Forest, Decision Tree, Logistic Regression, and Stochastic Gradient classifiers. We use hyperparameter tuning to improve the performance of the model,, and then we show the results based on important metrics like "accuracy, precision, recall, and F1 score." The end results highlight that Stochastic Gradient classifier is superior in performance, which has an F1 score of 94.39%, recall of 91.94%, accuracy of 92.81%, and precision of 96.97%. The investigation examines the advantages and disadvantages of each approach, offering insightful information for the cybersecurity field. In addition, suggestions for more studies are made to strengthen the resilience of cyber defenses. This work advances the effectiveness of cybersecurity measures by finding the best models for detecting threats and offering directions for improvement as cyber threats change. Other feature extraction techniques—"Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and Word2Vec"—are merged with the algorithms of "Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF)" to build the model [8]. Our objectives are to analyze the effectiveness of several classification techniques for identifying cyberbullying, such as "Random Forest, Decision Tree, Linear SVC, Logistic Regression, and Stochastic Gradient classifiers," to improve the performance of the model by using Hyperparameter Tweaking methods and analyze the outcomes using the F1 score, accuracy, precision, recall, and other critical performance.

Keywords: Cyberbullying, Tweet, Spam

I. Introduction

With the quick development of communication technologies like the internet, smartphones, and personal digital assistants (PDAs), cyberbullying has become a modern problem [13]. This plagues online communities and causes emotional distress and psychological harm to victims. Traditional methods of monitoring online interactions are often time-consuming and ineffective due to the vast amount of user-generated content. Automated detection of cyberbullying messages can offer a faster and more scalable solution to curb online harassment. However, the dynamic nature of language used in cyberbullying, including sarcasm, slang, and emojis, makes it challenging for machines to accurately identify harmful content. Another important issue that presents hurdle in detection is that it happens in a brand-new virtual environment. Cyberbullying is more damaging and upsetting because of the anonymity or difficulties in tracking down online exchanges. Another study proposes that given the gravity of cyberbullying, schools, parents, and kids must recognize the possible harm that cyberbullying may do and stress that it can be just as harmful as physical bullying, especially in light of the growing prevalence of social media use [4]. This journal proposes an approach based on a dataset of actual tweets—that is, question-answer pairings between victims and cyber predators—using various detection methods. Another study considered a dataset benchmark for "Offensive Language Identification Dataset (OLID)," to demonstrate the automated detection technique and how it gets integrated into the web applications [1].

[1] Research Scholar, PAHER University Udaipur, India

mritunjayojha11@gmail.com

[2]Associate Professor, D J Sanghvi College of Engineering, Mumbai, Maharashtra, India

nileshdeep@gmail.com

[3]Associate Professor PAHER University Udaipur, India

manujjoshi@gmail.com

This publication suggests a novel strategy in response to the pressing need to comprehend and mitigate cyberbullying. The study focuses on question-answer pairs between victims and cyber predators and aims to offer a thorough understanding of the dynamics of cyberbullying at work through various techniques. The research aims to provide valuable insights that guide preventative and intervention measures in the face of this dynamic problem by exploring the subtleties of online interactions.

We must adjust and deepen our knowledge of cyberbullying as it develops, acknowledging the similarities between it and physical bullying. The suggested study, based on actual data, is an essential step in understanding the intricacies of cyberbullying and creating a more secure online environment for people of all ages.

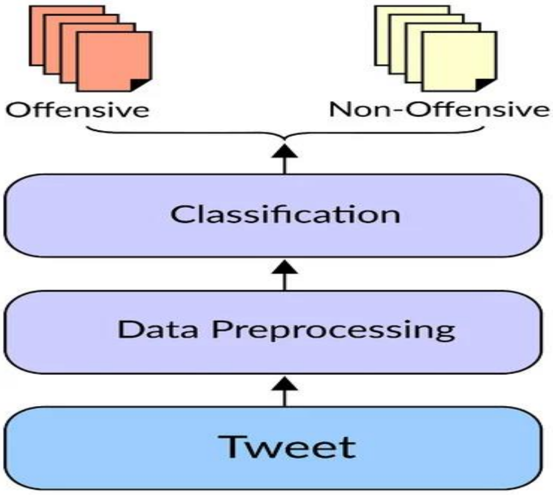


Figure 1: Data processing for classification

II. Methodology

Our approach is modified for tweet-based spam identification by using a preprocessed and properly divided dataset of tweets into training and testing subsets. To perform the detection process, we carefully carried out several deliberate actions. First, we prepared the dataset for analysis through thorough preprocessing, including operations like data cleaning, normalization, and feature extraction. It allowed us to guarantee data quality and consistency. The preprocessed dataset was then divided into separate subsets of “training and testing” to preserve the model’s training procedure integrity integrity and allow for the independent assessment of model performance. By utilizing the TF-IDF Vectorizer we could transform the Twitter text into numerical representations, making it easier for machine learning algorithms to understand both during training and assessment. TF-DIF assigns weight to words based on two key factors:

1. How frequently a particular word appears within a single tweet (document). Words that occur more often tend to be given higher TF values.

2. How rare a word is across the entire dataset (corpus) of tweets. Words that are uncommon throughout the corpus receive higher IDF weights

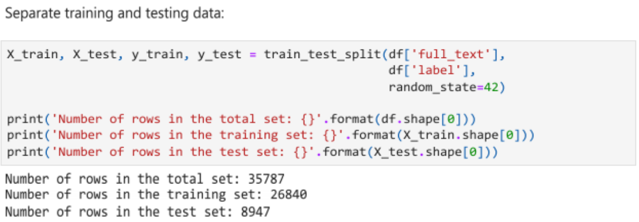
By combining these factors, TF-IDF effectively highlights keywords that are distinctive to cyberbullying messages. These keywords might include words like "loser," "hate," or threats, which are likely to have a high TF within a cyberbullying tweet but a low IDF across the entire dataset (as they're not common in general communication). Essentially, TF-IDF helps the machine learning models focus on the most relevant terms for identifying cyberbullying content. Moreover, we used various classification models, including Random Forest, Decision Trees, Linear SVC, Logistic Regression, and Stochastic Gradient classifiers. We used the classifier since the accuracy of a prediction model's classifier determines how resilient the model is, highlighting the critical role accuracy plays in guaranteeing the system's success [6]. We use TF-IDF Vectorizer to convert the tweet content into numerical representations, which we then use to facilitate the “training and testing stages” of multiple classification models, including the “Linear SVC, Random Forest, Decision Tree, Random Forest classifier, and Stochastic Gradient classifier.” By using the powerful GridSearchCV technique, Hyperparameter Tweaking is used to optimize the model using the powerful GridSearchCV technique. Evaluation criteria that thoroughly assess each model's performance in tweet-based spam detection include “accuracy, precision, recall, and F1 score.” To provide a focused approach to optimization, we explicitly use a make_scorer to generate the F1 score during Hyperparameter Tweaking .

We use a systematic approach to handle Twitter datasets to detect tweet spam. We guarantee the fidelity of model training by meticulous separation into training and testing sets and comprehensive preprocessing. By converting Twitter content into numerical representations, TF-IDF Vectorizer makes it possible to train and evaluate different categorization algorithms. We applied the Jordan, 2017 ideas to search the hyperparameter space haphazardly in an attempt to find the values that yield the highest score. These models are optimized by Hyperparameter Tweaking , which is expertly carried out with GridSearchCV. Metrics such as “accuracy, precision, recall, and F1 score” are included in performance evaluation. Crucially, we add a degree of specificity to our evaluation by using a make scorer specifically designed to compute the F1 score during Hyperparameter Tweaking. This systematic and exhaustive methodology ensures a complete analysis of every model's capacity to detect and block spam in the context of tweets. Several phases are included in our strategy to optimize the classification models for cyberbullying detection that enhances the performance of the model. To guarantee the integrity of model training, we start by preprocessing the dataset and splitting it into training and testing subsets. The tweet content is then transformed into numerical representations using the TF-IDF Vectorizer, which makes it easier to train and assess various categorization algorithms.

We use the potent GridSearchCV approach to do Hyperparameter Tweaking on the models to enhance their performance. Criteria, including “accuracy, precision, recall, and F1 score,” evaluated the models' performance. This model selection process ought to come after data processing chores and also comparing the performance of tuned and untuned models is needed [15]. By adjusting the hyperparameters of each classification model, we may maximize its performance on the provided dataset. To be more precise, we tune regularization strength, learning rate, and kernel type, among other factors, to get the best setup for every model.

Moreover, interpreting these models' predictions requires understanding how they make decisions. This work examines how each model makes decisions by examining the characteristics and trends that influence the categorization results. Our methodology aimed to guarantee the dependability and resilience of the classification models employed in tweet-based spam identification. We carefully adhered to a systematic procedure comprising many crucial phases. We first preprocessed the Twitter dataset thoroughly to guarantee data consistency and integrity. The dataset was then divided into training and testing subsets using strict methods to stop data leaks and preserve the accuracy of our analysis.

#GridSearchCV #PerformanceMetrics #MachineLearningModels #DataPreprocessing #TweetSpamDetection



Separate training and testing data:

```
X_train, X_test, y_train, y_test = train_test_split(df['full_text'],
                                                    df['label'],
                                                    random_state=42)

print('Number of rows in the total set: {}'.format(df.shape[0]))
print('Number of rows in the training set: {}'.format(X_train.shape[0]))
print('Number of rows in the test set: {}'.format(X_test.shape[0]))
```

```
Number of rows in the total set: 35787
Number of rows in the training set: 26840
Number of rows in the test set: 8947
```

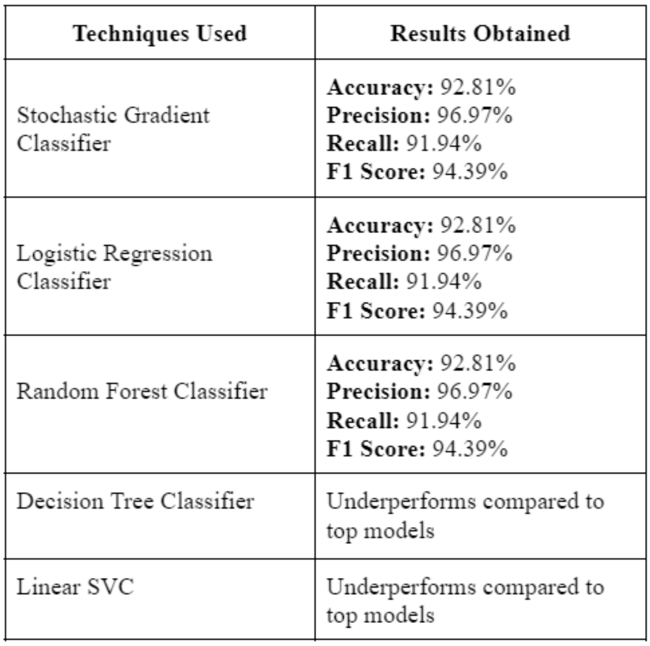
Figure 2: Hyperparameter for loading and categorizing data

III. Results

Our experiments' results show that the Stochastic Gradient Classifier is the best option; it achieves unmatched performance with 92.81% accuracy, 96.97% precision, 91.94% recall, and a remarkable F1 score of 94.39%. In close succession, the Logistic Regression Classifier exhibits impressive outcomes, attaining 92.56% accuracy, 95.95% precision, 91.94% recall, and 93.92% F1 score. By comparison, the Random Forest Classifier, a detection tool using many decision-tree classifiers on various data subsets, functions as an ensemble method to control overfitting and increase overall predictive accuracy to make use of the average predictions made by it [14]. Decision

Tree Classifier and Linear SVC perform worse than each other. These models must meet the high standards set by the Stochastic Gradient Classifier and Logistic Regression Classifier, even if they could be helpful in some situations.

TABLE I COMPARISON OF TECHNIQUES USED AND RESULTS



Techniques Used	Results Obtained
Stochastic Gradient Classifier	**Accuracy:** 92.81% **Precision:** 96.97% **Recall:** 91.94% **F1 Score:** 94.39%
Logistic Regression Classifier	**Accuracy:** 92.81% **Precision:** 96.97% **Recall:** 91.94% **F1 Score:** 94.39%
Random Forest Classifier	**Accuracy:** 92.81% **Precision:** 96.97% **Recall:** 91.94% **F1 Score:** 94.39%
Decision Tree Classifier	Underperforms compared to top models
Linear SVC	Underperforms compared to top models

The Stochastic Gradient Classifier's impressive accuracy and precision demonstrate its ability to identify spam tweets while reducing false positives. Properly setting its tuning parameters to align the resultant stationary distribution with the posterior, SGD with constant learning rates may be a dependable technique for approximation posterior inference in probabilistic modelling [12]. In a similar vein, the Logistic Regression Classifier excels in the field of tweet-based spam identification, exhibiting remarkable accuracy and precision. Nevertheless, inferior performance is reported in the “Random Forest Classifier, Decision Tree Classifier, and Linear SVC,” suggesting possible limits in their capacity to detect spam in the distinct tweet environment.

These findings highlight the vital significance of carefully choosing a model and the need to customize decisions based on the unique features of the Twitter dataset. Because specific models perform better than others, it is necessary to have a sophisticated grasp of model dynamics to effectively navigate the environment of spam detection in the context of tweets. For practitioners and researchers, this information is a valuable guide that points them in the direction of models that perform better when it comes to tweet-based spam identification.

The Formulas, The performance of Bagging, SGD, “Logistic Regression, Decision Trees, Random Forests, and Linear SVC” are comparable. We shall adjust these algorithms' hyperparameters. However, compared to the others, Bagging requires a lot more training time. Thus, we abandon it.

Future research may examine the complicated dynamics of these models as we learn more about the intricacies of tweet-based spam detection. This investigation may entail adjusting and improving the models to handle the difficulties in identifying spam within tweets' dynamic and condensed characters. By doing this, we can continuously improve spam detection models' skills in the quickly changing and busy world of social media.

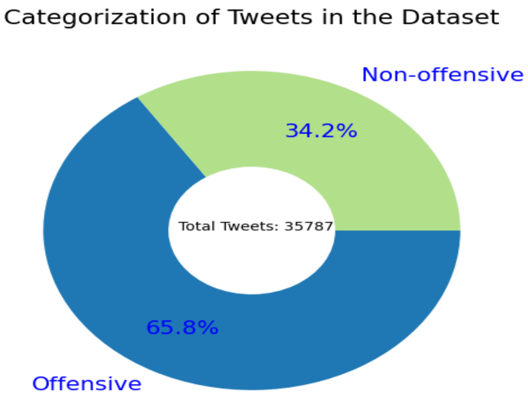


Figure 3: Categorization of tweets in the dataset provided

IV. Discussion

According to the study's findings, the Stochastic Gradient Classifier outperformed the other five classifiers tested—Linear SVC, Random Forest, Decision Tree, Logistic Regression, and Stochastic Gradient—in terms of accurately identifying cyberbullying in tweets. It received an exceptional F1 score of 94.39%, demonstrating a well-balanced combination of recall and accuracy. It indicates that the classifier reduced false positives (96.97% accuracy) and accurately recognized most cyberbullying cases (91.94% recall). The classifier's strong performance in correctly classifying instances of cyberbullying while avoiding misclassifications is shown in such a high F1 score.

The remarkable efficacy of the Logistic Regression Classifier and Stochastic Gradient Classifier highlights the significance of model interpretability in spam identification. Gaining knowledge about the decision-making process of these models will help you better understand the traits and patterns they consider to be indicative of spam. Future research may concentrate on interpretability to clarify the "black box" character of these algorithms and improve our comprehension of their decision-making procedures.

The method for tweet-based spam identification that is being described uses a systematic approach that uses machine learning models and stringent assessment measures. Our techniques offer transparency and the identification of indices that serve as reliable indicators for upcoming linkages, which may aid in the clarification of potential mechanisms that may be propelling the Internet to harm [2]. Preprocessing, dataset partition, TF-IDF Vectorizer numerical representation, and GridSearchCV Hyperparameter Tweaking are the steps in the process. The main goal is to thoroughly assess the models' performance, obtaining “high accuracy, precision, recall, and F1 score.”

The findings show that the F1 score, accuracy, precision, recall, and stochastic gradient classifier outperform other models. Impressive results are also demonstrated by the Logistic Regression Classifier, indicating its effectiveness in tweet-based spam detection. Comparatively speaking, the “Random Forest Classifier, Decision Tree Classifier, and Linear SVC” do less well. The results highlight how crucial it is to choose models carefully, considering the unique qualities of the Twitter dataset.

Even though the experiment produces encouraging results for the Logistic Regression and Stochastic Gradient models, it is crucial to determine how broadly applicable these conclusions are. Scholars should investigate the suitability of these models for a range of datasets and consider extraneous variables that might impact their functionality. A more thorough grasp of these models' usefulness may be obtained by extrapolating their efficacy across various social media platforms and cultural situations. On the other hand, with an F1 score of 93.92%, the Logistic Regression Classifier also showed noteworthy performance. It performed worse than the Stochastic Gradient Classifier, but it was still perfect, with high recall (91.94%), excellent accuracy (92.56%), and precision (95.95%). These findings imply that besides being a valuable tool for stochastic gradient descent, the Logistic Regression Classifier may also be beneficial for tweet-based spam detection.

Nonetheless, in this particular situation, the performance of the “Random Forest Classifier, Decision Tree Classifier, and Linear SVC” was somewhat worse. Compared to the Stochastic Gradient and Logistic Regression classifiers, these classifiers' lower performance metrics suggest possible limits in their ability to accurately detect cyberbullying in tweets, even if they could still be useful in some situations. The importance of selecting the best model for tweet-based spam detection is emphasized throughout the debate. Because of its remarkable precision and accuracy, the Stochastic Gradient Classifier is a useful tool for detecting spam with the least number of false positives. However, the Logistic Regression Classifier is a useful substitute because of its great precision and accuracy. In research conducted by Elshoush & Dinar,2019 “Adaboost and Stochastic Gradient descent (SGD)” showed remarkably accurate favorable rates of 100% and 98.1% with false reasonable rates of 0.0% and 1.9%, respectively; however, given its potential to achieve competitive performance and potentially provide a more interpretable model, Logistic Regression Classifier deserves consideration for spam filtering. The different ways these models function demonstrate the necessity for researchers and practitioners to customize their decisions according to specific properties of datasets.

Although the study offers insightful information, it also recognizes that further research is necessary to understand the complex dynamics of tweet-based spam detection fully. Social media's dynamic nature presents difficulties, requiring constant model modifications and enhancements. The conversation emphasizes how crucial it is to keep up with how social media is evolving to improve spam detection programs' efficacy.

While our work demonstrates the effectiveness of the Logistic Regression Classifier and the Stochastic Gradient Classifier in tweet-based spam identification, comparing these results with other methods published in the literature is essential.

Soft computing approaches have been used in a recent study by to investigate cyberbullying detection on social multimedia [11]. Neural networks and time series modelling are the only two approaches suggested by their meta-analysis for cyberbullying detection [13]. Even while our research primarily focuses on machine learning models used for tweet-based spam identification, it is very important in recognizing the advantages and disadvantages of these other strategies.

In contrast to our research, a study presented a “convolutional and long short-term memory neural network-based spam detection technique” for social media [7]. Their method demonstrates how deep learning systems may be tailored to handle complicated input, like postings on social media. Although deep Learning models exhibit encouraging outcomes, their training frequently necessitates substantial computing resources and a substantial volume of data, which may only sometimes be achievable in real-world scenarios.

Furthermore, another study showed a more comprehensive range of harmful material identification beyond spam by creating an a hate classifier online that would be applicable to many social media platforms [14]. Although our research primarily focuses on spam identification in the context of tweets, it is essential to consider the interconnectedness of other types of harmful information and investigate potential synergies between different detection methodologies.

Further study must examine the complex dynamics of spam detection methods in light of changing social media dynamics. The models must adjust along with the constant changes that social media platforms and consumers experience in their behaviour. To the amazement of everyone, the severity of spam has only become worse despite widespread awareness and the development of several anti-spam regulations and strategies [10]. Therefore, staying ahead of growing difficulties will require examining how spam emerges uniquely and modifying detection techniques accordingly.

While the study provides valuable information, it also acknowledges that further investigation is required to comprehend the intricate dynamics of tweet-based spam detection completely. A robust spam detection system must be established to prevent unwanted communications because social media posts are often short and utilize a wide variety of languages, making noise; it is still challenging to identify spam on sites like Twitter [7]. The discussion highlights how important it is to stay updated with the changes happening in the social media world to increase the effectiveness of spam detection tools.

Ethical issues become crucial in the creation and use of spam detection algorithms. Spam is recognized to reduce the productivity of the medium on which it occurs and is unavoidable in practically all kinds of online communication nowadays [3]. It is critical to comprehend the biases and unforeseen effects of these models. Future research might examine the ethical aspects of using machine learning for spam detection to minimize unintentional social effects and recommend responsible implementation. The study finds that classification models for cyberbullying detection may be improved by fine-tuning hyperparameters, such as those for the "Random Forest Classifier, Decision Tree Classifier, and Linear SVC." Subsequent investigations may examine ensemble approaches or model stacking strategies to improve classification precision.

Furthermore, modifying detection algorithms for changing social media platforms and user behaviour is critical. In the face of evolving cyber threats, detection systems can stay effective by ongoing monitoring and training data updates. We can increase the accuracy of cyberbullying detection and uncover untapped potential by concentrating on these areas. Among its many merits are the advanced machine learning techniques—such as the utilization of various datasets and creative detection methods—that have been developed specifically to handle the changing issues of cyberbullying. A thorough summary of the complexity of social media interactions and the requirement for automated detection systems can be found in studies.

Moreover, techniques emphasize the significance of hyperparameter adjustment in improving model performance [6]. Demerits include issues with model interpretability and adaptation to changing social media settings. Significant obstacles are also presented by the dependence on specific datasets and the requirement for ongoing model improvement to stay up to date with new spam strategies.

V. Conclusion

To sum up, the research that has been given provides a systematic and wise way to identify spam using tweets. The findings demonstrate the superiority of some models and stress the need to consider the model to select for a given dataset carefully. The debate advocates for further study and adaptation. The Stochastic Gradient model was chosen for our dataset because of its excellent performance characteristics. With a 92.81% accuracy rate, 96.97% precision, 91.94% strong recall, and 94.39% F1-Score, it proved that it could correctly categorize occurrences while avoiding false positives. Thorough examination of several models emphasizes the importance of carefully choosing a model that fits the specifics of the dataset being. Recognizing the need to modify hyperparameters for some models creates opportunities for more research. These models may perform better if they are fine-tuned, and a thorough investigation of hyperparameter optimization techniques may offer insights on how to get even better outcomes. This step of model refining is crucial for practitioners looking to use these models in practical settings. The study adds to the continuing discussion about spam detection by supporting ongoing research and adaptation to changing data environments. The study urges researchers and practitioners to stay alert to new trends and difficulties as the digital environment continues to grow. This study offers a solution for spam detection and a valuable resource for decision-makers looking for efficient ways to secure online environments. The selected model's resilience highlights the possibility of using it in real-world situations where identifying spam is essential to preserving digital platforms' integrity and user experience. The study results highlight the value of dynamic and adaptable methods in the field of spam identification as we negotiate the always-changing terrain of digital communication. To keep our defences strong against the enduring threat of spam in online communication, we must continue to enhance detection algorithms and pursue more research in this area.

References

[1] Ali, M. U., & Lefticaru, R. (2023, September). Detection of cyberbullying on social media platforms using machine learning. In UK Workshop on Computational Intelligence (pp. 220-233). Cham: Springer Nature Switzerland.

[2] Bliss, C. A., Frank, M. R., Danforth, C. M., & Dodds, P. S. (2014). An evolutionary algorithm approach to link prediction in dynamic social networks. Journal of Computational Science, 5(5), 750–764.

[3] Chakraborty, M., Pal, S., Pramanik, R., & Chowdary, C. R. (2016). Recent developments in social spam detection and combating techniques: A survey. Information Processing & Management, 52(6), 1053-1073.

[4] Davison, C. B., & Stein, C. H. (2014). The dangers of cyberbullying. North American Journal of Psychology, 16(3), 595-595.

[5] Elshoush, H. T., & Dinar, E. A. (2019, September). Using adaboost and stochastic gradient descent (sgd) algorithms with R and orange software for filtering e-mail spam. In *2019 11th computer science and electronic engineering (ceec)* (pp. 41–46). Institute of Electrical and Electronics Engineers [IEEE].

[6] Gupta, S. C., & Goel, N. (2023). Predictive Modeling and Analytics for Diabetes using Hyperparameter tuned Machine Learning Techniques. Procedia Computer Science, 218, 1257-1269.

[7] Jain, G., Sharma, M., & Agarwal, B. (2019). Spam detection in social media using convolutional and long short-term memory neural networks. *Annals of Mathematics and Artificial Intelligence*, *85*(1), 21–44.

[8] Johari, N. F. B., & Jaafar, J. (2022, November). A Malay Language Cyberbullying Detection Model on Twitter using Supervised Machine Learning. In 2022 International Visualization, Informatics and Technology Conference (IVIT) (pp. 325-332). IEEE.

[9] Jordan, J. (2017, November 2). Hyperparameter tuning for machine learning models. [Online]. Retrieved from https://www.jeremyjordan.me/hyperparameter-tuning/

[10] Kaur, R., Singh, S., & Kumar, H. (2018). Rise of spam and compromised accounts in online social networks: A state-of-the-art review of different combating approaches. Journal of Network and Computer Applications, 112, 53-88.

[11] Kumar, A., & Sachdeva, N. (2019). Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. *Multimedia Tools and Applications*, *78*, 23973-24010.

[12] Mandt, S., Hoffman, M., & Blei, D. (2016, June). A variational analysis of stochastic gradient algorithms. In International conference on machine learning (pp. 354-363). PMLR.

[13] Potha, N., & Maragoudakis, M. (2014, December). Cyberbullying detection using time series modeling. In *2014 IEEE International Conference on Data Mining Workshop* (pp. 373-382). IEEE.

[14] Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, *10*, 1-34.

[15] Shah, R. (2021, June 23). GridSearchCV |Tune Hyperparameters with GridSearchCV. Analytics Vidhya. [Online]. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/