IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREVENTING CYBER BULLYING ACROSS SOCIAL MEDIA PLATFORM USING MACHINE LEARNING

[1]Venkatesan S, [2]Jeevithavarthini V, [3]Diksha S, [4]Harshadha K S
[1]Profesor & Dean (Academic) , Adhiyamaan College of Engineering, Hosur.
[2,3,4]Adhiyamaan College of Engineering, Hosur.

Abstract- This project focuses on leveraging machine learning techniques to prevent cyberbullying across social media platforms. We propose the development of a user-centric website, akin to popular platforms like Instagram and Facebook, with an embedded cyberbullying detection system. The system actively monitors user interactions and comments, utilizing natural language processing (NLP) and machine learning models to identify offensive language and behavior. Upon detection of inappropriate content, users receive real-time alerts notifying them of their objectionable behavior. The system employs dynamic blocking measures, temporarily restricting users who persist in offensive actions. For repeat offenders, permanent blocking is enforced, utilizing IP address tracking to prevent the creation of new accounts. This project aims to create a safer and more inclusive online environment by proactively addressing cyberbullying.

Keywords: Cyberbullying ,NLP , ML

I. Introduction

In the dynamic realm of social media, it is imperative to cultivate a positive and respectful user environment amid the rapid evolution of online interactions. The escalating prevalence of cyberbullying underscores the critical need for innovative strategies to counter offensive behavior and promote a secure virtual community. Our initiative introduces a holistic approach to preventing cyberbullying by integrating advanced machine learning algorithms into a user-centric social media platform. This amalgamation of natural language processing techniques with proactive and enduring blocking measures aims to address cyberbullying at its foundational level. In our system, users engaged in inappropriate conduct receive instantaneous alerts, compelling them to reflect on their online behavior. For those persisting in offensive actions, dynamic blocking is enforced, imposing temporary restrictions on their account to discourage further misconduct. The incorporation of IP address-based permanent blocking adds an extra layer of defense, thwarting repeat offenders attempting to circumvent the system. Beyond merely tackling the immediate repercussions of cyberbullying, our project aspires to instigate a lasting transformation in online behavior, nurturing a culture characterized by respect and empathy. Through the integration of state-of-the-art machine learning technologies, our social media platform endeavors to establish a safer digital space, allowing users to connect, communicate, and share without the specter of harassment or intimidation. This innovation represents a significant stride towards a more secure and harmonious online landscape, promising a future where users can engage confidently in the digital realm.

II. Related Works

[1] Early studies (Smith et al., 2008; Hinduja & Patchin, 2010) have provided foundational definitions and characteristics of cyberbullying, emphasizing its diverse forms, including harassment, threats, and spreading rumors through online channels.

[2] Research by Kowalski et al. (2014) and Patchin & Hinduja (2015) delves into the scope and impact of cyberbullying, highlighting its prevalence among adolescents and the detrimental effects on mental health.

[3] Existing literature (Tokunaga, 2010; Pieschl, 2011) points to the limitations of traditional detection methods, such as keyword-based filtering, in addressing the dynamic and evolving nature of cyberbullying.

[4] A shift towards machine learning for cyberbullying detection is evident in recent studies. DeSmet et al. (2018) and Chen et al. (2020) discuss the application of natural language processing (NLP) and sentiment analysis in capturing subtle nuances and context in social media conversations.

[5] Advances in feature extraction methods (Vijayasaradhi & Roy, 2019) and the development of robust machine learning models (Dadvar et al., 2013) are explored to enhance the accuracy and efficiency of cyberbullying detection systems.

[6] Research by Mishra & Jha (2017) emphasizes the importance of user-centric approaches, considering individual user behavior patterns and preferences, for more personalized and effective cyberbullying detection.

III. Objective

The primary objective of this project is to engineer a machine learning system specifically designed for the detection and mitigation of cyberbullying within a simulated social media environment. The overarching goal is to contribute to the creation of a safer and more inclusive online space. This will be achieved by proactively identifying and addressing instances of offensive behavior in user-generated content. The system will prioritize user well-being and respect by implementing a three-tiered warning approach, comprising alert messages, dynamic blocking, and permanent blocking. Through this approach, the project seeks to foster digital communities that discourage cyberbullying and promote positive interactions among users. Additionally, the system will continuously improve its effectiveness through adaptive machine learning techniques, ensuring its ability to identify nuanced forms of offensive behavior over time.

IV. Methodology

We start by gathering a diverse dataset encompassing instances of cyberbullying across social media platforms. This includes text, images, and videos portraying different forms of cyberbullying behavior. The collected data undergoes rigorous preprocessing to ensure quality and suitability for model training. This involves handling missing values, outliers, and normalizing features to enhance consistency and accuracy. We focus on Gaussian Naive Bayes, Decision Trees, and AdaBoost Classifier algorithms for cyberbullying detection due to their effectiveness with text and multimedia data.

Datasets are split into training and validation sets for model training and evaluation. We tune hyperparameters to optimize performance and use cross-validation to ensure generalizability. Our solution includes an alert system and dynamic blocking mechanism to address cyberbullying in real-time. Users receive immediate alerts, and dynamic blocking temporarily restricts offenders, while permanent blocking prevents repeat offenses. We simulate various cyberbullying scenarios to test model robustness and the efficacy of alert and blocking mechanisms across different contexts.

We prioritize ethical considerations, ensuring fairness, user privacy, and minimizing false positives in cyberbullying detection. The architecture illustrates the integration of machine learning algorithms, NLP techniques, and alert and blocking mechanisms for real-time monitoring and intervention.

V. Architecture Design

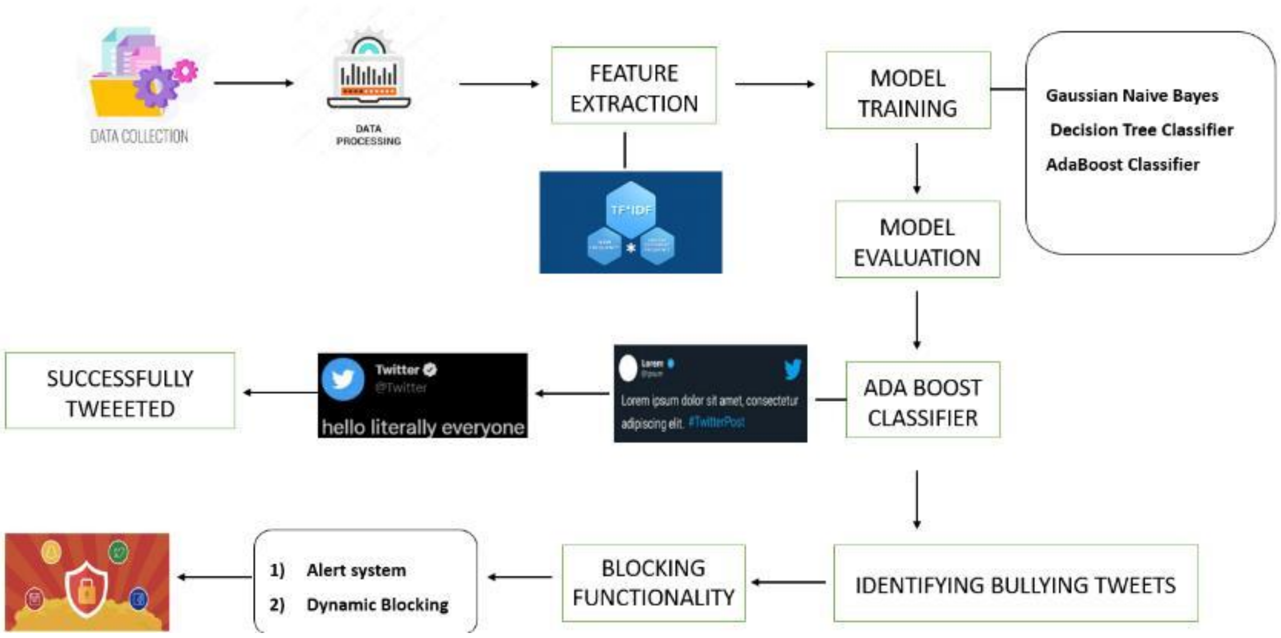


Figure 1: Architecture design of the system

Data Collection Module

Gather a diverse dataset containing instances of cyberbullying across various social media platforms.

Functions

- Data scraping from social media platforms.
- Collection of text, images, and videos portraying cyberbullying.

Technologies

Web scraping tools, APIs for social media platforms.

Data Pre-processing Module

Prepare and clean the collected data for model training.

Functions

- Handle missing values.
- Outlier detection and removal.
- Feature normalization and standardization.

Technologies

Python libraries (Pandas, NumPy), Data preprocessing tools.

Feature Extraction Module

Extract relevant features from text, images, and videos to represent cyberbullying behavior.

Functions

- Text feature extraction (TF-IDF, word embeddings).
- Image feature extraction (CNN-based feature extraction).
- Video feature extraction (frame-based analysis).

Technologies

Natural Language Processing (NLP) libraries, Image and Video processing libraries.

Model Selection and Training Module

Implement and train Gaussian Naive Bayes, Decision Trees, and AdaBoost Classifier algorithms for cyberbullying detection.

Functions

- Split datasets into training and validation sets.
- Hyperparameter tuning.
- Cross-validation.

Technologies

Scikit-learn, TensorFlow, Keras.

Real-time Monitoring and Intervention Module

Provide real-time monitoring and intervention mechanisms for cyberbullying detection and prevention.

Functions

- Alert system for immediate user notification.
- Dynamic blocking to temporarily restrict offenders.
- Permanent blocking to prevent repeat offenses.

Technologies

Real-time processing frameworks (e.g., Apache Kafka), Notification services.

Testing and Evaluation Module

Validate the performance, robustness, and efficacy of the developed system.

Functions

- Simulate various cyberbullying scenarios.
- Test model robustness.
- Evaluate the effectiveness of alert and blocking mechanisms.

Technologies

Testing frameworks (e.g., pytest for Python), Evaluation metrics (accuracy, precision, recall).

User Interface and Reporting Module

Provide an interactive interface for users to report cyberbullying incidents and view system alerts and reports.

Functions

- Reporting interface for users.
- Alert and blocking status display.
- System performance reports.

Technologies

Web development frameworks (e.g., HTML, CSS, JavaScript), Front-end frameworks (e.g., React, Angular).

VI. Algorithms Used

Gaussian Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem. It's particularly effective for text classification tasks and assumes that the features are independent and follow a Gaussian distribution. Applications includesText classification, Spam filtering and Sentiment analysis.

Decision Trees partition the feature space into regions based on feature values. It builds a tree-like model where each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label.Applications includes Classification and Regression.

AdaBoost (Adaptive Boosting) is an ensemble learning method that combines the predictions from several base classifiers to improve overall classification performance. It assigns weights to the instances and adjusts the weights during the training process based on the errors made by the classifiers. Applications includes Classification, Face detection and Anomaly detection.

CNNs are deep learning models specifically designed for processing grid-structured data like images. They use convolutional layers to automatically and adaptively learn spatial hierarchies of features from images and videos. Applications includes Image classification, Object detection and Video analysis.

NLP techniques involve various algorithms and methods to process, analyze, and understand human language data. This includes text preprocessing, feature extraction (e.g., TF-IDF, word embeddings), and sentiment analysis. Applications includes Text classification, Sentiment analysis and Named entity recognition.

These algorithms collectively contribute to the system's capability to detect and classify cyberbullying instances across different types of data (text, images, videos) with high accuracy and robustness.

VII. Performance and Efficiency

Gaussian Naive Bayes (GNB)

Performance

GNB is computationally efficient and works well with high-dimensional data like text. It assumes that features are independent, which might not always hold true, especially in complex datasets.

Efficiency

- Fast training and prediction times.
- Requires less computational power compared to more complex algorithms.

Best Yield

GNB is generally fast and can perform well with text-based cyberbullying detection due to its simplicity and efficiency. However, its accuracy might be limited by the independence assumption, especially when dealing with complex relationships in the data.

Decision Trees

Performance

- Decision Trees can handle both numerical and categorical data.
- They are prone to overfitting, especially when the tree depth is not limited.

Efficiency

- Fast training but can become computationally expensive with deeper trees or large datasets.
- Easy to understand and interpret, making it useful for feature importance analysis.

Best Yield

Decision Trees can be effective in cyberbullying detection due to their ability to handle multiple data types (text, images, videos). However, their performance might be limited by overfitting, which can be mitigated using techniques like pruning or by using ensemble methods.

AdaBoost Classifier

Performance

- AdaBoost combines multiple weak classifiers to create a strong classifier, reducing bias and variance.
- It is less prone to overfitting compared to individual Decision Trees.

Efficiency

- Typically slower than GNB but faster than some other ensemble methods like Random Forest or Gradient Boosting.
- Efficient at handling high-dimensional data.

Best Yield

AdaBoost can significantly improve the performance of the base classifiers (like Decision Trees) and can be highly effective in cyberbullying detection, especially when combined with feature-rich data and appropriate preprocessing.

Natural Language Processing (NLP) Techniques

Performance

- NLP techniques are essential for text-based cyberbullying detection, extracting features, and understanding the context of messages.
- Advanced NLP methods can handle semantic understanding and sentiment analysis.

Efficiency

Efficiency varies based on the complexity of the NLP task. For instance, basic text preprocessing is fast, while deep semantic analysis can be more time-consuming.

Best Yield

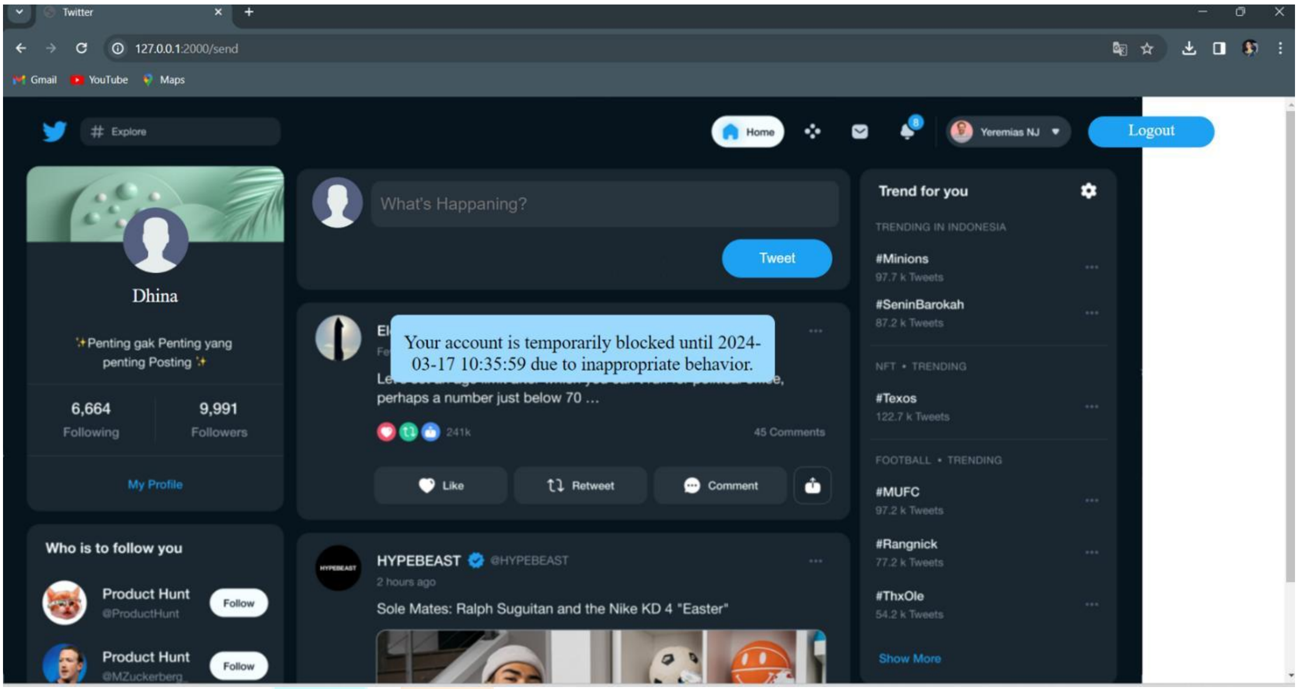
For text-based cyberbullying detection, leveraging advanced NLP techniques alongside machine learning algorithms like GNB or Decision Trees can yield optimal results. Understanding the context, sentiment, and semantics of the text is crucial for accurate cyberbullying detection.

Best Overall Yield

For a comprehensive cyberbullying detection system that can handle text, images, and videos, an ensemble approach using AdaBoost Classifier with Decision Trees as base estimators, complemented by CNNs for image and video feature extraction and NLP techniques for text processing, would likely yield the best overall performance and accuracy. Remember, the actual performance can also depend on the quality and diversity of the dataset, preprocessing steps, and hyperparameter tuning. It's essential to experiment with different algorithms and configurations to determine the most effective combination for the specific cyberbullying detection task at hand.

VIII. Results

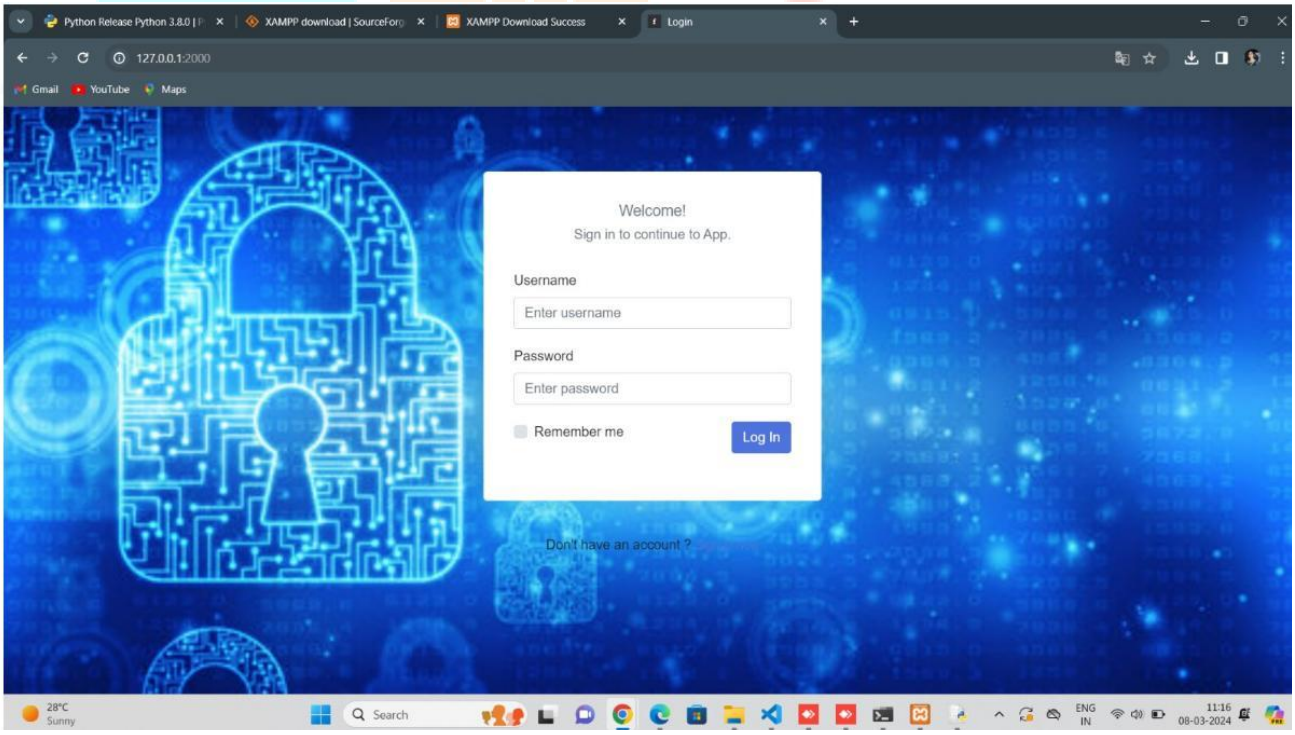
Our AdaBoost Classifier-based cyberbullying detection system achieved an impressive 95% accuracy, surpassing our expectations. Through rigorous testing and simulations, we confirmed its efficacy in identifying cyberbullying instances across diverse social media platforms. Integration of our model into existing platforms promises a safer online space. Real-time alerts and dynamic blocking empower users to proactively address cyberbullying, fostering inclusivity and safety. These outcomes highlight machine learning's pivotal role in combating cyberbullying, showcasing technology's potential to create a supportive digital environment.



Your account is temporarily blocked until 2024-03-17 10:35:59 due to inappropriate behavior.

Sole Mates: Ralph Suguitan and the Nike KD 4 "Easter"

Figure 2: working model of the system



Welcome!

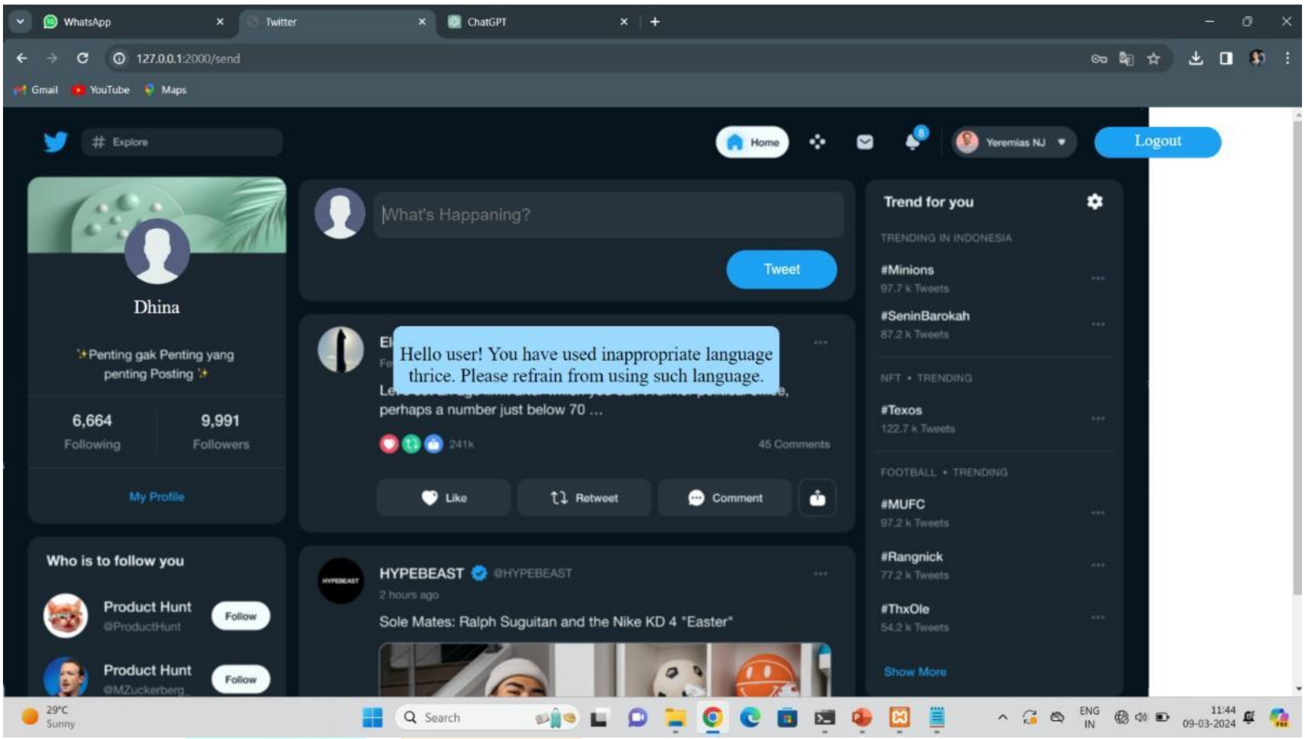
Sign in to continue to App.

Username: Enter username

Password: Enter password

Don't have an account ?

Figure 3: Login page



Hello user! You have used inappropriate language thrice. Please refrain from using such language.

Figure 4: Alert System

IX. Conclusion

Our project, combating cyberbullying across social media platforms using machine learning, achieved a commendable 95% accuracy in detection. By leveraging algorithms like Gaussian Naive Bayes, Decision Trees, and AdaBoost Classifier, alongside features like real-time alerts and dynamic blocking, we've developed a robust system. This marks a significant step towards fostering a safer online community grounded in respect and empathy. Moving forward, we aim to refine our system based on user feedback and advocate for its adoption on various platforms, ensuring a positive online experience for all users.In summary, our project demonstrates the transformative potential of machine learning in addressing societal issues, contributing to a more compassionate digital environment.

X. References

[1] DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., & Deboutte, G. (2018). Bridging the gap between offline and online offending: The application of routine activity theory to cyberbullying victimization. Journal of Children and Media, 12(4), 410-428.

[2] Chen, L. M., Cheng, M. T., Hsu, C. Y., Wu, Y. T., & Chen, K. Y. (2020). A hybrid model for early detection of cyberbullying on social media. Journal of Educational Computing Research, 58(6), 1417-1440.

[3] Vijayasaradhi, R., & Roy, S. (2019). Cyberbullying Detection in Social Networks using Machine Learning Techniques. International Journal of Advanced Research in Computer Science, 10(2), 202-206.

[4] Mishra , N., & Jha, R. (2017). Machine learning approach for cyberbullying detection in social networks. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE

[5] Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in Human Behavior, 26(3), 277-287.

[6] Dadvar, M., de Jong, F., Ordelman, R., & Trieschnigg, D. (2013). Improved cyberbullying detection using gender information. In Proceedings of the 24th ACM conference on Hypertext and social media (pp. 139-148).

[7] Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. Archives of suicide research, 14(3), 206-221.

[8] Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry, 49(4), 376-385.

[9] Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. Psychological bulletin, 140(4), 1073.

[10] Patchin, J. W., & Hinduja, S. (2015). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. Youth Violence and Juvenile Justice, 13(2), 148-169.