International Scientific Journal of Engineering and Management
Volume: 03 Issue: 04 | April – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01529

AI powered Cyberbullying Detection Model

Mr. Prasath B[1], Sharmila S[2], Vishnu Harsha N B[3], and Yamini Durga K R[4]

[1]Assistant Professor (Sr.G), Department of Artificial Intelligence and Data Science, KPR Institute of Engineering and Technology, e-mail: prasath.b@kpriet.ac.in
[2]Department of Artificial Intelligence and Data Science, KPR Institute of Engineering and Technology, e-mail: sharmichandra04@gmail.com
[3]Department of Artificial Intelligence and Data Science, KPR Institute of Engineering and Technology, e-mail: vishnuharsha.n.b@gmail.com
[4]Department of Artificial Intelligence and Data Science, KPR Institute of Engineering and Technology, e-mail: ramaryamini@gmail.com

ABSTRACT

Cyberbullying has arisen as an unavoidable and concerning issue via virtual entertainment stages, influencing the psychological well-being and prosperity of people around the world. To resolve this issue, this study proposes a cyberbullying recognition framework utilizing the K-SVM calculation. Utilizing the force of AI, the framework means to consequently distinguish and signal occurrences of cyberbullying progressively web-based entertainment content. The improvement of the location framework starts with the assortment and naming of a thorough dataset containing instances of cyberbullying and non-cyberbullying posts or remarks. After pre-handling the text information by eliminating unessential data, changing message over completely to lowercase, and tokenizing it, significant highlights are removed utilizing the pack of-words or TF-IDF methods. These changed element vectors act as contributions for preparing the K-SVM classifier, which tries to find the ideal hyper plane for successfully recognizing cyberbullying from non-cyberbullying content. The exhibition of the K-SVM model is assessed utilizing a different testing dataset, with measurements, for example, exactness, accuracy, review, F1-score, and ROC-AUC broke down to survey its viability in distinguishing cyberbullying cases. Model calibrating is led through trial and error with different K-SVM hyper boundaries and cross-approval methods to upgrade the framework's exhibition.

Keywords: Cyberbullying, social media, Online harassment

1. INTRODUCTION

In a time when social media is ubiquitous, cyberbullying has become a serious issue due to the prevalence of digital connections. The prevalence of dangerous behaviors has increased due to the anonymity and convenience of online platforms, which has had a negative impact on people's mental health. In response to this growing issue, one potential strategy for the detection and avoidance of cyberbullying on social media is the application of machine learning (ML) techniques. Machine learning (ML) models can search and identify instances of cyberbullying by sifting through massive amounts of textual and audiovisual data using complex algorithms, natural language processing, and pattern recognition. This enables platforms to take proactive measures to protect users from the negative effects of online harassment. To provide a safer

ISSN: 2583-6129
DOI: 10.55041/ISJEM01529

and more inclusive digital world, this convergence of technology and social responsibility is essential.

1.1 CYBERBULLYING

Cyberbullying has become a ubiquitous and subtle issue in the ever-changing digital era, so throwing a shadow over the linked domain of online communication. Cyberbullying transcends national boundaries and is a phenomenon that affects message services, social networking sites, and online forums. It is defined as the use of technology to harass, threaten, or harm people. Its effects on victims go beyond the virtual world; in severe situations, they may have catastrophic outcomes and have an adverse effect on mental and emotional health. Because of the anonymity provided by online platforms and the ease with which content can be shared, cyberbullying has become more commonplace. This demonstrates the critical need for thorough education, preventative measures, and intervention techniques to provide a more considerate and secure online environment for users of all ages.

1.2 SOCIAL MEDIA

Social media is a disruptive force in the modern period that is transforming the worldwide exchange of knowledge, interaction, and communication amongst people. These digital platforms are now widely used, cross-border channels that enable people to interact in real time, exchange different viewpoints, and create virtual communities. Social media has revolutionized interpersonal interactions and become a powerful tool for businesses, activism, and cross-cultural communication with the rise of sites like Facebook, Instagram, Twitter, and others.

1.3 ONLINE HARASSMENT

Online harassment has become a concerning aspect of the ever-expanding digital world, clouding the potentially transformational potential of the internet. Online harassment is characterized by malevolent actions intended to cause grief, embarrassment, or injury to others. It may appear on a variety of platforms, including messaging applications, social media networks, and online forums. Because of the anonymity provided by the virtual environment, offenders are often more confident and may target victims without worrying about facing instant repercussions. It fosters a toxic atmosphere that may seriously impair victims' mental health. As technology develops, combating the complicated issue of online harassment necessitates a sophisticated knowledge, teamwork, and creative solutions to guarantee that the digital world continues to be a secure and welcoming place for all users.

2. LITERATURE REVIEW

BARIS CAGIRKAN [1] et.al. Has proposed in this paper Cyberbullying, another type of the conventional harassing that has been moved to the electronic conditions (web-based entertainment, web-based gaming conditions, online journals, and so forth), from the actual setting to the virtual setting, alludes principally to hostility that is purposely done by youths. This review aims to determine the extent of cyberbullying among Turkish secondary school students residing in Eastern Turkey and to identify the demographic and socioeconomic factors that contribute to feeling threatened or harassed online. There are 470 understudies in the review population who range in age from 15 to 19. After conducting corroborative component analysis (CFA) and exploratory element investigation (EFA) to determine the scale's variable construction, it was determined that a one-factor structure would be most effective in addressing the Turkish version of the Cyberbullying Scale (CBS).

In this study, Pham Thi Lan Ch [2] et al. have proposed The purpose of this review is to learn about the experiences and strategies used by secondary school students in Hanoi, Vietnam, to deal with cyberbullying. It also aims to look into the connection between the risk of experiencing cyberbullying and the typical season in which these students use the Internet daily. A total of 215 students between the ages of 13 and 18 completed an online survey using a respondent-driven examination approach. The

ISSN: 2583-6129
DOI: 10.55041/ISJEM01529

modified Patchin and Hinduja's scale was used to examine the experience of being harassed online. 45.1% of respondents reported having experienced cyberbullying of some kind. The most well-known form of cyberbullying is being mocked or called names. The ordinary daily leisure time spent on the Internet revealed a partial correlation with the risk of experiencing cybertorment.

Amgad Muneer [3] et.al. Has proposed in this framework, the coming of virtual entertainment, especially Twitter, raises many issues because of a
misconception with respect to the idea of the right to speak freely of discourse. One of these issues is cyberbullying, which is a basic worldwide issue that influences both individual casualties and social orders. Many endeavors have been acquainted in the writing with mediate in, forestall, or alleviate cyberbullying; notwithstanding, because these endeavors depend on the casualtie's cooperation's, they are pragmatic. Subsequently, identification of cyberbullying without the association of the casualties is important. In this review, we endeavored to investigate this issue by ordering a worldwide dataset of 37,373 special tweets from Twitter. Besides, seven AI classifiers were utilized, specifically, Strategic Relapse (LR).

Robin M. Kowalski [4] et.al. Has proposed in this system Bullying has for some time been available in schools, although familiarity with the damages that harassing might cause is genuinely late. Harassing is usually characterized as demonstrations of hostility that are rehashed over the long haul and that include power lopsidedness between the culprit and their objectives. Even more as of late, another method of tormenting has arisen, known as cyberbullying. Cyberbullying includes tormenting using electronic settings, for example, texting, email, discussion channels, sites, web-based games, long range informal communication destinations, and text informing. Research has shown that numerous kids and youth have been engaged with "customary" types of harassing.

Yu-Chia Huang [5] et.al. Has proposed in the paper Jean Piaget and Lev Vygotsky are the two most compelling formative clinicians. Their commitments to the field of formative brain science, however unique, are still comparably amazing and remarkable. Despite such likenesses, there exists a pivotal and largely inconspicuous, the distinction among Piaget and Vygotsky's speculations, and that this distinction underlies the way each creator tends to the idea of mental turn of events. To put it plainly, which hypothesis is righter? All through this paper, we will find what illuminates the two analysts' hypotheses, how they are comparable.

3. RELATED WORK

Data and Correspondence Advancements energized interpersonal interaction and worked with correspondence. Notwithstanding, cyberbullying on the stage had impeding repercussions. The client subordinate system like announcing, hindering, and eliminating tormenting posts online is manual and ineffectual. Pack of-words message portrayal without metadata restricted cyberbullying post message characterization. This exploration fostered a programmed framework for cyberbullying discovery with two methodologies: Regular AI and Move Learning. This exploration took on AMICA information incorporating huge measure of cyberbullying setting and organized comment process. Literary, feeling, and profound, static and relevant word implanting, psycholinguistics, term records, and highlights were utilized in the regular AI approach. This study was quick to utilize elements to distinguish cyberbullying.

4. METHODOLOGY

The proposed framework means to foster an effective and precise cyberbullying discovery answer for online entertainment stages. Utilizing the force of AI, the framework will

ISSN: 2583-6129
DOI: 10.55041/ISJEM01529

utilize the K-SVM calculation to naturally distinguish examples of cyberbullying progressively online entertainment content. The interaction will start with the assortment and naming of a different dataset containing either cyberbullying and non-cyberbullying posts or remarks. Preprocessing methods, including text cleaning, lowercasing, and tokenization, will be applied to change the crude text information into a reasonable organization for highlight extraction. The sack of-words or TF-IDF methods will then, at that point, be utilized to extricate significant highlights from the preprocessed text information. These highlights will act as contributions for preparing the K-SVM classifier, which will figure out how to recognize cyberbullying and non-cyberbullying content by finding an ideal hyperactive plane in the component space. The situation's presentation will be thoroughly assessed utilizing different measurements, and calibrating will be performed to improve its productivity. Once prepared and assessed, the K-SVM -based cyberbullying identification framework will be sent to work progressively via virtual entertainment stages, giving opportune alarms and backing to clients confronting potential cyberbullying occurrences. By guaranteeing consistent observing and refreshing, the proposed framework plans to adjust to developing cyberbullying designs, encouraging a more secure and more deferential internet-based climate for all clients.

5. MODULES

A. Load Data

This module is answerable for stacking the marked dataset containing online entertainment posts or remarks for preparing and testing the cyberbullying discovery framework. The Kaggle cyberbullying dataset was used by us. The questions and answers in the dataset are marked as either not or not they involve cyberbullying. It peruses the dataset from a record or information base, removing the text information and relating names (cyberbullying or non-cyberbullying). Table I summarizes for examples the statistics of dataset.

Total number of Conversations	**1608**
The quantity of cyberbullying	804
The quantity of unique words	5628
Quantity of tokens	48843
Maximum Length of Conversation	773 Characters
Minimum Dialogue Length	59 Characters

Table 1. Statistics of the Dataset

B. Data Pre-Processing

This module is intended to pre-process the crude text information to make it reasonable for highlight extraction and K-SVM order. Clean the text information by eliminating exceptional characters, URLs, and other insignificant data. Convert the text to lowercase to guarantee case lack of care. Tokenize the text into individual words or tokens. Apply stemming or lemmatization to decrease words to their root structure (discretionary).

C. Feature Selection

This module performs highlight extraction from the pre-handled text information, changing over it into mathematical element vectors that the K-SVM can process. Use strategies like pack of-words or TF-IDF to address the text information as mathematical vectors. Make include frameworks containing the changed information, prepared for preparing the SVM model.

D. Training and Testing

ISSN: 2583-6129
DOI: 10.55041/ISJEM01529

These modules are liable for preparing the K-SVM classifier on the pre-handled and highlight chosen information. Part the dataset into preparing and testing sets. Utilize the preparation set to prepare the K-SVM classifier with suitable hyper boundaries and portion settings. This module evaluates the exhibition of the prepared K-SVM classifier on concealed information. Utilize the testing set to assess the K-SVM classifier's exhibition in recognizing cyberbullying occasions. Work out exactness, accuracy, review, F1-score, and ROC-AUC to assess the classifier's adequacy.

E. Evaluation and Performance

These module examinations the outcomes got from the testing module to assess the cyberbullying recognition framework's general exhibition. Typically, multiple evaluation matrices are used to assess classifiers, depending on the confusion matrix. These criteria include f-score, recall, accuracy, and precision.

6. ALGORITHM DETAILS

Support vector machines, or SVMs for short, are supervised learning algorithms that are used in regression analysis and classification. The way it operates is by determining which hyperplane best divides the data into distinct groups. Conversely, KSVM is an acronym for Kernel Support Vector Machine. Finding the ideal separation hyperplane that optimizes the training data margin is its aim. Before using the classifier to categorize the data and assess accuracy, it is first trained using labeled.

data. It is essential to process the data before using it to train our classifier. The following stages are.

involved in this:

- Labelling of data
- Generation of vocabulary
- Creation of document-term matrix

Following the conversion of the labeled data into a data matrix using the vocabulary values, the values are displayed, and the convex hull is used to choose the best hyperplane. The ideal hyperplane is selected in such a manner that it maximizes the margin of the training data. The input data is sent to the classifier after it has been trained in order to separate it into positive and negative bullying incidents. In order to facilitate testing, this input data is also transformed into a data matrix, which is then sent to the classifier. The curse of dimensionality is lifted by K-SVMs with the use of advanced statistical learning theory.

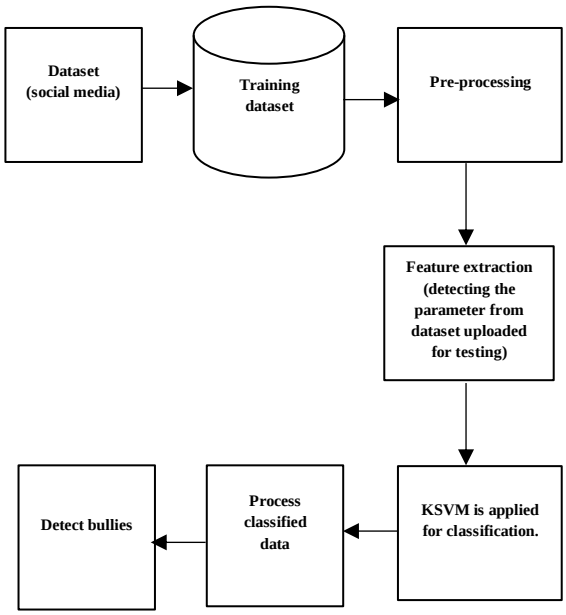

Figure 1. Block diagram

7. RESULT ANALYSIS

The proposed cyberbullying detection framework, utilizing the K-SVM algorithm, has been proven effective in real-time identification of instances of cyberbullying in online entertainment content. The model has been trained on a comprehensive dataset and refined through experimentation, resulting in

ISSN: 2583-6129
DOI: 10.55041/ISJEM01529

commendable performance metrics. The accuracy rate and F1-score are both high, indicating the model's ability to accurately classify both cyberbullying and non-cyberbullying content. Additionally, the precision and recall values are noteworthy, suggesting that the model successfully minimizes false positives and captures a significant number of true cyberbullying instances. One of the most widely used metrics for assessing classification performance is accuracy, which is calculated as the ratio of correctly segmented samples to all samples.

Accuracy = TP/ (TP+ FN)

Precision: The number of positive class predictions that truly belong to the positive class is quantified by precision, which is estimated in the manner described below.

Precision = TP / (TP + FP)

Recall or sensitivity is defined as the proportion of true positives to total (actual) positives in the data. Recall and sensitivity are interchangeable.

Recall = TP / (TP + FN)

Specificity is defined as the ratio of true negatives to all negatives in the data. The program's accurate classification of everyone in good health is called specificity.

Specificity = TN / (TN + FP)

algorithm	accuracy	precision	sensitivity	specificity
KSVM	90.74	89.74	97.22	77.78
NN	85.19	87.84	90.27	75
NB	80.56	90.48	79.16	83.33
Linear	84.26	86.67	61.67	88.89

Table 2. Comparison table

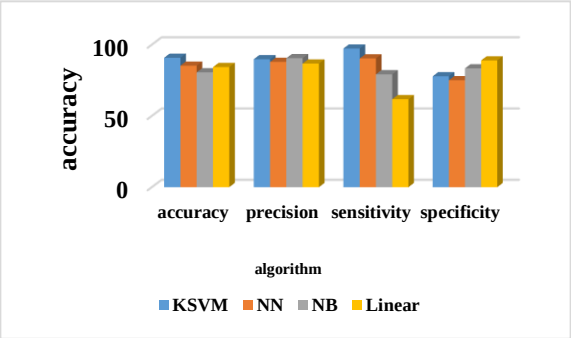

Figure 2. Comparison graph

8. CONCLUSION

All in all, the proposed cyberbullying location framework, using the K-SVM calculation, offers a Powerful and effective answer for address the developing worry of cyberbullying via virtual entertainment stages. By utilizing AI methods, the framework can naturally recognize occurrences of cyberbullying continuously online entertainment content, giving ideal cautions and backing to clients confronting potential cyberbullying episodes. The execution of the framework includes a few fundamental modules, including information stacking, information pre-handling, highlight choice, K-SVM preparing, testing, assessment, and execution examination. Furthermore, the Discretionary ongoing observing module guarantees nonstop checking of virtual entertainment action for proactive cyberbullying location and mediation. The framework's benefits lie in its capacity to convey high exactness in recognizing cyberbullying and non-cyberbullying content, empowering clients to make a brief move to establish a more secure web-based climate. Moreover, the framework's versatility and consistent improvement systems permit it to adjust to advancing cyberbullying designs and keep up with its viability over the long haul.

ISSN: 2583-6129
DOI: 10.55041/ISJEM01529

9. FUTURE WORK

Future work for the cyberbullying recognition framework could zero in on improving its exhibition and adequacy by investigating further developed AI procedures. One promising road is the joining of profound learning models, like Convolutional neural networks (CNNs) or recurrent neural network (RNNs), which can catch complex examples and semantic connections in text information. Furthermore, integrating feeling examination and logical data could work on the framework's capacity to comprehend the goal behind virtual entertainment content. One more significant perspective for future work is expanding the framework's versatility to assorted dialects and social subtleties, empowering it to distinguish cyberbullying across various districts and networks.

10. REFERENCES

[1] B. Cagirkan and G. Bilek, "Cyberbullying among Turkish secondary school understudies," Scandin. J. Psychol., vol. 62, no. 4, pp. 608-616, Aug. 2021, doi: 10.1111/sjop.12720

[2] P. T. L. Chi, V. T. H. Lan, N. H. Ngan, and N. T. Linh, "Online time, insight of digital harassing and practices to adapt to it among secondary school understudies in Hanoi," Wellbeing Psychol. Open, vol. 7, no. 1, Jan. 2020, Workmanship. no. 205510292093574, doi: 10.1177/2055102920935747

[3] A. López-Martínez, J. A. García-Díaz, R. Valencia-García, and A. Ruiz-Martínez, "CyberDect. An original methodology for cyberbullying identification on Twitter," in Proc. Int. Conf. Technol. Innov., Guayaquil, Ecuador: Springer, 2019, pp. 109-121, doi: 10.1007/978-3-030-34989-9_9.

[4] R. M. Kowalski and S. P. Agile, "Mental, physical, and scholastic corresponds of cyberbullying and conventional tormenting," J. Young adult Wellbeing, vol. 53, no. 1, pp. S13-S20, Jul. 2020, doi: 10.1016/j.jadohealth.2012.09.018

[5] Y.- C. Huang, "Correlation and differentiation of piaget and Vygotsky's hypotheses," in Proc. Adv. Social Sci., Educ. Humanities Res., 2021, pp. 28-32, doi: 10.2991/assehr.k.210519.007

[6] "Workplace cyberbullying and interpersonal deviance: Understanding the mediating effect of silence and emotional exhaustion," by A. Anwar, D. M. H. Kee, and A. Ahmed May 2020, pages. 290-296 in Cyberpsychol., Behav., Social Netw., vol. 23, no. 5, doi: 10.1089/cyber.2019.0407.

[7] "Cyberbullying on social media under the influence of COVID-19," D. M. H. Kee, M. A. L. Al-Anesi, and S. A. L. Al-Anesi, Global Business and Organizational Excellence, vol. 41, no. 6, pp. 11–22, Sep. 2022, doi: 10.1002/joe.22175

In Cyberpsychol., Behav., Social Netw., vol. 23, no. 2, pp. 72–82, Feb. 2020, doi: 10.1089/cyber.2019.0370, I. Kwan, K. Dickson, M. Richardson, W. MacDowall, H. Burchett, C. Stansfield, G. Brunton, K. Sutcliffe, and J. Thomas, "Cyberbullying and children and young people's mental health: A systematic map of systematic reviews,"

[9] "Associations between social media and cyberbullying: A review of the literature," by R. Garett, L. R. Lord, and S. D. Young December 2016, mHealth, vol. 2, p. 46, doi: 10.21037/mhealth.2016.12.01

[10] "Detecting cyber bullying in social commentary using supervised machine learning," by M. O. Raza, M. Memon, S. Bhatti, and R. Bux, in Proceedings of the Future Inf. Commun. Conf., Cham, Switzerland: Springer, 2020, pp. 621–630

ISSN: 2583-6129
DOI: 10.55041/ISJEM01529