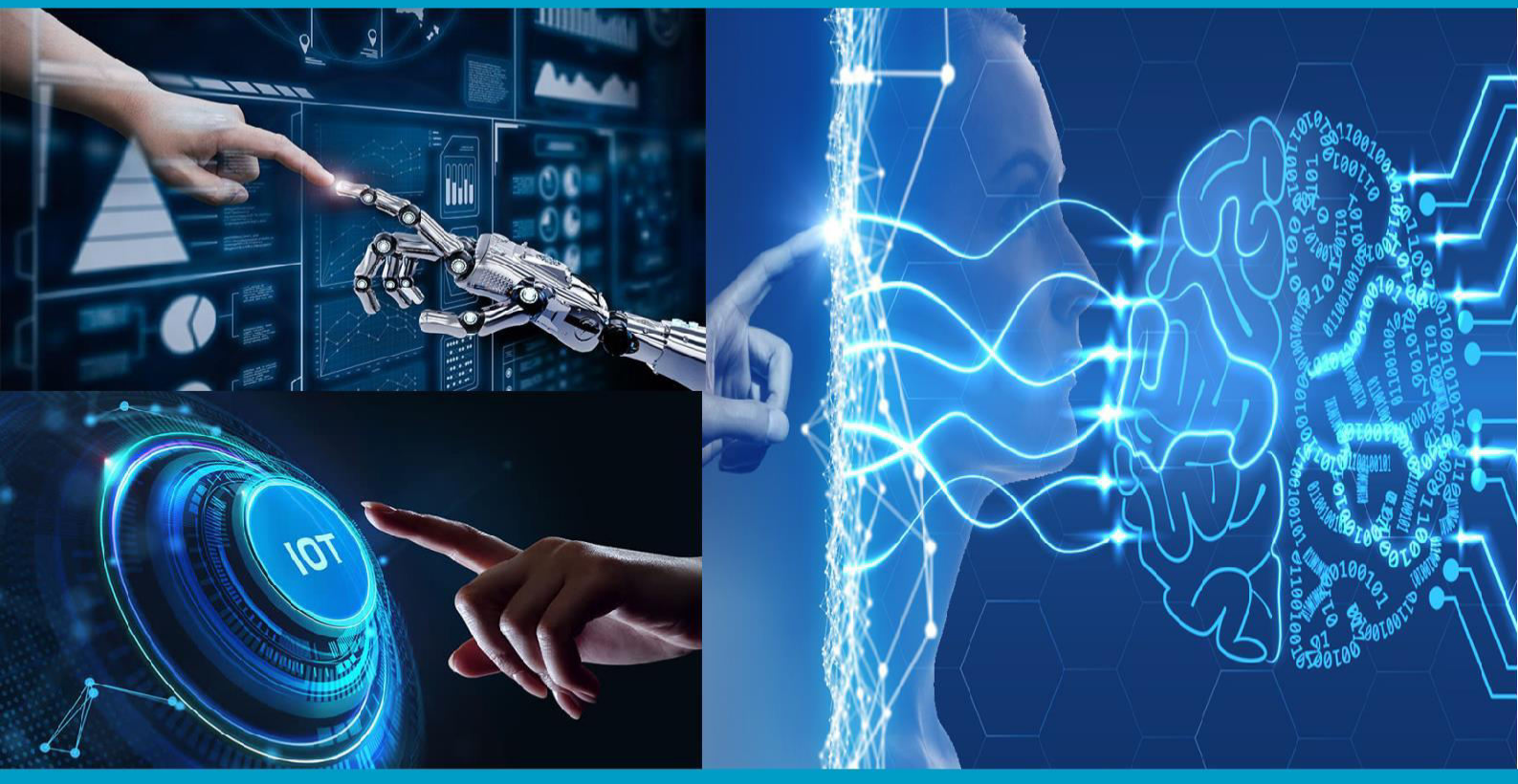
IJMRSETM

e-ISSN: 2395 - 7639



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 12, Issue 3, March 2025

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.214

+91 99405 72462 +9163819 07438 ijmrsetm@gmail.com www.ijmrsetm.com

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 3, March 2025 |

Detection of Cyberbullying on Social Media

Dr. S. Suganyadevi[1], M. Madhumitha[2]

Assistant Professor, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, India[1]

U.G Student, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, India[2]

ABSTRACT: Cyberbullying is a growing concern with the rapid rise of social media platforms. Traditional methods of monitoring and mitigating cyberbullying are often ineffective due to the vast amount of data generated daily. This paper presents an approach utilizing machine learning techniques to detect cyberbullying incidents in social media content. Various classification models, including Support Vector Machines (SVM), Naïve Bayes, and deep learning models such as Recurrent Neural Networks (RNN) and Transformers, are evaluated for their effectiveness. The study highlights the importance of feature extraction techniques, dataset selection, and model performance metrics in achieving accurate cyberbullying detection.

KEYWORDS: Cyberbullying, Machine Learning, Social Media, Sentiment Analysis, NLP, Deep Learning

I. INTRODUCTION

Social media platforms' explosive expansion has changed how people express their thoughts, communicate, and share information. Cyberbullying, an aggressive and persistent online conduct that damages people's psychological and emotional well-being, has become more prevalent as a result of this digital revolution. Cyberbullying is a serious problem for social media companies and law enforcement organizations since it can take many different forms, including threats, hate speech, harassment, and defamation.

To make the internet a safer place, cyberbullying must be identified and addressed. Because of the large amount of social media content, traditional techniques for detecting cyberbullying, like manual moderation and user reports, are ineffective. Consequently, automated methods that use machine learning (ML), deep learning (DL), and natural language processing (NLP) have become more popular for identifying dangerous discussions.

II. OBJECTIVES

1. **To look at the characteristics and effects of cyberbullying:**
   Examine various types of cyberbullying, such as hate speech, harassment, and abusive language, in order to establish detection criteria.
2. **To gather and pre-process social media data:** Use natural language processing (NLP) methods like tokenization, stop word removal, stemming, and lemmatization to improve text representation in datasets curated and cleaned from social media sites.
3. **To put machine learning models into use for detecting cyberbullying**: Develop and assess a variety of supervised and unsupervised machine learning methods, such as Random Forest, Naïve Bayes, Support Vector Machines (SVM),
   and deep learning models like transformer- based architectures (e.g., BERT) and Long Short-Term Memory (LSTM).PROPOSED METHODOLOGY

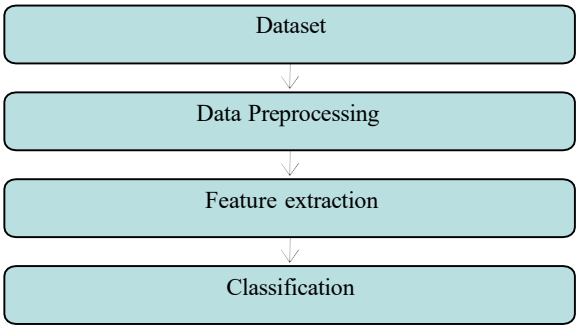


Figure:01

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 3, March 2025 |

III. DATASET

A. TWITTER DATASET:

Two datasets containing hate speech are combined to create the Twitter Dataset:

- Waseem, Zeerak, and Hovy, Dirk's [11] Hate Speech Twitter Dataset, which includes 17,000 tweets classified as racist or sexist. The annotations are used to mine the tweets.5900 tweets are lost as a result of deleted tweets or deactivated accounts.

The Hate Speech Language Dataset was created by Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber [12].It had 25,000 tweets that were gathered through crowdsourcing.

This results in 35787 tweets overall for the task distribution that is displayed in Figure 3.Thirty percent of the dataset (10,737) is utilized for testing, while seventy percent (25,050) is used for training.

B. WIKIPEDIA DATASET:

One million comments are marked as personal assaults in the Wikipedia dataset created by Wulczyn, Thais, and Dixon [13].Thirteen thousands of the forty thousand comments in the sample are classified as cyberbullying because they contain personal attacks. These remarks are taken from discussions between editors of Wikipedia pages that have been annotated by ten people using Crowd Flower. The same split—70 percent, or 28,000, for training data and 30percent, or 12,000, for testing data—is applied to this dataset.

IV. ALGORITHM USED

A. SUPPORT VECTOR MACHINE:

SVM, or support vector machine for applications involving text classification, SVM is a popular supervised learning technique. Finding the best hyper plane to optimize the margin between classes is how it operates. Because text data is high-dimensional, SVM can effectively identify cyberbullying by separating bullying from non-bullying content and mapping words into feature spaces.

Steps for SVM classification:

1. **Data Collection**: Gather social media text that has been categorized as either non-cyberbullying or cyberbullying.
2. **Pre-processing:** Make text cleaner by tokenizing, deleting stop words, and deleting special characters.
3. **Feature Extraction**: Use word embedding's or TF-IDF to transform text into numerical representation.
4. **Dataset splitting**: Separate the data into sets for testing and training.
5. **Train SVM Model**: Select an appropriate kernel (RBF, linear), then train the model.

B. RANDOM FOREST CLASSIFIER:

Several decision trees are constructed using the Random Forest ensemble learning technique, which then aggregates the results to enhance classification performance. When paired with feature extraction methods like Term Frequency-Inverse Document Frequency (TF-IDF) or word embedding's, it can detect cyberbullying since it is resistant to over fitting and able to handle big datasets.

Steps for RF classification:

1. **Data Collection**: Gather labeled social media text (cyberbullying vs. non-cyberbullying).
2. **Preprocessing**: Clean text (tokenization, stop word removal, lemmatization).
3. **Feature Extraction**: Convert text into numerical format using **TF-IDF**:

   TF−IDF=TF×log⁡f0 (NDF)TF-IDF = TF \times \log\left(\frac{N}{DF}\right)

4. **Dataset Splitting** – Divide into training and testing sets.
5. **Train Random Forest** – Construct **N** decision trees, where each tree predicts a class:
   y^=1N∑i=1NTi(x)\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x)
6. **Prediction** – Majority voting for classification.
7. **Evaluation** – Use **Accuracy**, **Precision**, **Recall**, **F1-score**:
   F1=2×Precision×RecallPrecision+RecallF1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}
8. **Optimization** – Tune hyper parameters (tree count, depth) using GridSearchCV.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 3, March 2025 |

V. DATA VISUALIZATION

Understanding patterns, connections, and insights in datasets is aided by data visualization. Here are some typical visualization methods along with some

1. **Bar Chart:**

For Comparing Categorical Data used to contrast various groups or classifications.
For instance, comparing the frequency of communications that are cyberbullying vs those that are not.

2. **Histogram:** A dataset's frequency distribution is displayed using a histogram (data distribution).
Example: Message length in the detection of cyberbullying.

3. **Pie chart:** Pie charts, or proportions of categories, are used to display percentages or proportions.
For instance, the proportion of communications that are cyberbullying and those that are not.
Confusion **matrix:** Machine learning uses the Confusion Matrix (Model Performance) to assess categorization performance.
Cyberbullying detection and categorization performance, for instance

4. **Line chart:** Data trends over time are displayed using a line chart (Trends over Time).
For instance, a rise in instances of cyberbullying over several months.

VI. CONCLUSION

Controlling the spread of cyberbullying is necessary since it is risky and can result in negative outcomes like despair and suicide.

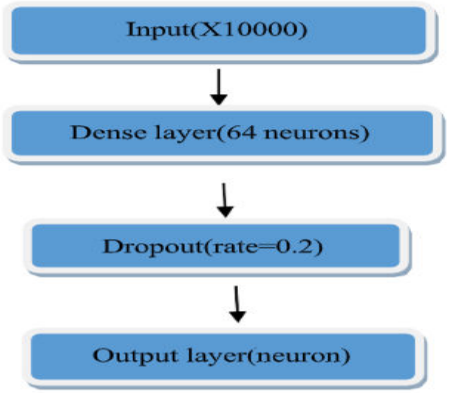


Figure:02

For numerous additional types of cyber-attacks Social media platforms can use cyberbullying detection to ban users who attempt to engage in such behaviour. In order to address the issue, we suggested an architecture in this study for detecting cyberbullying. We spoke about the architecture for two different kinds of data: personal attacks on Wikipedia and hate speech data on Twitter. Because hate speech was easily identifiable due to the usage of profanity in tweets, Natural Language Processing approaches proved effective with accuracies of over 90 percent utilizing basic machine learning algorithms. As a result, Bow and Tf-Idf models produce superior outcomes than Word2Vec

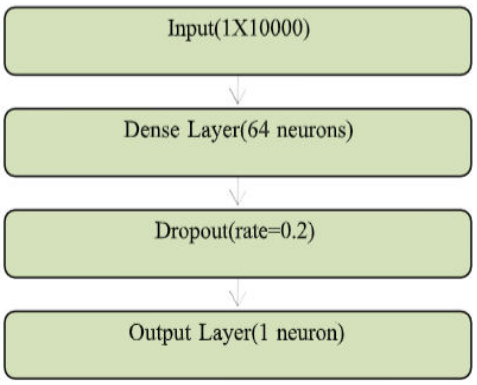


Figure:03

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 3, March 2025 |

By displaying true positives (bullying communications that were correctly identified) and false negatives (missing instances of cyberbullying), the confusion matrix offers information about the accuracy of the model. Techniques like feature engineering, data balancing, and hyper parameter tuning can be used to improve performance.
For a more thorough approach to cyberbullying detection, future developments might use multimodal analysis (text, photos, and videos), real-time detection systems, and deep learning models (LSTMs, Transformers).

REFRENCES

[1] C. L. Gómez, I. Santos, P. G. Bringas, J. G. de la Puerta, and P. Galán-García, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," 2014, doi: 10.1007/978-3- 319-01854-6_43. 10.1109/BESC.2017.8256403.

[2] "Collaborative detection of cyberbullying activity in Twitter data," by A. Mangaonkar, A. Hayrapetian, and R. Raje 10.1109/EIT.2015.7293405, doi.

[3] "Automatic detection of cyberbullying on social networks based on bullying traits," by R. Zhao, A. Zhou, and K. Mao, 2016, doi: 10.1145/2833312.2849567.

[4] "Detection of Cyberbullying Using Deep Neural Network," by V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, 2019, doi: 10.1109/ICACCS.2019.8728378.

[5] K. Reynolds, A. Kontostathis, and L. Edwards, "Detecting cyberbullying with machine learning," 10.1109/ICMLA.2011.152, 2011.

[6] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, "Towards the detection of cyberbullying based on social network mining techniques," in Proceedings of 4th International Conference on Behavioral, Economic, and Socio

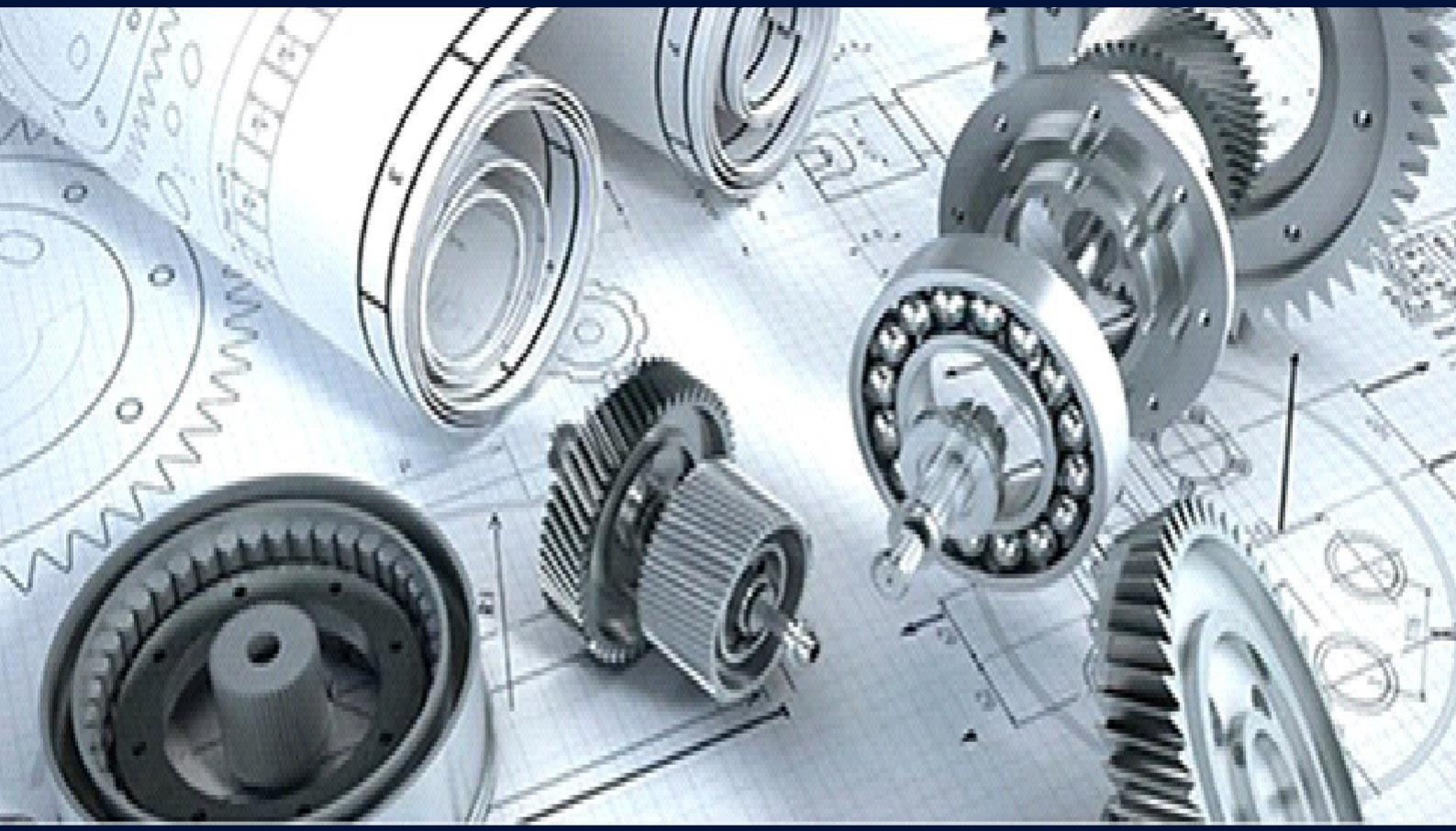
[7] In 2020, Cheng, Li, Li, Silva, Y. N., Hall, D. L., and Liu, H. published "XBullying: Cyberbullyin g Detection within a Multi-Modal Context."*Web Conference 2020 Proceedings (WWW '20)*, 845-856.
(https://doi.org/10.1145/3366423.3380074) DOI: [10.1145/3366423.3380074]

[8] In 2018, Fortuna and Nunes conducted a study titled "A Survey on Automatic Detection of Hate Speech in Text." 85:1–85:30 in *ACM Computing Surveys*, **51**(4). DOI: [10.1145/3232676](https://doi.org/10.1145/32326 76)

[9] In their 2017 paper, "A Survey on Hate Speech Detection using Natural Language Processing," S chmidt, A., and Wiegand, M.*Conference Proceedings on Natural Language Pr ocessing for Social Media, Fifth International Wor kshop, 1–10.[10.18653/v1/W17-1101] is the DOI.

[10] Schmidt, A., & Wiegand, M. (2017). "A Survey on Hate Speech Detection using Natural Language Processing." *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1-10. DOI: 10.18653/v1/W17-1101

[11] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., & Hoste, V. (2018**).** "Automatic Detection of Cyberbullying in Social Media Text." *PLoS ONE*, **13**(10), e0203794. DOI:10.1371/journal.pone.0203794



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

+91 99405 72462 +91 63819 07438 ijmrsetm@gmail.com