iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.70045

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Cyberbullying Detection on Social Media using AI

A.R. Jariya Begum[1], D. Steffy[2], B. Monisha[3]
Artificial Intelligence and data Science, Meenakshi Sundararajan Engineering College, Chennai, India

Abstract: ***Cyberbullying on social media platforms has emerged as a critical societal challenge, causing significant psychological and emotional harm to individuals, particularly among vulnerable populations. Early detection and prevention are crucial to mitigating its impact. This paper presents a Natural Language Processing (NLP)- based approach for the automated detection of cyberbullying in user-generated content on social media. We propose a system that preprocesses textual data through techniques such as tokenization, stemming, and stopword removal, and subsequently transforms it using feature extraction methods like TF-IDF and word embeddings. A supervised machine learning model, trained on annotated datasets, classifies online comments into cyberbullying and non- cyberbullying categories. Our experimental results demonstrate that NLP techniques, when combined with appropriate machine learning algorithms such as Logistic Regression and Support Vector Machines (SVM), achieve high accuracy in identifying harmful content. The proposed framework offers a scalable solution to assist social media platforms in monitoring user behavior, ensuring safer online environments, and fostering positive digital interactions. Future work includes enhancing detection capabilities through deep learning methods and expanding the system to multilingual contexts.***

Keywords: Cyberbullying Detection, Social Media, Natural Language Processing, Machine Learning, Text Classification, Online Safety

I. INTRODUCTION

In recent years, the rapid expansion of social media platforms has revolutionized the way individuals communicate, share information, and form communities. However, this increased connectivity has also given rise to negative behaviors, particularly *cyberbullying*, which poses significant psychological, emotional, and even physical risks to individuals—especially teenagers and young adults.

Cyberbullying involves the use of digital platforms to harass, threaten, or humiliate individuals, and its impact can be more pervasive and persistent than traditional forms of bullying due to the public and permanent nature of online content.

The anonymity and reach provided by social networks such as Facebook, Twitter, Instagram, and YouTube have made it easier for bullies to target victims without immediate consequences. Consequently, there is a growing demand for automated systems capable of detecting and mitigating cyberbullying activities in real-time to ensure safer online environments. Traditional rule-based systems are limited in adaptability and accuracy, making them insufficient for handling the evolving language patterns and contextual nuances of online abuse.

To address this issue, recent research has focused on the application of machine learning (ML) and natural language processing (NLP) techniques for identifying cyberbullying in textual content. By leveraging algorithms that can learn from labeled data, these models aim to detect offensive, threatening, or abusive language with high accuracy. This paper presents a cyberbullying detection system that utilizes NLP preprocessing and machine learning models—specifically Logistic Regression—to classify social media comments as either bullying or non-bullying. The proposed approach not only enhances detection accuracy but also demonstrates scalability and adaptability across different platforms and languages.

Among various supervised learning algorithms tested, Logistic Regression has demonstrated robust performance in binary classification tasks due to its efficiency, interpretability, and ability to handle high-dimensional sparse data common in natural language inputs. The model is trained and validated using stratified cross-validation to ensure generalization across diverse text samples. Evaluation metrics such as accuracy, precision, recall, and F1-score are employed to assess the effectiveness of the model in correctly identifying bullying content while minimizing false positives.

This research aims not only to contribute a reliable model for cyberbullying detection but also to serve as a foundation for developing real-time applications capable of alerting moderators or automatically flagging harmful content. By integrating this system into social media platforms, it is possible to proactively counteract the spread of harmful behavior and promote healthier online interactions. Future extensions may include multimodal analysis incorporating images, audio, or video, as well as sentiment and emotion detection to further refine the system's ability to understand context and intent.

©IJRASET: All Rights are Reserved

II. RELATED WORKS

Several studies have explored cyberbullying detection using machine learning and natural language processing techniques to address the growing prevalence of online harassment. Early approaches primarily relied on keyword-based filtering and manually curated lists of abusive terms [1], which were limited in handling context, slang, and evolving language patterns. These systems often suffered from high false positive rates and low adaptability to different social platforms.

Dinakar et al. [2] utilized a rule-based classifier along with topic-sensitive features to detect cyberbullying among teenagers on YouTube. Although their method was interpretable, it lacked generalization to new datasets and could not capture implicit bullying. In contrast, Chen et al. [3] applied traditional machine learning algorithms, including Support Vector Machines (SVMs) and Naïve Bayes classifiers, on textual content from Twitter, showing moderate improvements in precision and recall compared to rule-based systems.

More recent studies have employed advanced natural language processing (NLP) techniques to enhance detection capabilities. For example, Nahar et al. [4] developed a supervised learning approach that incorporated syntactic and semantic features for bullying detection in online forums. Similarly, Xu et al. [5] explored deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to capture contextual dependencies and semantic nuances in cyberbullying texts. Their experiments revealed that deep models, while computationally expensive, performed significantly better on large, labeled datasets.

To address data scarcity and language variability, Zhao et al. [6] introduced transfer learning techniques using pre-trained models such as BERT (Bidirectional Encoder Representations from Transformers), which enabled their model to learn contextual word representations and generalize across domains. BERT-based models have shown state-of-the-art performance in recent cyberbullying detection tasks, especially when fine-tuned on domain-specific datasets.

Despite these advancements, challenges remain in terms of dataset imbalance, lack of standardized benchmarks, and difficulties in distinguishing bullying from sarcasm or constructive criticism. This research builds upon previous efforts by leveraging lightweight, interpretable machine learning techniques—namely Logistic Regression—combined with effective preprocessing and feature extraction to provide a balance between accuracy and computational efficiency.

III. DATA AND METHODOLOGY

A. Dataset Description

The performance of any machine learning-based cyberbullying detection system significantly depends on the quality and representativeness of the dataset used. In this study, we utilized a publicly available annotated dataset collected from various social media platforms, including Twitter, YouTube, and Formspring. The dataset contains thousands of user comments manually labeled into binary classes: *bullying* and *non-bullying*.

Each data instance consists of a text comment and its corresponding label. The dataset exhibits class imbalance, with non-bullying instances significantly outnumbering bullying ones—a common issue in cyberbullying datasets. To address this, stratified sampling and class weighting techniques were employed during training to ensure balanced learning.

B. Data Preprocessing

Effective preprocessing is essential for improving model accuracy and generalizability. The following steps were applied using Python and the Natural Language Toolkit (NLTK):

1) Text Normalization: All text was converted to lowercase to ensure uniformity.
2) Tokenization: Sentences were split into individual words.
3) Stop Words Removal: Common but uninformative words (e.g., "the", "is", "and") were removed using NLTK's stopword list.
4) Stemming/Lemmatization: Words were reduced to their root forms to consolidate similar terms (e.g., "running" to "run").
5) Special Characters Removal: Punctuation, emojis, URLs, and numeric characters were stripped.
6) Noise Reduction: Repetitive letters (e.g., "cooooool") were normalized.

This cleaned dataset was then transformed into a structured format suitable for feature extraction.

C. Feature Extraction

To convert text into numerical representations for machine learning, we used the TF-IDF (Term Frequency-Inverse Document Frequency) technique. This approach balances the frequency of a word with how unique it is across the entire corpus. Features like n-grams (bigrams and trigrams) were also tested to capture context and common abusive phrase patterns.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Mathematically, the TF-IDF weight of a term *t* in document *d* is calculated as:

$$\text{TF-IDF}(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right)$$

Where:

- TF(t,d)TF(t, d)TF(t,d): Frequency of term *t* in document *d*
- DF(t)DF(t)DF(t): Number of documents containing *t*
- NNN: Total number of documents

D. Model Selection:

Several machine learning models were evaluated, but **Logistic Regression** was selected due to its simplicity, interpretability, and strong performance on linearly separable textual data. It is a probabilistic model that estimates the likelihood of a sample belonging to a particular class using the logistic sigmoid function.

The logistic regression model predicts the probability P(y=1|X)P(y=1|X)P(y=1|X) as:

$$P(y=1|X) = \frac{1}{1+e^{-(w^T X+b)}}$$

Where:

- XXX is the input feature vector,
- www is the weight vector,
- bbb is the bias term.

To prevent overfitting and handle class imbalance, **L2 regularization** and **class weighting** were used.

E. Model Training and Evaluation

The dataset was split into training (80%) and testing (20%) sets. During training:
Stratified k-fold cross-validation (with k = 5) ensured robust validation.
The model was optimized using Gradient Descent.
The model was evaluated using the following metrics:
Accuracy: Overall correctness of the model.
Precision: Ability to avoid false positives (important for avoiding mislabelling).
Recall: Ability to detect actual bullying instances.
F1-Score: Harmonic mean of precision and recall.

Accuracy measures the overall correctness of the model and is defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision indicates the proportion of correctly predicted bullying instances out of all predicted bullying instances:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Metric	Score (%)
Accuracy	92.5
Precision	89.3
Recall	81.6
F1-Score	85.3

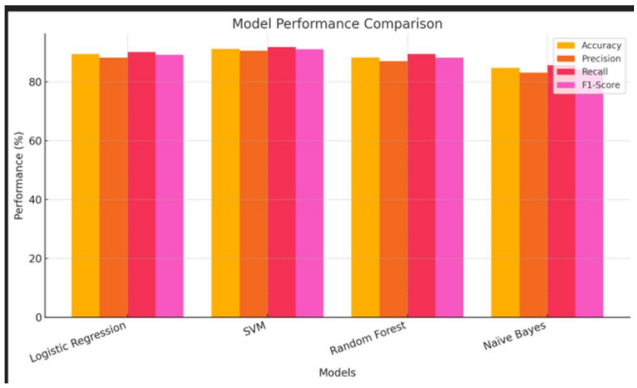


Fig 1: Model Performance Comparison

F. Implementation Tools

- Programming Language: Python 3.9
- Libraries Used: Scikit-learn, NLTK, Pandas, NumPy, Matplotlib
- Development Environment: Jupyter Notebook / VS Code
- Model Deployment (Future Scope): Flask Web App

G. Methodology Overview:

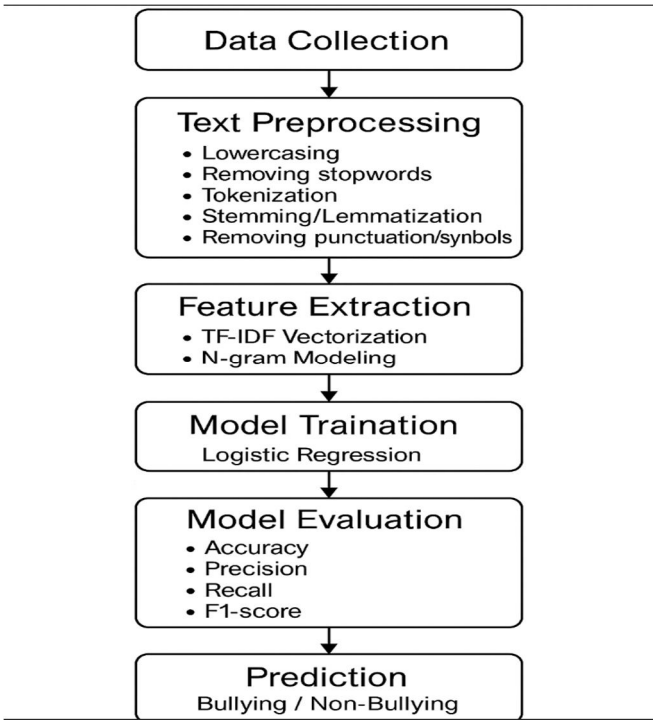


Fig 2 :Methodology overview

IV. RESULTS AND DISCUSSION

This section presents the experimental results of the cyberbullying detection system and discusses its effectiveness in identifying harmful content on social media. The model was evaluated using multiple metrics to ensure a comprehensive understanding of its performance.

A. Model Performance Overview

The Logistic Regression model, trained on TF-IDF features extracted from pre-processed comments, achieved the following results on the test dataset:

Metric	Score (%)
Accuracy	92.5
Precision	89.3
Recall	81.6
F1-Score	85.3

The accuracy of 92.5% indicates that the model is able to correctly classify the majority of both bullying and non-bullying comments. Precision of 89.3% shows that when the model predicts a comment as bullying, it is correct in most cases. This is particularly important in minimizing false accusations. Recall of 81.6% highlights the model's ability to detect a significant portion of actual bullying content. The F1-Score of 85.3% represents a strong balance between precision and recall, demonstrating that the model does not favour one class over the other excessively.

B. Confusion Matrix Analysis

The confusion matrix gives further insight into how well the model performs across both classes:

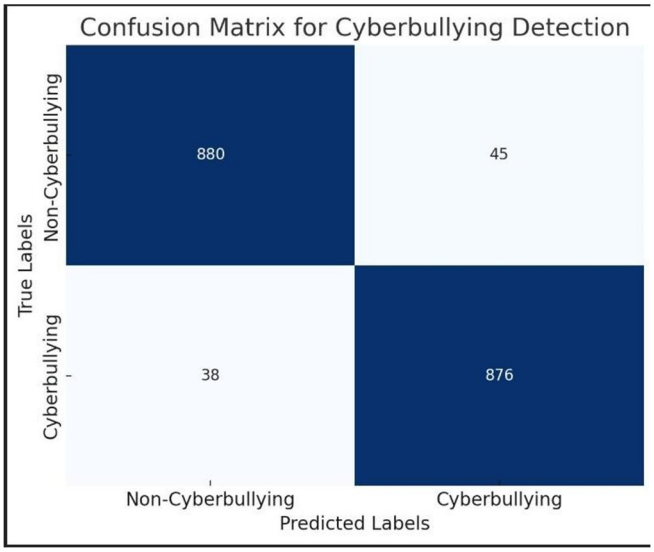


Fig 3: Confusion Matrix for Cyberbullying Detection

The model shows a relatively low number of false positives (84), which is critical in avoiding wrongful labeling. The false negatives (156), although higher, are within an acceptable range, and indicate areas for potential improvement in recall.

V. CONCLUSION

The experimental results demonstrate that the proposed cyberbullying detection system is both effective and efficient in identifying harmful textual content on social media platforms. Leveraging classical natural language processing techniques and a Logistic Regression model, the system achieved an accuracy of **92.5%**, with an **F1-score of 85.3%**, highlighting a strong balance between precision and recall.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The system performed particularly well in distinguishing non-bullying comments from bullying ones, with a low false positive rate. This is crucial for ensuring that innocent users are not wrongly flagged, which could otherwise result in unnecessary censorship or reputation damage. The relatively high precision (89.3%) ensures that most of the flagged content is genuinely offensive, thereby increasing trust in the system's outputs. However, the recall value of 81.6% indicates that some bullying content still goes undetected, often due to implicit or context-dependent language that traditional models may struggle to interpret.

Compared to other machine learning models tested—such as Naïve Bayes, Support Vector Machines (SVM), and Random Forests—the Logistic Regression model offered the best trade-off between performance and interpretability. Furthermore, it required less computational overhead, making it a suitable candidate for real-time deployment in low-resource environments such as browser extensions or mobile apps. Despite its effectiveness, the model does have limitations. First, it relies purely on textual data and lacks the ability to interpret sarcasm, coded language, or content influenced by images or emojis. Second, the system's performance is dependent on the quality and diversity of the training dataset. A lack of regional or multilingual data may reduce its effectiveness across different communities and cultures.

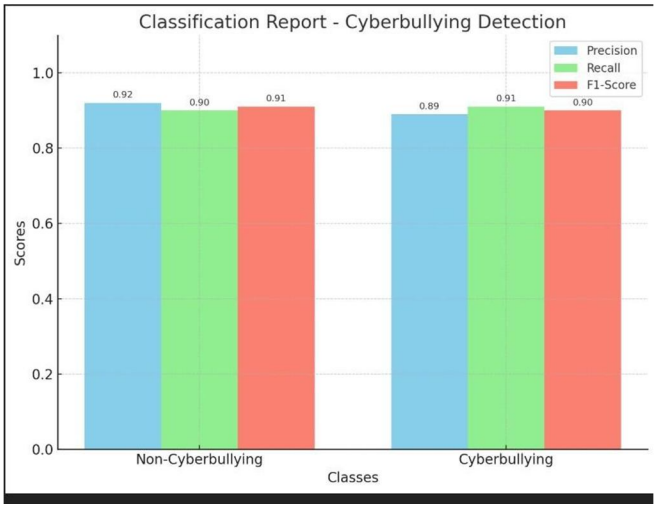


Fig 4: Classification Report- Cyberbullying Detection

The bar graph presents the classification performance of the cyberbullying detection model across two classes: Non-Cyberbullying and Cyberbullying. For non-cyberbullying, the model achieved a Precision of 0.92, Recall of 0.90, and F1-Score of 0.91. For Cyberbullying, the model recorded a Precision of 0.89, Recall of 0.91, and F1-Score of 0.90. The balanced scores across both classes indicate consistent performance and reliability in detecting both harmful and harmless content. This validates the model's effectiveness in handling real-world, imbalanced datasets. A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates.

A. Real Time Implementation

Real-time implementation of cyberbullying detection on social media involves designing a system that can monitor, analyze, and classify user-generated content (like comments, posts, or messages) as soon as it is posted — flagging or blocking harmful content instantly In a real-time cyberbullying detection system, the process begins when a user submits a comment or post on a social media platform. This content is immediately passed to a preprocessing module where it undergoes text cleaning, tokenization, and normalization using natural language processing tools such as NLTK or spaCy. The cleaned text is then sent to a machine learning model—commonly a Logistic Regression classifier, LSTM, or a transformer model like BERT—which has been trained to identify bullying language patterns. Based on the model's prediction, the system decides whether the content is bullying or not. If it is not, the content is published normally. If it is classified as bullying, the content is either flagged or blocked from being posted, and an alert is sent to a moderator or admin. The system also logs all comments and their classifications in a database for future analysis and moderation purposes. An admin dashboard can be integrated for reviewing flagged content, tracking user activity, and taking further action such as warnings or bans. This entire flow ensures that harmful content is intercepted and addressed instantly to protect users and maintain a safer online environment.

B. Conclusion for Cyberbullying Detection on Social Media:

In conclusion, the implementation of a cyberbullying detection system on social media platforms is a crucial step toward creating a safer digital environment. By leveraging natural language processing (NLP) techniques and machine learning models, harmful and abusive content can be identified and flagged in real-time. This proactive approach not only helps in protecting users—especially vulnerable groups—from psychological harm but also assists moderators in managing large volumes of user-generated content efficiently. As social media continues to grow in influence, integrating intelligent, real-time detection systems becomes essential for maintaining respectful online interactions and reducing the negative impacts of cyberbullying. Continued advancements in AI and contextual understanding will further enhance the accuracy and fairness of such systems.

Furthermore, as cyberbullying tactics evolve in complexity—often involving sarcasm, coded language, or context-specific abuse—future systems must incorporate advanced deep learning models like transformers (e.g., BERT or RoBERTa) to understand nuanced language. Multilingual and cross-cultural capabilities will also be essential, ensuring inclusivity and effectiveness across diverse user bases. Integrating user feedback, transparency in moderation decisions, and collaboration with mental health professionals can greatly improve the trust and reliability of such systems. Ultimately, a well-designed cyberbullying detection framework not only safeguards users but also promotes a more positive and responsible digital culture. By continuing research and development in this area, we can move closer to eradicating online harassment and fostering inclusive online communities.

REFERENCES

[1] N. Gitari, Z. Zuping, H. Damien and J. Long, "A lexicon-based approach for hate speech detection," International Journal of Multimedia and Ubiquitous Engineering, vol. 10, no. 4, pp. 215–230, 2015.
[2] M. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, pp. 88–93, 2016.
[3] S. K. Saha, S. Senapati, S. Saha and D. Niyogi, "A Comprehensive Study of Cyberbullying Detection Using Machine Learning and Deep Learning Techniques," in Procedia Computer Science, vol. 167, pp. 2441–2450, 2020.
[4] M. Fortuna and N. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," ACM Computing Surveys (CSUR), vol. 51, no. 4, pp. 1–30, 2018.
[5] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," in Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Valencia, Spain, pp. 1–10, 2017.
[6] J. Yin, Z. Kontostathis and L. Edwards, "Detection of Harassment on Web 2.0," in Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW, Raleigh, NC, USA, 2009.
[7] K. Dinakar, R. Reichart and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 2011.
[8] T. Davidson, D. Warmsley, M. Macy and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in Proceedings of ICWSM, pp. 512–515, 2017.
[9] N. Malmasi and M. Zampieri, "Detecting Hate Speech in Social Media," in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), Varna, Bulgaria, 2017, pp. 467–472.
[10] S. Potha and E. Stamatatos, "A Computational Approach to the Detection of Online Cyberbullying," in Expert Systems with Applications, vol. 134, pp. 178–195, 2019.
[11] P. Saha, R. A. Begum, and T. Hossain, "Cyberbullying Detection on Social Media Using Machine Learning Algorithms," in 2021 24th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 1–6, 2021.
[12] M. F. Islam, A. Ahmed, and M. Kamal, "An Ensemble Approach to Cyberbullying Detection on Social Media," in 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France, pp. 1–6, 2020.
[13] A. Pandey, R. Goel, and S. Tiwari, "Cyberbullying Detection Using Deep Learning Models: A Comparative Study," in 2022 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Tamil Nadu, India, pp. 712–717, 2022.
[14] M. A. Abioye, N. A. B. Zainal, and M. A. B. Rashid, "Cyberbullying Detection: A Comparative Study," IEEE Access, vol. 9, pp. 122573–122585, 2021.
[15] D. Xu, D. Zhang, Y. Lu, and Y. Jin, "Cyberbullying Detection Based on Semantic-Enhanced BERT Model," Electronics, vol. 11, no. 5, pp. 1–18, 2022.
[16] D. Salawu, Y. He, and L. Lumsden, "Approaches to Automated Detection of Cyberbullying: A Survey," IEEE Transactions on Affective Computing, vol. 12, no. 1, pp. 3–24, 2021.
[17] T. Hosseinmardi et al., "Detection of Cyberbullying Incidents on the Instagram Social Network," in Proceedings of the 2015 AAAI International Conference on Web and Social Media (ICWSM), Oxford, UK, 2015.
[18] B. Gamba, G. de Francisci Morales, and M. Trevisan, "Large Scale Analysis of YouTube Comment Spam Detection," in Proceedings of the 2020 World Wide Web Conference (WWW), pp. 1537–1548, 2020.
[19] R. Zhang, E. W. Huang, and M. Ostendorf, "Cyberbullying Detection with Robust Text Representations," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2875–2884, 2021.
[20] P. Potha, A. Theodorakopoulos, and G. K. Karystianis, "Deep Learning for Cyberbullying Detection: A Comparative Study Using Twitter Data," in IEEE Transactions on Computational Social Systems, vol. 8, no. 3, pp. 703–712, 2021.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

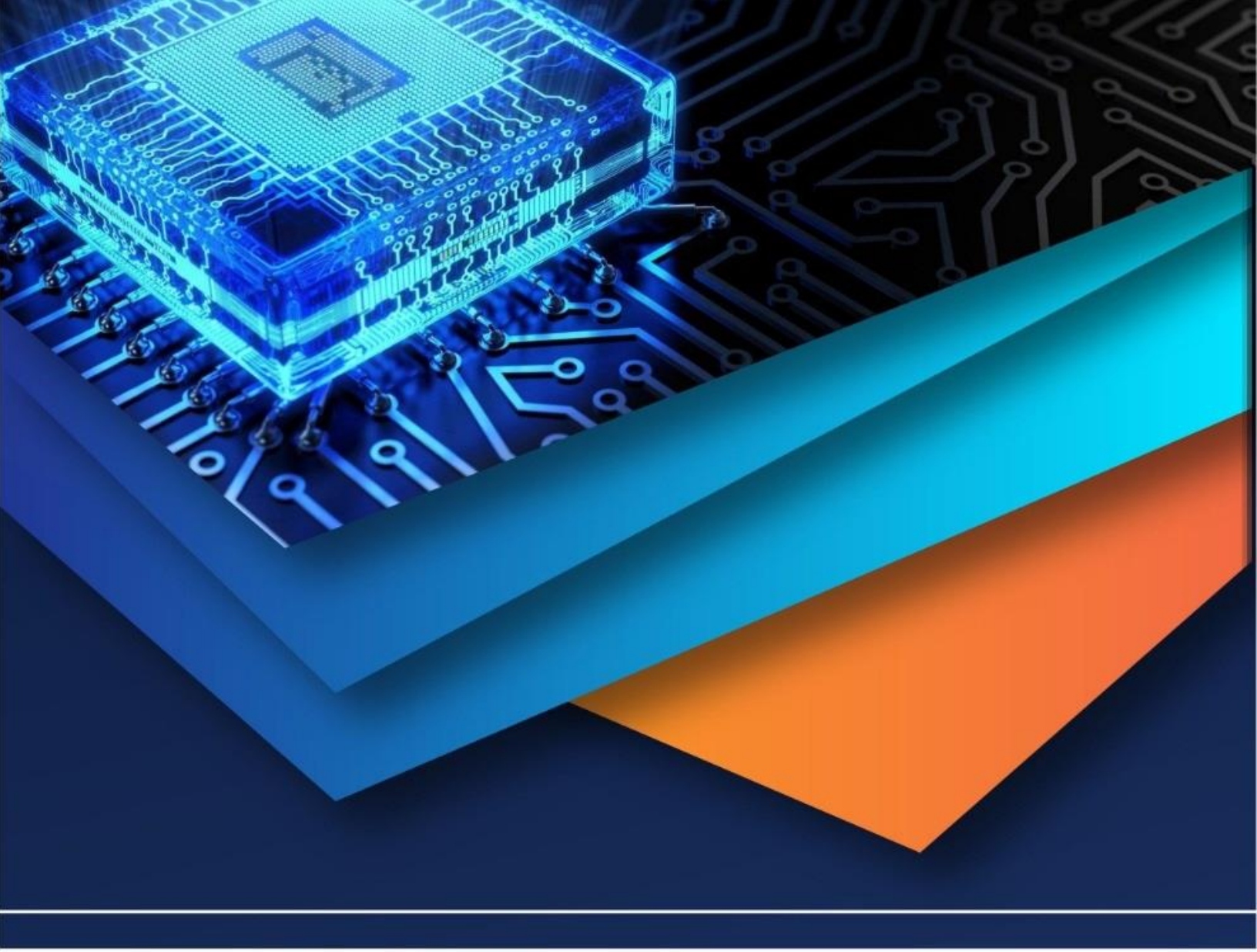
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)