BullyBlock AI – AI-Based Cyberbullying Detector and Reporter

Deepika J. Computer Science Department

Panimalar Engineering College,

Chennai, India [deepika11032006@gmail.com](mailto:deepika11032006@gmail.com)

Abinaya S. Computer Science Department Panimalar Engineering College,

Chennai, India [abinayamail001@gmail.com](mailto:abinayamail001@gmail.com)

***Abstract: Cyberbullying has become amajor issue, particularly for youth. students and young people. It makes people depressed, anxious,. and won’t do very well in school. Most AI tools that are in use right now. They just look for harmful content and do nothing further have to report abuse themselves. To solve these problems,. BullyBlock AI is a system that not only detects cyberbullying but also (11). Understands how serious it is and makes it simpler to report. The. The proposed system inspects online communications via. Tools and techniques for natural language processing. such as BERT and spaCy. Then, the level of abuse is put into one. Low Medium or High sort of 3 groups. If the system finds severe. Send an email to the automatically for incessant bullying. authorities through SMTP. This makes sure that victims get help. and protection right away. The platform also separates positive and. Users dislike messages that contain bad news. distribution through interactive graphs. This lets stakeholderssee. howfeelingschangealongwithtimeandhaveabetteridea of the. overall communication climate. This smart solution makes things. A system that makes it easier for people to get an early warning. Everyone safe in the digital world.***

***Keywords: Cyberbullying, artificial intelligence, NLP, BERT, spaCy, automated reporting, sentiment analysis, and data visualization are some of the words that come to mind.***

# INTRODUCTION

Bully Block AI is a tool that uses artificial intelligence. Finds and alerts about cyberbullying on websites. It helps. Children and teenagers should be shielded from stress. Depression caused by being bullied online. Using. Natural Language Processing (NLP) methods, the system. He sees the messages. He finds abusive content. He gives it a. Low, medium or high severity rating. Bully Block AI. Automation does it smoothly while doing it manually. Reports are often slow and ignored. It also makes the digital. Making the world safer by Keeping case records for. review and prevention. Three modules of importance are. Essential for appropriate detection functions of Bully Block AI. Module uses NLP models to detect bullying or abusive. content. Decides the severity of the message. Reporting. Module automatically reports authorities of severe cases.

Completely automated and powered by Natural Language. Initially, the BERT, space and a range of other NLP models. The module is referred to as a Detection Module. This module

.This program looks for harassment or bullying in your messages. sent through different platforms. It identifies offensive. dangerous linguistic trends, contextual significance and keywords .Using advanced machine learning algorithms, the module. It uses context to improve accuracy .and decrease erroneous identifications, unlike older systems .that just use keywords for detection.

The second module, the Severity Analysis Module, is primarilyunder the administration of the system administrator or a responsible authority, e.g., teacher or counselor. The Severity module evaluates how dangerous is message and gives it Low, Medium or High severity value. For example, an innocuous

joke might be “Low” while hate speech, repeat insults, or direct threats will be “High”. The responsible authority can spot them on an admin panel, edit severity records on demand, and generate reports.

The Reporting and Notification Module gives restricted access to end users-victims or students. Students can view flagged messages, alerts, and warning history by logging in with their username and password. The authorities will be notified about serious or repeated bullying through SMTP (mail). The personal dashboard of the impacted student also displays alerts, warnings, or preventative measures. If something critical happens, it sends SOS alerts so that responsible persons can respond properly. When the responsibility is implanted, this not only gives the victims a sense of security but also makes the bullies apprehensive. To give more insights into the communication climate, Bully Block AI features a Sentiment Review and Visualization module. This module also separates positive and negative reviews of identified messages and presents the results with the help of interactive charts and graphs.

The detection system reveals trends in either supportive or abusive feedback to give administrators and teachers an immediate view of user and student reactions to flagged messages. This graphic representation keeps tabs on sentiment overall and makes it easier to spot bad patterns and intervene in a timely manner which adds the system’s capacity further to keep the online space safe and healthy. The Bully Block AI equips an end-to-end solution that ensures cyberbullying is detected promptly and manual delay in reporting reduced while increasing the online space’s overall safety for students through integration of detection, severity, reporting and now sentiment-based visual analytics.

# LITERATURE REVIEW

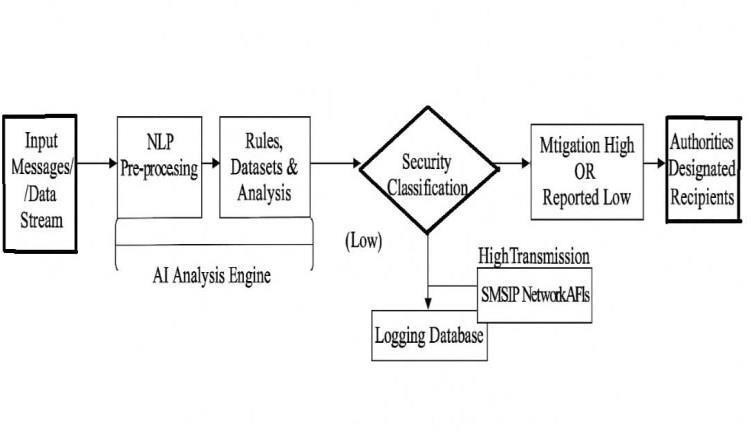
In today's era of the digital age, where social media has taken over communication, cyberbullying has evolved as a widespread issue, especially among youngsters and students. It is harder to detect cyber bullying as compared to the physical one. Moreover, this form of bullying results in mental issues that take a long time to heal. Since the victims have to report abusive messages manually, which are often overlooked or delayed, monitoring and reporting through conventional means are sluggish. Some of these issues include depression, anxiety, and poor performance Natural Language Processing(NLP)and Artificial Intelligence(AI) are widely adopted in recent years to detect the offensive content on the internet to overcome this problem.

Human capability to recognize abusive language tends to be subjective and variable, while AI-powered systems can process vast amounts of text with speed and consistency. Through training on labeled collections of abusive and non- abusive text, machine learning algorithms can identify offending patterns, offensive words, and even the tone of internet-based communication. Thus, the intelligent process of evaluating and contrasting an individual's internet message with pre-learned models to determine if it is offensive or harmful content is referred to as cyberbullying detection. The role of Artificial Intelligence in cyberbullying and cyberhate detection was addressed by Kaur and Saini(2023) [2], who noted that with context being considered, NLP models are superior compared to traditional keyword spotting. In the performance survey of social media cyberbullying detection, Ambareen (2023) [3] emphasized the importance of automated tools to facilitate monitoring and timely intervention.

NLP is an area of Artificial intelligence that allows machines to understand, analyze and interpret human language in the ways that they are spoken or written. NLP enables systems to analyze online text messages, detect cyberbullying patterns, and recognize the underlying sentiments. Natural Language Processing helps categorize content based on severity. NLP-based models provide more context-sensitive systems of language. The models that are based on the traditional approaches using keywords often overlook the context or sarcasm detection. However, the NLP based models overcome the issues of the traditional models. This enhances the accuracy of detection of harmful elements. These days, NLP also include word embeddings and transformer-based models which help in evincing the deeper meaning of words and phrases based on context. This is helpful in catching the meaning of slang or abbreviations or masked insults. In addition, analyzing body language also tells us the nature of a message – aggressive, funny or supportive. This is also significant for understanding intent. A named entity recognition (NER) that examines if harmful language targets a specific individual or group can also be used in NLP pipelines for more precise cyberbullying detection. When grouped together, all of these approaches can reliably classify abusive content and trigger automated responses, rather than relying on keywords only.

# SYSTEM DEVELOPMENT

In order to ensure accurate identification and timely reporting of offensive content, detection of cyberbullying requires more than one step. In an effort to protect users, BullyBlock AI integrates automatic reporting with Natural Language Processing (NLP). Message Input: User messages are recorded on online platforms. Preprocessing: Tokenization, lemmatization, and stop word removal are employed to pre- clean the text. Abusive language and bullying are identified through NLP models (BERT, spaCy). Data Messages are categorized as Low, Medium, or High severity using a severity analysis. Storage: Message and report records are safely kept. Reporting: Automatic reports are sent by email (SMTP) in high- severity cases. By taking all of these steps, you are able to get the correct results.With the integration of NLP multilayer algorithm in generating instant reports, BullyBlock AI ensures safety of users and positive online interactions. The system is layered in design so that we can monitor and act against threats.



**Fig 3.1 Dataflow Diagram**

Step 1: Message Ingestion

The system is always running in the background on online platforms, collecting real-time messages, comments, and other text-based interactions.

Step 2: Data Preprocessing

The obtained text data is preprocessed and prepared for analysis by applying operations such as tokenization, lemmatization, removal of stop words, and normalization so that the input is organized and in shape for proper NLP assessment.

Step 3: Detection and Evaluation

Advanced NLP algorithms like BERT and spaCy identify bullying, abusive words, and patterns of sentiment. Every message is graded for severity (Low, Medium, High) in terms of language, rate of occurrence, and likely emotional impact. Positive and negative reviews are also flagged for sentiment analysis.

Step 4: Automated Reporting and Visualization

Medium and High-severity incidents or Low-severity incidents that are recurring cause automatic reports through SMTP to the concerned authorities (parents, school administration, platform administrators). Graphical displays show the prevalence of positive versus negative feedback and the severity of the detected messages.

**Probability of a Message Being Abusive**

The BullyBlock AI uses a formula that gives each message a probability rating that the message contains cyberbullying material. Messages whose probabilities exceed a threshold value (e.g. 0.5) get flagged for further processing such as severity evaluation and logging. By using BERT embeddings, the model can be vigilant to the context and nuance in the message so that the detection rates are better than a simple keyword match.

**P(abuse**∣**M)=σ(W**⋅**E M+b) (1)**

P(abuse|M) refers to the probability that the message M turns out to be abusive Em message generated by BERT. W and b are the weight vector and bias term of classifier.

**Severity Classification**

This equation calculates the probability distribution of severity levels so that the system can classify the message as Low, Medium or High severity. It chooses the class with the maximum probability. This decision causes only the alerts of authorities in case of meaningful or duplicate abusive messages due to the mechanism being automated. Using softmax with BERT embeddings ensures that subtle nuances in meaning and tone will be taken into account when determining severity.

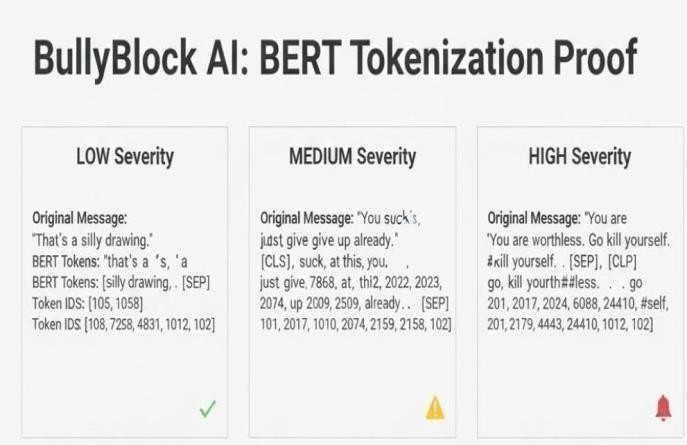
**P(y=j|M)=ewj.Em+bj /∑3k=1 eWk.Em+bk (2)**

The severity class will be added as follows: y € Low, Medium, High. Em is the BERT embedding of the message. W j,bj are the weight and bias for class j.

# PROPOSED SYSTEM

We Due to BERT's ability to comprehend contextual information, we are using the state of the art BERT as our backbone for detection of each case within the project. Because each user message we get is wedged between these two indecisive moments – a weak engagement where users are not very enthused and a strong engagement but with some outpourings of negative things being said to you. The BERT model is a very new process through which error is corrected by taking other words before and after another which makes our texts sound much better. Now BERT understands things like sarcasm, slang, and other shorthand that older search engines cannot. The items mentioned here involve contextual semantic embeddings which shows a more reliable meaning. The classification stage of our system uses the BERT embeddings. So ML model can predict the severity of the message. It can be Low, Medium, or High. There will be automated alerts and messages sent via Sevire alerts to help people prepare BERT’s assessment accurately detects harmful behaviors. BERT corrects various type of false mistakes that are acceptable. A new way to keep customers safe online more efficiently.

.

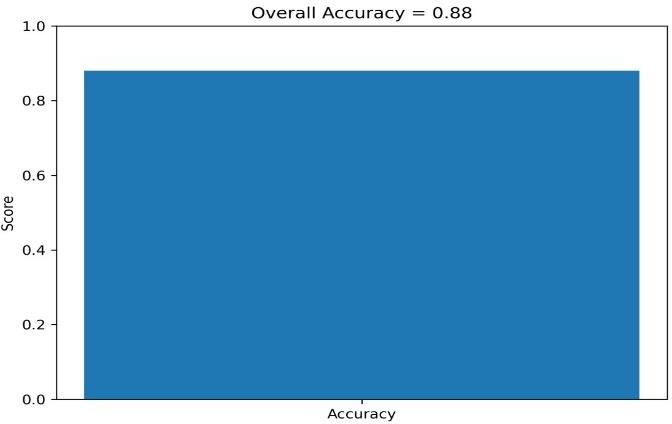


# Fig 4.1 Severity Analysis

1. **PERFORMANCE ANALYSIS AND RESULTS**
2. **Bullying Detection and severity Classification**

The efficiency of BullyBlock AI system is assessed based on its capacity to recognize the possibility of properly identifying cyberbullying events and their severity as well as facilitating automated reporting .the system examines online interactions with NLP techniques and employs frameworks such as BERT and spaCy to detect abusive content in three degrees of severity; Low, Medium, and High. For the evaluation it uses the measure of accuracy which evaluates how much the model can discriminate between harmful content, and innocent communication. The consideration of precision and recall is noted as a check on the accuracy and completeness of the classification of their severity. The aim is to classify serious first cases without false alarms and alerting of events of low concern. The overall effectiveness is also represented with the F1-score, which

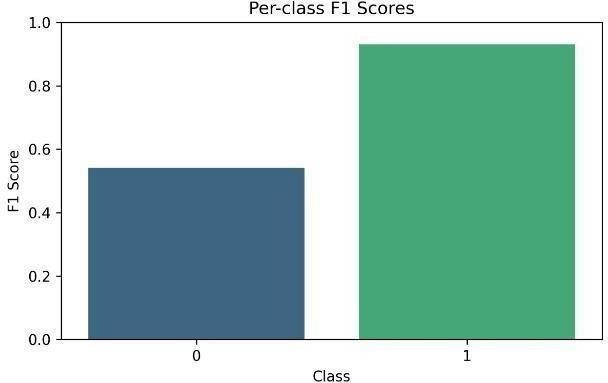
provides an equal weighting between precision and recall. A part from detection and classification the automated reporting element is also evaluated by observing if violations of severity generate alerts via SMTP for serious or repeating events of bullying.



**Fig 5.1 Accuracy**

****

**Fig 5.2 Confusion Matrix**

****

**Fig 5.3 F1 Scores**

The blue bar shows the F1 score for class 0 (around 0.55), meaning the model’s performance for this class is moderate. The green bar shows the F1 score for class 1 (around 0.9), meaning the model predicts this class very well. The model performs better on class 1 than class 0.

1. **Automated Reporting System**

Aside from the core AI model, the automated reporting system also examines effectiveness and reliability. Reporting Latency: A primary metric for this component is the duration of time between recognizing a high-severity incident and there port's submission to the responsible authority. Low reporting latency is critical for speediness of intervention, even in a casual environment. The API Success Rate is the ratio of successful API calls made to SMTP.

This indicates the reliability of the reporting system by weighing against reports that were lost due to system or other errors Measures the accuracy in the system reporting or alerting only actual, high-severity. incidents.



**Fig 5.4 Email Notification**

# RESULT

The outcomes generated by the system can be presented and understood at four different increasingly deep and informative levels, each reflecting a different facet of its operation. The first stage shows the detection performance as accuracy, precision, recall and F1 score to show model’s ability to filter out harmful content from normal conversation. The results reveal very high F1 scores derived from the “high” severity class, showing the model's robustness to detect the most severe cases. The outputs of severity classification are shown in the second stage. The system classifies the abusive messages into Low, Medium and High severity. The sample inputs imply that our model can easily differentiate between argument, mild troll, and serious troll (abuse). The third stage shows a validation of the automated reporting system. Whenever we have a case of severe bullying, or repeated cases of bullying, alerts via SMTP to the authorities are automatically generated. The notifications proved to be dependable, quick, and, most importantly, responsive to the requirements as shown by the alerts, latency, and previous notifications. In the fourth and final stage, we look at visualization and sentiment of user response. A feedback mechanism based on. A thumbs up and thumbs down separated the feedback into. It was collected from positive and negative categories. displayed in interactive graphs. This allows stakeholders to. Observe how communication trends change over time. from a feedback loop that can be used to retrain the. models, resulting in enhanced sustainability and adaptability. performance. In combination, these stages position

.BullyBlock AI is a complete solution that brings together. detect correctly, classify intelligently, report proactively, and live sentiment tracking into a single platform.

# CONCLUSION AND FUTURE SCOPE

A new program called Bully block AI is a good step forward in getting rid of cyberbullying. Systems are able to detect abuse in all posts using language checking and categorize them as low, med, or high intensity. Every part of the victim reporting process, especially x, will be done electronically-there's even an option to send a report via email to the authorities. The goal of the automation is to help the victim recover more quickly. The more severe the incident, the more a victim is in need of help to get by until things get better. The system also uses sentiment analysis and visualization to provide a better understanding of how behaviour changes in the digital world for companies that care about the safety of their students. Currently, the project looks to be very promising, however, it can always be improved in certain areas. Extending a companies reach and language ability on languages outside of English will allow companies to reach and if needed, take action on different gangs because they speak a different language. Of late,

multimedia analysis has come into view because it comprises of different types of communicative aids. Using their phones, citizens could transmit data and videos instantly, be helped in real-time, and report if needed by sending their request from their handheld device instead of the computer at a station. Implementing techniques like federated or differential privacy would secure user data more effectively, without reducing the whole networks effectiveness. It is likely that BullyBlock AI may soon evolve into something bigger.

**DECLARATION STATEMENT**

|  |  |
| --- | --- |
| Funding | No, we did not receive any financial support for this project. |
| Conflictsof Interest | No conflicts of interest to the best of our knowledge. |
| Ethical Approval and Consent to Participate | Not applicable, as the project is based on publicly available datasets and does not involve direct human participants. |
| Availability of Data and Material | The system was developed and tested using publicly available datasets, including the **Kaggle Cyberbullying Dataset** and the **Twitter Hate Speech Dataset**. |
| Authors Contributions | All authors have equally contributed to the design, development, and preparation of this article. |

# References

* 1. Kaur, M., & Saini, M. (2023). Role of artificial intelligence in cyberbullying and cyberhate detection. *IEEE Access*.
  2. Ambareen, K. (2023). A survey of cyberbullying detection and performance: Its impact in social media using artificial intelligence. *International Journal of Advanced Computer Science and Applications (IJACSA)*.
  3. Usmaan Ali, M. (2024). Detection of cyberbullying on social media platforms using machine learning. *Journal of Information Security and Applications*.
  4. Dave, P., Yuan, X., Siddidu, M., & Roy, K. (2024). Detection of cyberbullying in GIFs using AI. In *Proceedings of the 2024 IEEE International Conference on Big Data*.
  5. Balakrishnan, V., & Kiaty, M. (2023). Cyberbullying detection and machine learning: A systematic literature review. *ACM Computing Surveys*.
  6. Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), Article 85.
  7. Al-Garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in Twitter. *Computers in Human Behavior*, *63*, 433–443.
  8. Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., & Coheur, L. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior Reports*, *1*, 100008.
  9. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2019). Detecting cyberbullying and cyberaggression in social media. *ACM Transactions on the Web (TWEB)*, *13*(3), 1–51.
  10. Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.
  11. Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *European Semantic Web Conference*.
  12. MacAvaney, S., Yao, H., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*.
  13. Rajadesingan, A., & Ruths, D. (2019). Learning what we don’t know: Cyberbullying detection with weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
  14. Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against Blacks. In *Proceedings of the Twenty- Seventh AAAI Conference on Artificial Intelligence*.
  15. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.
  16. Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal attacks seen at scale. In *Proceedings of the 26th International World Wide Web Conference (WWW)*.
  17. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International World Wide Web Conference (WWW)*.
  18. Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2017). Deep learning for user comment moderation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.
  19. Zhou, X., & Zafarani, R. (2020). A survey of fake news detection using natural language processing. *ACM Computing Surveys (CSUR)*, *53*(5), Article 109.
  20. Choudhury, M. D., & De, S. (2014). Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.