ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov532

Investigating the Use of AI to Detect Cyberbullying

[1]Agastya Desai

Publication Date: 2025/11/28

Abstract: This paper investigates the effectiveness of artificial intelligence in detecting cyberbullying across multiple platforms. Using a set of simulated chat logs containing signs of bullying, borderline and safe interactions, five widely used AI models were tested and compared. Each system's ability to identify harmful language was measured taking into consideration false positives and negatives. These findings demonstrate the progress of AI moderation tools but also emphasize the importance of human involvement and ethical oversight in preventing harm online.

How to Cite: Agastya Desai (2025) Investigating the Use of AI to Detect Cyberbullying. *International Journal of Innovative Science and Research Technology*, 10(11), 1720-1723. https://doi.org/10.38124/ijisrt/25nov532

I. INTRODUCTION

Cyberbullying has become a global issue, especially among teenagers. As digital communication grows, young people are spending more time on social media platforms. While these tools can be positive spaces for connection, they have also created new ways for individuals to be targeted and harassed.

Cyberbullying is especially damaging because it happens anonymously, at any time and in front of a wider audience than traditional bullying. Studies show that victims of online bullying often experience anxiety, depression and even long term effects on their mental health [1].

Because of the vast amount of content shared and sent online every minute, it is nearly impossible for human moderators to review everything. Many platforms rely on users to report harmful content but this system is slow and inconsistent. To improve safety, companies have started to turn to artificial intelligence to detect harmful messages in real time. AI tools can scan large amounts of content and flag potential bullying, often before a human even recognizes it.

However, the effectiveness of these AI systems remains uncertain. Some models are good at catching obvious forms of abuse, but struggle with more subtle cases. For example, sarcasm, coded language or jokes that depend on context might go undetected while harmless messages might be mistakenly flagged. Understanding how well AI can detect different levels of cyberbullying is important if we want to create safer online spaces without censoring everything.

This research explores the current strengths and weaknesses of AI systems that are used to detect cyberbullying. By creating sample chat logs with clear bullying, borderline messages and safe content and then testing them across several popular AI models, this study aims to measure how accurately these systems respond. The goal is not only to identify how well the AI performs but also to reflect on where it may be making mistakes and the reasoning behind it. Through this analysis, we can better understand how to improve these tools and ensure that they support users without introducing new risks.

II. WHAT COUNTS AS CYBERBULLYING

Cyberbullying is commonly defined as the use of digital platforms to harass, intimidate, or harm others. It can include direct attacks like insults, threats, or exclusion, but it also includes more subtle forms such as spreading rumors, using sarcasm, or making indirect comments meant to hurt someone.

Teenagers are especially vulnerable to cyberbullying due to the amount of time they spend on social media and messaging platforms, where interactions are often casual and fast-paced.

One major challenge in detecting cyberbullying is that harmful messages do not always appear aggressive at first glance. For example, a message like "Nice job, genius" could be supportive or mocking, depending on the tone and context. AI systems often struggle to understand these nuances because they rely on literal interpretation of words, not human intuition or social context. In some cases, even messages that appear neutral or positive might be used to bully someone if they are repeated in a certain way or targeted at a particular individual. This is known as covert bullying.

Sarcasm further complicates the issue. A sarcastic comment like "Wow, you're so popular" might actually be meant to hurt someone who feels isolated or rejected. Detecting sarcasm is difficult even for humans, and current AI models are still far from accurately interpreting these kinds of messages.

Research has found that most natural language processing models struggled to detect sarcasm without additional context, such as voice tone or previous messages in a conversation [2]. This poses a serious limitation when trying to detect bullying using AI.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov532

In addition, bullying is not always about a single message. Often, it is a pattern of behavior that occurs over time. A message that seems harmless on its own may be part of a longer series of posts that together create a hostile environment. For example, repeatedly commenting "lol" on someone's posts might seem harmless, but if used to mock every post made by that person, it becomes targeted harassment. Without access to a full conversation history, many AI systems may miss the deeper intent behind such actions.

Therefore, it is important to understand that cyberbullying is not only about extreme language or threats. It includes a wide range of communication styles and patterns that may not be obvious at first. This complexity makes it essential to carefully evaluate how AI systems classify messages and whether they are capable of handling the grey areas that lie between clear bullying and casual teasing.

III. HOW AI TRIES TO DETECT BULLYING

Artificial intelligence systems use a combination of language processing techniques to detect online bullying and harmful content. These methods are constantly improving, but the core approach relies on analyzing patterns in language that may indicate harassment, threats, or emotional harm. This section outlines three of the most common techniques used in AI moderation: keyword detection, sentiment analysis, and contextual filters.

The most basic form of content moderation is keyword detection. In this approach, AI tools search for specific words or phrases that have been previously associated with bullying. These include insults, slurs, threats, or words related to violence. For example, words like "kill yourself," "fat," or "loser" may trigger a warning or automated response. While this technique is simple and fast, it is not always reliable. Many users who intend to bully someone online avoid using obvious language, or they might spell harmful words incorrectly on purpose to bypass filters. Studies have shown that keyword-based detection systems have a high false positive rate, often flagging messages that are not actually harmful, especially when the words are used in jokes or informal conversations [3].

To improve accuracy, many AI models also rely on sentiment analysis. This method involves analyzing the emotional tone of a message to determine whether it expresses anger, aggression, or negativity. Sentiment analysis uses machine learning algorithms trained on large datasets to recognize whether a message sounds friendly, neutral, or hostile. For example, a message saying "I hate you" may be flagged as negative even if it does not contain any obvious bullying keywords. However, sentiment analysis can be inaccurate when applied to sarcasm or humor. A message might sound neutral but carry a harmful meaning in context. Research has pointed out that sentiment-based classifiers often struggle with messages that appear harmless on the surface but are threatening when combined with previous messages [4].

Contextual moderation tools aim to address this limitation. These systems attempt to analyze not just a single message, but also the conversation before and after it. Some advanced models try to track how a conversation escalates over time or whether one person is repeatedly targeting another. This is closer to how human moderators evaluate bullying, but it requires much more computing power and large-scale data. Context-aware moderation is still in development, and few public AI tools implement it fully. However, it is considered one of the most promising methods for identifying subtle or persistent harassment online.

IV. TESTING THE AI MODELS

To understand how well modern AI systems detect cyberbullying, I conducted a controlled evaluation of five widely used conversational AI tools: ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google), Microsoft Copilot (powered by OpenAI), and Perplexity AI. The aim of this evaluation was to compare their performance in identifying harmful, borderline, and safe messages using a consistent scoring system and realistic data.

The dataset used for testing consisted of 30 chat logs. These logs were created manually and divided into three categories: clear bullying, borderline cases, and safe messages. Each category included approximately 10 examples. The messages were designed to reflect real-world language, incorporating sarcasm, informal tone, and coded language, which are often present in online interactions. The clear bullying messages included direct insults, aggressive commands, or targeted humiliation. Borderline messages were more subtle, often containing sarcasm, exclusion, or passive-aggressive behavior. Safe messages consisted of normal conversations without harmful intent.

Each AI model was presented with the same set of messages individually. For every chat log, I asked the model a standard question, such as: "Is this message likely to be considered bullying or harassment?" or "Does this message contain any harmful or inappropriate content?" The phrasing was kept consistent across all tools, with minor adjustments depending on the platform's requirements. The responses were then rated on a scale of 1 to 10. A score of 10 indicated that the model clearly identified the message as bullying. A score of 1 meant the model did not recognize any harmful content. Scores between 4 and 6 were interpreted as uncertainty or partial recognition.

To reduce bias, the scoring was done based on how explicitly the model identified harmful behavior, whether it provided justification, and how confident it appeared in its assessment. For example, if a model said, "This is clearly bullying due to the aggressive language used," it received a high score. If it said, "This might be rude but not necessarily bullying," it received a mid-range score. If it failed to recognize the issue or claimed the message was acceptable without justification, it received a low score.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov532

This evaluation method aimed to test the AI models' practical effectiveness rather than theoretical accuracy. By using a manually constructed dataset and standardized scoring rubric, the testing process remained consistent and allowed for fair comparison.

V. RESULTS AND WHAT THEY MEAN

To evaluate the performance of the five AI models, each message from the dataset was classified as either bullying or non-bullying. These predictions were then compared with the intended labels to calculate true positives, true negatives, false positives, and false negatives. A true positive occurs when the model correctly identifies a bullying message. A true negative occurs when the model correctly recognizes a safe message. A false positive occurs when the model mistakenly flags a safe message as bullying. A false negative occurs when a bullying message is missed or incorrectly classified as safe.

Table 1 The Results are Summarized in a Confusion Matrix for Clarity. An Example of the Combined Results across All Models

	Predicted Bullying	Predicted Safe
Actual Bullying	72%	18%
Actual Safe	15%	75%

Analysis of the results reveals some clear patterns. The models performed well on explicit bullying messages, such as direct insults or threats, achieving high true positive rates. However, borderline messages, including sarcasm, passive-aggressive comments, or covert harassment, were more difficult for the models to detect. These cases accounted for the majority of false negatives. Moreover, some safe messages that contained strong language or joking insults were sometimes flagged incorrectly, resulting in false positives.

The differences between the AI models were also noteworthy. ChatGPT and Claude generally had higher recall for borderline messages but produced slightly more false positives. Perplexity and Copilot were more conservative, producing fewer false positives but missing subtle bullying. Gemini showed balanced performance but occasionally misclassified sarcastic comments.

These observations indicate that each AI system has distinct strengths and weaknesses, and no single model can be considered fully reliable on its own.

Overall, the results support the conclusion that AI can be a useful tool in identifying cyberbullying, particularly in clear-cut cases, but human oversight remains critical for interpreting subtle or context-dependent interactions.

VI. PROBLEMS AND ETHICAL ISSUES

While this study offers insights into how AI models detect cyberbullying, several limitations need to be considered. First, the dataset used in this research was created with simulated chat logs. Although care was taken to make the conversations realistic, they cannot fully capture the nuance and unpredictability of real-world online interactions [5]. Without authentic data, there is always a risk that the results may not reflect the full complexity of cyberbullying cases.

Another challenge lies in context. Many AI systems rely heavily on isolated sentences or short message fragments, which can lead to misinterpretations. A phrase that appears harmful on its own might be harmless when seen in the broader conversation, and vice versa [6]. This problem can result in both false positives and false negatives, making it difficult to achieve consistent accuracy.

Beyond methodological issues, there are significant ethical concerns. One is privacy. For AI to effectively monitor and detect bullying, it often requires access to personal messages and conversations, which raises questions about how data is stored, who can access it, and how consent is obtained [7]. Another issue is bias in AI models. If the datasets used to train these systems are not diverse enough, the models may unfairly target certain groups or fail to recognize subtle forms of bullying [8]. Lastly, false reporting remains a risk.

Overly sensitive detection systems could incorrectly flag normal conversations, which may frustrate users or even result in unfair punishments [9]. Addressing these problems is crucial to ensure that AI-based detection systems are not only effective but also fair and ethical.

VII. CONCLUSION

This research examined how five different AI models detect cyberbullying by analyzing simulated conversations with clear, borderline, and safe messages. The results showed that while AI is capable of identifying harmful content, there are significant variations in accuracy and consistency across different platforms [6]. These findings suggest that no single approach is sufficient, and that combining multiple methods, such as keyword recognition, sentiment analysis, and contextual evaluation, may improve overall reliability [7].

However, effectiveness alone is not enough. Issues such as privacy, fairness, and accountability must also be addressed if AI is to play a sustainable role in combating online harassment [8]. Cyberbullying is a complex and evolving problem, and while AI offers valuable tools to help reduce its impact, it cannot replace human judgment. Teachers, parents, and platform moderators remain essential in interpreting results, providing emotional support, and creating safe online environments [9].

In conclusion, this study demonstrates both the potential and the limits of AI in detecting bullying. By recognizing current weaknesses and working toward more transparent and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov532

ethical systems, researchers and developers can ensure that AI continues to evolve as a helpful ally in protecting individuals from online harm.

REFERENCES

[1]. 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.
[2]. Zhang, Z., Robinson, D., & Tepper, J. (2016). Detecting sarcasm on Twitter: A contrastive approach. P
[3]. Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of cyberbullying incidents on the Instagram social network.
[4]. Ptaszynski, M., Masui, F., Kimura, Y., Rzepka, R., & Araki, K. (2016). Towards context-aware cyberbullying detection.
[5]. Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying.
[6]. Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection.
[7]. Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale.