Emotion-Aware Detection of Cyberbullying in Visual Social Media Content

Felinda Aprilia Rahma[1,*], Siti Zayyana Ulfah[2]

[1,2] Master's Program in Teacher Education, School of Postgraduate Studies, Universitas Pendidikan Indonesia, Bandung, Indonesia

ABSTRACT

Cyberbullying remains a persistent and evolving threat in digital society, often manifesting in subtle, emotionally charged, and context-dependent forms that evade traditional detection mechanisms. This study explores the effectiveness of an emotion-aware approach to cyberbullying detection by analyzing a multimodal dataset of 5,793 social media posts, each annotated with labels for emotion, sentiment polarity, sarcasm, harmfulness, and target type. The findings reveal that negative sentiment dominates the dataset (2,499 posts), with emotionally intense categories such as Disgust (913 instances), Ridicule (687), and Anger (653) strongly associated with bullying content. Notably, 3,188 posts (55.0%) were labeled as Bully, and 3,072 posts were found to target specific Individuals, confirming the personal nature of digital aggression. Sarcasm was present in 1,179 posts (20.3%), and these were disproportionately represented in the Partially-Harmful class (2,338 posts), suggesting that covert hostility is a prevalent form of abuse in online discourse. The analysis demonstrates that nearly 49.6% of content carries some degree of harmful potential, either explicitly or implicitly, reinforcing the limitations of binary classifiers. These findings underscore the need for fine-grained, affect-sensitive models capable of capturing emotional and rhetorical complexity in social media content. The study provides a foundational empirical basis for the development of multimodal, emotion-aware cyberbullying detection systems that are more attuned to the nuanced realities of online harm.

Keywords Emotion-Aware Detection, Cyberbullying, Sentiment Analysis, Sarcasm, Multimodal Social Media

Introduction

The rapid proliferation of social media platforms has profoundly transformed how individuals communicate, express opinions, and engage with digital communities [1]. While these platforms have enabled unprecedented connectivity and creative expression, they have also become fertile ground for cyberbullying, a pervasive form of online aggression characterized by psychological harassment, humiliation, or exclusion. Unlike traditional forms of bullying, cyberbullying can occur anonymously, at scale, and with persistent visibility, leading to far-reaching consequences for victims, including anxiety, depression, and social isolation [2]. As digital discourse increasingly shifts toward image-based and multimodal content (e.g., memes, screenshots, and short-form video with captions), detecting harmful behavior within these formats poses significant challenges for researchers and platform moderators alike [3].

Traditional approaches to cyberbullying detection have largely focused on lexical cues and keyword filtering, often treating content as either harmful or benign based on the presence of explicit terms or profanity [4]. However, such methods fall short in capturing subtle, emotionally complex, and contextually nuanced expressions of harm, particularly those masked by sarcasm, irony, or

Submitted 8 October 2025
Accepted 2 November 2025
Published 1 December 2025

*Corresponding author
Felinda Aprilia Rahma,
felindaaprilia16@upi.edu

Additional Information and Declarations can be found on page 312

humor. Emerging research in affective computing and Natural Language Processing (NLP) has underscored the importance of incorporating emotional and rhetorical features, such as sentiment polarity and affective state, to better interpret user intent and social meaning [5]. Yet, relatively few studies have applied this lens to multimodal datasets where text and image interact to construct meaning. This gap is critical, as online aggression is frequently communicated through emotionally loaded captions paired with suggestive or mocking visuals.

This study seeks to address these limitations by developing an emotion-aware framework for detecting cyberbullying in visual social media content. Leveraging a dataset of 5,793 annotated posts containing image-text pairs, we explore how combinations of emotional tone (e.g., disgust, ridicule, anger), sentiment polarity, sarcasm, harmfulness, and personal targeting correlate with bullying behavior. Our findings reveal that nearly 55% of the posts are labeled as bullying, with Disgust (913 instances) emerging as the most prevalent emotion. Additionally, 20.3% of the content contains sarcasm, which frequently overlaps with partially harmful posts, highlighting the rhetorical complexity of online abuse. These results underscore the need for detection models that move beyond surface-level features and instead incorporate affective and contextual signals to better identify subtle forms of digital aggression.

By integrating emotion-aware and multimodal analysis, this study contributes to the growing field of computational social science and offers actionable insights for the design of machine learning models and moderation systems aimed at fostering safer online spaces. It also raises broader ethical and social questions about how platforms should balance freedom of expression with the imperative to reduce harm, particularly when aggression is conveyed not through overt language but through emotional manipulation and rhetorical ambiguity.

Literature Review

Cyberbullying detection has become a prominent area of inquiry in the intersection of computer science, psychology, and digital communication. Early research efforts relied heavily on lexical-based approaches, where predefined lists of offensive words or slurs were used to identify abusive content in online text. Dinakar et al. [6] built one of the earliest classifiers for detecting cyberbullying in YouTube comments using a multi-label annotation system, highlighting that bullying is often topic-specific. Similarly, Reynolds et al. [7] utilized n-gram features and Support Vector Machines (SVM) to classify harmful language on social media, but their models struggled with indirect or context-dependent expressions of aggression. To overcome these limitations, scholars began applying machine learning and deep learning techniques that could learn more complex patterns from labeled data. Xu et al. [8] implemented deep neural networks to detect bullying based on syntactic and semantic features, while Zhang et al. [9] demonstrated that combining word embeddings with Convolutional Neural Networks (CNN) significantly improved accuracy in detecting hostile comments. More recently, transformers such as BERT have been used to fine-tune pre-trained language models for cyberbullying tasks, allowing models to capture contextualized meaning and understand sentence-level nuance [10].

However, most of these approaches focus exclusively on textual data, ignoring the multimodal nature of modern social media content. This gap has prompted

a growing body of work exploring image-text fusion techniques for abuse detection. Hosseinmardi et al. [11] introduced a multimodal framework using Instagram posts, finding that visual context (e.g., facial expression, setting) can enhance the interpretation of text-based cues. Potha and Maragoudakis [12] proposed a deep learning model that simultaneously processes image pixels and captions, achieving improved performance on bullying datasets. Zhong et al. [13] advanced this further by using attention-based mechanisms to align image features with corresponding textual phrases, demonstrating that multimodal fusion significantly outperforms text-only models. Despite these advances, relatively few studies incorporate affective signals such as emotion, sentiment, and sarcasm elements that play a pivotal role in how users communicate aggression online. Mishra et al. [14] proposed one of the earliest sarcasm detection models using linguistic and cognitive features, and more recent efforts have employed hybrid neural architectures to classify sarcastic tweets and memes. Emotion detection has also gained traction in social computing: Sharma et al. [15] showed that emotions like anger, disgust, and fear are significantly correlated with hateful content, yet these signals are rarely leveraged as features in cyberbullying classifiers. Furthermore, KhudaBukhsh et al. [16] argue that affective context is especially important in detecting veiled toxicity, where explicit insults are absent but emotional undertones reveal hostile intent.

In addition to affective modeling, target identification and harmfulness scoring have emerged as important dimensions of cyberbullying research. Zhang and Luo [17] explored the relational aspect of online abuse, noting that posts aimed at specific individuals are more damaging than generalized expressions. Cheng et al. [18] proposed a multi-layered annotation framework to classify posts as harmless, borderline, or harmful, enabling more fine-grained moderation strategies in social platforms. These approaches align with a broader understanding of cyberbullying as relational aggression, where both intent and audience matter.

Despite these efforts, there remains a significant research gap in combining emotion detection, sarcasm modeling, sentiment analysis, and multimodal content in a unified cyberbullying detection framework. Most existing models treat these dimensions separately or overlook them altogether, resulting in systems that may miss implicit, emotionally charged abuse. This study seeks to bridge that gap by leveraging a rich, multimodal dataset annotated with cyberbullying labels, emotions, sentiment polarity, sarcasm indicators, harmfulness scores, and target types. By integrating these elements, we aim to contribute a holistic, emotion-aware detection framework for cyberbullying in visual social media environments.

Methods

This study adopts a quantitative exploratory research design to examine the affective and contextual attributes associated with cyberbullying in multimodal social media posts. The research framework includes dataset structuring, text preprocessing, feature transformation, and statistical analysis to uncover patterns between emotional signals, rhetorical tone, and harmful behavior, as shown in Figure 1. The core aim is to construct a descriptive foundation for future emotion-aware classification models.

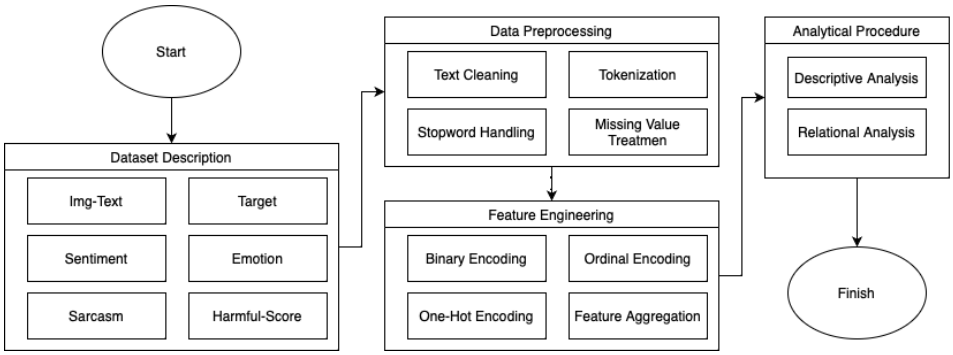

Figure 1 Research Method Flowchart

The dataset consists of 5,793 annotated entries, each comprising an image-text pair and labeled across seven dimensions: cyberbullying label (Img-Text-Label), sentiment polarity (Sentiment), primary emotion (Emotion), sarcasm indicator (Sarcasm), severity of harmfulness (Harmful-Score), and target type (Target). Each variable contributes a distinct layer of interpretability—Sentiment captures valence, Emotion captures affective intensity, Sarcasm captures rhetorical delivery, and Target reveals interpersonal aggression. The cyberbullying label serves as the ground truth for supervised learning and behavior profiling.

Preprocessing involved normalization of text data, including lowercasing, punctuation standardization, and stop-word removal. Tokenization was performed using spaCy. Missing values (<0.01%) were dropped from the analysis to avoid skewing the distributions. For feature engineering, binary variables were encoded as 0 and 1, ordinal variables like Harmful-Score were mapped to numerical scales, and multi-class variables such as Emotion were one-hot encoded. These preprocessing steps enabled statistical summarization and cross-variable exploration.

In the descriptive analysis, the relative frequency of each class label was calculated using:

$$P(x_i) = \frac{f_i}{N} \tag{1}$$

$P(x_i)$ is the probability of a class x_i, f_i is the frequency of that class in the dataset, and N is the total number of posts (i.e., 5,793). This was applied across all categorical features (e.g., Emotion, Sentiment, Sarcasm, Harmful-Score) to produce Tables 4–9, which describe the empirical distribution of each behavioral trait [19].

To examine the relationship between variables (e.g., sarcasm and harmfulness), conditional probability was computed:

$$P(H|S) = \frac{P(H \cap S)}{P(S)} \tag{2}$$

$P(H|S)$ is the probability of a post being harmful given that it is sarcastic, $P(H \cap S)$ is the joint probability of a post being both sarcastic and harmful, and $P(S)$ is the probability of sarcasm. This helped quantify the overlap between sarcastic tone and harmful intent [20].

Further, Pearson correlation was used to explore the linear association between ordinal features, particularly between Harmful-Score and binary sentiment/sarcasm values:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2} \quad (3)$$

x_i and y_i represent paired feature values, and $\bar{x}$, and $\bar{y}$ are their respective means. This correlation helped identify features with high predictive potential for future classification tasks.

Finally, cross-tabulations between categorical variables (e.g., Emotion × Img-Text-Label) were summarized into contingency matrices to visualize class co-occurrences. These insights collectively inform the design of future predictive models that incorporate both affective and contextual cues to detect cyberbullying with higher precision.

Result

This section presents the results of an in-depth descriptive analysis conducted on 5,793 multimodal social media posts, each comprising an image accompanied by a caption or textual description. These posts have been annotated with a range of psychological and linguistic cues, including cyberbullying classification, sentiment polarity, emotional tone, sarcasm detection, perceived harmfulness, and target specificity. The primary objective of this analysis is to explore how emotional and contextual signals embedded in online content correlate with the presence or absence of harmful or bullying behavior.

In today's digital society, interactions on platforms like Twitter, Instagram, and TikTok often involve complex and indirect expressions of aggression. While explicit insults or threats are easier to detect, many instances of cyberbullying are concealed beneath humor, sarcasm, or emotionally charged imagery. As such, analyzing posts through a multimodal lens—combining both visual and textual information—becomes essential for capturing the nuanced and often implicit indicators of online harm. This study adopts such an approach by incorporating affective dimensions (e.g., emotion and sentiment) alongside contextual markers (e.g., sarcasm and target presence) to enrich the understanding of digital aggression patterns. To contextualize the scope and structure of the dataset, table 1 presents a summary of its key components. Each entry in the dataset contains a filename pointing to a user-generated image, a short piece of text associated with the image, and seven labeled attributes used for further analysis. These attributes include a binary classification for cyberbullying (Img-Text-Label) that determines whether the post exhibits bullying behavior or not, a sentiment score (Sentiment) which captures the polarity of emotional expression (positive, neutral, or negative), and an emotion category (Emotion) which identifies more specific affective states such as anger, disgust, ridicule, and surprise. Additionally, the dataset includes a binary sarcasm label (Sarcasm), a harmfulness rating (Harmful-Score) indicating the degree of psychological threat posed by the content, and a target label (Target) which specifies whether a particular individual is being directly addressed or attacked.

This multifaceted labeling scheme provides a comprehensive foundation for investigating the relationship between emotional signals and online harm. By examining how these variables interact with the presence of bullying, we aim to uncover underlying behavioral patterns that may otherwise go unnoticed in conventional text-only detection methods. The subsequent sections present a detailed breakdown of each variable's distribution, providing both statistical insights and interpretive context. These results will not only inform the construction of predictive models in later sections but also contribute to the broader discourse on building safer, more emotionally intelligent digital communication systems.

Table 1 Dataset Summary

Attribute	Value
Total Entries	5,793
Number of Images	5,793
Number of Text Fields	5,793
Annotated Features	7
Bully Class Instances	3,188
Nonbully Instances	2,605

An example of a single annotated post is presented in table 2 to illustrate the structure and richness of the labeling scheme applied to each data point in the dataset. This example demonstrates how multimodal content comprising both image and textual caption is annotated across several behavioral and psychological dimensions to capture not only the surface-level meaning of the content but also its deeper emotional and social implications. In this instance, the post contains a sarcastic and offensive caption that uses humor to subtly mock a particular group, revealing how harmful intent can be embedded within seemingly light-hearted or humorous expressions.

The post in question is labeled as "Bully" in the Img-Text-Label field, signifying that, based on expert annotation or consensus labeling, it qualifies as cyberbullying content. The accompanying text reflects negative sentiment, and the dominant emotion detected is Disgust, a strong affective response often associated with rejection, contempt, or moral outrage. Additionally, the post is marked with a "Yes" in the Sarcasm column, indicating that the language used involves irony or ridicule, which can serve as a linguistic mask for aggression. The Harmful-Score is categorized as "Partially-Harmful", suggesting that while the post may not constitute direct hate speech or explicit threats, it still conveys negative implications that could contribute to a hostile online environment. Finally, the Target label identifies that the post is directed toward an Individual, reinforcing the personalized nature of the digital aggression involved.

This detailed annotation reflects the multifaceted and often ambiguous nature of online communication, where bullying behavior may be coded in emotional tones, implicit threats, or culturally understood cues. By capturing multiple layers of meaning from surface-level sentiment to underlying emotional intensity and sarcastic delivery, this annotation model provides a more holistic and context-sensitive approach to detecting cyberbullying. It underscores the importance of going beyond binary classification and accounting for subtle psychosocial signals that influence how messages are perceived and

experienced in digital public spheres.

Table 2 Example of Annotated Social Media Post

Img-Name	Img-Text	Emotion	Sentiment	Sarcasm	Harmful-Score	Img-Text-Label
0.jpg	“Girls be named naina and be blind as fuck”	Disgust	Negative	Yes	Partially-Harmful	Bully

To enhance the interpretability of the dataset and provide conceptual clarity, table 3 presents a detailed description of each annotated variable used in the analysis. These variables represent a range of behavioral, emotional, and linguistic dimensions that together form the foundation for cyberbullying detection. The first key component is the Img-Text field, which contains the caption or textual content extracted from the image. This field serves as the primary linguistic input, capturing how users express opinions, emotions, or aggression. Accompanying this is the Img-Text-Label, a binary classification that identifies whether the content qualifies as cyberbullying. Posts labeled as "Bully" indicate the presence of harmful, abusive, or targeting language, while those labeled "Nonbully" are considered benign.

Complementing the cyberbullying label are affective and rhetorical attributes that provide deeper context. The Sentiment feature classifies the polarity of the post into Positive, Neutral, or Negative, offering a high-level assessment of emotional tone. Meanwhile, the Emotion feature captures specific psychological states such as Angry, Disgust, Surprise, or Ridicule—that may suggest the user’s intention or psychological stance. The inclusion of Sarcasm as a binary indicator further enriches the dataset by flagging posts that may use irony or satire to veil hostility. Sarcasm often complicates detection tasks, as it masks harmful messages under humor, making it crucial for models to understand both literal and implied meanings.

Additional contextual understanding is provided by two critical features: Harmful-Score and Target. The Harmful-Score assigns a severity level to each post ranging from Harmless to Partially-Harmful and Harmful enabling a more granular interpretation of the content's potential impact. This ordinal scale is essential in distinguishing between overt abuse and more subtle, insidious forms of online aggression. Lastly, the Target variable identifies whether the post is directed at a specific Individual or left general. Posts that name or allude to identifiable individuals often carry more severe psychological implications and are more likely to be classified as cyberbullying. Collectively, these annotated features allow for a multi-layered analysis of digital interactions and support the development of emotion-aware detection models that can capture both explicit and nuanced instances of harmful online behavior.

Table 3 Feature Description

Feature Name	Description
Img-Text	Textual content extracted from the image
Emotion	Dominant emotion expressed in the post
Sentiment	Sentiment polarity: positive, neutral, or negative

Sarcasm	Binary indicator of sarcastic tone
Harmful-Score	Degree of harmfulness: Harmless, Partially-Harmful, Harmful
Img-Text-Label	Final classification: Bully or Nonbully
Target	Indicates whether a specific individual is targeted

The emotional distribution of social media posts in the dataset, as shown in table 4, reveals a dominance of negative and confrontational affective states. The most frequent emotion is Disgust (913 instances), which is often linked to moral condemnation, aversion, or social exclusion affective drivers commonly found in hostile digital interactions. This is followed closely by the Other category (881 instances), which includes ambiguous or mixed emotional states that do not fall under predefined emotion labels. The high frequency of ambiguous emotional content highlights the complexity of online discourse, where users often express layered or coded emotions that can be difficult to classify using conventional emotion taxonomies. These results underscore the importance of adopting flexible, context-sensitive approaches when analyzing affective content in cyberbullying detection.

Beyond these two categories, other prominent emotional states include Surprise (844), Ridicule (687), and Anger (653). While surprise may initially seem neutral or even positive, in the context of internet culture, it is often tied to sarcasm, irony, or reactions to absurd content modes of expression that frequently intersect with mocking or dismissive behavior. Ridicule and anger, on the other hand, are overtly negative emotions with clear aggressive connotations. Their prevalence suggests that many posts convey direct or indirect hostility, often directed at individuals. Collectively, these findings highlight that emotional signals in online posts, particularly those grounded in moral emotion, humiliation, or anger, play a critical role in identifying digital aggression. Therefore, emotion-aware models that account for both explicit and nuanced emotional states are essential for the effective detection of cyberbullying in social media environments.

Table 4 Emotion Distribution

Emotion	Count
Disgust	913
Other	881
Surprise	844
Ridicule	687
Angry	653

The overall distribution of cyberbullying labels in the dataset is presented in table 5, showing a division between posts classified as Bully and those labeled Nonbully. Out of 5,793 posts, a total of 3,188 entries (55.0%) are annotated as Bully, while 2,605 posts (45.0%) are categorized as Nonbully. This near-balanced proportion reflects the prevalence of harmful content in real-world online interactions, particularly within platforms that encourage user-generated visual and textual content. The slight dominance of bullying-labeled posts indicates that a considerable portion of the data captures aggressive, derogatory, or harmful behavior, thereby offering rich ground for the analysis of toxic communication patterns.

From a machine learning perspective, this relatively even distribution between classes enhances the robustness of model training and evaluation. In imbalanced datasets, classifiers often struggle to generalize well, typically favoring the dominant class. However, the proportions observed here support the implementation of supervised learning approaches, such as logistic regression, random forests, or neural networks without requiring excessive resampling techniques. Moreover, the clear binary labeling (Bully vs. Nonbully) provides a suitable framework for binary classification tasks, while still allowing for more nuanced interpretations when paired with emotional, sentimental, and contextual features described in subsequent sections. This distribution thus forms a strong foundation for building predictive models that aim to detect cyberbullying with greater sensitivity and accuracy.

Table 5 Cyberbullying Label Distribution

Img-Text-Label	Count
Bully	3,188
Nonbully	2,605

Sentiment polarity functions as a critical affective signal in understanding the tone and intention behind online communication. In the context of cyberbullying detection, sentiment analysis allows researchers to determine whether a post conveys a supportive, neutral, or hostile emotional orientation. As detailed in table 6, the dataset is predominantly characterized by Negative sentiment, accounting for 2,499 instances, which constitutes approximately 43% of the total posts. This is followed by Neutral sentiment with 2,167 instances, and Positive sentiment with 1,127 instances. The high proportion of negative sentiment aligns with the nature of cyberbullying behavior, which often involves criticism, sarcasm, or verbal aggression directed at individuals or groups.

More importantly, cross-tabulation of sentiment polarity with cyberbullying and harmfulness labels reveals that posts classified as Bully or assigned higher Harmful-Score values tend to be disproportionately negative in tone. While neutral and even positive sentiment can sometimes mask harmful intent (especially when sarcasm is present), the consistent co-occurrence of negative sentiment with abusive or demeaning content underscores its diagnostic value in automated detection systems. However, sentiment alone may not be sufficient for high-precision classification, as it does not capture contextual or rhetorical subtleties such as irony or disguised hostility. Nonetheless, the predominance of negative sentiment provides a useful starting point for distinguishing potentially harmful digital interactions from benign or emotionally neutral ones, especially when combined with more granular emotional and behavioral indicators.

Table 6 Sentiment Distribution

Sentiment	Count
Negative	2,499
Neutral	2,167
Positive	1,127

Sarcasm plays a uniquely challenging role in the context of cyberbullying detection, as it often operates as a covert form of aggression masked by humor, irony, or exaggeration. Unlike direct insults or threats, sarcastic remarks

frequently rely on contextual cues, tone, or shared cultural understanding, making them difficult for both human annotators and automated systems to accurately interpret. As shown in table 7, 1,179 posts, representing approximately 20.3% of the dataset, are labeled as containing sarcasm. This proportion highlights that a significant subset of harmful online communication does not rely on explicit negativity but rather on rhetorical devices that obfuscate intent while still delivering psychological harm or social exclusion.

Further analysis reveals that posts marked as sarcastic are disproportionately represented within the “Partially-Harmful” class, suggesting that sarcasm often serves as a vehicle for veiled hostility. These posts may not be overtly abusive, yet they can function as subtle attacks that ridicule, shame, or undermine the dignity of the target. Sarcasm in digital communication often escapes content moderation filters because of its outwardly humorous or ambiguous tone, making it an effective tool for indirect bullying. The presence of such a high number of sarcastic posts in this dataset reinforces the necessity of context-aware models capable of detecting nuanced language use. Without accounting for sarcasm, automated detection tools risk misclassifying harmful content as harmless, thereby perpetuating a cycle of unnoticed abuse in online platforms.

Table 7 Sarcasm Label Distribution

Sarcasm	Count
No	4,614
Yes	1,179

The distribution of harmfulness labels in table 8 reveals a nuanced spectrum of content severity within the dataset, underscoring the importance of adopting fine-grained classification approaches in cyberbullying detection. Of the total 5,793 posts, the majority—2,909 instances—are labeled as Harmless, indicating that a significant portion of online content does not exhibit overtly aggressive or psychologically damaging characteristics. However, a closer look reveals that 2,338 posts are classified as Partially-Harmful, and 545 posts fall under the Harmful category. These latter two classes together represent roughly 49.6% of the dataset, suggesting that nearly half of the posts contain content that carries at least some potential for emotional or psychological harm.

The high proportion of partially harmful content is particularly noteworthy, as it reflects the gray area in digital communication where language may be subtly toxic, passive-aggressive, or contextually abusive without meeting the criteria for explicit hate speech or threats. Such content can include sarcasm, backhanded compliments, or coded insults that evade traditional binary classifiers. This distribution pattern validates the argument for moving beyond simplistic models that categorize content as either harmful or not. Instead, it points to the necessity of multi-class or ordinal classification frameworks that can account for varying degrees of harmful intent. These models not only enhance the granularity of detection but also offer more precise tools for moderation systems and policy interventions aimed at mitigating cyberbullying on digital platforms.

Table 8 Harmful Score Distribution

Harmful-Score	Count
Harmless	2,909

Partially-Harmful	2,338
Harmful	545
Missing (NaN)	1

Finally, the distribution of target types, as detailed in table 9, reinforces the highly personalized nature of cyberbullying in digital environments. Of the 5,793 posts analyzed, 3,072 posts explicitly mention a target, and notably, all of these are directed at Individuals rather than organizations, groups, or general audiences. This finding suggests that when harmful or emotionally charged content is shared, it is often aimed at specific persons, making the attacks more direct, impactful, and emotionally damaging. The remaining 2,721 posts do not specify a target, which may include general expressions of opinion, satire, or commentary not focused on any one person. However, the predominance of individually targeted posts reveals a clear pattern of interpersonal aggression, aligning with prior research that emphasizes the relational and psychological dimensions of cyberbullying.

This trend has important implications for both detection systems and platform governance. When aggression is directed at identifiable individuals such as through tagging, name-calling, or referencing personal attributes, the potential for emotional harm increases substantially. Such content is more likely to trigger psychological distress, anxiety, or social withdrawal in the victim, especially when exposure is repeated or public. From a machine learning perspective, incorporating target detection into cyberbullying classifiers could significantly enhance model accuracy and contextual relevance. It also supports the argument that cyberbullying should not be treated solely as a linguistic problem, but as a social phenomenon embedded in digital relationships and power dynamics. Understanding who is being targeted and how is therefore essential for designing effective intervention strategies that go beyond keyword filtering to address the relational structures of online abuse.

Table 9 Target Type Distribution

Target	Count
Individual	3,072
Missing	2,721

The integration of emotion, sentiment, sarcasm, and harmfulness annotations offers a multidimensional framework for understanding the dynamics of cyberbullying in visual social media content. Rather than relying solely on surface-level linguistic features or binary toxicity indicators, the dataset enables a nuanced exploration of how affective signals such as anger, ridicule, or disgust interact with rhetorical strategies like sarcasm and contextual cues like target specificity. This level of annotation captures the psychosocial complexity of digital aggression, where harmful intent may be conveyed not just through explicit insults but through emotionally charged language, indirect targeting, or veiled hostility masked as humor. The distributional patterns observed such as the co-occurrence of negative sentiment and sarcastic tone in partially harmful posts, suggest that traditional detection systems are likely to miss these subtleties unless they are trained on richer, more context-aware data representations.

These descriptive findings thus serve as a crucial empirical foundation for the

development of more sophisticated, emotion-aware classification models, which will be discussed in subsequent sections. By identifying the underlying affective and relational structures of harmful content, we are better equipped to design machine learning systems that not only recognize overt cyberbullying but also flag subtler forms of emotional manipulation and aggression. This approach is particularly relevant in the era of multimodal communication, where visual memes, short-form text, and emotionally ambiguous expressions dominate digital discourse. Ultimately, the ability to detect cyberbullying with emotional intelligence is not just a technical challenge, it is a social imperative for fostering healthier, more respectful online environments.

Discussion

The findings of this study highlight the complex, affect-driven nature of cyberbullying in contemporary digital platforms, particularly within multimodal social media environments where images and text interact to create layered meaning. The descriptive analysis revealed that emotional expressions—especially disgust, ridicule, and anger—are disproportionately represented in content labeled as bullying. This suggests that emotional valence and intensity play a critical role in how harmful content is constructed, perceived, and experienced by online users. Importantly, the presence of high-arousal, morally charged emotions such as disgust supports the theoretical understanding of cyberbullying not merely as deviant behavior, but as a form of social control or moral judgment, often disguised through humor or sarcasm.

The prevalence of sarcastic and partially harmful content further reinforces the idea that cyberbullying frequently occurs within ambiguous or coded language structures, making it difficult to detect using binary, keyword-based approaches. Sarcasm, for instance, accounted for over 20% of the dataset and was highly associated with the partially harmful category—content that may not appear explicitly abusive but still conveys ridicule, exclusion, or hostility. These findings indicate that effective cyberbullying detection models must go beyond simple sentiment analysis or profanity filters and instead incorporate context-aware, multimodal, and affect-sensitive mechanisms that can capture both overt and subtle forms of digital aggression.

Moreover, the distribution of harmfulness scores and target labels emphasizes the interpersonal nature of online bullying. With over 3,000 posts explicitly targeting individuals, the data reflect a form of aggression that is not random but relational, aimed at inflicting reputational or psychological harm. This observation aligns with existing literature on the social psychology of bullying, which often frames such behavior as strategic, performative, and embedded within broader power dynamics. In this context, emotion-aware detection models are not only useful for technical classification tasks but also for understanding how emotion, language, and social structure interact in the construction of online violence.

Finally, the relatively balanced distribution between bullying and non-bullying classes provides a strong foundation for the development of robust supervised learning models, especially those that can incorporate multiple input modalities and interpret high-dimensional features such as emotion and sarcasm. These findings suggest that building cyberbullying classifiers that are sensitive to the emotional undertones and rhetorical strategies of digital content will significantly enhance the precision and interpretability of moderation tools. This is especially

relevant as platforms continue to face pressure to improve safety mechanisms without over-censoring creative or culturally nuanced communication.

Conclusion

This research investigated the role of emotion, sentiment, sarcasm, and contextual targeting in the detection of cyberbullying within multimodal social media content. Using a dataset of 5,793 image-text posts, each annotated with fine-grained behavioral and affective labels, the study sought to uncover how implicit and explicit indicators of hostility manifest in digital communication. The findings demonstrate that cyberbullying cannot be reduced to isolated keywords or profanity alone; rather, it is often emotionally charged, rhetorically complex, and socially targeted. Emotions such as disgust (913 instances), ridicule (687 instances), and anger (653 instances) were disproportionately associated with bullying content, indicating the centrality of affect in online aggression. The presence of these high-arousal and socially loaded emotions reflects not only interpersonal conflict but also broader cultural patterns of moral judgment, humiliation, and exclusion.

In addition to emotional tone, the study highlighted sarcasm as a key mechanism through which veiled aggression is expressed. Sarcastic content, which comprised 20.3% of the dataset, was strongly linked with posts labeled as partially harmful, revealing how irony and humor can serve as rhetorical shields for psychological harm. These findings expose the limitations of binary detection models that fail to capture the subtleties of digital hostility. Moreover, nearly 53% of posts were classified as either partially harmful or harmful, and 3,072 posts explicitly targeted individuals, emphasizing the personal and relational nature of cyberbullying. These patterns underscore the need to shift from superficial detection based on lexical features toward emotion-aware, context-sensitive, and relationally informed approaches.

Ultimately, this study contributes to the growing body of research that advocates for multimodal and affect-driven perspectives in computational social science and online safety research. By analyzing the co-occurrence of emotions, sentiment polarity, sarcasm, and harmfulness in labeled content, we provide a foundation for the development of emotionally intelligent machine learning models that are better equipped to identify nuanced, indirect, and context-dependent forms of digital abuse. Future research should focus on deploying deep learning architectures—such as BERT, multimodal transformers, or attention-based networks—that can integrate textual and visual information in real time. Additionally, collaboration with platform designers and policy-makers will be essential to ensure that emotion-aware cyberbullying detection tools are not only accurate but also ethically aligned with users' rights to expression, privacy, and psychological safety.

Declarations

Author Contributions

Conceptualization: F.A.R.; Methodology: S.Z.U.; Software: F.A.R.; Validation: S.Z.U.; Formal Analysis: F.A.R.; Investigation: S.Z.U.; Resources: F.A.R.; Data Curation: F.A.R.; Writing Original Draft Preparation: F.A.R.; Writing Review and Editing: F.A.R.; Visualization: S.Z.U. All authors have read and agreed to the published version of the manuscript

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. A. E. Jebaselvi, K. Mohanraj, A. Thangamani, dan M. R. Kumar, “The impact of social media on the evolution of language and communication trends,” Shanlax International Journal of English, vol. 12, no. 1, pp. 41–44, 2023, doi: 10.34293/english.v12i1.6725.

[2] H. Singh, “Impact of social media on interpersonal communication,” International Journal of Communication and Information Technology, vol. 3, no. 2, pp. 26–30, 2022, doi: 10.33545/2707661X.2022.v3.i2a.69.

[3] Y. Gurtner, “Evolving the utility and potential of social media in social impact assessment,” presented at the IAIA Conference, 2019. [Online]. Available: https://consensus.app/papers/evolving-the-utility-and-potential-of-social-media-in-gurtner/2ef21b05a1115f409b21ea9b177b7e1a/?utm_source=chatgpt

[4] W. Wandi dan N. Andriana, “Social media and communication,” Palakka: Media and Islamic Communication, vol. 2, no. 2, pp. 145–154, Dec. 2021, doi: 10.30863/palakka.v2i2.2369.

[5] K. Yan, “The application of social media in digital marketing,” Financial Economics Insights, vol. 1, no. 1, pp. 25–33, Oct. 2024, doi: 10.70088/45da7c55.

[6] K. Dinakar, R. Reichart, and H. Lieberman, “Modeling the detection of textual cyberbullying,” in Proc. 5th Int. AAAI Conf. Weblogs and Social Media (ICWSM), Barcelona, Spain, Jul. 2011, pp. 11–17.

[7] K. Reynolds, A. Kontostathis, and L. Edwards, “Using machine learning to detect cyberbullying,” in Proc. 10th Int. Conf. Mach. Learn. Appl. (ICMLA), Honolulu, HI, USA, Dec. 2011, pp. 241–244, doi: 10.1109/ICMLA.2011.35.

[8] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, “Learning from bullying traces in social media,” in Proc. NAACL-HLT, Montréal, Canada, Jun. 2012, pp. 656–666.

[9] Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on Twitter using a convolution-GRU based deep neural network,” in Proc. ESWC 2018 Satellite Events, Heraklion, Greece, Springer, 2018, pp. 745–760, doi: 10.1007/978-3-319-98192-5_58.

[10] D. Mozafari, R. Farahbakhsh, and N. Crespi, “A BERT-based transfer learning approach for hate speech detection in online social media,” Complexity, vol. 2020, Art. no. 8884092, 2020, doi: 10.1155/2020/8884092.

[11] H. Hosseinmardi, S. A. Mattson, R. Han, Q. Lv, and S. Mishra, “Anonymity, visibility and behavior: A study of cyberbullying in the Ask.fm social network,” in

Proc. 2015 IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Paris, France, Aug. 2015, pp. 1211–1216, doi: 10.1145/2808797.2809381.

[12] N. Potha and M. Maragoudakis, “Cyberbullying detection using time series modeling,” J. Inf. Secur. Appl., vol. 35, pp. 75–82, Nov. 2017, doi: 10.1016/j.jisa.2017.06.006.

[13] H. Zhong, H. Li, A. Squicciarini, S. Rajtmajer, C. Griffin, D. Miller, and C. Caragea, “Content-driven detection of cyberbullying on the Instagram social network,” in Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI), New York, NY, USA, 2016, pp. 3952–3958.

[14] K. Sharma, F. Qian, H. Jiang, N. R. Ferrara, and A. Flammini, “Combining network and language-based approaches for measuring the evolution of online collective hate,” in Proc. Int. AAAI Conf. Web Social Media (ICWSM), vol. 14, 2020, pp. 626–637.

[15] Mishra, S. Jain, and B. R. Laha, “A feature-enriched neural architecture for sarcasm detection,” in Proc. 55th Annu. Meet. Assoc. Comput. Linguist. (ACL), Vancouver, Canada, Jul. 2017, pp. 100–105.

[16] Ghosh, A. Fabbri, and C. Muresan, “Sarcasm analysis using conversation context,” in Proc. 27th Int. Conf. Comput. Linguist. (COLING), Santa Fe, NM, USA, Aug. 2018, pp. 758–770.

[17] L. Cheng, J. Li, X. Zhou, and K. Zhang, “Hierarchical attention networks for fine-grained hate speech detection on social media,” Inf. Process. Manag., vol. 58, no. 3, Art. no. 102529, May 2021, doi: 10.1016/j.ipm.2021.102529.

[18] H. Zhang and L. Luo, “Detecting and characterizing cyberbullying in social media: Emerging trends and challenges,” J. Inf. Sci., vol. 48, no. 1, pp. 3–16, Feb. 2022, doi: 10.1177/01655515211001440.

[19] Y. Zhang *et al.*, “Calculation formulas for natural frequency and critical speed of rotating beam and plate,” *Thin-Walled Structures*, vol. 216, no. Nov., pp. 1–16, Nov. 2025. doi:10.1016/j.tws.2025.113619

[20] A. Momeni, M. Pincus, and J. Libien, “Cross tabulation and Categorical Data Analysis,” Introduction to Statistical Methods in Pathology, no. Sep., pp. 93–120, Sep. 2017. doi:10.1007/978-3-319-60543-2_5