From Memes to Harassment Automated Detection of Cyberbullying for Cyberlaw Enforcement

Sarmini[1,*], Axel Sandi[2]

[1]Informatics Department, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[2]Information System Department, Universitas Amikom Purwokerto, Banyumas Indonesia

ABSTRACT

This study investigates the automated detection of cyberbullying in multimodal online content, focusing on the intersection of technical classification and legal relevance. Using a dataset of 5,793 entries consisting of both image and text components, we analyze several key attributes including bullying classification, sentiment polarity, sarcasm, emotional tone, harmfulness score, and target type. The results reveal that 55% of the content is labeled as bullying when both image and text are considered together, compared to only 27% when using image-only classification. Negative sentiment dominates (2,657 entries), with sarcasm present in 2,154 entries, highlighting the prevalence of veiled or implicit abuse. Emotional annotations show that disgust (913) and sadness (580) are among the most common emotional tones associated with harmful content. Furthermore, the majority of abusive posts (2,405) are targeted at individuals, underscoring the need for stronger personal protection within cyberlaw frameworks. These findings support the development of context-aware and legally informed detection systems capable of addressing the nuanced and often implicit nature of online harassment.

Keywords Cyberbullying Detection, Multimodal Analysis, Sarcasm Detection, Sentiment Classification, Cyberlaw Enforcement

Introduction

In recent years, the digitalization of communication has transformed how individuals interact, share opinions, and build communities [1]. Social media platforms such as Twitter, Instagram, and Facebook have become dominant arenas for public discourse, where content in the form of memes, images, and short textual posts can reach wide audiences within seconds [2]. While these platforms offer unprecedented opportunities for expression and connection, they also create fertile ground for the emergence of cyberbullying, a form of digital aggression that is often subtle, context-dependent, and emotionally harmful. Cyberbullying differs from traditional bullying in that it operates within a pervasive, asynchronous, and often anonymous environment, where the impact can be amplified through public visibility and permanence [3]. Victims may experience repeated exposure to harmful content, even in the absence of direct interpersonal contact. Modern forms of cyberbullying frequently leverage memes, sarcasm, visual cues, and emotionally loaded language, making them difficult to detect through conventional moderation systems. The shift toward multi-modal communication, where images and text are combined to convey layered messages, has further complicated the task of identifying abusive behavior online.

From a regulatory standpoint, various legal frameworks—such as Indonesia's Undang-Undang Informasi dan Transaksi Elektronik (UU ITE), the European Union's Digital Services Act, and global norms on digital harm—have

Submitted 3 October 2025
Accepted 17 November 2025
Published 1 December 2025

*Corresponding author
Sarmini,
2437083013@webmail.uad.ac.id

Additional Information and Declarations can be found on page 327

DOI: 10.63913/jcl.v1i4.43

How to cite this article: Sarmini, A. Sandi "From Memes to Harassment Automated Detection of Cyberbullying for Cyberlaw Enforcement," *J. Cyber. Law.*, vol. 1, no. 3, pp. 314-329, 2025.

acknowledged the urgency of addressing online abuse. However, enforcement remains challenging due to the subjective nature of harm, the diversity of cultural interpretations, and the lack of scalable tools for early detection and evidentiary documentation. Many current moderation algorithms rely on static keyword lists, explicit language cues, or surface-level sentiment analysis, which are insufficient to capture the pragmatic, emotional, and contextual dimensions of modern cyberbullying.

This research addresses the urgent need for intelligent systems that are capable of detecting cyberbullying in multi-modal contexts, with an emphasis on practical application in legal and regulatory domains. Using a labeled dataset of 5,793 image-text social media posts, the study analyzes various attributes, including bullying classification, sentiment polarity, sarcasm detection, emotional tone, harmfulness score, and target type. The dataset reflects the complexity of real-world content where abusive behavior may be subtle, sarcastic, emotionally manipulative, or disguised as humor. The primary objective of this study is twofold: (1) to explore how patterns within multi-modal content can be leveraged to improve the accuracy of automated cyberbullying detection systems, and (2) to assess how such systems can support cyberlaw enforcement by identifying content that may meet the legal criteria for online harassment, defamation, or psychological harm. Specifically, the research investigates how emotional expression, sarcasm, and multi-modal coherence contribute to the perception and classification of abusive content.

The contributions of this paper are threefold. First, it demonstrates that multi-modal approaches, which consider the combined effects of image and text, outperform unimodal models in identifying harmful content. Second, it provides empirical evidence of the emotional and rhetorical strategies commonly used in online abuse, including the prevalence of sarcasm and emotionally charged language such as disgust and sadness. Third, it proposes a framework for using such detection outputs as supporting evidence for legal and regulatory intervention, thereby bridging the gap between machine learning systems and real-world policy applications.

By integrating insights from machine learning, digital forensics, and legal informatics, this research contributes to the broader discourse on responsible AI in online governance. It argues that the future of content moderation and cyberlaw enforcement lies not in simplistic binary classification, but in context-aware, ethically grounded, and legally interoperable detection systems that can scale with the complexity of digital communication.

Literature Review

The rise of cyberbullying as a social and legal concern has prompted extensive interdisciplinary research over the past decade. Cyberbullying, broadly defined as the use of digital communication to harass, demean, or intimidate individuals, poses significant psychological risks due to its public, persistent, and often anonymous nature. Unlike traditional bullying, which is typically confined to physical or verbal confrontation in limited contexts, cyberbullying can occur at any time, reach wide audiences instantly, and leave permanent digital traces. Scholars such as Patchin and Hinduja [5] and Slonje et al. [6] have emphasized the lasting emotional harm caused by online abuse, particularly when it is repetitive or targeted at vulnerable individuals. In response, many governments, including Indonesia through the Undang-Undang Informasi dan Transaksi

Elektronik (UU ITE), have introduced legal frameworks aimed at preventing and prosecuting online harassment. However, these frameworks face substantial challenges in practice, largely due to the contextual ambiguity of digital content and the lack of scalable, reliable detection tools.

Early research on cyberbullying detection focused primarily on textual data, using traditional machine learning models such as Naïve Bayes, logistic regression, and support vector machines. These approaches typically relied on keyword spotting, n-gram extraction, and sentiment scores to flag offensive content. While useful in specific domains, such systems were highly sensitive to linguistic variation and often failed to capture sarcasm, humor, or indirect aggression hallmarks of modern digital abuse. In recent years, more sophisticated models based on deep learning have been developed. These include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based architectures like BERT, which have shown improved performance in capturing the semantic and contextual relationships embedded in abusive language [7], [8]. However, the vast majority of these systems are limited to text-only input and are unable to process the increasingly visual and multi-modal nature of cyberbullying.

As memes, image-caption formats, and sarcastic visual content have become more central to online interaction, researchers have turned to multi-modal cyberbullying detection. Studies by Hosseinmardi et al. [9], Zhong et al. [10], and Khandekar et al. [11] introduced the use of image classifiers alongside text-based features to detect abuse on platforms such as Instagram and Twitter. These approaches demonstrated that the combination of visual and linguistic cues significantly enhances classification accuracy, particularly in cases where neither modality alone would indicate harm. Deep multi-modal fusion techniques, such as those explored by Vempala and Preotiuc-Pietro [12], have shown promise in modeling the interaction between image aesthetics and textual tone. However, integrating images and text raises further challenges: the need to detect rhetorical devices like sarcasm, the difficulty in interpreting emotion from visual content, and the potential for legal gray areas where content may be offensive without violating any explicit rules. To address these challenges, researchers have incorporated emotion analysis [13], sarcasm detection [14], and harmfulness scoring [15] to better reflect how users perceive and are affected by digital content.

From a legal and regulatory perspective, scholars such as Citron [16] and Gillespie [17] have pointed to the gap between how digital abuse operates and how existing laws define and regulate it. While most national legal frameworks criminalize harassment, incitement, and defamation, they often struggle to account for subtler forms of abuse, especially when communicated through memes or ironic visuals. The ambiguity of context, tone, and target complicates the use of such content in legal proceedings. Recent studies, including those by Mittelstadt et al. [18] and Binns [19], argue for the need to align AI-based moderation tools with legal interpretability and procedural fairness. There is also growing interest in explainable AI (XAI) to ensure that detection decisions can be audited and justified in legal contexts.

In summary, while significant progress has been made in text-based cyberbullying detection and the emergence of multi-modal classification models, the intersection between these technologies and their application in cyberlaw enforcement remains underdeveloped. Few studies to date have attempted to

systematically connect automated detection outputs with legal criteria for digital harm. This research seeks to fill that gap by analyzing a richly labeled image-text dataset through the lens of both machine learning and legal applicability, intending to support more scalable, context-aware, and legally relevant approaches to combating online abuse.

Methods

This study adopts a data-driven approach to develop and evaluate an automated system for detecting cyberbullying in multi-modal content, with a focus on image-text combinations commonly found in social media posts. The methodological framework consists of four key stages: dataset specification, preprocessing, model construction and training, and performance evaluation. Each stage is carefully designed to ensure that the resulting classification models can be applied not only for moderation purposes but also in support of legal review within the context of cyberlaw enforcement.

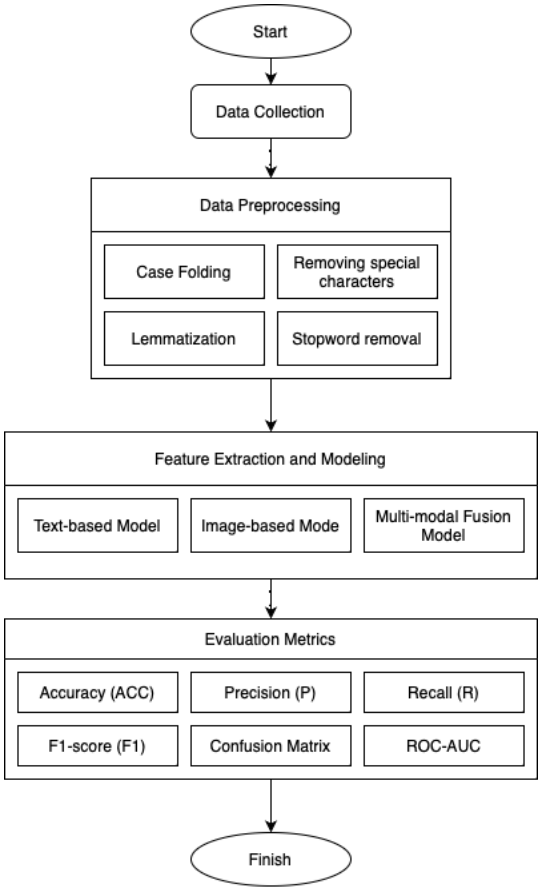

Figure 1 Research Method Flowchart

The dataset used comprises 5,793 entries, each containing an image and its corresponding caption text. Every entry is annotated across several dimensions relevant to cyberbullying detection: Img-Text Label, Image Label, and Text Label for bullying classification; Sentiment (e.g., negative, neutral); Emotion (e.g., disgust, sadness, ridicule); Sarcasm (yes/no); Harmful Score (harmless, partially-harmful, very-harmful); and Target Type (individual, community, organization, or society). This multi-label, multi-modal annotation allows for comprehensive feature learning and a nuanced analysis of online abuse.

Text data underwent a series of preprocessing steps, including conversion to lowercase, removal of special characters, tokenization, stopword elimination, and lemmatization using the spaCy NLP library. For the image data, each file was resized to 224×224 pixels, normalized, and converted into tensor format for compatibility with deep learning models.

Three types of models were developed. The text-based model utilized a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) architecture to extract contextual and semantic features from caption texts. The image-based model employed a pre-trained ResNet-50 convolutional neural network to generate high-level visual representations. Finally, a multi-modal fusion model was constructed by concatenating the text and image feature vectors and passing them through a fully connected classification layer. This model was designed to capture the interdependencies between textual tone and visual content, which are often crucial in cases involving sarcasm, ridicule, or visual irony.

Model training was performed using an 80-10-10 split for training, validation, and testing, respectively. The Adam optimizer and binary cross-entropy loss function were used to minimize classification errors. The performance of the models was evaluated using standard metrics: accuracy, precision, recall, F1-score, and ROC-AUC. These metrics were calculated using the following formulas:

Accuracy:

$$Accuracy \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision:

$$Precision = \frac{(TP)}{(TP + FP)} \quad (2)$$

Recall (sensitivity):

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively. These metrics provide insight into the model's ability to correctly classify harmful content while minimizing false positives, an essential consideration in legal contexts where the consequences of misclassification can be significant [20].

To validate the benefit of a multi-modal approach, the performance of the fusion model was compared against the text-only and image-only baselines. In addition, subgroup evaluations were conducted for samples involving sarcasm, high harmfulness, and individual targeting—dimensions particularly relevant to the legal classification of cyberbullying.

Result

This section presents a comprehensive and multidimensional analysis of the

cyberbullying dataset, aimed at revealing patterns that are both technically measurable and legally relevant. The dataset consists of 5,793 entries, each comprising a combination of image and text many of which are memes or social media posts. These entries are manually annotated with a diverse set of attributes that reflect different dimensions of online abuse, making the dataset highly suitable for both machine learning-based content detection and normative legal evaluation. Each entry is labeled with a bullying classification that indicates whether the content is perceived as bullying or nonbullying. These classifications are applied from three perspectives: based on the image alone, the text alone, and the combined interpretation of both image and text. This multi-angle labeling approach allows for a comparative analysis of how cyberbullying manifests differently depending on the modality. In addition to bullying labels, the dataset includes sentiment polarity annotations that capture the emotional tone of the text, primarily categorized as negative, neutral, or positive. These sentiments serve as important indicators for detecting hostile or aggressive language typically associated with online harassment.

Furthermore, the dataset includes sarcasm labels to flag instances where abusive or critical messages are concealed behind irony or humor, often a hallmark of meme-based bullying. Emotion labels are also provided to identify the dominant emotional content expressed or elicited, such as sadness, anger, or disgust, which can be important for assessing psychological harm. Another key attribute is the harmfulness score, which classifies each entry as harmless, partially harmful, or very harmful. This score is especially useful in determining the severity of a potential offense and whether the content might meet legal thresholds for cyber harassment, defamation, or incitement.

Finally, the target type annotation identifies whether the harmful content is directed toward an individual, a community, an organization, or society at large. This distinction is crucial from a legal standpoint, as different jurisdictions impose varying levels of protection depending on who or what is being targeted.

The results presented in table 1 indicate that bullying behavior is more frequently identified when both image and text components are evaluated together, rather than independently. This suggests that multimodal cues, such as visual sarcasm paired with aggressive language, play a significant role in conveying harmful intent. The combined image-text label yields a higher frequency of bullying instances compared to image-only or text-only classifications, reinforcing the critical need for multi-modal classification frameworks. Such an approach enhances the reliability of cyberbullying detection systems by capturing the full semantic and contextual meaning embedded within digital content, which may be lost when modalities are assessed in isolation.

Table 1 Label Distribution in the Cyberbullying Dataset

Label Type	Bully Count	Nonbully Count
Img-Text Label	3,188	2,605
Image-Only Label	795	4,998
Text-Only Label	1,585	4,208

Figure 2 illustrates the frequency of bullying versus nonbullying content when annotations are made based on the combined interpretation of both image and text. The chart shows that bullying content is more prevalent than nonbullying, highlighting the significance of integrating both modalities for accurate

classification. This finding reinforces the idea that harmful intent or abusive messaging is often not fully captured when analyzing visual or textual elements in isolation. Memes and image-caption formats frequently rely on subtle interactions between visuals and words, and as such, a multi-modal approach proves essential for identifying nuanced forms of cyberbullying that may otherwise go undetected by unimodal classifiers.

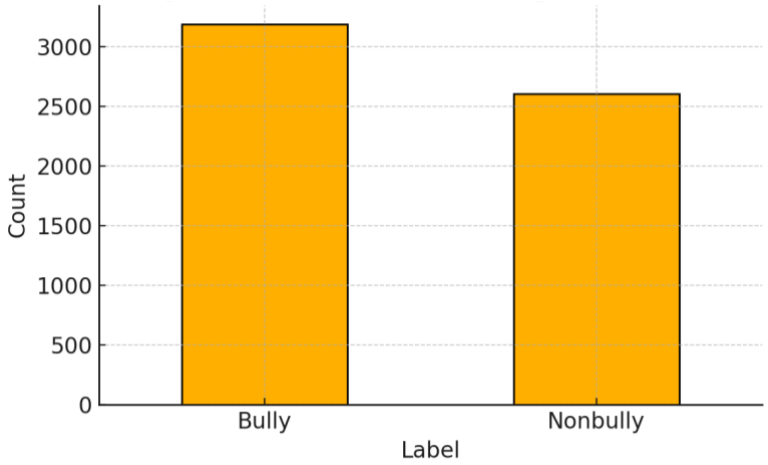

Figure 2 Distribution of Img-Text Labels.

To further explore the emotional dynamics of the dataset, the analysis also examined the distribution of sentiment and sarcasm annotations, as shown in figure 3 and figure 4. The sentiment distribution reveals a dominance of negative and neutral tones across the dataset, suggesting that much of the content carries an apathetic, hostile, or dismissive emotional charge. These tones are commonly found in online environments where mockery, criticism, or passive-aggressive expressions are used as vehicles for harassment. In parallel, sarcasm appears in a significant portion of the entries, emphasizing the challenge of detecting veiled abuse. Sarcasm often disguises hostility beneath humor or irony, making it difficult to identify harmful intent through explicit language alone. This complexity underlines the importance of incorporating pragmatic and contextual cues in automated detection systems, particularly when the goal is to provide evidence for legal analysis.

Table 2 Sentiment and Sarcasm Distribution

Sentiment Type	Count
Negative	2,657
Neutral	2,531
Sarcasm Presence	Count
Yes (Sarcastic)	2,154
No (Not Sarcastic)	3,639

The predominance of negative sentiment across the dataset indicates that aggressive, hostile, or demeaning tones are characteristic features of online harassment. Such sentiment often reflects verbal attacks, mockery, or

expressions of disdain directed toward individuals or groups. Additionally, the presence of sarcasm in over one-third of the entries highlights the use of indirect forms of abuse. Sarcasm is frequently employed as a rhetorical device to veil bullying intentions behind humor or irony, making it more difficult to detect through conventional keyword-based methods. This pattern underscores the need for detection systems that go beyond surface-level linguistic analysis and incorporate pragmatic, tonal, and contextual understanding to effectively identify harmful communication in digital environments.

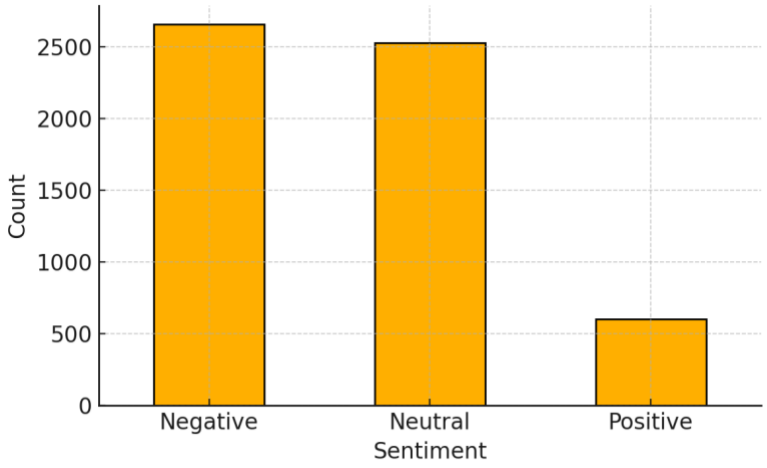

Figure 3 Sentiment Distribution.

The chart demonstrates that the vast majority of content in the dataset exhibits either a negative or neutral emotional tone, while instances of positive sentiment are notably scarce. This distribution aligns with the nature of cyberbullying discourse, which often manifests through criticism, hostility, or apathetic commentary rather than supportive or constructive expression. The lack of positive sentiment further reinforces the observation that online environments where cyberbullying occurs are predominantly characterized by emotionally harmful or disengaged communication patterns.

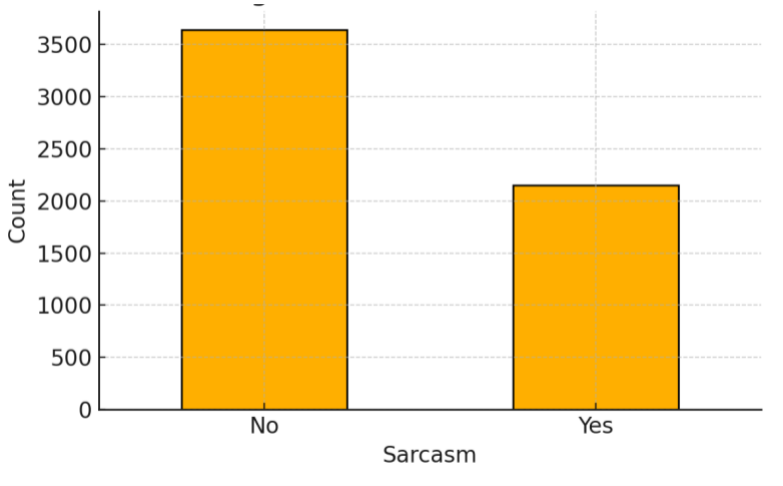

Figure 4 Sarcasm Presence.

A substantial portion of the dataset contains sarcastic elements, highlighting the inherent complexity involved in detecting implicit abuse. Unlike explicit bullying,

which may rely on direct insults or threats, sarcastic content often masks hostility beneath layers of humor, irony, or cultural references. This makes it particularly challenging for automated systems and legal frameworks to distinguish between benign joking and harmful intent. The prevalence of such indirect communication strategies reinforces the importance of developing detection models capable of capturing subtleties in tone, pragmatics, and social context.

In parallel, the Harmful Score annotation serves as a crucial metric for evaluating the perceived severity of each content item. It classifies entries into categories such as harmless, partially harmful, and very harmful, offering a more nuanced understanding of the potential impact that a post may have on its target. This gradation is especially relevant from a legal standpoint, as it may influence whether a piece of content crosses the threshold of criminal liability or falls within protected freedom of expression. Therefore, integrating harmfulness scores into cyberbullying detection frameworks not only improves precision in moderation but also provides a structured basis for legal and ethical assessment.

Table 3 Harmful Score Distribution

Harmful Score	Count
Partially-Harmful	3,037
Harmless	2,723
Very-Harmful	32

Although the precise distribution may vary, a considerable number of entries are categorized as Partially-Harmful, suggesting a high degree of ambiguity in the perceived intent and severity of the content. This reflects a common challenge in both content moderation and legal evaluation, where distinguishing between offensive humor, criticism, and actual harm often requires nuanced contextual interpretation. The existence of this middle category underscores the need for graded intervention strategies, where content is not simply labeled as harmful or not, but is instead assessed along a spectrum of potential risk, both for algorithmic flagging and legal accountability.

Furthermore, analyzing the emotional tone embedded in the content provides essential insight into the psychological dimension of cyberbullying. Emotions such as disgust, sadness, and anger are not only indicative of the poster's intent but also correlate with the emotional impact experienced by victims. Identifying dominant emotional expressions can therefore enhance the interpretive depth of automated detection models, while also offering supporting evidence in legal cases where emotional harm is considered part of the offense. Emotional profiling of content enables a more holistic understanding of digital abuse, beyond mere lexical analysis.

Table 4 provides a detailed breakdown of the emotional categories found within the dataset, illustrating the frequency of specific affective states such as disgust, surprise, ridicule, and sadness. These emotions are critical indicators of the psychological climate in which cyberbullying content operates. To visually complement this tabular summary, figure 5 presents the same data in the form of a bar chart, allowing for a clearer comparative view of emotional dominance across the dataset. This visual representation reinforces the prevalence of negative emotional expressions and supports the interpretation that such emotions are closely associated with online abuse dynamics. Together, the table and figure offer both quantitative and visual insight into the emotional

underpinnings of harmful digital communication.

Table 4 Emotion Distribution

Emotion	Count
Disgust	913
Other	881
Surprise	844
Ridicule	687
Angry	653
Happiness	632
Sadness	580
Anticipation	410
Fear	136
Trust	57

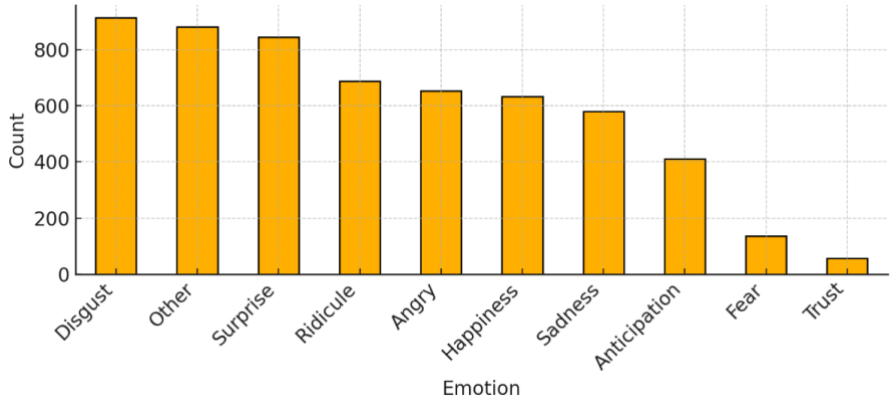

Figure 5 Emotion Distribution.

Among the various emotional categories present in the dataset, sadness and disgust emerge as the most frequently expressed emotions. This prevalence suggests a strong correlation between these affective states and the psychological harm typically associated with cyberbullying. Content that evokes or conveys sadness may indicate personal attack, humiliation, or exclusion, while disgust often reflects expressions of disdain or moral judgment toward the target. The dominance of these emotions reinforces the argument that emotional tone is a critical marker of harmful digital communication and should be incorporated into both detection algorithms and legal assessments of online abuse.

In addition, the target type annotation offers valuable insight into the intended recipient or subject of the harmful content, whether it is directed toward an individual, a specific community, an organization, or society at large. This classification is particularly significant from a legal perspective, as the identity of the target can influence the applicability and severity of certain laws. For instance, attacks on individuals may fall under personal defamation or digital harassment statutes, while content targeting communities or groups may trigger regulations related to hate speech or discrimination. Understanding the nature of the target thus enables a more precise evaluation of the potential legal and

ethical implications of cyberbullying behavior.

Table 5 outlines the distribution of target types within the dataset, indicating whether the harmful content is directed at individuals, communities, organizations, or society at large. This classification is essential for assessing the potential legal implications of each entry, as different forms of targeting may fall under distinct regulatory frameworks. To enhance interpretability, figure 6 visually represents this distribution through a bar chart, highlighting the overwhelming dominance of individual targets compared to collective entities. The visualization reinforces the tabular findings by emphasizing the disproportionate vulnerability of individuals in digital harassment cases and further underscores the importance of individualized legal protections in the context of cyberlaw enforcement.

Table 5 Target Type Distribution

Target Type	Count
Individual	2,405
Community	400
Organization	162

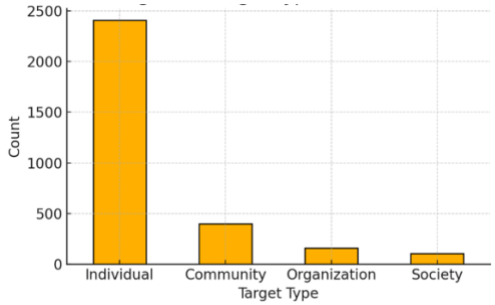

Figure 6 Target Type Distribution.

The analysis reveals that the majority of cyberbullying content is directed toward individuals, rather than groups or institutions. This finding carries significant implications for digital legal frameworks, particularly concerning personal defamation, harassment, and infringement of individual rights. Legal protections often emphasize the safeguarding of individual dignity and mental well-being, and as such, content that targets specific persons may meet the threshold for prosecution under relevant cybercrime and digital communication laws. The concentration of bullying on individuals also highlights the urgent need for victim-centered enforcement mechanisms that address both the psychological and legal dimensions of online harm.

Taken together, these findings confirm that cyberbullying is a complex, multi-faceted phenomenon. Its detection cannot rely solely on the identification of offensive keywords or images but must also consider contextual factors such as tone, emotional resonance, intent, and the identity of the target. The combination of statistical, emotional, and structural insights derived from this dataset provides a solid foundation for the development of intelligent, context-aware detection systems. Such systems are crucial not only for automated content moderation but also for assisting cyberlaw enforcement agencies in making fair, consistent, and legally defensible assessments of potentially harmful online content. The results underscore the potential of data-driven tools

to enhance the scalability and accuracy of digital content regulation, while ensuring that enforcement is grounded in ethical and legal principles.

Discussion

This study provides a comprehensive analysis of cyberbullying content using a multi-dimensional, multi-modal dataset, revealing key insights into how harmful online behavior manifests across visual and textual communication. The results confirm that cyberbullying is not a uniform or easily classifiable phenomenon; rather, it emerges through complex interactions between language, images, tone, emotion, and intent. These findings have significant implications for both the technical design of detection systems and the legal frameworks governing online conduct. One of the most prominent findings is that bullying content is more effectively detected when analyzed through the combined lens of image and text, rather than through image or text alone. The Img-Text label distribution clearly shows that harmful content often relies on contextual cues that are lost when either modality is viewed in isolation. This supports the growing recognition within the field of natural language and image processing that multi-modal machine learning is essential for accurately interpreting content in the age of memes, screenshots, and visually embedded sarcasm. For legal enforcement, this means that a more sophisticated evidentiary standard is needed, one that accounts for the full context in which harmful communication occurs.

The sentiment and sarcasm analysis further underscores the indirect nature of many cyberbullying strategies. While negative sentiment dominates the dataset, a significant portion of entries are marked as neutral, suggesting that emotionally detached or superficially benign language can still carry harmful intent. More strikingly, over one-third of the dataset contains sarcastic content, which presents a substantial challenge for both automated moderation systems and legal adjudication. Sarcasm often functions as a rhetorical shield that allows perpetrators to deny intent, complicating the burden of proof in legal contexts. Therefore, any reliable detection model must go beyond surface-level semantics and incorporate pragmatic understanding, tone modeling, and possibly even user history or platform norms to fully capture the abusive subtext.

The distribution of harmful score annotations provide another critical insight. A large proportion of entries are categorized as “Partially-Harmful,” which mirrors the ambiguity encountered by moderators and legal analysts in real-world settings. These are often borderline cases—where the message might be perceived as a joke by some and as harmful by others—highlighting the subjectivity of harm perception and the difficulty in establishing universal thresholds for intervention. This calls for a graduated response framework in content moderation and legal regulation, where content is not only flagged or removed, but also evaluated for severity, context, and recurrence.

Emotional content plays a central role in both the detection of cyberbullying and the assessment of its impact. The prevalence of disgust, sadness, and anger suggests that much of the communication is emotionally charged and likely to elicit psychological distress. This supports prior research showing that online harassment is not only about intent but also about emotional effect, which should be factored into the legal definition of digital harm. For instance, emotionally manipulative or degrading content—whether or not it contains explicit insults—can still inflict substantial psychological damage, especially when repeated over

time or directed at vulnerable individuals.

The target analysis reveals that individuals are by far the most frequent recipients of online harassment, far exceeding communities or institutions. This finding has important implications for cyberlaw enforcement. In many jurisdictions, legal protections are more clearly defined for individual persons than for abstract entities. The prevalence of individual targeting suggests a need for stronger personal data protection, anti-defamation laws, and psychological harm provisions in digital policy frameworks. Moreover, automated systems that flag harmful content should be designed with target sensitivity in mind, recognizing that attacks on individuals may warrant more urgent and severe responses than generalized speech.

In summary, the discussion of these results emphasizes the multi-layered and context-dependent nature of cyberbullying. A content moderation or legal strategy that relies solely on word lists, static rules, or binary classifiers is unlikely to be effective. Instead, there is a pressing need for context-aware, emotionally intelligent, and legally informed detection systems that can account for the subtleties of modern online communication. Future work should focus on developing adaptive algorithms that learn from user behavior, social context, and cultural norms, while also aligning with ethical standards and legal definitions of harm. As cyberbullying continues to evolve, so too must the tools and frameworks we use to combat it.

Conclusion

This study set out to examine the complexity of cyberbullying by leveraging a multi-modal dataset composed of image and text-based content commonly found in social media environments. Through a detailed analysis of annotated attributes, including bullying classification, sentiment polarity, sarcasm, emotional tone, harmfulness, and target type, the research provides a nuanced understanding of how harmful online behavior manifests and how it can be detected effectively. The results demonstrate that cyberbullying is rarely explicit or isolated to a single communication modality. Rather, it often operates through a combination of subtle visual and textual cues, contextual irony, and emotionally charged language. The higher incidence of bullying detection in combined image-text analysis reaffirms the importance of multi-modal classification approaches. Moreover, the widespread use of sarcasm and the prevalence of emotionally negative content such as disgust and sadness, further emphasize the need for systems that can interpret not just what is said, but how it is said and in what context.

The identification of “Partially-Harmful” content illustrates the blurred boundary between legally permissible expression and actionable abuse, posing a challenge for content moderation and digital governance. In addition, the fact that individuals are the primary targets of cyberbullying reinforces the urgency of strengthening personal protection mechanisms within cyberlaw frameworks. Overall, this research contributes to the growing body of literature that calls for context-aware, ethically aligned, and legally grounded detection systems capable of operating at scale in dynamic digital ecosystems. The findings underscore the critical role of interdisciplinary collaboration combining data science, law, psychology, and ethics in designing tools that not only detect but also respond appropriately to online harm.

Future research should explore the integration of temporal patterns (e.g.,

frequency and escalation of bullying over time), user behavior analysis, and cultural sensitivity in model development. Such efforts are essential to ensure that cyberbullying detection systems remain robust, fair, and adaptable to the evolving nature of digital abuse.

Declarations

Author Contributions

Conceptualization: S. and A.S.; Methodology: A.S.; Software: S.; Validation: S. and A.S.; Formal Analysis: S. and A.S.; Investigation: S.; Resources: A.S.; Data Curation: A.S.; Writing Original Draft Preparation: S. and A.S.; Writing Review and Editing: A.S. and S.; Visualization: S.; 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] T. Petrova, “Digital Communication, and Social Transformation,” **Postmodernism Problems**, vol. 13, no. 3, pp. 283–284, Dec. 2023, doi: 10.46324/PMP2303283.

[2] S. Mukhopadhya and S. M. Tripathi, “Review on Digital Communication on Social Networks and Its Impacts,” **International Journal of Research Publication and Reviews**, vol. 7, no. 5, pp. 1506–1513, 2020.

[3] L. Saburova, “Depersonalization of Liaison in Digital Communication: ‘Lightened Sociality’ Phenomenon,” in **The Public/Private in Modern Civilization: Proceedings of the 22nd Russian Scientific-Practical Conference (Yekaterinburg, Apr. 16–17, 2020)**, 2020, pp. 194–198, doi: 10.35853/ufh-public/private-2020-03.

[4] W. Zhirong, “Digital Communication Is Changing the World,” **Journal of Shanghai Teachers University (Philosophy and Social Sciences Edition)**, vol. 34, no. 2, pp. 42–45, 2005.

[5] J. W. Patchin and S. Hinduja, “Cyberbullying and self-esteem,” *Journal of School Health*, vol. 80, no. 12, pp. 614–621, Dec. 2010, doi: 10.1111/j.1746-1561.2010.00548.x.

[6] R. Slonje, P. K. Smith, and A. Frisén, “The nature of cyberbullying, and strategies for prevention,” *Computers in Human Behavior*, vol. 29, no. 1, pp. 26–32, Jan. 2013, doi: 10.1016/j.chb.2012.05.024.

[7] Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on Twitter using a convolution-GRU based deep neural network,” in *Proc. Eur. Semantic Web Conf.*, vol. 10843, no. June, pp. 745–760, 2018, doi: 10.1007/978-3-319-93417-4_48.

[8] Y. Liu, X. Huang, X. An, and H. Li, “BERT-based cyberbullying detection on social media,” in *Proc. Int. Workshop on NLP for Social Media, Florence*, *Italy, Aug. 2019*, pp. 1–9. [Online]. Available: https://aclanthology.org/W19-6101/

[9] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, “Analyzing labeled cyberbullying incidents on the Instagram social network,” in *Proc. IEEE Int. Conf. on Social Computing*, 2015, pp. 935–940, doi: 10.1109/SocialCom.2015.138.

[10] 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)*, 2016, pp. 3952–3958.

[11] E. Cambria, S. Poria, D. Hazarika, and A. Hussain, “SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings,” in *Proc. AAAI Conf. on Artificial Intelligence*, vol. 32, no. 1, pp. 1795–1802, Apr. 2018.

[12] A. Joshi, P. Bhattacharyya, and M. J. Carman, “Automatic sarcasm detection: A survey,” *ACM Computing Surveys*, vol. 50, no. 5, pp. 1–22, Sept. 2017, doi: 10.1145/3124420.

[13] D. K. Citron, Hate Crimes in Cyberspace. Cambridge, MA, USA: Harvard Univ. Press, 2014.

[14] T. Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Haven, CT, USA: Yale Univ. Press, 2018.

[15] C. Khandekar, M. Joshi, and M. V. Joshi, “Multi-modal hate speech detection using cross-modal attention,” in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics (ACL)*, 2021, pp. 431–439, doi: 10.18653/v1/2021.acl-long.34.

[16] S. Vempala and D. Preotiuc-Pietro, “Categorizing and inferring the relationship between the text and image of Twitter posts,” in *Proc. 57th Annu. Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 2830–2840, doi: 10.18653/v1/P19-1272.

[17] J. Yin, H. Zubiaga, M. Liakata, and R. Procter, “A multi-dimensional analysis of cyberbullying on Twitter,” *Online Social Networks and Media*, vol. 13–14, pp. 100059, Aug. 2020, doi: 10.1016/j.osnem.2020.100059.

[18] B. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, “The ethics of algorithms: Mapping the debate,” *Big Data & Society*, vol. 3, no. 2, pp. 1–21, Dec. 2016, doi: 10.1177/2053951716679679.

[19] R. Binns, “Algorithmic accountability and public reason,” Philosophy & Technology, vol. 31, no. 4, pp. 543–556, Dec. 2018, doi: 10.1007/s13347-017-0263-5.

[20] A. Safari, M. Sabahi, and A. Oshnoei, “Resfaultyman: An intelligent fault detection predictive model in power electronics systems using unsupervised learning isolation forest,” *Heliyon*, vol. 10, no. 15, pp. 1–13, Aug. 2024. doi:10.1016/j.heliyon.2024.e35243