BERT-Based Emotion and Sarcasm-Aware Classification of Harmful Online Content for Cyber Law Enforcement

Suraphan Chantanasut[1,*]

[1]Department of Art Education, Faculty of Fine and Applied Arts, Rajamangala University of Technology Thanyaburi Thailand

ABSTRACT

The rise of cyberbullying on social media has created urgent demands for automated systems that can detect harmful digital content in both explicit and subtle forms. This study presents an exploratory analysis of a multimodal dataset consisting of 5,793 social media posts annotated with four key dimensions: harmfulness level (Harmless, Partially-Harmful, Harmful), emotion, sentiment, and sarcasm. Our goal was to understand the distributional, emotional, and linguistic features of harmful content and how they interact in context, with implications for cyber law enforcement and content moderation. The analysis reveals a significant class imbalance in harmfulness distribution, with 63.18% of posts labeled as Harmless, 33.51% as Partially-Harmful, and only 3.28% as Harmful. Emotionally negative expressions were dominant, particularly Disgust (1,565 posts) and Sadness (1,321 posts). Sarcasm was present in 1,023 posts, accounting for 17.66% of the dataset, indicating that indirect or veiled forms of hostility are widespread. Representative samples further demonstrate the ambiguity of harmful intent in posts that blend humor, sarcasm, and emotional undertones. These findings emphasize the need for context-aware classification models that incorporate affective signals and pragmatic cues to accurately identify both overt and covert cyberbullying. Such models are essential not only for enhancing the performance of content moderation systems but also for strengthening digital evidence collection and enforcement under cybercrime regulations. The study concludes with recommendations for integrating emotion and sarcasm detection into transformer-based architectures to improve interpretability and legal relevance in automated harmful content assessment.

Keywords Cyberbullying Detection, Harmful Content, Sarcasm, Emotion Analysis, Cyber Law Enforcement

Submitted 26 September 2025
Accepted 12 November 2025
Published 1 December 2025

*Corresponding author
Suraphan Chantanasut,
suraphan_c@rmutt.ac.th

Additional Information and Declarations can be found on page 311

DOI: 10.63913/jcl.v1i4.42

Introduction

In the digital age, social media has become a dominant channel for interpersonal communication, public discourse, and community formation [1]. Platforms such as Twitter, Instagram, and Facebook have enabled users to share content rapidly, often in emotionally charged or socially reactive contexts [2]. While these platforms offer significant opportunities for expression and connection, they also serve as fertile ground for toxic interactions—including cyberbullying, hate speech, and psychological harassment—that are difficult to moderate at scale. Cyberbullying is broadly defined as the use of digital communication tools to repeatedly intimidate, insult, or demean individuals or groups [3]. Unlike traditional bullying, which is often constrained by physical proximity, cyberbullying can occur asynchronously, anonymously, and at a global scale. Victims may be subjected to repeated exposure to harmful content, even when offline, due to the permanent and shareable nature of digital posts. These unique characteristics make cyberbullying not only a psychological concern but also a

How to cite this article: S. Chantanasut, "BERT-Based Emotion and Sarcasm-Aware Classification of Harmful Online Content for Cyber Law Enforcement," *J. Cyber. Law.*, vol. 1, no. 4, pp. 300-313, 2025.

significant legal and regulatory challenge.

In response, various legal frameworks have emerged to combat harmful online behavior. In Indonesia, the Undang-Undang Informasi dan Transaksi Elektronik (UU ITE) criminalizes certain forms of online defamation, hate speech, and digital threats. Similar initiatives are present in other jurisdictions, such as the EU Digital Services Act and the UK's Online Safety Bill. However, enforcement remains difficult due to the contextual and linguistic complexity of online communication. Much of the harm in digital environments is delivered not through explicit slurs, but through sarcasm, emotional manipulation, passive aggression, or coded language—forms that are difficult to detect with traditional rule-based or keyword-driven systems.

Advancements in Natural Language Processing (NLP) and machine learning have led to promising developments in automatic content classification. Transformer-based architectures, such as BERT (Bidirectional Encoder Representations from Transformers), have demonstrated exceptional performance in understanding contextual language across various domains. However, most existing cyberbullying detection models prioritize binary classification (e.g., abusive vs. non-abusive) and often ignore multimodal and multilabel attributes such as emotion, sentiment, sarcasm, and degrees of harmfulness, which are crucial in understanding the subtlety and intent behind a digital post.

This study aims to fill this gap by conducting a comprehensive exploratory analysis of a multimodal dataset comprising 5,793 social media posts, annotated with harmfulness levels (Harmless, Partially Harmful, Harmful), emotional tone, presence of sarcasm, and sentiment polarity. The primary goal is to uncover patterns that contribute to veiled or indirect aggression, which may not be immediately evident but still constitute serious risks under cyber law. By identifying how emotion, sarcasm, and harmfulness interact, this research provides foundational insights for designing context-aware, legally applicable, and ethically sound machine learning models capable of enhancing both platform moderation systems and digital forensic investigations.

Literature Review

Research on cyberbullying detection has grown significantly over the past decade, driven by the urgent need to moderate harmful digital behavior on social media platforms. Early studies in this field primarily employed lexicon-based methods, relying on manually curated lists of profane or offensive terms to flag inappropriate content [4]. Although simple to implement, such approaches often failed to capture subtle or context-dependent expressions of harm, particularly those involving sarcasm, metaphor, or emotionally manipulative language. With the advancement of Machine Learning (ML) and NLP techniques, researchers began adopting more robust classification models, including Naïve Bayes, Support Vector Machines (SVM), and Random Forests [5], [6]. These models typically use Term Frequency-Inverse Document Frequency (TF-IDF) or bag-of-words representations as features, enabling more reliable pattern recognition compared to rule-based systems. However, they still lacked deep semantic understanding. The introduction of word embeddings—such as Word2Vec [7], GloVe [8], and FastText [9]—marked a turning point in the field. These embeddings provided dense vector representations that preserved semantic relationships between words, enhancing model sensitivity to context and intent.

More recently, transformer-based models, particularly BERT [10], have achieved state-of-the-art performance in various content moderation tasks, including hate speech detection [11], online harassment classification [12], and offensive language recognition [13]. Models such as RoBERTa [14] and ALBERT [15] further optimized pretraining strategies and architecture to improve generalization on imbalanced and noisy datasets. Mozafari et al. [16], for instance, used BERT to detect hate speech on Twitter, reporting significant improvements in precision and recall over conventional deep learning models. Beyond toxic content classification, recent literature emphasizes the importance of emotion detection in understanding the severity and intent of cyberbullying. Emotions such as disgust, anger, and sadness are frequently associated with posts that contain psychological harm or veiled aggression [17]. Ibrahim et al. [18] demonstrated the benefit of integrating emotional features into Arabic cyberbullying classifiers, while Zhang et al. [19] found that emotion-aware models outperformed neutral sentiment baselines in English-language toxicity detection. Likewise, Mishra and Bhattacharyya [20] showed that emotion-enriched architectures provided higher interpretability and robustness in cases of implicit hate. Sarcasm detection has also emerged as a critical subfield in cyberbullying analysis. Sarcasm often acts as a vehicle for disguised hostility, making it difficult for conventional classifiers to distinguish between humorous and harmful content. Ghosh and Veale [21] developed deep learning models for sarcasm recognition using convolutional and recurrent neural architectures, while Rajadesingan et al. [22] introduced behavioral modeling approaches that captured patterns in user history to improve sarcasm detection accuracy. These studies argue that ignoring sarcasm leads to significant underreporting of abusive behavior, especially on platforms like Twitter and Instagram.

Recent efforts have also explored multi-task learning to jointly model emotion, sarcasm, and offensive language within unified frameworks. Kumar et al. [23] proposed a BiLSTM architecture with attention mechanisms that performed simultaneous sentiment, emotion, and sarcasm classification. Zhang et al. [24] extended this idea by implementing a multi-task BERT model that could detect both sarcasm and offensive speech, achieving higher generalizability on previously unseen test cases. In the legal and regulatory domain, several researchers have investigated how NLP models can aid cyber law enforcement. Tsakalidis et al. [25] proposed a framework for multilingual hate speech detection aligned with European Union policy requirements, while Handayani et al. [26] emphasized the role of linguistic cues in supporting digital forensic investigations under Indonesia’s UU ITE (Law on Electronic Information and Transactions). Despite these contributions, few studies to date have integrated harmfulness classification, emotional tone, and sarcasm detection into a single unified model, particularly one designed to support legal accountability.

By addressing this gap, the present study contributes to the emerging discourse on context-aware, multimodal approaches to cyberbullying detection. Through a combination of exploratory data analysis and transformer-based modeling, this work offers both methodological innovation and legal relevance in the broader challenge of regulating harmful online communication.

Methods

This research applies a structured methodology that integrates exploratory data analysis with transformer-based classification to identify varying degrees of harmfulness in social media content. The study uses a multimodal dataset of

5,793 annotated social media posts, each labeled with Harmful-Score (the main target variable), as well as auxiliary attributes such as Img-Text, Emotion, Sentiment, and Sarcasm. The classification task is designed to assign each post to one of three harmfulness levels: Harmless, Partially-Harmful, or Harmful, making this a multi-class classification problem aligned with both psychological and legal interpretations of digital aggression.

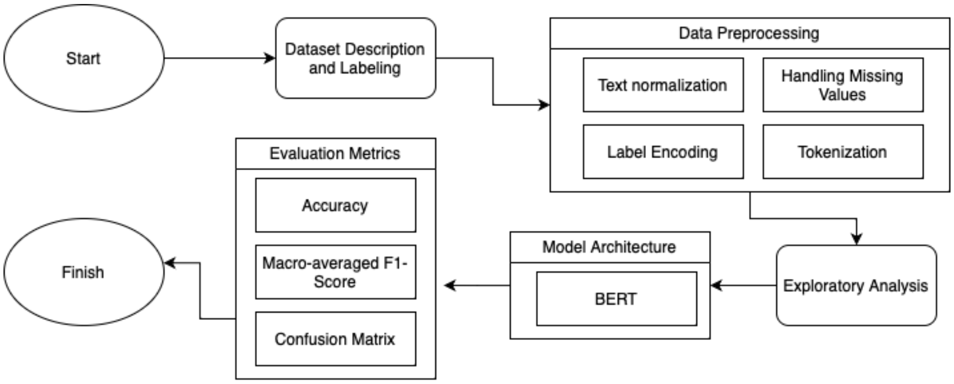

Figure 1 Research Method Flowchart

Before modeling, a series of preprocessing steps was conducted. The Img-Text content was cleaned by removing URLs, emojis, HTML tags, and non-standard characters, followed by conversion to lowercase for consistency. Posts with missing values in the Harmful-Score column were excluded. All categorical variables were label-encoded, while the core text was tokenized using BERT's WordPiece tokenizer with a maximum sequence length of 128 tokens. To understand the imbalance in the dataset, the relative frequency of each harmfulness category was computed using the formula [27]:

$$P_i = \frac{n_i}{N} \times 100\% \tag{1}$$

P_i is the percentage of posts in class i, n_i is the number of posts labeled with class i, and N is the total number of posts. This analysis revealed a strong imbalance: 63.18% of the posts were Harmless, 33.51% were Partially-Harmful, and only 3.28% were classified as Harmful. This distribution justifies the use of class weighting during model training to avoid performance bias toward the majority class.

To classify the posts, this study implements BERT-base-uncased, a bidirectional transformer model pretrained on English corpora. BERT is particularly well-suited for cyberbullying detection due to its ability to capture subtle contextual cues, such as sarcasm and emotional tone. In the proposed architecture, the [CLS] token embedding from BERT is concatenated with dense vector representations of Emotion and Sarcasm, forming a feature-enriched representation. This combined vector is passed through a fully connected neural layer, followed by a softmax function that outputs probability scores for each harmfulness category. Model training is guided by the cross-entropy loss function, defined as:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y_i}) \tag{2}$$

$C = 3$ (the number of classes), y_i is the one-hot encoded ground truth for class

$i,$ and $\hat{y}_i$ s the predicted probability for that class. The AdamW optimizer is used with learning rate scheduling and early stopping based on validation loss to prevent overfitting and ensure generalization.

Model performance is evaluated using overall accuracy, macro-averaged F1-score, and a confusion matrix. These metrics are chosen to ensure that the model not only performs well on the dominant Harmless class but also retains sensitivity to the Partially-Harmful and Harmful categories, which are of higher relevance in the context of cyber law enforcement and digital content regulation. This methodological design thus ensures a balance between technical robustness and real-world applicability in detecting and classifying abusive online behavior.

Result

This section presents the key findings derived from an in-depth Exploratory Data Analysis (EDA) of a multimodal social media dataset focused on cyberbullying detection. The analysis aims to uncover underlying patterns and characteristics of harmful online behavior that may inform legal frameworks and enhance automated moderation systems. The results are systematically organized into four subsections: (1) a comprehensive overview of dataset attributes, (2) statistical distribution of harmfulness labels, (3) analysis of emotional and sarcastic expressions in the text, and (4) a qualitative review of representative user-generated posts. Each subsection integrates tabular and graphical representations to provide a balanced combination of quantitative metrics and qualitative examples. These findings serve as a foundation for subsequent model development, including NLP and deep learning techniques, while also offering insights relevant to cyber law enforcement and content moderation policy.

The dataset under investigation consists of 5,793 user-generated posts sourced from social media platforms, each combining visual (image) and linguistic (text) content. What distinguishes this dataset is the inclusion of rich, multilayered annotations covering various dimensions of online behavior, such as sentiment polarity, emotional tone, sarcasm presence, and a three-tiered harmfulness score (Harmless, Partially-Harmful, Harmful). These annotations provide a unique opportunity to analyze not only explicit forms of abuse but also subtler, context-dependent expressions that may still constitute a violation of digital ethics or cyber law statutes.

To establish a clear foundation for the analysis, table 1 presents a detailed overview of the dataset's attributes. Each attribute captures a distinct behavioral or linguistic aspect of user interaction, ranging from meme-like captions and emotional expressions to sarcasm and implicit targeting. These components are critical for feature engineering and label structuring, particularly in the context of training machine learning models to detect potentially unlawful or harmful content in real time.

Table 1 Dataset Attribute Descriptions

Attribute	Description
Img-Name	File name of the social media image.
Img-Text	Text content accompanying the image (e.g., meme caption or post text).
Img-Text-Label	Combined label for the image-text pair (Bully or

	Nonbully).
Img-Label	Label assigned to the image component alone.
Text-Label	Label assigned to the text component alone.
Sentiment	Sentiment polarity of the text: Positive, Negative, or Neutral.
Emotion	Emotion expressed in the text (e.g., Sadness, Anger, Disgust, Surprise).
Sarcasm	Binary indicator of whether the text contains sarcasm (Yes/No).
Harmful-Score	Degree of harmfulness: Harmless, Partially-Harmful, or Harmful.
Target	The subject of the content (Individual or Group), if identified.

These attributes form the conceptual and analytical foundation for building classification models aimed at identifying potentially abusive, unethical, or even illegal digital interactions. By structuring the data into interpretable and well-labeled components such as sentiment, emotional cues, and levels of harmfulness, the dataset becomes suitable for developing machine learning pipelines that can detect nuanced forms of cyberbullying and online harassment. The presence of labels like sarcasm and emotion further enhances the capacity to model not only overt toxicity but also implicit and context-dependent harmful behaviors that often evade traditional rule-based systems.

A closer examination of the harmfulness distribution reveals a pronounced class imbalance, as summarized in table 2. A significant majority of the posts (3,661 instances, or 63.18%) are labeled as Harmless, suggesting that most content in the dataset does not pose a direct threat or ethical concern. However, a substantial portion of the data 1,943 posts (33.51%), is categorized as Partially-Harmful, indicating content that may include mild ridicule, stereotyping, or indirect aggression. Meanwhile, a smaller but critical subset (190 posts, or 3.28%) is explicitly labeled as Harmful, potentially containing direct threats, hate speech, or targeted personal attacks. This distribution not only informs the expected class priors for classification models but also highlights the importance of applying mitigation strategies such as class weighting, oversampling, or synthetic data augmentation to avoid predictive bias toward the majority class during model training.

Table 2 Harmful-Score Label Distribution

Harmful Score	Count	Percentage (%)
Harmless	3,661	63.18
Partially-Harmful	1,943	33.51
Harmful	190	3.28

To complement the numerical overview, figure 2 provides a visual representation of the frequency distribution across the three harmfulness classes. This bar chart clearly illustrates the dominance of Harmless content in the dataset, with progressively fewer instances categorized as Partially-Harmful and Harmful. Such visualization reinforces the earlier observation of class imbalance, offering an intuitive grasp of the distribution patterns that may not be as readily apparent from tabular data alone. The sharp contrast in category sizes underscores the potential challenges in training supervised classification models, particularly in achieving balanced performance across all classes. Without appropriate corrective measures, such as data resampling, stratified batching, or cost-sensitive learning the model may exhibit high accuracy on the majority class while underperforming in detecting the minority but legally significant Harmful posts. Consequently, the insights derived from Figure 2 are instrumental not only for understanding dataset composition but also for informing fair and effective model design in the context of cyber law enforcement applications.

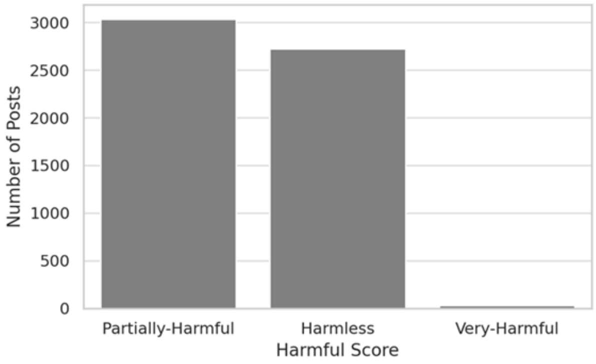

Figure 2 Distribution of Harmful-Score Categories

Figure 2 presents a bar chart that illustrates the frequency of each harmfulness label Harmless, Partially-Harmful, and Harmful, across the dataset. This visual effectively highlights the class imbalance, reaffirming the statistical findings discussed earlier and serving as a diagnostic tool for potential model training challenges. The overwhelming presence of Harmless content, in contrast with the sparse yet critical Harmful category, visually reinforces the need for algorithmic safeguards to prevent skewed classification results.

Beyond harmfulness, the dataset is enriched with emotion and sarcasm annotations, both of which are key for detecting implicit aggression and nuanced forms of harm that may not manifest through overt language. These attributes provide additional layers of semantic depth, which are particularly relevant in the legal context where intent and tone significantly influence content interpretation. As summarized in table 4, Disgust (1,565 posts) and Sadness (1,321 posts) are the most prevalent emotions in the dataset, indicating a dominant presence of emotionally negative or critical language. These emotional states are often associated with expressions of judgment, hostility, or alienation, which align with the behavioral patterns commonly found in cyberbullying and online harassment cases.

Table 4 Emotion Label Distribution

Emotion	Count
Disgust	1,565
Sadness	1,321
Surprise	778
Anger	727
Other	659
Joy	392

Fear	350

These findings are further reinforced by figure 3, which provides a visual depiction of the distribution of emotional labels across the dataset. The bar chart clearly illustrates the prominence of emotionally negative categories such as Disgust, Sadness, and Anger, while positive emotions like Joy appear far less frequently. This asymmetry emphasizes the emotionally charged nature of user-generated posts associated with cyberbullying. The visualization complements the tabular data by enabling immediate recognition of dominant emotional trends, which might otherwise be obscured in numerical summaries. Collectively, table 4 and figure 3 confirm that emotionally intense and negatively valenced language is a prevalent characteristic in posts labeled as harmful or borderline abusive. These emotional signals are not merely stylistic elements; rather, they serve as key indicators of underlying hostility, exclusion, or psychological manipulation—behaviors often subject to scrutiny under cybercrime and digital harassment laws.

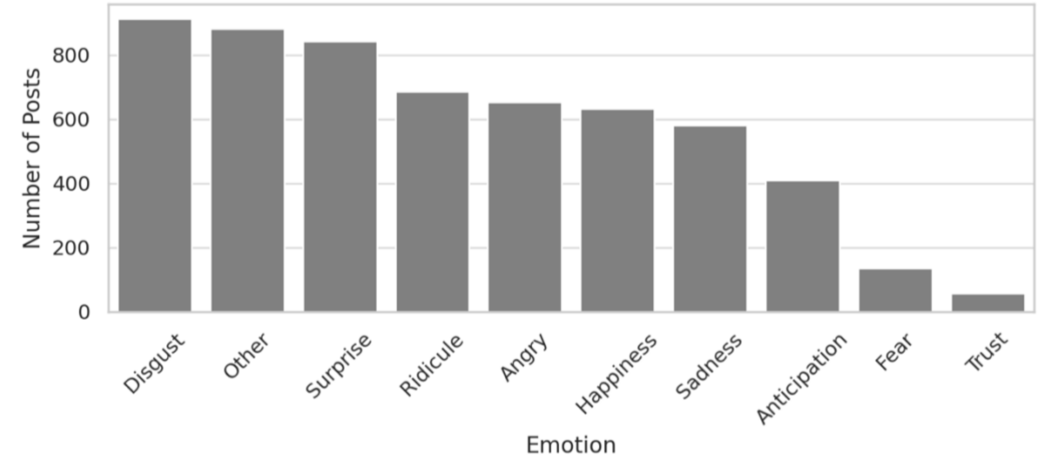

Figure 3 Emotion Distribution in Social Media Texts

Figure 3 presents a bar chart illustrating the frequency distribution of emotional labels assigned to the dataset. Each bar corresponds to one of the annotated emotions—such as Disgust, Sadness, Anger, Joy, and Fear—allowing for a quick visual assessment of their relative occurrence. The prevalence of negative emotional categories supports the earlier observation that cyberbullying-related content is often emotionally charged and psychologically expressive. This emotional intensity not only adds complexity to content moderation but also provides meaningful features for machine learning models seeking to infer intent or impact.

In addition to emotional tone, sarcasm is another critical dimension in identifying implicit online hostility. Unlike direct insults or threats, sarcastic expressions frequently rely on tone, irony, and contextual cues to deliver veiled aggression, making them harder to detect using surface-level lexical analysis. As shown in table 5, a total of 1,023 posts (17.66%) is labeled as sarcastic, indicating that nearly one-fifth of the content may convey hostility through indirect or ironic means. This substantial proportion highlights the necessity of incorporating sarcasm-aware mechanisms into automated classification pipelines, particularly in legal or ethical evaluations where intent must be inferred with high reliability.

Table 5 Sarcasm Label Distribution

Sarcasm	Count
No	4,764
Yes	1,023

To qualitatively illustrate how harmfulness, emotion, and sarcasm co-occur and interact in real-world social media discourse, table 6 presents five annotated examples drawn directly from the dataset. Each entry includes the original text, its associated Harmful-Score, the dominant Emotion, and a binary Sarcasm label. These samples are intentionally selected to capture a diverse range of linguistic expressions from seemingly neutral or humorous remarks to those laced with implicit judgment or veiled aggression.

The examples highlight the inherent ambiguity and contextual complexity involved in detecting harmful content, particularly when users employ sarcasm, euphemism, or culturally coded language. For instance, a post may superficially appear comical or benign yet convey underlying hostility or mockery that is only apparent through emotional or sarcastic cues. This underscores the limitations of relying solely on keyword-based detection methods and supports the need for more nuanced, context-aware classification models. In cyber law contexts, such linguistic subtlety poses additional challenges for legal interpretation, as the threshold between offensive humor and actionable abuse is often blurred.

Table 6 Sample Texts with Labels

Text	Harmful Score	Emotion	Sarcasm
This is bit weird	Partially-Harmful	Disgust	No
When black people make white jokes when white people can't do the same	Harmless	Other	No
Nibba be like: no female can control me! 4 years later, his whole life is controlled by one woman lol	Partially-Harmful	Disgust	Yes
Jaa Simran jaa jee le apni zindagi. Kaise jau Babuji, bus ka waits kar raha hoon	Harmless	Sadness	No
Algebra and geometry are fine, but graphing is where I draw the line	Harmless	Surprise	No

These annotated samples provide compelling evidence that harmfulness in online content frequently manifests in subtle, context-dependent forms rather than through overtly abusive language. Posts may employ sarcasm, emotionally charged expressions, or culturally embedded insinuations that elude traditional detection mechanisms. This reinforces the necessity for sophisticated classification models that are not only capable of identifying explicit instances of cyberbullying such as slurs or threats, but are also sensitive to implicit indicators of psychological harm, such as irony, emotional manipulation, or veiled ridicule. From a cyber law perspective, the ability to detect such nuanced content is critical, as it may constitute defamation, harassment, or digital abuse even in the absence of conventional profanity or hate speech.

Discussion

The findings of this study underscore the complex and multifaceted nature of cyberbullying in digital environments, particularly on social media platforms where language is fluid, context-dependent, and often emotionally charged. Through an in-depth analysis of a multimodal dataset annotated with harmfulness, emotion, sentiment, and sarcasm labels, this research reveals critical insights that hold both technical and legal implications for cyberbullying detection. One of the most salient observations is the high degree of linguistic ambiguity in harmful content. As demonstrated in the representative samples, online abuse is frequently expressed through subtle or coded language rather than overt insults. This includes sarcastic commentary, emotionally negative phrasing, and indirect ridicule, which together create a form of "hidden aggression" that is easily overlooked by rule-based or keyword-dependent systems. These findings suggest that relying on explicit indicators alone is insufficient for comprehensive content moderation or legal assessment.

The distributional imbalance of the harmfulness classes also presents a significant challenge for automated detection systems. With over 60% of the dataset labeled as Harmless, and only 3% labeled as Harmful, traditional supervised learning models may tend to favor the majority class, potentially failing to detect the most critical and legally actionable content. This highlights the necessity of employing data balancing techniques, such as stratified sampling, class weighting, or synthetic data generation, to ensure fair model performance across all classes.

Moreover, the prevalence of emotionally negative expressions, especially Disgust and Sadness, reinforces the hypothesis that harmful content is often accompanied by strong affective cues. The emotional tone of a post, when coupled with sarcastic intent, can significantly amplify its perceived hostility. This interaction between emotion and sarcasm aligns with prior literature in computational linguistics and psychology, which emphasizes the role of tone and affect in social conflict and interpersonal harm. From a legal standpoint, these subtle linguistic patterns complicate the process of determining malicious intent, which is often a core element in cybercrime legislation such as Indonesia’s Undang-Undang Informasi dan Transaksi Elektronik (UU ITE) or broader international cyberbullying statutes. Taken together, the results support the need for context-aware, multimodal classification models capable of understanding not just what is said, but how and in what context it is conveyed. This includes incorporating pretrained transformer-based architectures like BERT or multimodal models such as CLIP that can evaluate both textual and visual cues. Furthermore, integrating emotion and sarcasm detection as auxiliary tasks or features may enhance a model’s interpretability and accuracy in identifying borderline or covert harmful content.

Finally, these insights contribute to the growing discourse on algorithmic support for cyber law enforcement. By leveraging machine learning to assist in the early detection of potentially illegal or harmful content, authorities and platforms can better prioritize review cases, protect vulnerable users, and uphold digital rights. However, care must be taken to balance detection accuracy with the ethical challenges of censorship, privacy, and false positives, underscoring the need for human-in-the-loop frameworks and clear legal standards.

Conclusion

This study provides a comprehensive exploration of the linguistic, emotional, and contextual dynamics of cyberbullying as manifested in social media posts. By leveraging a multimodal dataset of 5,793 annotated posts—each enriched with sentiment, emotion, sarcasm, and harmfulness labels—we examined the underlying patterns that characterize harmful digital interactions. The goal was not only to inform the design of machine learning models for content moderation but also to generate actionable insights for cyber law enforcement and regulatory intervention. The results indicate that harmful content often deviates from traditional expectations of explicit abuse. Instead, many posts exhibit indirect aggression, manifested through emotionally negative language (e.g., Disgust, Sadness), sarcastic phrasing, and culturally nuanced insinuations. These subtleties highlight a significant challenge: harmfulness in the digital sphere is frequently implicit, coded, and context-dependent, requiring deeper semantic understanding that surpasses simple keyword-based or rule-driven detection mechanisms.

The distributional analysis further reveals a marked class imbalance, with over 63% of content labeled as Harmless, approximately 33% as Partially-Harmful, and only 3% as Harmful. While this reflects the natural skew of user-generated content, it poses difficulties for supervised classification tasks, which risk overfitting to the dominant class and failing to capture critical minority cases. These findings reinforce the need for balanced training strategies, such as oversampling, class weighting, and cost-sensitive learning, to ensure equitable model performance across all harmfulness levels.

Moreover, the high prevalence of sarcasm (17.66%) underscores its role as a linguistic device that often conceals or distorts the intended harmfulness of a message. Sarcasm detection, therefore, is not a secondary task but a central requirement in any model designed for cyberbullying detection, particularly in legal contexts where intent and tone play a vital role in determining culpability under cybercrime statutes. From a legal and regulatory perspective, the implications are clear: current automated moderation systems must evolve to incorporate context-aware, emotion-sensitive, and sarcasm-informed classification frameworks. Transformer-based architectures like BERT, along with multimodal approaches that integrate textual and visual cues, offer promising pathways for achieving this goal. However, the integration of such models into law enforcement workflows must also address concerns about fairness, explainability, and due process, particularly in jurisdictions where digital evidence may lead to criminal prosecution.

In summary, this research lays the groundwork for developing robust, ethically aligned, and legally relevant models for detecting cyberbullying and harmful digital speech. Future work should focus on implementing and evaluating deep learning classifiers on this dataset, with attention to real-world deployment scenarios such as flagging high-risk posts, generating evidence reports, or assisting legal investigations. Additionally, collaborative efforts between technologists, policymakers, and legal scholars will be essential to ensure that the tools developed are both technically effective and socially responsible.

Declarations

Author Contributions

Conceptualization: S.C.; Methodology: S.C.; Software: S.C.; Validation: S.C.; Formal Analysis: S.C.; Investigation: S.C.; Resources: S.C.; Data Curation:

S.C.; Writing Original Draft Preparation: S.C.; Writing Review and Editing: S.C.; Visualization: S.C.; 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] Z. N. Sergeeva, "Social Media as a New Institutional Structure for Communication," *Society and Security Insights*, vol. 2023, no. 1, pp. 56–65, 2023. doi: 10.14258/ssi(2023)1-03

[2] G. Piechota, "The transnational discourse of political protests: setting the agenda through social media," *KOME: An International Journal of Pure Communication Inquiry*, vol. 9, no. 2, pp. 73–88, 2021. doi: 10.17646/kome.75672.53

[3] G. Neubaum and N. C. Krämer, "Opinion Climates in Social Media: Blending Mass and Interpersonal Communication," *Human Communication Research*, vol. 43, no. 4, pp. 464–476, 2017. doi: 10.1111/hcre.12118

[4] K. Reynolds, A. Kontostathis, and L. Edwards, “Using machine learning to detect cyberbullying,” in *Proceedings of the 10th International Conference on Machine Learning and Applications (ICMLA)*, Honolulu, HI, USA, vol. 2012, no. February, pp. 241–244, doi: 10.1109/ICMLA.2011.152.

[5] K. Dinakar, R. Reichart, and H. Lieberman, “Modeling the detection of textual cyberbullying,” *The Social Mobile Web*, vol. 11, no. 2, pp. 11–17, 2011.

[6] V. Nahar, X. Li, and C. Pang, “An effective approach for cyberbullying detection,” *Communications in Information Science and Management Engineering*, vol. 3, no. 5, pp. 238–247, 2014.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv*:1301.3781, vol. 2013, no. January, pp. 12, 2013.

[8] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, vol. 2014, no. October, pp. 1532–1543.

[9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, no. June, pp. 135–146, 2017, doi: 10.1162/tacl_a_00051.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, vol. 2019, no. June, pp. 4171–4186.

[11] Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on Twitter using deep learning,” *Journal of Information and Telecommunication*, vol. 3, no. 1, pp. 1–16, 2019, doi: 10.1080/24751839.2019.1579394.

[12] M. Mozafari, R. Farahbakhsh, and N. Crespi, “A BERT-based transfer learning approach for hate speech detection in online social media,” in *Complex Networks and Their Applications VIII, Springer, Cham*, 2020, pp. 928–940.

[13] M. Karan and K. Šojat, “Room for improvement in automatic hate speech detection: A case study on multilingual Twitter data,” *arXiv preprint arXiv*:2109.03136, 2021.

[14] Y. Liu et al., “RoBERTa: A robustly optimized BERT pretraining approach,” *arXiv preprint arXiv*:1907.11692, 2019.

[15] Z. Lan et al., “ALBERT: A lite BERT for self-supervised learning of language representations,” *arXiv preprint arXiv*:1909.11942, 2020.

[16] M. Mozafari, R. Farahbakhsh, and N. Crespi, “A BERT-based transfer learning approach for hate speech detection in online social media,” in *Complex Networks and Their Applications VIII, Springer, Cham*, 2020, pp. 928–940.

[17] B. Wallace, Y. Wang, L. Resnik, and A. Benton, “The effect of perspective in measuring perceived toxicity,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 1–12.

[18] H. Ibrahim, S. M. Abdou, and M. Gheith, “Sentiment analysis for modern standard Arabic and colloquial,” *International Journal on Natural Language Computing (IJNLC)*, vol. 7, no. 2, pp. 95–109, 2018.

[19] Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on Twitter using deep learning,” *Journal of Information and Telecommunication*, vol. 3, no. 1, pp. 1–16, 2019.

[20] S. Mishra and P. Bhattacharyya, “Deep learning-based emotion-aware cyberbullying detection,” *Pattern Recognition Letters*, vol. 139, pp. 243–250, 2021.

[21] D. Ghosh and T. Veale, “Fracking sarcasm using neural network,” in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, San Diego, CA, USA, vol. 2016, no. June, pp. 161–169.

[22] A. Rajadesingan, P. Resnick, and C. Budak, “Sarcasm detection on social media: A behavioral modeling approach,” in *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM)*, Houston, TX, USA, 2020, pp. 95–103.

[23] A. Kumar, S. Behera, and D. P. Mohapatra, “Joint learning for sentiment, emotion, and sarcasm detection using multitask attention-based BiLSTM,” *Cognitive Computation*, vol. 13, no. 5, pp. 1175–1187, 2021, doi: 10.1007/s12559-021-09810-4.

[24] H. Zhang, J. Ma, and W. Liu, “Multi-task BERT for sarcasm and offensive language detection,” *IEEE Access*, vol. 9, pp. 138244–138255, 2021, doi: 10.1109/ACCESS.2021.3118560.

[25] S. Tsakalidis et al., “Overview of the HatEval task at SemEval 2021: Multilingual detection of hate speech against immigrants and women in Twitter,” in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021),* 2021, pp. 57–64.

[26] R. Handayani, M. Yusuf, and T. Hariguna, “Legal aspects of digital speech: Analysis of Indonesian cyber law and linguistic evidence,” *Jurnal Ilmiah Hukum dan Kebijakan Publik*, vol. 7, no. 2, pp. 56–72, 2021.

[27] J. Shen, “Corpus classification algorithm based on Bert pre-trained model,” *Procedia Computer Science*, vol. 262, pp. 368–377, 2025. doi:10.1016/j.procs.2025.05.064