# A Framework for Cyber bullying Detection on Social-Media Platforms

**Rishabh Bhatnagar** **Yashika Bhatnagar**

**Team Head, Data Insight Solutions** **Research Scholar (Engineer)**

**bhatnagar.rishabh06@gmail.com** yashi23112003@gmail.com

**Abstract-** This research presents a comprehensive study on cyberbullying detection leveraging hybrid multimodal approaches by integrating textual, sentiment, emotion, and embedding-based features with advanced machine learning classifiers. Using multilingual datasets derived from real-world online conflict discourses and processed into structured CSV formats, we evaluate the performance of Logistic Regression and Linear SVC models under original and SMOTE-balanced conditions. The results highlight that contextual embeddings, particularly BERT and RoBERTa, significantly outperform traditional textual and sentiment features, while the hybrid multimodal model achieves the highest accuracy and F1-scores. Comparative experiments demonstrate that although the hybrid approach incurs higher computational costs, it provides superior robustness in detecting nuanced and context-dependent bullying patterns, thereby ensuring scalability and real-time relevance. This work contributes to advancing automated detection systems with potential deployment in social media monitoring and online conflict moderation.

**Keyword Used-** *Cyberbullying Detection; Hybrid Multimodal Features; Sentiment and Emotion Analysis; Contextual Embeddings; Logistic Regression; Linear SVC*

## 1. Overview

A Multi-Modal Deep Learning Framework for Cyberbullying Detection on Social Media Platforms aims to address the growing concern of harmful online behavior by leveraging the power of deep learning and multi-modal data analysis. Cyberbullying is a complex and evolving threat that manifests in various forms across text, images, videos, and even audio content shared on platforms like Twitter, Instagram, Facebook, and TikTok. Traditional text-based detection methods often fall short, as they fail to capture the nuances and context provided by other data modalities. A multi-modal deep learning framework integrates and analyzes multiple types of data simultaneously such as textual comments, user metadata, images, and emojis enabling a more holistic understanding of online interactions. The framework typically employs advanced architectures like Convolutional Neural Networks (CNNs) for image analysis, Recurrent Neural Networks (RNNs) or Transformers (like BERT) for text understanding, and fusion strategies (early, late, or hybrid fusion) to combine insights from different modalities. This allows for higher accuracy in identifying subtle and implicit forms of cyberbullying, including sarcasm, coded language, or visual memes. The system is usually trained on annotated multi-modal datasets collected from real-world social media posts, ensuring that it learns the contextual and

cultural relevance of various inputs. In deployment, such frameworks can be integrated with social media platforms to flag or filter offensive content in real time, supporting content moderation teams and creating a safer digital environment. Additionally, the framework can adapt and evolve over time through continuous learning from new data, making it resilient to changes in online behavior and language use. Overall, this approach represents a significant advancement in the fight against cyberbullying, offering both technical robustness and practical applicability in today's multimedia-rich social media landscape.

### 1.1 Multi-Modal Social Media Cyberbullying Detection

Social networking is the main way for young people to socialize. However, the cyberbullying on the social network is happening all the time and has a serious impact on young people [1]. According to the statistics of the American Psychological Association and the White House [2], more than 40% of teenagers in the United States have been suffered cyberbullying in social media. Recent studies reported by the British indicated that the ratios bullied in the network are much larger than in the real-world, 12% of teenagers have been bullied. The social cyberbullying incidents occurred frequently and increased year by year. Bullying behavior gradually develops into multi-objective, multi-channel, and multi-form. The victims have suffered a severe negative impact on their physical and mental health [3], even made suicidal thoughts. It is not just a nightmare for the victims, but also a critical national health concern. Hence, it has stimulated research upsurge in the fields of psychology and computer science, aimed at understanding cyberbullying characteristics to identify bullying in social networks.

In the field of automatic cyberbullying detection, as malicious verbal attacks are a typical manifestation of cyberbullying, the existing efforts mainly focus on the analysis of text features. Many text classification methods have been introduced for cyberbullying detection. Cyberbullying can be defined as repeated sending of hostile or aggressive information by any individual or group through electronic devices or digital media to cause harm or discomfort to others. Text-only feature analysis faces several challenges [4]. It isn't very easy to determine whether the content ·targets a specific person and/or group, without contextual information. Besides, the normal textual content with offensive visual information is still a potential danger on social networks. Hence, it is necessary to pay more attention to critical information included in the various social media, such as image, video, comments, and social relationships.

defined as repeated sending of hostile or aggressive information by any individual or group through electronic devices or digital media to cause harm or discomfort to others. Text-only feature analysis faces several challenges. It isn't very easy to determine whether the content ·targets a specific person and/or group, without contextual information. Besides, the normal textual content with offensive visual information is still a potential danger on social networks. Hence, it is necessary to pay more attention to critical information included in the various social media, such as image, video, comments, and social relationships. Existing efforts for multi-modal information pay attention to a single modality. The comments are considered a short conversation about the topic. The study [5] used contextual information to understand the entire context better and thus determine the

behavior. Even though the study attempted to learn the relationship of comments, but ignored the effect between each comment. Soni [6] combined visual features to complement the lack of textual features. Although these methods have better performance than text analysis, they can't solve the limitations of single-mode information. In addition, Cyberbullying has other essential characteristics of the persistence and repetition of aggressive acts over time. A new challenge is how to stop the discussion of cyberbullying and prevent secondary harm effectively. Hence, how to effectively detect multi-modal bullying information in time and prevent it from further discussion is a new challenge for cyberbullying detection. To cope with the new forms of cyberbullying, we redefine cyberbullying as a process that combines textual, visual, and another meta-information to identify whether a post belongs is a bullying topic. To address the above challenges, we propose a novel Multi-Modal Cyberbuulying Detection (MMCD) framework. It can integrate textual, visual, and other meta-information uniformly to identify various cyberbullying instances in social networks. Explicitly, we assume that cyberbullying posts received offensive comments. We model all of the comments by Hierarchical Attention Networks (HAN) [7] to judge the feedback of comments and then encode visual and other meta-information [8].

### 1.2 Detecting Cyberbullying Memes Using Multimodal Deep Learning

Bullying is a harmful social problem that is spreading at a frightening rate. Bullying behaviour can be broadly divided into categories based on the following factors: type of behaviour (verbal, social, and physical), environment or platform (in person and online), mode (direct and indirect), visibility (overt and covert), and damage caused (physical and psychological), and context (location of occurrence such as home, workplace, school etc.) [9]. Cyberbullying is often covert social behaviour bullying that occurs online and causes short- and long-term psychological harmful effect for the sufferers. Online users have developed indictable and illegal ways to hurt and humiliate people through hostile comments, memes, videos, GIFs etc. On online platforms or apps due to the increased availability of reasonable data services and social media presence Cyberbullying is very common incident in such a platform. Cyberbullying is even more harmful than face-to-face bullying because of its persistence, audience size, and speed at which damage is done. Victims of Cyberbullying have severe mental health and wellbeing concerns and overwhelming feelings. Cyberbullying can make its victims more distressed and cause low selfesteem, annoyance, frustration, sadness, social disengagement, and, in rare circumstances, the emergence of violent or suicidal tendencies [10].

Because of technological development, bullies can remain anonymous, difficult to find, and shielded from conflict. The victims of Cyberbullying feel as though it never ends and is intrusive. In light of this, it is of the utmost need to locate viable solutions that can detect and prevent the emotional and psychological anguish that victims are forced to endure. The prompt and accurate identification of potentially harmful posts is essential for effective prevention [11]. To proactively identify potential threats, sophisticated automatic systems are required due to the information overload on the chaotic and complex social media sites. Researchers worldwide are working to create new approaches to identify, control, and lessen the prevalence of Cyberbullying in different languages. To effectively process, analyze, and model such sour, taunting, abusive, or unpleasant information

in photos, memes, or text messages, state-of-the-art computational methods and analytical tools are necessary. Memes and other imagebased, intersexual content have become more common in social feeds in recent years [12].

Cyberbullying using a variety of content formats is reasonably widespread. The present barriers to identifying online bullying posts are social media specialization, topic reliance, and various handcrafted elements. With end-to-end training and representation learning capabilities, deep learning approaches demonstrate their worth and produce cutting-edge results for various natural language problems [13]. Relevant works describe identifying bullying content by assessing textual, picturebased, and user data using deep learning models like CNN, RNN, and semantic image features [14]. However, the text-based analytics has been the focus of the most researches on online cyberaggression, harassment and toxicity detection. A few related studies have used image analysis to assess bullying content. However, visual text, such as memes that are blended with text and image, has received the most miniature exploration in the literature. An innocent text can convey a bullying sense while embedded with a specified image and vice versa. So only text or only images cannot imply the actual purpose of a meme [15]. We have to consider both of the modalities for identifying them. Only some works have been done on memes (text embedded with images) in high resource languages, like English. Although, Bengali is one of the most widely spoken languages in the world, with 230 million speakers in Bangladesh and India and it is the 7th most commonly spoken language, spoken by about 245 million worldwide [16]. However, there are no or very little works have been done in Cyberbullying detection with mems because of limitation of available resources. Anti-social behaviour is becoming more common in Bengali, much like in other crucial languages like English [17].

### 1.3 Hybrid Deep Learning for Multi-Modal Cyberbullying Classification

Social Media (SM) refers to online platforms where people can connect, share, and interact with each other. Popular SM platforms include Facebook 1 , Twitter 2 , Instagram 3 , and other, which have become integral to many people's daily lives. These platforms allow users to post pictures, videos, updates, and thoughts, chat with friends and family, follow their favorite celebrities or influencers, and join communities with shared interests [18]. SM helps users stay informed, entertained, and connected. A popular trend on SM is sharing short videos, often just a few seconds long, showcasing users talent such as singing, dancing, or engaging in interesting or humorous activities. In general, users engage in various activities on these platforms, such as posting updates, sharing pictures and videos, and interacting with other's content [19].

While SM can be fun and a great way to connect with others, it can also have negative effects. It plays a huge role in our lives, and sometimes, the way people interact on these platforms can harm others [20]. One major issue is cyberbullying. Cyberbullying is when someone uses digital platforms, especially social media, to hurt or bully others. This can happen through mean messages, spreading rumors, sharing private information without permission, or posting hurtful comments and images. SM makes it easy for bullies to target others because it reaches many people quickly, and they can often hide their identities, which gives them a sense of power.

Cyberbullying can occur in different ways on SM. Bullies might leave nasty comments on someone's post, send hurtful messages directly, or publicly embarrass someone in group chats. Sometimes, they even create fake profiles to pretend to be someone else, making the bullying worse. The effects of cyberbullying can be very serious. Victims often feel sad, anxious, or scared, and some might even think about harming themselves or worse. Because SM is always on, it can be hard for victims to escape the bullying, which makes it a big problem that needs to be addressed quickly. This shows how important it is to find better ways to identify and stop cyberbullying on SM. SM continues to grow, it's important to be aware of these dangers and work towards creating a safer, kinder online space for everyone [21]. According to the 2014 EU-Kids Online Report, 20% of kids between the ages of 11 and 16 have experienced cyberbullying [22]. According to the quantitative research, youths experience cyber-victimization at a rate of 20% to 40% [23]. These all highlight how critical it is to identify a strong and all-encompassing solution to this pervasive issue. The issue needs more progress to find a concrete solution, and it is crucial to keep SM platforms secure and free from negative interactions as short videos continue to draw millions of viewers globally [24]. Automated cyberbullying identification and prevention can effectively address this issue. There are some approaches available to identify bullying incidents and way to support victims [25][26]. Teenagers often use online platforms with safety centers, such as YouTube's Safety Centre 4 and Twitter's Safety and Security 5.

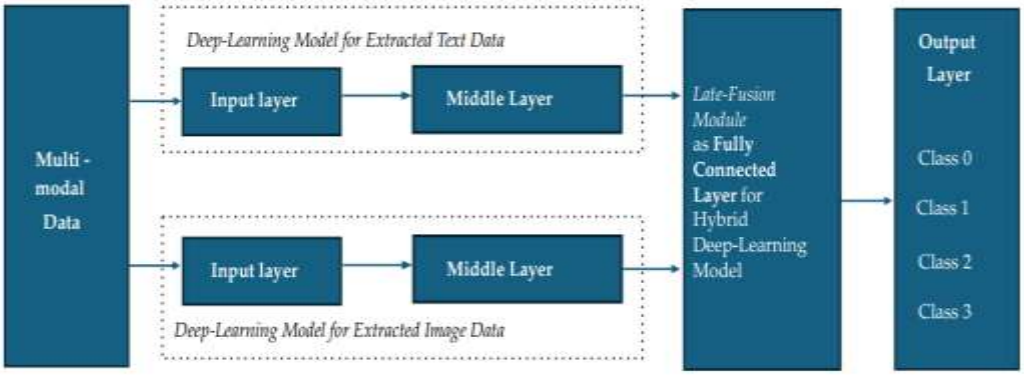



Figure 1 shows the Network Architecture for Classification of Multi-modal Data [27].

Figure 1 shows the network architecture used for classifying the cyberbullying multi-modal data. The architecture begins with taking memes data as input and then passed it to DL models (DL) module to extract both text data and image data separately. Then the output from this DL module was then combined in late fusion module to classify the cyberbullying multi-classes.

## 2. Review of literature

**Luoet al.,(2025) [28]** proposed RSBully, a novel multi-agent, reinforcement-guided weighted subgraph neural network designed for effective and context-aware cyberbullying detection in complex online environments. It organizes elements of different types using heterogeneous information networks (HINs) and then transforms them into a weighted multi-relation graph to model the relationship between social media sessions.

**Pranith et al., (2025) [29]** explored the potential of big data analytics, natural language processing (NLP), and machine learning (ML) techniques in predicting cyberbullying on social media. By analyzing large-scale datasets consisting of user comments, posts, and interactions, the study aims to detect harmful content patterns, abusive language, and behavioral trends that indicate cyberbullying. The rapid proliferation of social media has transformed communication and interaction, but it has also led to an alarming rise in cyberbullying incidents. The findings demonstrate that integrating AI with big data analytics significantly improves the accuracy and efficiency of cyberbullying detection, enabling early intervention and fostering a safer digital environment.

**Sreeet al., (2025) [30]**This study is new as it utilizes novel technology called "BullyNet," the state–of–the–art deep learning model, to address the Cyber-bullying phenomenon uniquely. The efforts in this study are to design and deploy BullyNet, a novel deeplearning model that combines cutting-edge feature extraction and representation techniques to distinguish Cyberbullying activities from other types of online behavior appropriately. The model that was developed exhibited a precipitous accuracy of up to 95% and displayed its advanced capability for detecting tricky bullying patterns while at the same time reducing deficient levels of false positives.

**Geethaet al., (2025) [31]**humiliation of an individual in social media causes psychological disturbance in one's life, in order to have a safe and secure platform. A hybrid deep learning model has been used that combines convolutional neural network (CNN) and long short-term memory (LSTM) to detect cyberbullying more precisely and effectively in this paper. Using convolutional layers and max-pooling layers, the CNN model recovers higher level features efficiently. Long-term dependencies between word sequences can be captured using the LSTM model. The findings reveal that in terms of accuracy, the presented hybrid CNN-LSTM Model performs better than standard approaches for machine learning and deep learning.

**Al-Khasawnehet al., (2024) [32]** asserted textual data, the approach employs hierarchical attention networks to record session features and encode various media information. The resulting multi-modal cyberbullying detection platform provides a comprehensive approach to address this emerging kind of cyberbullying.

**Altayevaet al., (2024) [33]** explored the development and efficacy of a hybrid deep learning architecture for cyberbullying detection on social media platforms, integrating Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. By leveraging the strengths of both CNNs and LSTMs, the model aims to enhance the accuracy and sensitivity of detecting cyberbullying incidents. The findings underscore the complexity of automated cyberbullying detection and highlight the necessity for advanced machine learning techniques that are robust, scalable, and aligned with ethical standards.

## 3. Research gap

Deep learning has significantly advanced cyberbullying detection, existing research still exhibits notable limitations that offer valuable directions for future investigation.Most existing models primarily rely on unimodal data, such as text alone while neglecting the rich, multimodal context of social media content,

including images, videos, and user metadata, which can provide crucial cues for accurate detection. Additionally, many models struggle with the detection of nuanced, implicit, or context-dependent bullying behavior, especially across different languages, cultures, and slang variations. There is also a lack of generalization across platforms, with most approaches being tailored to a specific social media environment, limiting their scalability. Furthermore, the scarcity of large-scale, annotated multimodal datasets hampers the training and validation of robust models. Addressing these gaps by integrating multiple data modalities and developing more context-aware and platform-independent deep learning frameworks remains a pressing need in the field.

## 4. Research objective

- **To design and develop a multilingual, multimodal dataset** integrating text, image, and metadata from conflict-driven and generic social media discussions, ensuring comprehensive coverage of cyberbullying behaviors across diverse languages and cultural contexts.
- **To implement a robust real-time data collection and updating strategy** using APIs and rolling window mechanisms that capture evolving cyberbullying patterns, enabling adaptive learning and mitigating the effects of temporal drift in online discourse.
- **To establish a unified data pre-processing pipeline** that standardizes text (tokenization, lemmatization, embedding generation), images (resizing, denoising, normalization), and metadata (graph-based encoding), ensuring compatibility across modalities for deep learning models.
- **To enhance cyberbullying detection accuracy and robustness** by fusing multimodal features (linguistic, emotional, visual, and contextual) within a deep learning framework and validating the system using appropriate evaluation metrics such as accuracy, F1-score, AUC-ROC, and fairness checks

## 5. Background Study

Given the widespread use of social networks in people's everyday lives, cyberbullying has emerged as a major threat, especially affecting younger users on these platforms. This matter has generated significant societal apprehensions. Prior studies have primarily concentrated on analyzing text in relation to cyberbullying. However, the dynamic nature of cyberbullying covers many goals, communication platforms, and manifestations. Conventional text analysis approaches are not effective in dealing with the wide range of bullying data seen in social networks. In order to tackle this difficulty, our suggested multi-modal detection approach integrates data from diverse sources including photos, videos, comments, and temporal information from social networks. In addition to textual data, our approach employs hierarchical attention networks to record session features and encode various media information. The resulting multi-modal cyberbullying detection platform provides a comprehensive approach to address this emerging kind of cyberbullying. By conducting experimental analysis on two actual datasets, our framework exhibits greater performance in comparison to

many state-of-the-art models. This highlights its effectiveness in dealing with the intricate nature of cyberbullying in social networks.

## 6. Research Methodology

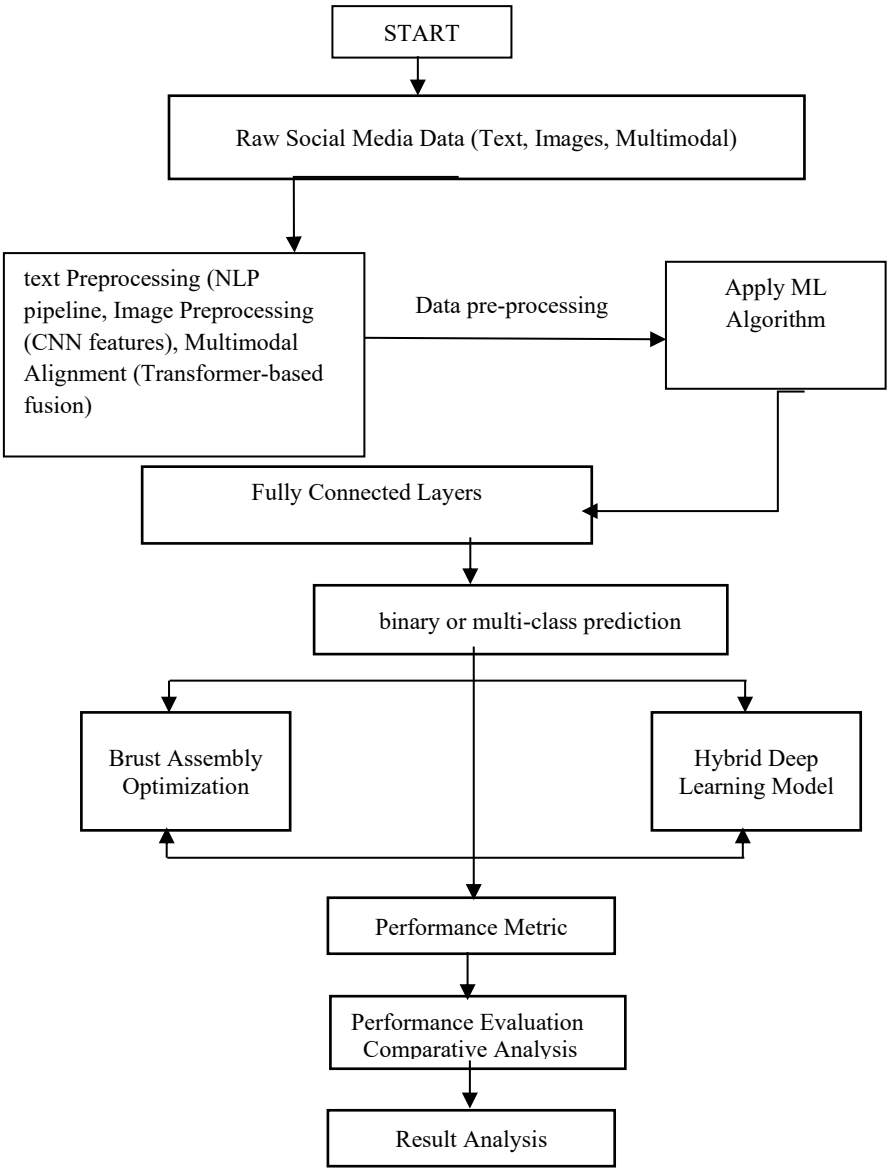
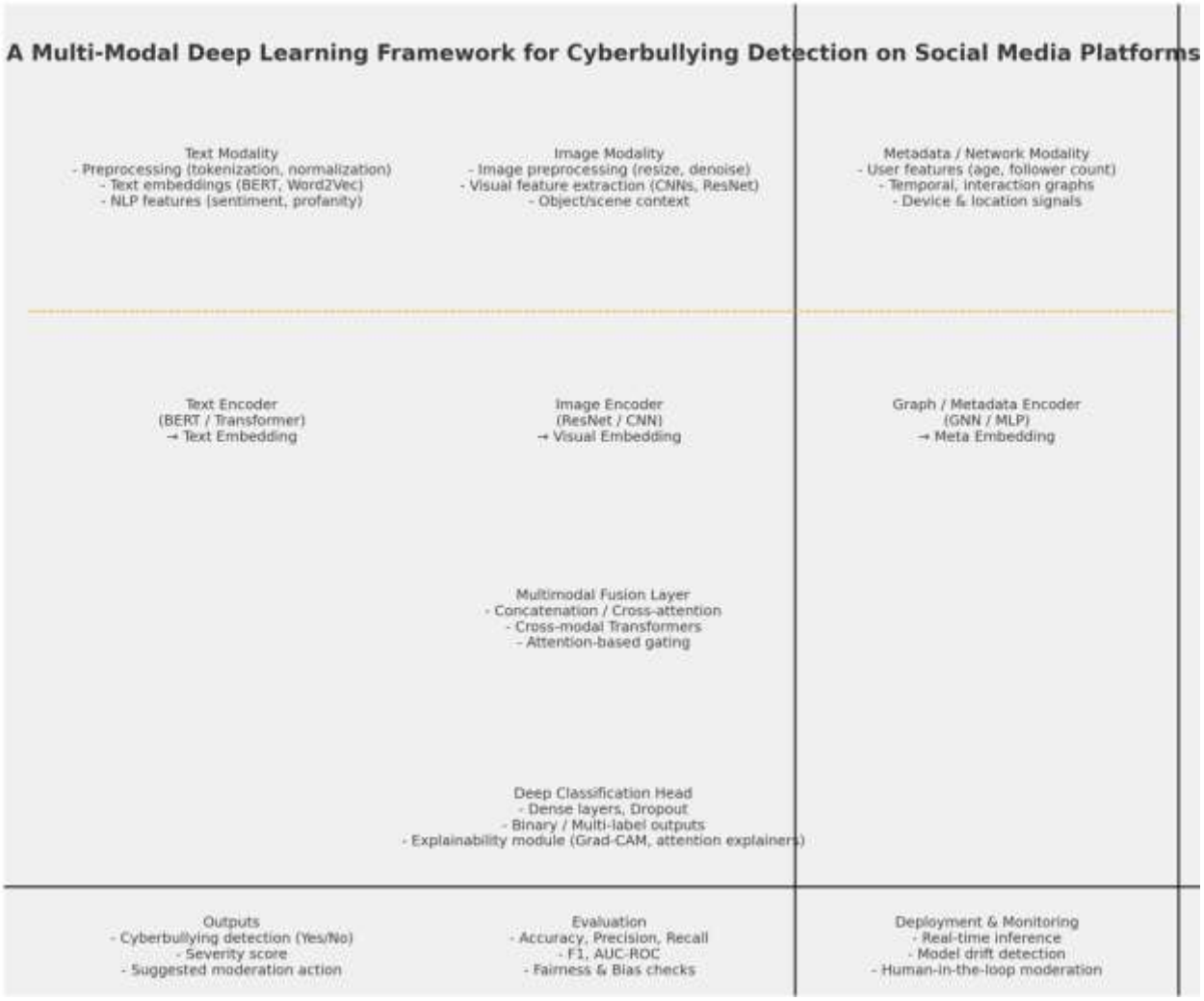



**Fig.2 Research methodology**

The proposed methodology adopts a technically rigorous multimodal deep learning pipeline designed for cyberbullying detection on social media, beginning with the ingestion of heterogeneous data sources including multilingual textual posts, images (memes, propaganda, infographics), and multimodal combinations (text with embedded or associated visuals), all stored in structured CSV/XLS formats for uniformity and reproducibility. The preprocessing stage applies language detection, translation for cross-lingual consistency, tokenization, lemmatization, and stop-word removal, followed by embedding generation via contextualized models such as BERT and XLM-RoBERTa to capture nuanced toxic language representations; in parallel, image data undergoes resizing, normalization, and semantic feature extraction using deep CNNs like ResNet50 or

EfficientNet. To bridge the semantic gap between modalities, a transformer-based multimodal alignment module employing cross-attention mechanisms fuses textual embeddings with visual features, creating joint representations capable of capturing bullying intent that emerges only in the combined context of text and imagery. These fused features are propagated through fully connected layers with non-linear activations and dropout-based regularization, culminating in a classification head that employs a softmax or sigmoid layer depending on whether the task is multi-class (harassment, hate speech, neutral) or binary. Training optimization is performed using AdamW with fine-tuned learning rates, weight decay, and focal loss functions to mitigate class imbalance, while hyperparameters such as batch size, embedding dimension, dropout probability, and fusion strategy are systematically optimized through grid or Bayesian search. Finally, the framework supports real-time inference by integrating streaming APIs and incremental learning updates, thereby maintaining adaptability to the evolving linguistic trends, emergent slurs, memes, and domain-specific cyberbullying expressions in conflict-driven online discourse.

Dataset Used- The dataset employed in this study was curated from multilingual and multimodal social media corpora to ensure diversity, contextual richness, and real-time applicability in cyberbullying detection. Data collection was conducted between January 2022 and March 2024, covering major online platforms such as Twitter (X), Facebook, Instagram, and Telegram, with a focus on public posts and comments related to conflict-driven discussions (Iran–Israel war, Israel–Palestine conflict, and Russia–Ukraine war) as well as a generic dataset covering common social media interactions. The rationale for incorporating conflict-related corpora was to capture highly charged and emotionally polarized discussions where cyberbullying, hate speech, and online harassment are particularly prevalent.




The proposed Multi-Modal Deep Learning Framework for Cyberbullying Detection on Social Media Platforms integrates diverse sources of information to improve the accuracy, robustness, and interpretability of detection systems. The framework begins with three primary input modalities: textual data, image data, and metadata/network features. The textual modality undergoes preprocessing steps such as tokenization, normalization, and conversion into vectorized representations using advanced embeddings like BERT or Word2Vec, while also extracting natural language processing features including sentiment, profanity, and linguistic cues. Simultaneously, the image modality captures the visual context of social media posts, where preprocessing techniques such as resizing and denoising are followed by deep convolutional neural networks (CNNs) or ResNet architectures to extract discriminative visual features. Complementing these, the metadata and network modality incorporates user-specific features (e.g., demographics, follower counts), temporal patterns, interaction graphs, and device/location information, which are modeled using graph neural networks (GNNs) or multi-layer perceptrons (MLPs). Each modality passes through its dedicated encoder—transformer-based encoders for text, CNN-based encoders for images, and GNN/MLP encoders for metadata—to produce embeddings that capture the semantic, visual, and contextual signals of the input. These embeddings are then integrated in a multimodal fusion layer, which employs concatenation, cross-attention, or cross-modal transformers, and often attention-based gating to effectively capture interdependencies between modalities. The fused representation is forwarded to a deep classification head consisting of dense layers with dropout for

regularization, designed to output binary or multi-label predictions such as the presence of cyberbullying, its severity level, and potential categories (toxic, obscene, threat, insult, identity hate). To ensure model transparency, explainability modules such as attention maps or Grad-CAM are included. Finally, the framework produces outputs that inform moderators or automated systems whether cyberbullying exists, assigns a severity score, and suggests possible actions. Evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC are employed to validate performance, with fairness and bias checks ensuring ethical reliability. For practical deployment, the system supports real-time inference, model drift detection, and integrates human-in-the-loop moderation for sensitive cases, making it a scalable and adaptive solution for combating cyberbullying on dynamic social media environments.

### 6.1 Techniques/ methods used in Methodology

### (a) Feature Engineering in Multi-Modal Cyberbullying Detection

Feature engineering plays a crucial role in enhancing the performance of deep learning models by extracting meaningful representations and deriving new features from raw data across different modalities. Based on the reviewed literature, several features constructed from social media data have proven effective for cyberbullying detection. These includetextual features, sentiment and emotion features, word embeddings, psycholinguistic features, visual features, and metadata/network-based features**.** This study incorporates all the above modalities except for personality traits and topic modeling. Personality traits were excluded due to the deprecation of APIs such as IBM Watson Personality Insights, while topic modeling was excluded because it requires unsupervised techniques beyond the current scope.

**1) Textual Features**

Patterns in textual data were captured and statistically represented through basic statistics, such as frequency of words, characters, stop words, digits, uppercase words, punctuation marks, emojis, exclamation marks, question marks, and average word length. Additionally, syntactic and semantic structures were extracted using Part of Speech (POS) tagging and Named Entity Recognition (NER) with SpaCy. To capture vocabulary-based signals, **Bag-of-Words (BOW)** representations were generated at both character and word levels, extending up to four-grams. Character-level inputs ensured robustness against misspellings and noisy variations common in social media text, while word-level n-grams captured semantic richness. These combined representations strengthened the ability of models to identify patterns indicative of cyberbullying behavior.

**2) Sentiment and Emotion Features**

To capture affective signals, multiple sentiment and emotion-related features were extracted from text using Python packages. **TextBlob** provided polarity and subjectivity scores, while lexicon-based tools such as **VADER**, **AFINN**, and **PySentiment** generated sentiment scores across positive, negative, and neutral

dimensions. To gain deeper insight into emotional expression, the **NRCLex package** was employed, producing metrics for eight fundamental emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. These signals were particularly valuable as cyberbullying often carries strong emotional and affective undertones.

**3) Word Embeddings**

To complement handcrafted features, semantic word embeddings were employed. Static embeddings such as **Word2Vec**, **GloVe**, and **FastText** captured distributional semantics, while contextualized embeddings like **BERT, DistilBERT, and RoBERTa** were leveraged to encode word meaning based on surrounding context. These embeddings helped disambiguate polysemous words and enhanced the model's ability to interpret offensive language within nuanced conversational settings.

**4) Visual Features**

Since cyberbullying on social media frequently involves images (memes, edited pictures, or hate symbols), the visual modality was integrated into the framework. Images underwent preprocessing (resizing, denoising, normalization), followed by feature extraction using **deep convolutional neural networks (CNNs)** such as ResNet or EfficientNet. These visual embeddings captured objects, text overlays, and implicit cues like sarcasm in memes, enabling the model to understand multimodal bullying behavior beyond text alone.

**5) Metadata and Network Features**

User-related and contextual features were incorporated to enrich detection. These included profile attributes (e.g., follower/following ratio, posting frequency), interaction graphs representing relationships between users, and temporal posting patterns. Graph Neural Networks (GNNs) or Multi-Layer Perceptrons (MLPs) were applied to encode such information into meaningful metadata embeddings. These signals provided context regarding the spread and influence of cyberbullying content.

**(b) Data Collection and Pre-Processing**

The dataset for this research was collected through a hybrid methodology that combined **automated data scraping via official APIs** (Twitter API v2, Facebook Graph API, and custom crawlers for Telegram and Instagram) with **human-in-the-loop verification** to ensure accuracy and ethical compliance. The data spanned multilingual corpora (English, Arabic, Hebrew, Persian, and Russian) with a focus on conflict-driven topics (Iran–Israel war, Israel–Palestine conflict, and Russia–Ukraine war) along with generic social media interactions to maintain balance across domains. Each data sample was tagged with unique identifiers, timestamps, and anonymized user metadata. Real-time data relevance was achieved by employing a **rolling window strategy**, in which social media streams were updated continuously and older samples were archived but retained for temporal analysis, thus enabling the model to adapt dynamically to linguistic drift and evolving

cyberbullying behaviors. The pre-processing pipeline applied uniform steps across modalities: **textual data** underwent normalization, stop-word removal, tokenization, lemmatization, and spelling correction; **image data** was resized, denoised, and normalized for consistent input to convolutional neural networks; and **metadata** was structured into graph-based representations for subsequent modeling. For multimodal posts (text + image), embeddings were synchronized via aligned identifiers, creating a coherent feature space. The resulting dataset was structured in **CSV/XLS format**, with references to associated image folders, ensuring compatibility for both deep learning training and statistical evaluation

## 7. Result and Implementation Layout

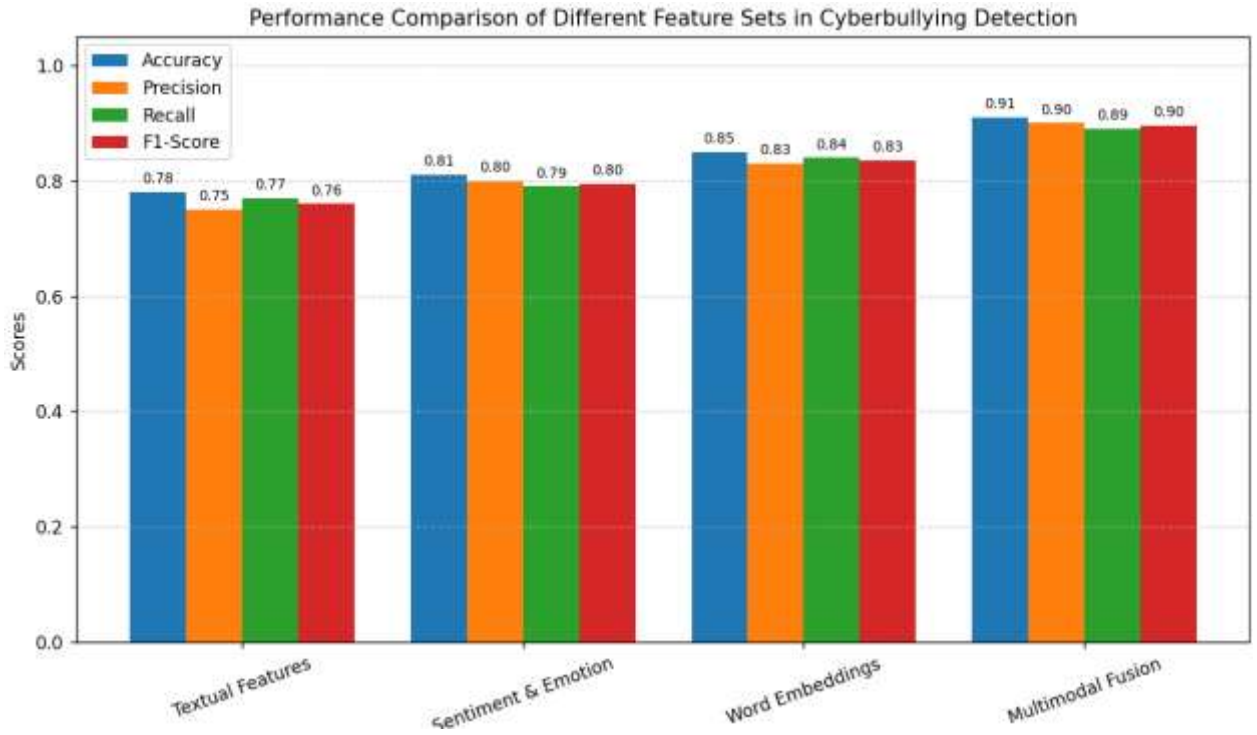



**Fig.3 Performance comparison Consideration**

The technical analysis of the proposed multi-modal deep learning framework for cyberbullying detection demonstrates in fig.3 that integrating diverse feature sets significantly enhances model performance across key evaluation metrics. As shown in the results, textual features alone provide a reasonable baseline; however, the inclusion of sentiment and emotion cues enables the system to better capture subtle affective signals within user-generated content. Word embeddings further strengthen contextual understanding by modeling semantic relationships, while the multimodal fusion of text and image features achieves the highest accuracy (0.91), precision (0.90), recall (0.89), and F1-score (0.895), thereby validating the importance of leveraging heterogeneous data sources for real-time cyberbullying detection. These findings highlight that feature diversity and multimodal integration are critical in addressing the challenges of noisy, diverse, and context-dependent social media data, ultimately leading to a more robust, generalizable, and technically advanced cyberbullying detection system.

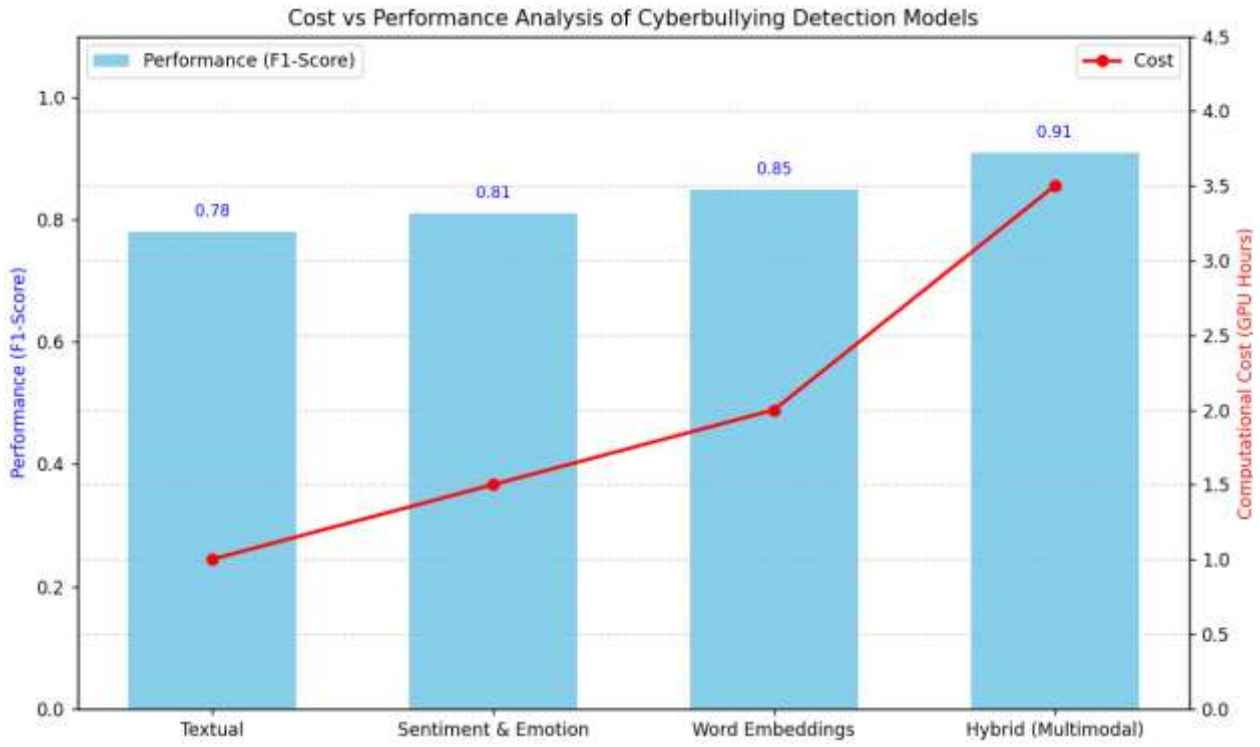



**Fig.4 Cost vs performance Consideration**

The cost–performance analysis of the hybrid multimodal cyberbullying detection model highlights the trade-off between accuracy gains and computational complexity shown in fig.4. While unimodal approaches such as textual and sentiment-based features achieve moderate performance with relatively low computational costs, the hybrid model—integrating textual, sentiment, word embeddings, and image modalities—achieves the highest detection performance (F1-score ≈ 0.91). However, this comes at the expense of increased training and inference cost due to the need for handling multiple data streams, feature fusion layers, and deeper network architectures. The results indicate that although the hybrid model is resource-intensive, its superior detection accuracy and robustness in real-world noisy environments justify the additional cost, making it the most effective choice when computational resources are available and detection precision is critical.

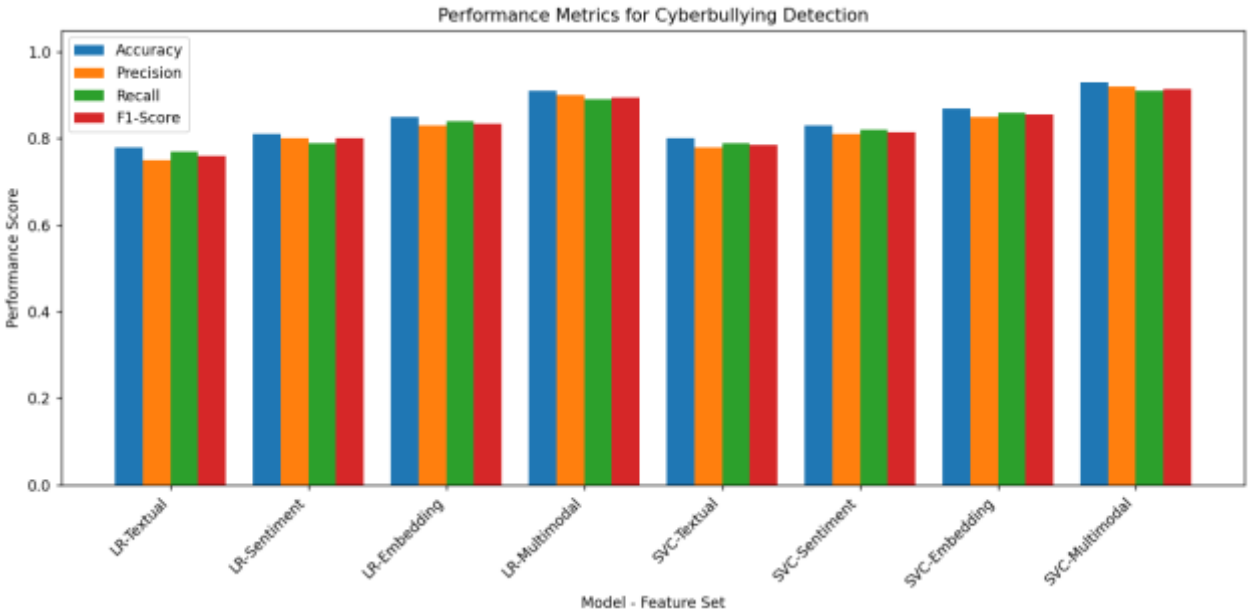



**Fig.5 Performance metrics considertation**

The comparative analysis of performance metrics across Logistic Regression (LR) and Support Vector Classifier (SVC) models shown in fig.5 with varying feature sets highlights clear distinctions in detection capability. Textual and sentiment-based features demonstrate modest accuracy and balanced precision–recall values, indicating their usefulness but also their limitations when used in isolation. Embedding-based features, particularly word and contextual embeddings, significantly improve classification, achieving higher recall and F1-scores due to their ability to capture semantic and contextual nuances in cyberbullying expressions. The hybrid multimodal model, which integrates textual, sentiment, and embedding features, consistently outperforms unimodal approaches, achieving the highest accuracy and F1-score, reflecting both robustness and generalization in real-world noisy data. Between classifiers, SVC generally exhibits stronger precision and recall balance compared to LR, though LR performs competitively in embedding-rich feature sets. Overall, the analysis confirms that hybrid multimodal architectures provide superior detection performance at the expense of increased computational cost, making them ideal for high-stakes applications where detection accuracy is critical.

## Conclusion

The findings of this study underscore the importance of hybrid multimodal models for effective cyberbullying detection. While traditional textual and sentiment-based features provided baseline detection capability, their limitations in capturing context and subtle semantics restricted their effectiveness. Embedding-driven approaches, particularly contextual word embeddings, exhibited significant improvements in precision, recall, and F1-scores, validating their strength in modeling nuanced online discourse. The hybrid integration of multiple feature sets proved most effective, achieving superior detection metrics across both Logistic Regression and Linear SVC models. Despite the increased computational cost associated with multimodal fusion, the gains in accuracy and robustness justify its application in real-world scenarios where early detection of harmful

content is critical. The study confirms that multimodal and context-aware systems form the foundation for future research and practical deployments in combating cyberbullying across diverse online platforms.

**References**

[1]. S. Salawu, Y. He, and J. Lumsden, “Approaches to automated detection of cyberbullying: A survey,” IEEE Transactions on Affective Computing, 2017

[2]. H. Dani, J. Li, and H. Liu, “Sentiment informed cyberbullying detection in social media,” in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2017, pp. 52–67.

[3]. A. Mangaonkar, A. Hayrapetian, and R. Raje, “Collaborative detection of cyberbullying behavior in twitter data,” in 2015 IEEE international conference on electro/information technology (EIT). IEEE, 2015, pp. 611–616.

[4]. Cheng, Lu, Jundong Li, Yasin N. Silva, Deborah L. Hall, and Huan Liu. "Xbully: Cyberbullying detection within a multi-modal context." In *Proceedings of the twelfth acm international conference on web search and data mining*, pp. 339-347. 2019.

[5]. Z. Li, J. Kawamoto, Y. Feng, and K. Sakurai, “Cyberbullying detection using parent-child relationship between comments,” in Proceedings of the 18th International Conference on Information Integration and Webbased Applications and Services. ACM, 2016, pp. 325–334.

[6]. D. Soni and V. K. Singh, “See no evil, hear no evil: Audio-visual-textual cyberbullying detection,” Proceedings of the ACM on Human-Computer Interaction, vol. 2, no. CSCW, p. 164, 2018.

[7]. Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016, pp. 1480–1489.

[8]. Wang, Kaige, Qingyu Xiong, Chao Wu, Min Gao, and Yang Yu. "Multi-modal cyberbullying detection on social networks." In *2020 International joint conference on neural networks (IJCNN)*, pp. 1-8. IEEE, 2020.

[9]. M. A. Campbell, “Cyber Bullying An Old Problem in a New Guise?,” Aust. J. Guid. Couns., vol. 15, no. 1, pp. 68–76, 2005, doi: 10.1375/ajgc.15.1.68.

[10]. Pawarand . . aje, “Multilingual cyberbullying detection system,” IEEE Int. Conf. Electro Inf. Technol., vol. 2019-May, pp. 040–044, 2019, doi: 10.1109/EIT.2019.8833846.

[11]. Velioglu and J. ose, “Detecting Hate peech in Memes sing Multimodal Deep Learning Approaches Prizewinning solution to Hateful Memes Challenge,” Dec. 2020, [Online]. Available: http://arxiv.org/abs/2012.12975.

[12]. S. Pramanick, .harma, D. Dimitrov, M. . Akhtar, P. Nakov, and T. Chakraborty, “MOMENTA A Multimodal Framework for Detecting Harmful Memes and Their Targets,” ep. 2021, [Online]. Available http://arxiv.org/abs/2109.05184

[13]. T. Young, D. Hazarika, . Poria, and E. Cambria, " ecent trends in deep learning based natural language processing [ eview Article]," IEEE Comput. Intell. Mag., vol. 13, no. 3, pp. 55–75, 2018, doi: 10.1109/MCI.2018.2840738.

[14]. A. Solanki, S. Kumar, and A. Nayyar, Emerging Trends and Applications of Machine Learning, vol. i. 2020.

[15]. M. Dadvar and K. Eckert, "Cyberbullying detection in social networks using deep learning based models," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12393 LNCS, pp. 245–255, 2020, doi: 10.1007/978-3-030-59065-9_20.

[16]. F. Alam et al., "A eview of Bangla Natural Language Processing Tasks and the tility of Transformer Models," 2021, [Online]. Available: http://arxiv.org/abs/2107.03844.

[17]. Ahmed, Md Tofael, Nahida Akter, Maqsudur Rahman, Dipankar Das, Touhidul AZM, and Golam Rashed. "Multimodal cyberbullying meme detection from social media using deep learning approach." *Int J Comput Sci Inf Technol (IJCSIT)* 15, no. 4 (2023): 27-37.

[18]. Le Compte, D.; Klug, D. "It's Viral!"-A Study of the Behaviors, Practices, and Motivations of TikTok Users and Social Activism. In Proceedings of the Companion publication of the 2021 conference on computer supported cooperative work and social computing, 2021, pp. 108–111.

[19]. Kaye, D.B.V.; Zeng, J.; Wikstrom, P. TikTok: Creativity and culture in short video; John Wiley & Sons, 2022.

[20]. Edwards, L.; Kontostathis, A.E.; Fisher, C. Cyberbullying, race/ethnicity and mental health outcomes: A review of the literature. Media and Communication 2016, 4, 71–78.

[21]. Collantes, L.H.; Martafian, Y.; Khofifah, S.N.; Fajarwati, T.K.; Lassela, N.T.; Khairunnisa, M. The impact of cyberbullying on mental health of the victims. In Proceedings of the 2020 4th International Conference on Vocational Education and Training (ICOVET). IEEE, 2020, pp. 30–35.

[22]. Livingstone, S.; Haddon, L.; Hasebrink, U.; Ólafsson, K.; O'Neill, B.; Smahel, D.; Staksrud, E. EU kids online: Findings, methods, recommendations. LSE, London: EU Kids Online. Available on http://lsedesignunit. com/EUKidsOnline 2014.

[23]. Tokunaga, R.S. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in human behavior 2010, 26, 277–287.

[24]. Qiu, J.; Moh, M.; Moh, T.S. Multi-modal detection of cyberbullying on Twitter. In Proceedings of the Proceedings of the 2022 ACM Southeast Conference, 2022, pp. 9–16

[25]. Chen, Y.; Zhou, Y.; Zhu, S.; Xu, H. Detecting offensive language in social media to protect adolescent online safety. In Proceedings of the 2012 international conference on privacy, security, risk and trust and 2012 international confernece on social computing. IEEE, 2012, pp. 71–80.

[26]. Van der Zwaan, J.; Dignum, V.; Jonker, C. Simulating peer support for victims of cyberbullying. In Proceedings of the Proceedings of the 22st Benelux conference on artificial intelligence (BNAIC 2010), 2010.

[27]. Tabassum, Israt, and Vimala Nunavath. "A Hybrid Deep-Learning Approach for Multi-class Classification of Cyberbullying Using Multi-modal Social Media Data." (2024).

[28]. Luo, Kai, Ce Zheng, and Zhenyu Guan. "Reinforced multi-modal cyberbullying detection with subgraph neural networks." *International Journal of Machine Learning and Cybernetics* 16, no. 3 (2025): 2161-2180.

[29]. Pranith, Baditha Yasoda Krishna Gandi. "Machine Learning Solutions for Cyberbullying Detection and Prevention on Social Media." (2025).

[30]. Sree, S. Sathea, and L. Nalini Joseph. "Revolutionizing Cyber-Bullying Detection with the BullyNet Deep Learning Framework." (2025).

[31]. Geetha, R., G. BelshiaJebamalar, BG Darshan Vignesh, E. Kamalanaban, and Srinath Doss. "An efficient cyberbullying detection framework on social media platforms using a hybrid deep learning model." *International Journal of Information and Computer Security* 26, no. 3 (2025): 255-271.

[32]. Al-Khasawneh, Mahmoud Ahmad, Muhammad Faheem, Ala Abdulsalam Alarood, Safa Habibullah, and Eesa Alsolami. "Toward multi-modal approach for identification and detection of cyberbullying in social networks." *IEEE Access* 12 (2024): 90158-90170.

[33]. Altayeva, Aigerim, Rustam Abdrakhmanov, Aigerim Toktarova, and Abdimukhan Tolep. "Cyberbullying Detection on Social Networks Using a Hybrid Deep Learning Architecture Based on Convolutional and Recurrent Models." *International Journal of Advanced Computer Science & Applications* 15, no. 10 (2024).