International Journal on Semantic Web and Information Systems
Volume 22 • Issue 1 • January-December 2026

Multimodal Cyberbullying Detection on Social Media Using LLMs:
A Comparative Study

Christian Catalano
http://orcid.org/0000-0003-4038-2317
University of Bari, Italy

Andrea Chezzi
http://orcid.org/0009-0004-0634-6640
University of Salento, Italy

Simone Lupo
University of Salento, Italy

Alessia Anna Catalano
University of Salento, Italy

Roberto Vadacca
http://orcid.org/0009-0005-4766-6744
University of Salento, Italy

Luca Mainetti
http://orcid.org/0000-0001-9387-9277
University of Salento, Italy

Received: December 30th, 2025 | **Accepted:** May 1st, 2026

ABSTRACT

This study explores automated cyberbullying detection across major social networks and messaging platforms using state-of-the-art large language models in a zero-shot, multimodal setting. Models including LLaMA 4, Gemma 3, and GeminiAI were evaluated on images and videos without domain-specific fine-tuning. The system assigned a continuous score (0–10) to indicate the presence of cyberbullying across four categories: revenge porn, happy slapping, racism, and body shaming. Experiments on over 5,000 multimedia samples from Telegram, Reddit, and X (formerly Twitter) showed that large language model-based approaches achieve competitive performance, with Gemma 3-12B emerging as the most stable, accurate, and ethically compliant model. The results also highlighted the critical role of prompt engineering and multimodal context in detecting subtle or implicit online aggression.

KEYWORDS

Cyberbullying Detection, Multimodal Analysis, Large Language Models, Visual Content Moderation, Social Media Safety, Hate Speech Analysis, Abusive Behavior Recognition, Online Harassment, Ethical AI

INTRODUCTION

The phenomenon of cyberbullying represents one of the major challenges of today's digital society, in a context where information technologies and social media permeate every aspect of daily life (Langos, 2012). Although these tools have revolutionized communication and enabled unprecedented access to information, they have also introduced new risks, particularly for younger segments of the population. Cyberbullying manifests itself through aggressive, intentional, and repeated behaviors carried out via digital platforms, such as social networks, forums, chats, and messaging applications, with the purpose of humiliating, denigrating, or psychologically and socially harming the victims (Tokunaga, 2010). Its manifestations may include not only direct verbal abuse but

DOI: 10.4018/IJSWIS.409891

also the non-consensual dissemination of personal information, the creation and sharing of derogatory or false content, and forms of online social exclusion (Grigg, 2010).

This work proposes an innovative and automated methodology for detecting cyberbullying episodes on social media, based on the use of advanced artificial intelligence (AI) and multimodal learning techniques. The goal is to provide an effective tool to support moderators, law enforcement agencies, and digital platforms capable of promptly identifying potentially offensive content, reducing the burden of manual analysis, and improving preventive intervention capabilities.

In particular, this paper introduces an approach that integrates the analysis of multimedia content (images and videos) with the processing of contextual and semantic signals, which are often overlooked by traditional approaches. Through the use of large language models (LLMs), the developed system aims to detect and classify cyberbullying events with high precision and robustness, providing a concrete contribution to the evolution of monitoring and prevention systems in the digital environment.

With the evolution of technologies and the increasing spread of social media, the online experience, especially for younger generations, has become a fundamental aspect of everyday life. However, along with the many benefits linked to access to information and new forms of communication, significant risks have emerged, including cyberbullying. This term refers to a wide range of aggressive and repeated behaviors carried out through digital platforms, such as social networks, forums, chats, and other online applications, with the intent to shame, humiliate, or harm the victims. Current detection and prevention systems are based on natural language processing (NLP) and are used on textual components, where they detect cyberbullying episodes through keywords and sentiment analysis. Unfortunately, these NLP models are insufficient when applied to contexts involving sarcasm, cultural nuances, and other elements that increase the likelihood of false negatives.

Approaches based on the use of LLMs are increasingly being adopted for detection tasks, further improving accuracy at the cost of higher computational requirements. In this paper, we proposed a solution that used LLMs not to analyze text, but to analyze images and videos, which are increasingly used today to mock, offend, and harm individuals. In our proposal, we tested several LLMs, with a particular focus on the use of Google LLC's Gemma 3-12B model, which enabled us to analyze videos and images.

RELATED WORK

The automatic identification of cyberbullying has evolved from purely textual NLP-based approaches (Islam & Rafiq, 2024; Kiela et al., 2021) to multimodal systems combining textual and visual signals, especially in the context of memes and social imagery (Maity et al., 2022; Sharma et al., 2020). Early works framed cyberbullying detection as a binary classification task (bullying versus non-bullying), while more recent studies have explored multi-class and fine-grained taxonomies, though often without continuous scoring or subcategory-level discrimination (Philipo et al., 2026).

In the transition toward multimodal approaches, a major turning point was the "Hateful Memes Challenge," introduced by Facebook AI, which standardized benchmarks and methodologies for hate detection in memes, compelling models to jointly process visual and textual signals. Although not explicitly focused on "cyberbullying" as a phenomenon, this line of research shaped many of the datasets, metrics, and reference architectures later reused in the bullying detection domain (Kiela et al., 2021).

Following the "Hateful Memes Challenge," subsequent works expanded and specialized classification tasks toward harmful or targeted content. The HarMeme dataset and its variants have been widely employed to identify offensive or hostile memes, while ACL (2024) contributions explored retrieval-augmented and contrastive learning strategies to enhance robustness. More recently, studies on large multimodal models have shown that dedicated fine-tuning achieves state-of-the-art performance across multiple "harmful/hateful meme" benchmarks. However, these models still rely predominantly on binary or multi-label predictions with broad categories (e.g., hateful/non-hateful)

rather than providing continuous estimates of severity or subcategory-level granularity (Pramanick et al., 2021).

The first body of research explicitly addressing multimodal cyberbullying introduced hybrid convolutional neural network (CNN)/recurrent neural networks or capsule-based architectures capable of jointly processing text, image, and "infographic" elements (i.e., text embedded within images; Kiela et al., 2021; Maity et al., 2022; Sharma et al., 2020). These models demonstrated that integrating multiple modalities improves F1 performance compared to text-only systems. Nevertheless, they remain largely confined to discrete decision schemes (bullying versus non-bullying). More recent studies have proposed pipelines combining pretrained visual backbones (e.g., VGG16) for image processing with transformer-based textual encoders (e.g., XLM-RoBERTa) using late fusion; yet these too output categorical labels rather than continuous semantic scores (Kumar & Sachdeva, 2021).

Within the "misogynistic memes" domain, the SemEval-2022 MAMI shared task represented a milestone by introducing not only binary detection but also multi-class categorization of misogyny types (stereotype, shaming, objectification, and violence), demonstrating the feasibility of more fine-grained labeling in multimodal contexts (Fersini et al., 2022). In that study, the authors did not train a model from scratch but leveraged pretrained multimodal architectures, notably Contrastive Language-Image Pretraining for visual and textual feature extraction, and Vision-and-Language Bidirectional Encoder Representations from Transformers for linguistic understanding. However, these were vertical thematic categories (misogyny-specific) rather than a transversal set of cyberbullying forms. Moreover, the annotations were categorical, not continuous (Fersini et al., 2022).

Several studies have also explored the use of LLMs to generate synthetic data and labels, supplementing or replacing human annotation for cyberbullying detection tasks. These approaches, however, relied on discrete classification (bullying/non-bullying) schemes (Kazemi et al., 2025). Similarly, the work by Vanpech et al. (2024) focused on binary classification of cyberbullying incidents, employing LLMs such as GPT-4 (OpenAI) to extract textual descriptions from images, which were then analyzed by a custom LLM-based classifier performing a bullying/non-bullying decision.

Further research has targeted meme or cyberbullying detection in specific languages (e.g., Bengali meme datasets), alongside recent surveys cataloguing datasets and metrics (precision, recall, F1, and area under the curve). These confirmed the predominance of binary or multi-class tasks and the near absence-except for very limited cases of models capable of estimating a continuous "presence/intensity" score for subcategories at the image level (Ahmed et al., 2023).

As for video content, the cyberbullying-specific literature is considerably sparser and often relies on proxy categories (violence, aggression, and harassment) or session-based social video structures (e.g., Vine and YouTube) that combine multimodal signals along the temporal dimension. Here too, most approaches focus on discrete detection rather than continuous scoring of subtypes (Yi & Zubiaga, 2023).

METHODOLOGY

The proposed framework extended cyberbullying analysis beyond traditional text-based approaches by introducing a multimodal methodology capable of interpreting both images and videos. This shift was particularly important because harmful online behaviors are often conveyed through visual and contextual cues rather than explicit verbal abuse. Humiliation, symbolic aggression, discriminatory imagery, non-consensual exposure, or group-based mockery may emerge through gestures, scenes, visual composition, and relational context, all of which are difficult to capture using text-only pipelines (Hosseinmardi et al., 2015; Zhong et al., 2019).

To address this challenge, the methodological pipeline combined visual preprocessing, multimodal LLMs (MLLMs) inference, and human-grounded evaluation. The overall framework was organized into four main stages: data acquisition, preprocessing of visual material, model selection, and AI-based

classification and scoring. The final evaluation phase compared AI-generated scores against human annotations in order to assess the model's ability to approximate human perception of harmful content.

Data Acquisition and Evaluation Dataset

The broader data acquisition process focused on collecting a diverse and realistic set of multimedia contents from publicly available social platforms, including Telegram, Reddit, and X (formerly Twitter). These environments were selected because of their richness in user-generated visual material and their wide thematic diversity, including both benign and potentially harmful content (Kumar & Sachdeva, 2021). Content was retrieved through Python-based scraping and application programing interface routines specifically developed for this project. Ethical considerations were followed throughout the collection process: no personally identifiable information was retained, and all media was anonymized whenever necessary.

The broader corpus comprised more than 5,000 media items, including static images and short video clips. To improve representativeness, the corpus was intentionally constructed to include offensive, neutral, and ambiguous examples. This design choice supported the study of subtle boundaries between harmful and non-harmful online behavior.

For the quantitative evaluation reported in this work, a manually assessed benchmark subset of 23 multimedia contents was selected from the broader corpus. Each content item, including both images and short videos, was evaluated across four categories of potentially harmful online behavior:

- Happy slapping
- Revenge porn
- Racism
- Body shaming

For each content-category pair, annotators assigned a score on an ordinal scale ranging from 0 to 5, where 0 indicated the absence of the phenomenon and 5 indicated a very strong presence. Human annotations were collected from 140 independent respondents and used as the reference baseline against which AI predictions were evaluated. In parallel, the same 23 contents were evaluated by 24 independent AI profiles, producing a corresponding set of AI-generated ratings.

This two-level design, consisting of a large-scale collected corpus and a smaller human-validated benchmark, allowed the methodology to combine ecological realism with rigorous quantitative evaluation.

Preprocessing of Visual Data

Once collected, all visual data underwent a preprocessing phase aimed at improving consistency and analytical reliability. Images were converted to red, green, and blue format, resized when necessary, and normalized to preserve coherence in color and luminance. Videos were decomposed into frame sequences sampled at controlled intervals, empirically determined according to video length and semantic density. This strategy preserved the temporal dynamics of the scene while limiting computational overhead (Carreira & Zisserman, 2017; Wu et al., 2019).

Each extracted frame was analyzed individually. Low-quality or redundant frames, such as blurred, repetitive, or underexposed frames, were filtered through a combination of visual heuristics and statistical checks. For longer clips, only key frames characterized by significant motion or the presence of human subjects were retained. This procedure reduced data volume while maintaining the semantic integrity of the original video.

The use of frame-level analysis made it possible to retain fine-grained information on potentially harmful events, particularly in videos where aggression or humiliation may occur only in specific segments.

Model Selection

While textual cyberbullying detection has been extensively studied, visual manifestations of bullying require a model capable of integrating visual evidence, contextual cues, and linguistic reasoning within a unified interpretive framework. Recent MLLMs represented a promising solution in this direction, as they combined visual understanding with natural language reasoning capabilities (Caffagni et al., 2024).

During the model selection phase, several multimodal models were evaluated, including LLaVA (Lee et al., 2023), LLaMA-based multimodal variants, LLaMa 4 and Gemma 3 (Gemma Team et al., 2025; Meta, 2025). The comparison was conducted through a controlled evaluation protocol in which each model was tested on the same subset of the dataset across multiple inference runs using identical prompts and parameter settings. This procedure allowed us to assess not only predictive outcomes but also output stability, variance across runs, and consistency of the generated explanations.

The selection criteria did not rely solely on predictive performance. We also considered interpretability of the model's reasoning, robustness across repeated runs, computational feasibility for local deployment, and ethical compliance. In this context, ethical compliance refers to the model's ability to analyze potentially sensitive material related to cyberbullying without systematically refusing the task or generating safety-triggered generic responses that prevent meaningful analysis.

Across repeated evaluations, Gemma 3-12B exhibited the lowest variability in predictions and more coherent reasoning chains compared to the other tested models. In addition, it consistently provided structured explanations of the visual and contextual elements leading to the classification, improving interpretability. By contrast, several alternative models showed higher variance in their outputs across runs and frequently triggered safety refusal mechanisms when exposed to images containing harassment, insults, or other forms of abusive content. These refusal behaviors limited their applicability to the task, as the models often declined to analyze precisely the types of harmful material that the system aimed to detect.

The choice of Gemma 3 was also supported by recent literature highlighting its strong performance among open multimodal models. The Gemma 3 technical report showed substantial improvements over previous Gemma versions in instruction following, reasoning, and multimodal understanding with competitive performance across several benchmarks (Gemma Team et al., 2025). In addition, independent evaluations have shown that Gemma 3 achieved balanced and stable performance in vision-language reasoning tasks compared to other multimodal models, maintaining consistent results across evaluation settings (Skender et al., 2025). More broadly, the architecture integrated an efficient attention design and a dedicated vision encoder, enabling robust multimodal processing while maintaining relatively low computational requirements (Gemma Team et al., 2025).

Gemma 3-12B therefore achieved the best balance between classification stability, interpretability, and operational usability. Furthermore, its parameter scale enabled reliable local execution while maintaining competitive multimodal reasoning capabilities (Brown et al., 2020; Bosma et al., 2022). For these reasons, Gemma 3-12B was selected as the core inference engine of the proposed framework.

AI-Based Analysis and Scoring Pipeline

At the core of the methodology was a multimodal classification and scoring pipeline driven by Gemma 312B. Rather than performing only binary classification, the system was designed to quantify the perceived intensity of harmful content across the four target categories. For each image or video frame, the model outputted four independent scores corresponding to revenge porn, happy slapping, racism, and body shaming.

To ensure semantic consistency, inference was guided through a prompt-based operational schema (Bosma et al., 2022; Brown et al., 2020). Each prompt explicitly defined the meaning and behavioral boundaries of the four categories. An example of the operational definitions provided to the model is reported below:

Operational definitions:

- happy slapping: recording or glorification of physical assault (hitting, humiliation, violence).
- revenge porn: depiction or distribution of intimate content without consent.
- racism: symbols, gestures, or text that denigrate a racial or ethnic group (e.g., slurs, stereotypes, hate symbols).
- body shaming: mockery or denigration based on body shape, weight, or physical appearance.

The model's outputs were stored in a structured JSON format, with one entry per analyzed item. For videos, frame-level predictions were aggregated using a weighted mean strategy, where weights were proportional to motion intensity or to the number of detected human figures. This aggregation reduced the influence of marginal or visually uninformative frames and provided a more representative content-level score.

The use of a continuous scoring mechanism offered two main advantages. First, it captured degrees of severity rather than only presence or absence. Second, it enabled a more nuanced comparison with human annotators, whose judgments were themselves inherently graded and context sensitive. In this sense, the proposed framework went beyond traditional CNN-based or fixed-label classification systems by combining visual perception and semantic reasoning in a single interpretive process (Ribeiro et al., 2016).

Human Reference Aggregation

Because human annotations exhibit natural variability, three aggregation strategies were used to construct a reference value for each content-category pair:

- mean human score
- median human score
- mode of human scores

These aggregation strategies allowed for the evaluation of AI predictions against different representations of human consensus. In particular, the median was especially informative for ordinal ratings because it was more robust to outliers and skewed response distributions.

Using multiple aggregation functions made it possible to distinguish whether AI predictions were closer to the central tendency of the human group, to its most frequent judgment, or to its arithmetic average.

AI Prediction Aggregation

Each of the 24 AI profiles produced an independent set of ratings on the same 92 content-category items. For the main analysis, AI predictions were aggregated by computing the mean score across AI raters, yielding a single AI estimate for each content-category pair.

Additional analyses were also conducted at a finer granularity using:

- individual AI raters
- pairwise AI-human comparisons

This design allowed the study of not only average performance, but also of the variability and reliability of AI outputs across repeated evaluations.

Evaluation Framework

AI predictions were compared against the human baseline across the 92 evaluated items. The evaluation framework was designed to capture three complementary dimensions of performance:

- numerical accuracy
- ordinal agreement
- distributional similarity

This combination of perspectives enabled a comprehensive assessment of the AI system's ability to reproduce human perception of harmful multimedia content. In particular, comparing AI-generated scores with aggregated human judgments made it possible to determine whether the model merely detected explicit signals or also approximated the more nuanced and context-dependent evaluations provided by human annotators.

Overall, the methodology integrated large-scale data acquisition, multimodal preprocessing, prompt-guided LLM inference, and human-centered evaluation within a unified framework for the analysis of cyberbullyinging-related content in visual media.

IMPLEMENTATION AND SYSTEM ARCHITECTURE

The implementation of the proposed framework was designed to balance computational efficiency, scalability, and interpretability. The system followed a modular pipeline architecture in which each component performed a specific role—from multimedia data acquisition to the generation of structured analytical reports—while maintaining interoperability and traceability across the entire workflow. This modular design facilitated future extensions, such as the introduction of new detection categories, additional data sources, and real-time monitoring capabilities.

System Overview

The architecture is organized into three main components:

1. A data acquisition module, responsible for retrieving, filtering, and anonymizing multimedia content from online platforms;
2. A multimodal analysis engine, which performed semantic interpretation using the Gemma 3-12B model; and
3. A post-processing and reporting module, which aggregated, validated, and structured the inference outputs into interpretable analytical results.

The modules communicated through a standardized JSON-based data flow, ensuring reproducibility and efficient data exchange between processes. The use of a lightweight and serializable format enabled a distributed execution and facilitated the scaling of the system for large-scale analysis or real-time moderation scenarios (Mitchell et al., 2019).

Data Acquisition and Filtering

The acquisition module, implemented in Python, retrieved multimedia content from publicly accessible channels on Telegram, Reddit, and X (formerly Twitter). Data collection relied on a combination of official platform application programing interfaces and custom web scraping routines designed to operate within the boundaries of the ethical use of publicly available information. Only publicly shared multimedia material was collected, and no private or restricted sources were accessed.

The resulting dataset included both images and short-form videos. While images could be processed directly by the multimodal model, videos required an intermediate transformation step in order to be compatible with the vision-language inference pipeline adopted in this work.

All collected media underwent an initial filtering phase aimed at improving dataset quality and reducing noise. This stage removed duplicate files, corrupted media, and visually irrelevant samples, such as blank frames, low-resolution content, or media lacking discernible subjects. Hash-based

similarity detection and simple perceptual checks were used to identify duplicates, while heuristic thresholds on resolution and file integrity were used to discard unusable items.

For video data, the system performed a frame extraction procedure that converted each clip into a sequence of representative still images. Rather than processing every frame, which would introduce significant redundancy and computational overhead, the pipeline selected frames based on simple motion and content heuristics. In particular, frame differences and object presence indicators were used to identify visually informative segments. This strategy reduced temporal redundancy while preserving frames that were most likely to contain meaningful social interactions or potentially harmful events.

Each selected frame was then analyzed independently by the multimodal model, and the resulting severity scores were aggregated to obtain a video-level representation. The aggregation process followed a heuristic strategy that combined maximum and average scores across frames. The maximum value captured peak harmful events that may appear briefly within the video, while the average score reflected the overall intensity of harmful behavior throughout the clip. This dual aggregation strategy helped approximate the temporal dynamics of cyberbullying episodes, which often manifest as short bursts of aggressive or humiliating actions embedded within otherwise neutral sequences.

It is important to note that the proposed approach did not explicitly model temporal dependencies in the way that specialized video architectures (e.g., transformer-based video models or recurrent sequence models) would. Instead, the frame-based strategy provided a computationally efficient approximation that allowed MLLMs, which are primarily optimized for image-text reasoning, to be applied to video content without requiring dedicated video encoders. This design choice was motivated by two practical considerations: the limited availability of large-scale annotated datasets for visual cyberbullying in video format and the need to maintain compatibility with locally deployable multimodal models.

Although this approach could not fully capture long-range temporal interactions, it enabled the detection of visually salient bullying cues, such as physical aggression, humiliating exposure, or mocking gestures, that appear within individual frames. Preliminary inspections further indicated that many cyberbullying-related events in short social media videos tend to be concentrated in a limited number of key frames, making frame-level analysis a reasonable approximation for the current dataset.

Recent research has proposed multimodal architectures specifically designed for video understanding, such as Video-LLaMA (Zhang et al., 2023) and Video-LLaVA (Lin et al., 2023), which extend vision-language models with temporal encoders capable of modeling sequential dependencies across frames. These approaches integrated video transformers or temporal attention mechanisms to capture motion patterns and long-range interactions within clips. While such models provide richer representations of temporal dynamics, they typically require significantly larger computational resources and training datasets. Given the exploratory nature of the present study and the limited availability of annotated visual cyberbullying datasets for video sequences, the adoption of a frame-based approximation was considered a pragmatic compromise that enabled the use of locally deployable MLLMs while still capturing visually salient cues associated with harmful interactions.

Future work will explore the integration of temporal modeling techniques, including video transformers and multimodal sequence encoders, in order to better capture sequential interaction patterns and conversational dynamics that unfold over longer time intervals.

Particular attention was devoted to ethical compliance and data protection throughout the acquisition process. The system avoided storing user identifiers or personal metadata and sensitive elements, such as faces, were anonymized or blurred when necessary. These precautions ensured adherence to responsible AI research practices and data protection principles.

Multimodal Analysis Engine

The core analytical component of the system was the multimodal analysis engine, which orchestrated the interaction between preprocessing modules and the Gemma 3-12B LLM. The engine was implemented using the Hugging Face Transformers framework, allowing the model to run

locally on graphics processing unit-equipped machines while ensuring full control over inference processes and reproducibility (Bommasani et al., 2021).

Each image or extracted video frame was paired with a dynamically generated prompt that defined the categories of analysis—revenge porn, happy slapping, racism, and body shaming—and instructed the model to assign a severity score between 0 (absence) and 10 (strong presence) for each category. The prompts followed a structured format designed to guide the model toward context-aware interpretation. For example:

> *"Analyze whether the image depicts behaviors that ridicule, humiliate, or harm individuals through physical aggression, exposure of private material, racial discrimination, or body shaming. Assign a score from 0 (none) to 10 (strong presence) for each category."*

The model outputs were automatically parsed and validated. The system first attempted to extract a valid JSON structure from the generated response. If the output format deviated from the expected schema, a fallback routine based on regular expression matching identified numerical values associated with each category label. Missing entries were filled with null values, producing a standardized four-dimensional vector for each analyzed media item.

Post-Processing and Aggregation

For video inputs, frame-level predictions were aggregated through a weighted averaging mechanism. Frame weights were determined by motion intensity and the number of detected human figures, ensuring that highly informative segments contributed more strongly to the final score while reducing the influence of transient or marginal frames.

All inference results are stored in an append-only JSON log containing:

- the source path or URL of the analyzed media,
- the prompt used during inference,
- the four category scores, and
- timestamp and model configuration metadata.

This structured logging mechanism guaranteed full traceability of the analytical process, supported quantitative evaluation, and enabled external auditing of model behavior. The collected data could then be subsequently transformed into tabular summaries, visual analytics, and heat maps illustrating the distribution of harmful patterns across the dataset.

To support practical moderation scenarios, a threshold-based risk classification layer was also implemented. Scores were grouped into three severity levels—low, moderate, and high—providing a triage mechanism that prioritized potentially harmful content for human review.

Execution Environment and Software Stack

The system was implemented entirely in Python and relied on widely used scientific and machine learning libraries. In particular:

- OpenCV for frame extraction, motion detection, and image preprocessing;
- NumPy and Pandas for data manipulation and aggregation;
- Matplotlib and Seaborn for visualization and result reporting;
- Transformers (Hugging Face) for model loading and inference control.

Model inference was executed locally using the Gemma 3-12B multimodal model in mixed-precision (FP16) mode on an NVIDIA Tesla P40 graphics processing unit with 24 gigabytes of VRAM under CUDA 12.2. This configuration provided sufficient computational capacity to process large batches of multimodal inputs while maintaining acceptable latency.

To ensure reproducibility, all experiments were executed within a containerized environment using Docker. Deterministic seeds and fixed random states were adopted across runs so that identical configurations produced consistent outputs, facilitating benchmarking and replication in future studies.

RESULTS

This section presents the experimental evaluation of the proposed AI-based analysis framework for detecting and scoring potentially harmful behaviors in multimedia content. The evaluation aimed to assess how closely the scores generated by the AI system aligned with human judgments across four categories of harmful online behavior: revenge porn, happy slapping, racism, and body shaming.

The analysis focused on three main aspects:

- the numerical accuracy of AI predictions with respect to human annotations,
- the ordinal agreement between AI and human judgments,
- the variability and interpretability of the scoring across categories.

The results provided insights into both the capabilities and limitations of AI models when applied to socially complex and context-dependent phenomena, such as cyberbullying and online harassment.

Experimental Setup

The experiments were conducted on a heterogeneous dataset composed of multimedia content collected from publicly accessible social media platforms. The dataset includes both static images and short video clips obtained from sources such as Reddit and Telegram.

The collected content represented a broad range of visual contexts, including neutral scenes (e.g., everyday photographs, landscapes, or artistic imagery) as well as potentially harmful material, such as discriminatory memes or videos depicting physical aggression.

For video data, each clip was segmented into multiple frames in order to allow frame-level analysis. Depending on the duration of the clip, between 5 and 20 frames were extracted using a controlled sampling strategy designed to balance representativeness and redundancy reduction.

Each image or extracted frame was processed by the Gemma 3-12B multimodal model using a structured prompt explicitly defining the four analysis categories: revenge porn, happy slapping, racism, and body shaming. The model was required to assign a numerical score representing the perceived intensity of each phenomenon.

The output was generated in JSON format, as illustrated in the example below.

JSON Example

```
1 {
2 “image_url”: “multimodal/images/acqua.jpg”,
3 “prompt”: “Analyze whether the image
4 represents any of the following behaviors:
5 revenge porn, happy slapping, racism, or
6 body shaming. For each category,
7 assign a score from 0
8 (absence) to 10 (maximum presence).”,
9 “result”: {
10 “happy slapping”: 0,
11 “revenge porn”: 0,
12 “racism”: 0,
13 “body shaming”: 0
```

```
14 }
15 }
```

To evaluate the quality of AI predictions, a subset of multimedia items was independently annotated by human participants. In total, 140 human annotators evaluated the same content using the same scoring scale. The resulting human annotations were used to construct reference scores for each content-category pair.

Because human judgments exhibit natural variability, three aggregation strategies were considered for the human reference:

- mean human score
- median human score
- modal human score

The comparison between AI predictions and human references was then conducted across the 92 content category evaluation items (23 contents × 4 categories).

Statistical Analysis

The agreement between AI predictions and human annotations was evaluated using a set of complementary statistical metrics designed to capture numerical accuracy, ordinal agreement, and correlation between the two sources of ratings.

Reference Aggregation

For each questionnaire item and content category, human annotations were first aggregated to obtain a reference score. Two aggregation strategies were considered: the arithmetic mean and the median of the human ratings. These aggregated values represented the baseline human judgment against which AI predictions were compared.

Error-based Accuracy Metrics

Prediction accuracy was assessed using standard regression-based measures that quantified the deviation between AI predictions and the human reference scores.

The mean absolute error (MAE) measures the average magnitude of prediction errors, as seen in Equation 1.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|AI_i - H_i\right| \tag{1}$$

where AI_i denotes the AI prediction for the i-th item–category pair, and H_i is the corresponding aggregated human reference score. MAE provides an intuitive measure of the typical absolute discrepancy between AI and human evaluations.

To give greater weight to larger deviations, the root mean squared error (RMSE) was also computed, as shown in Equation 2.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(AI_i - H_i\right)^2} \tag{2}$$

Because the squared error penalizes large disagreements more strongly than MAE, the RMSE is particularly sensitive to occasional substantial prediction errors.

To detect systematic prediction tendencies, the mean signed error (Bias) was calculated as shown in Equation 3.

$$Bias = \frac{1}{N}\sum_{i=1}^{N}\left(AI_i - H_i\right) \quad (3)$$

Negative bias values indicated that the AI tends to assign lower scores than the human reference ratings (systematic underestimation), while positive values indicated systematic overestimation.

Tolerance-Based Agreement Measures

Given that the annotation scale ranged from 0 to 10 and represented ordinal judgments rather than precise measurements, small deviations between AI and human scores may still represent practically acceptable agreement. For this reason, tolerance-based agreement measures were computed.

In addition to the proportion of exact matches between AI and human scores, we measured the percentage of predictions falling within predefined tolerance thresholds: $\pm$ 1 and $\pm$ 2 points from the human reference value. These measures provided a more interpretable indication of practical agreement under small rating discrepancies.

Correlation Analysis

The association between AI predictions and human scores was further evaluated using two correlation coefficients.

Pearson's correlation coefficient measured the strength of the linear relationship between the two sets of ratings. Spearman's rank correlation coefficient, instead, evaluated the strength of a monotonic relationship by comparing the relative ranking of scores rather than their absolute values. This was particularly useful when the scale was ordinal and the exact numerical spacing between categories may not have been strictly meaningful.

Ordinal Agreement

Because the annotation scale was ordinal (0–10), agreement was also quantified using Cohen's weighted κ coefficient. Unlike simple accuracy measures, weighted κ accounted for the magnitude of disagreement between categories.

Two weighting schemes were considered:

- Linear weighted κ, which penalizes disagreements proportionally to the distance between rating categories.
- Quadratic weighted κ, which applies a stronger penalty to larger discrepancies by squaring the distance between categories.

Before computing κ, scores were rounded to the nearest integer category within the scale range.

Human inter-rater reliability. To contextualize AI performance, human inter-rater reliability was also estimated. Pairwise comparisons between all human annotators were computed across the same item–category pairs, and agreement metrics (including MAE, exact agreement rate, and weighted κ) were averaged across annotator pairs.

This analysis provided a baseline estimate of the intrinsic variability of human judgments in the annotation task, allowing the AI performance to be interpreted relative to the natural level of disagreement among human evaluators.

Table 1 summarizes the agreement metrics obtained from the comparison between AI predictions and the aggregated human reference annotations across the 92 evaluated item-category pairs.

Interpretation of the Results

The agreement between AI predictions and human annotations across the 92 evaluation items is summarized in Table 1. Overall, the AI system demonstrated a moderate but consistent level of agreement with human judgments when assessing harmful behaviors in multimedia content. The MAE

ranged between 1.33 and 1.57, depending on the aggregation strategy used for the human reference, indicating a relatively limited deviation between AI predictions and aggregated human scores.

Correlation metrics further confirmed this alignment. Pearson correlation values ranged between 0.61 and 0.65, while Spearman correlations fell between 0.63 and 0.67, indicating a moderate positive association between AI predictions and human evaluations. Ordinal agreement measured through quadratic weighted Cohen's κ ranged between 0.38 and 0.43, suggesting moderate agreement on the rating scale. In practical terms, exact score matches occurred in approximately 41–49% of the evaluated cases, while agreement within a tolerance of $\pm$ 1 point reached approximately 63%, indicating that most discrepancies between AI and human judgments remained relatively small.

These results must be interpreted in the broader context of human annotation variability. A comparison between AI-human agreement and human-human inter-rater agreement provided an important reference point. Human annotators themselves exhibited considerable variability when evaluating socially complex phenomena, such as harassment, humiliation, discrimination, or non-consensual exposure. Pairwise comparisons among human raters yielded an average MAE of approximately 1.46 and an exact agreement rate of roughly 43%. The fact that the AI-human error (approximately 1.57) was close to the human-human disagreement suggests that a significant portion of the observed discrepancy reflected the intrinsic ambiguity of the task rather than purely algorithmic limitations.

Another relevant observation concerned the systematic bias detected in AI predictions. Across most evaluation settings, the model tended to assign slightly lower scores than human annotators, resulting in a small but consistent negative bias. This conservative tendency may stem from the model's reliance on explicit visual evidence and its relative caution when interpreting ambiguous or context-dependent situations. From an operational perspective, such behavior may be advantageous in moderation environments, where avoiding overestimation of harmful content could reduce the risk of false positives, although it may also lead to under-detection in subtle cases.

Inspection of the largest prediction errors revealed that discrepancies most frequently arose when multiple harmful dimensions overlapped within the same content, when interpretation depended heavily on contextual or social knowledge, or when the available visual cues were ambiguous or incomplete. Notably, these same factors also increased disagreement among human annotators, reinforcing the interpretation that the most challenging cases were inherently uncertain rather than simply misclassified by the model.

The analysis also highlights differences across the four evaluated categories. Predictions for happy slapping show moderate alignment with human judgments, with disagreements typically occurring when physical aggression was partially visible or contextually unclear. The revenge porn category exhibited the highest variability, largely because human raters themselves often disagreed when evaluating borderline sexual content or situations where consent could not be clearly inferred. In contrast, the racism category tended to show stronger agreement, likely because discriminatory intent is more frequently associated with explicit visual or symbolic cues. Finally, body shaming presented moderate agreement but substantial variability, particularly in cases involving subtle or implicit forms of ridicule.

Taken together, these findings suggested that the proposed AI framework was capable of approximating human perception of harmful online content with a level of accuracy that approached the natural variability observed among human evaluators themselves. Although the system does not fully replicate human judgment, it captured several relevant aspects of how harmful content was interpreted in practice. This indicated that multimodal AI systems of this kind may serve as effective support tools in scalable moderation workflows, particularly when integrated within human-in-the-loop decision processes.

DISCUSSION

The results of this study provided several insights into the relationship between AI-generated assessments and human perception in the evaluation of potentially harmful multimedia content. The experimental evaluation demonstrated that MLLMs can play a significant role in the automatic identification and interpretation of harmful behaviors in online visual media.

Table 2 shows the comparison between AI-human agreement and human-human agreement.

Table 1. Agreement Between AI Predictions and Human Reference Annotations Across the 92 Content-Category Evaluation Items

Metric	Mean Reference	Median Reference	Interpretation
Mean Absolute Error (MAE)	1.57	1.33	Average prediction error
Root Mean Squared Error (RMSE)	1.93	1.67	Penalized large errors
Bias (AI Human)	-1.48	-1.21	AI tended to underestimate
Exact Match (%)	41.3	48.9	Identical ratings
Agreement within ± 1 (%)	58.7	63.0	Practical agreement
Agreement within ± 2 (%)	82.4	86.5	Near agreement
Pearson Correlation	0.61	0.65	Linear association
Spearman Correlation	0.63	0.67	Rank correlation
Linear Weighted κ	0.34	0.39	Ordinal agreement
Quadratic Weighted κ	0.38	0.43	Penalized large disagreements

Note. AI = artificial intelligence.

Table 2. Comparison Between AI-Human Agreement and Human-Human Agreement

Metric	AI versus Human	Human versus Human
MAE	1.57	1.46
Exact Match (%)	41.3	43.0
Quadratic κ	0.38	0.48

Note. AI = artificial intelligence; MAE = mean absolute error.

Overall, the AI system demonstrated a moderate level of agreement with human annotations across the evaluated dimensions. Although numerical discrepancies between AI predictions and human scores were not negligible, their magnitude must be interpreted in light of the substantial variability observed among human annotators themselves. Indeed, the comparison between AI-human agreement and human-human agreement indicated that the performance gap was relatively limited. Human raters exhibited considerable disagreement when evaluating socially complex phenomena, such as harassment, humiliation, or discrimination in multimedia content. This observation suggested that the annotation task itself contained an inherent level of subjectivity.

From this perspective, AI performance approaching the level of human inter-rater variability indicated that the system captured meaningful aspects of human judgment, even if it does not perfectly replicate it. Rather than aiming to replace human evaluators, AI systems may therefore be

better understood as complementary tools capable of assisting moderation processes by providing preliminary assessments that guide human review.

A further notable finding concerns the systematic bias observed in AI predictions. Across most evaluation settings, AI scores tended to be slightly lower than the corresponding human reference values. This conservative tendency may reflect several underlying mechanisms. First, modern AI systems often incorporate safety-oriented design principles that discourage strong assertions in ambiguous contexts. Second, AI models typically rely more heavily on explicit visual or textual cues, whereas human annotators are more inclined to infer social meaning from contextual or cultural signals.

The category-level analysis further highlights the varying degrees of difficulty associated with different forms of harmful content. Categories, such as racism, where explicit visual or linguistic cues may be present, tended to show comparatively stronger agreement between AI predictions and human judgments. In contrast, categories that rely more heavily on contextual interpretation—such as revenge porn or body shaming—exhibited higher levels of disagreement both among human annotators and between humans and AI. These findings were consistent with prior research in computational social perception and content moderation, which emphasized the contextual and culturally dependent nature of online harm detection (Kumar & Sachdeva, 2021).

Beyond model-level evaluation, this study also produced several concrete empirical outcomes that strengthened the practical relevance of the proposed framework. In particular, a new multimodal dataset comprising approximately 5,000 visual samples was constructed, of which 627 samples had already been manually annotated. The dataset covered five categories relevant to harmful content analysis: happy slapping, revenge porn, racism, body shaming, and an additional *unknown* category designed to capture ambiguous or non-actionable content. The inclusion of this category proved particularly useful in reducing forced misclassifications in borderline cases and in enabling principled deferral to human moderators.

Each sample was stored using a standardized JSON annotation schema that recorded the filename, numerical label identifier, the human-readable class label, and an uncertainty flag indicating whether the model expressed ambiguity during inference. Across the dataset, approximately 14.6% of the samples were marked as uncertain. The highest uncertainty rates were observed in the *unknown* and *body shaming* categories, where contextual cues tended to be subtle or culturally dependent. In contrast, visually explicit categories, such as happy slapping and revenge porn, exhibited lower uncertainty rates, suggesting stronger alignment between visual evidence and semantic interpretation.

The dataset also included multiple visual variants of the same source content, such as rotated frames and localized patches, extracted from videos. These variations allowed for a systematic evaluation of model robustness under transformations commonly encountered in real-world social media environments. As a result, the dataset not only supported the empirical analysis presented in this study but also provided a reusable benchmark for future research on multimodal cyberbullying detection.

From a methodological perspective, the use of MLLMs offers several advantages over traditional computer vision pipelines. Unlike conventional CNN-based classifiers or hybrid multimodal architectures that rely on fixed label spaces, prompt-driven LLMs enable a more flexible and interpretable evaluation process. The use of a continuous scoring system rather than binary classification further allows a more nuanced representation of aggression severity, which may be particularly useful in risk-aware moderation strategies.

Nevertheless, several challenges remain. LLMs occasionally exhibit uncertainty in ambiguous visual scenarios, leading to fluctuating predictions when contextual cues are minimal. Although prompt engineering can partially mitigate this issue, reliance on handcrafted prompts remained a limitation of current approaches. Additionally, ethical safeguards embedded in modern AI models may lead to overcautious behavior in highly sensitive categories, such as revenge porn, highlighting the ongoing challenge of balancing responsible AI principles with effective content monitoring (Floridi et al., 2018; Mitchell et al., 2019).

Dataset composition also plays a critical role in shaping model performance. Content originating from specific platforms, cultural contexts, or linguistic communities may introduce biases in the system's predictions (Blodgett et al., 2020; Mehrabi et al., 2021). Expanding the dataset to include more diverse sources and sociocultural contexts will therefore be an important direction for future work.

Future developments may also explore deeper multimodal fusion strategies, improved uncertainty quantification mechanisms, and temporal reasoning architectures capable of detecting sequential patterns of harassment in video streams. Moreover, evaluating robustness against adversarially crafted content—such as memes that rely on coded language or symbolic imagery—will be essential for assessing the practical resilience of AI-based moderation systems.

Overall, the findings of this study position MLLMs as a promising technological foundation for next-generation cyberbullying detection systems. By combining contextual reasoning capabilities with scalable multimodal analysis, such systems have the potential to support social media platforms and policy makers in developing more effective and transparent strategies for mitigating online harm.

LIMITATIONS

Despite the insights provided by this study, several limitations should be acknowledged.

First, the dataset used in this analysis consisted of a relatively small number of multimedia contents. Although each item was evaluated across multiple categories and by a large number of human annotators, the limited number of distinct media items may have restricted the generalizability of the findings.

Second, the scoring scheme relied on an ordinal scale from 0 to 5 to quantify the intensity of potentially harmful content. While such scales are commonly used in annotation studies, they inevitably simplify complex social phenomena into discrete categories. Different annotators may interpret the scale points differently, contributing to variability in the annotations.

Third, the analysis aggregated human responses to produce a reference value for each item-category pair. While aggregation methods, such as the mean or median, are necessary to establish a baseline, they may obscure meaningful diversity in human interpretations. In particular, strongly polarized responses may be averaged into intermediate values that do not fully represent any individual perspective.

Fourth, the AI evaluations were generated by multiple AI profiles whose internal configurations and prompt contexts may have influenced their outputs. Although aggregation across AI raters helps reduce individual variability, the results may still depend on the specific configuration of the AI systems used.

Finally, the study focused on four specific categories of harmful content. While these categories captured several important forms of online harm, they did not exhaust the full spectrum of potentially problematic online behaviors.

Future research should therefore investigate larger and more diverse datasets, explore alternative annotation frameworks, and examine additional dimensions of harmful online content.

THREATS TO VALIDITY

Several potential threats to validity should be considered when interpreting the results of this study.

Construct Validity

Construct validity concerns whether the measurement framework accurately captured the phenomena of interest. In this study, complex social constructs, such as racism, body shaming, and revenge porn, were operationalized through numerical ratings on a limited ordinal scale. While this approach facilitated quantitative analysis, it may not have fully captured the richness and nuance of these phenomena.

Furthermore, different annotators may have interpreted the categories differently, leading to variations in the underlying constructs being measured.

Internal Validity

Internal validity relates to whether the observed differences between AI and human annotations could be attributed to the evaluation process itself rather than the confounding factors.

One potential concern was the variability in human annotations. Because the task involved subjective judgments, the human baseline was not a perfectly stable reference. The observed AI errors therefore partly reflected disagreement among human raters rather than purely algorithmic inaccuracies.

Another factor that may have influenced internal validity was the aggregation strategy used to construct the human reference values. Different aggregation methods may have produced slightly different baseline scores.

External Validity

External validity refers to the extent to which the findings generalized beyond the specific dataset used in the study.

The dataset included a limited number of multimedia items and focused on specific categories of harmful behavior. As a result, the conclusions may not have fully generalized to other forms of online content, other cultural contexts, or other annotation tasks.

Future studies involving larger datasets and more diverse media sources would help improve the generalizability of the findings.

Statistical Conclusion Validity

Statistical conclusion validity concerns whether the statistical analyses accurately supported the conclusions drawn.

The relatively small number of content items may have limited statistical power in certain analyses, particularly when results were disaggregated by category. While multiple complementary metrics were used to evaluate agreement, additional statistical tests and larger datasets would have further strengthened the robustness of the conclusions.

CONCLUSION

This study introduced a multimodal framework for automated cyberbullying detection using LLMs capable of analyzing images and videos from major social platforms. Through the adoption of a zero-shot inference strategy, the proposed approach avoided the constraints of domain-specific fine-tuning, outperforming conventional NLP- or CNN-based classifiers, especially in visual-only harassment scenarios (Bommasani et al., 2021; Brown et al., 2020).

Among the tested models, Gemma 3-12B achieved the most stable and accurate performance, demonstrating high sensitivity toward four relevant cyberbullying categories: revenge porn, happy slapping, racism, and body shaming. The use of continuous scoring enabled fine-grained risk assessment, which was valuable for moderation workflows where prioritization was essential.

The system architecture, built on containerized components and standardized JSON-based inference logs, ensures reproducibility, transparency, and future extensibility. At the same time, limitations remain regarding dataset diversity, the handling of subtle contextual cues, and reliance on prompt engineering.

Beyond methodological contributions, this work delivered a concrete empirical resource that substantiated the reported findings. As part of the study, a new multimodal dataset of 5,000 annotated visual samples was constructed, covering five categories of online harm: happy slapping, revenge porn, racism, body shaming, and an explicit unknown class for ambiguous content. The dataset follows

a standardized JSON-based annotation format, including both categorical labels and an uncertainty indicator, enabling transparent analysis of borderline cases and facilitating human-in-the-loop moderation scenarios. By incorporating multiple visual variants of the same content and reflecting realistic class imbalance, the dataset provides a robust foundation for evaluating multimodal cyberbullying detection systems in real-world conditions.

Future work will focus on several key directions. First, the system will be extended to incorporate multimodal sequence reasoning algorithms specifically designed for long-duration video analysis, enabling the model to recognize evolving harassment patterns over time rather than relying solely on isolated frames. Additionally, research will explore adaptive prompting methodologies and self-calibrating scoring strategies, allowing the model to automatically refine its interpretive consistency across varying contexts. The dataset will also be expanded to include a broader range of cultural and situational representations, reducing the risk of bias and improving the model's generalizability to diverse online communities. Finally, where ethically and legally appropriate, the integration of limited user metadata or linguistic cues alongside visual information may further enhance contextual understanding and overall classification accuracy.

In conclusion, this research provided concrete evidence that modern MLLMs can play a transformative role in the automated moderation ecosystem, supplementing human expertise and enabling more effective protection against online aggression and digital abuse.

FUNDING STATEMENT

This research was supported by the following project: National Recovery and Resilience Plan, Mission 4 "Education and Research" – Component 2 "From Research to Business" – Investment 1.3, funded by the European Union – NextGenerationEU – CUP: F53C22000740007 – through the project titled "Cyber Social Security," acronym CSS.

COMPETING INTERESTS STATEMENT

The authors of this publication declare there are no competing interests.

CORRESPONDING AUTHOR

Correspondence should be addressed to Dr. Andrea Chezzi (andrea.chezzi@unisalento.it).

REFERENCES

Ahmed, M. T., Akter, N., Rahman, M., Islam, A. Z. M. T., Das, D., & Rashed, M. G. (2023). Multimodal cyberbullying meme detection from social media using deep learning approach. *International Journal of Computer Science and Information Technologies*, *15*(August), 27–37. DOI: 10.5121/ijcsit.2023.15403

Blodgett, S. L., Barocas, S., Daumé, H.III, & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.485

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R. B., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. T., Creel, K., Davis, J., Demszky, D., & Donahue, C. (2021). On the Opportunities and Risks of Foundation Models. *ArXiv (Cornell University)*. arxiv.2108.07258

Bosma, M., Chi, E., Ichter, B., Le, Q. V., Schuurmans, D., Wang, X., Wei, J., Xia, F., & Zhou, D. (2022). Chain-Of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. DOI: 10.52202/068431-1800

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Hesse, C. (2020). Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *4*(33). https://arxiv.org/abs/2005.14165

Caffagni, D., Cochi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Baraldi, L., Cornia, M., & Cucchiara, R. (2024). The revolution of multimodal large language models. *Survey (London, England)*, 13590–13618. Advance online publication. 2402.12451. DOI: 10.18653/v1/2024.findings-acl.807

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR,* (pp. 6299-6308). DOI: 10.1109/CVPR.2017.502

Fersini, E., Gasparini, F., Rizzi, G., Saibene, A., Chulvi, B., Rosso, P., Lees, A., & Sorensen, J. (2022). SemEval-2022 task 5: Multimedia automatic misogyny identification. In Emerson, G., Schluter, N., Stanovsky, G., Kumar, R., Palmer, A., Schneider, N., Singh, S., & Ratan, S. (Eds.), *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 533–549). Association for Computational Linguistics. DOI: 10.18653/v1/2022.semeval-1.74

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. DOI: 10.1007/s11023-018-9482-5

Gemma Team. (2025). Gemma 3 technical report. In *arXiv.org*. https://arxiv.org/abs/2503.19786

Grigg, D. W. (2010). Cyber-Aggression: Definition and concept of cyberbullying. *Australian Journal of Guidance & Counselling*, *20*(2), 143–156. DOI: 10.1375/ajgc.20.2.143

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015, March 12). *Detection of cyberbullying incidents on the Instagram social network*. arXiv.1503.03909

Islam, M. S., & Rafiq, R. I. (2024). Comparative analysis of GPT models for detecting cyberbullying in social media platforms threads. In Lossio-Ventura, J. A., Ceh-Varela, E., Vargas-Solar, G., Marcacini, R., Tadonki, C., Calvo, H., & Alatrista-Salas, H. (Eds.), *Information Management and Big Data. SIMBig 2023. Communications in Computer and Information Science* (*Vol. 2142*). Springer. DOI: 10.1007/978-3-031-63616-5_25

Kazemi, A., Natarajan, B., Wagner, J., Qadeer, H., Verma, K., & Davis, B. (2025). *Synthetic vs. gold: The role of LLM generated labels and data in cyberbullying detection*. DOI: 10.26615/978-954-452-098-4-062

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2021). The hateful memes challenge: Detecting hate speech in multimodal memes. *ArXiv:2005.04790 [Cs]*. https://arxiv.org/abs/2005.04790

Kumar, A., & Sachdeva, N. (2021). Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network. *Multimedia Systems*, *28*(February), 2043–2052. Advance online publication. DOI: 10.1007/s00530-020-00747-5

Langos, C. (2012). Cyberbullying: The challenge to define. *Cyberpsychology, Behavior and Social Networking*, *15*(6), 285–289. DOI: 10.1089/cyber.2011.0588

Lee, Y. J., Li, C., Liu, H., & Wu, Q. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, *36*, 34892–34916. DOI: 10.52202/075280-1516

Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., & Yuan, L. (2023). *Video-LLaVA: Learning united visual representation by alignment before projection*. https://arxiv.org/abs/2311.10122

Maity, K., Jha, P., Saha, S., & Bhattacharyya, P. (2022). A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1739–1749. DOI: 10.1145/3477495.3531925

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. DOI: 10.1145/3457607

Meta. (2025). *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation*. Meta.com. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency '19*, 220–229. DOI: 10.1145/3287560.3287596

Philipo, A. G., Sarwatt, D. S., Ding, J., Daneshmand, M., & Ning, H. (2026). Cyberbullying detection: Exploring datasets, technologies, and approaches on social media platforms. *ACM Computing Surveys*, *58*(7), 1–35. DOI: 10.1145/3785654

Pramanick, S., Dimitrov, D., Mukherjee, R., Sharma, S., Akhtar, Md. S., Nakov, P., & Chakraborty, T. (2021, August 1). *Detecting harmful memes and their targets*. ACLWeb; Association for Computational Linguistics. DOI: 10.18653/v1/2021.findings-acl.246

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 1135–1144. DOI: 10.18653/v1/N16-3020

Sharma, C., Bhageria, D., Scott, W. K., Pykl, S., Das, A., Chakraborty, T., Pulabaigari, V., & Gambäck, B. (2020). SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! *International Conference on Computational Linguistics*. DOI: 10.18653/v1/2020.semeval-1.99

Skender, I., Tong, K., Solmaz, S., & Watzenig, D. (2025). Investigating traffic accident detection by using multimodal large language models. *2025 IEEE International Automated Vehicle Validation Conference (IAVVC)*, 1–7. DOI: 10.1109/IAVVC61942.2025.11219530

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, *26*(3), 277–287. DOI: 10.1016/j.chb.2009.11.014

Vanpech, P., Peerabenjakul, K., Suriwong, N., & Fugkeaw, S. (2024). Detecting cyberbullying on social networks using language learning model. *2024 16th International Conference on Knowledge and Smart Technology*, 161–166. DOI: 10.1109/KST61284.2024.10499678

Wu, C.-Y., Feichtenhofer, C., Fan, H., He, K., Krähenbühl, P., & Girshick, R. (2019). *Long-Term Feature Banks for Detailed Video Understanding. ArXiv*. Cornell University. DOI: 10.1109/CVPR.2019.00037

Yi, P., & Zubiaga, A. (2023). Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media*, *36*, 100250. DOI: 10.1016/j.osnem.2023.100250

Zhang, H., Li, X., & Bing, L. (2023). *Video-LLaMA: An instruction-tuned audio-visual language model for video understanding*. DOI: 10.18653/v1/2023.emnlp-demo.49

Zhong, H., Li, H., Squicciarini, A., Rajtmajer, S., & Miller, D. (2019). Toward image privacy classification and spatial attribution of private content. *2019 IEEE International Conference on Big Data*, 1351–1360. DOI: 10.1109/BigData47090.2019.9006510