Article

A Hybrid Deep-Learning Approach for Multi-class Classification of Cyberbullying Using Multi-modal Social Media Data

Israt Tabassum * and Vimala Nunavath

Posted Date: 7 November 2024

doi: 10.20944/preprints202411.0392.v1

Keywords: Cyberbullying; Multi-modal Data; Multi-class Classification; Deep-Learning; Hybrid (RoBERTa+ViT) Model; Late Fusion; Social Media

Article

A Hybrid Deep-Learning Approach for Multi-Class Classification of Cyberbullying Using Multi-Modal Social Media Data

Israt Tabassum * and Vimala Nunavath

Department of Science and Industry Systems, University of South-Eastern Norway, Kongsberg, Norway
* Correspondence: israt.tabassum34@gmail.com

Abstract: Cyberbullying is defined as the use of social media platforms to hurt or humiliate people online. The anonymity on these platforms makes it easy to spread hurtful content, sometimes leading victims to self-harm. This highlights the urgent need for efficient methods to identify to prevent cyberbullying. Numerous studies have addressed this issue by focusing on cyberbullying classification, primarily through binary classification using multi-modal data or multi-class approaches targeting either text or image data. While deep-learning advancements have improved cyberbullying identification, a gap remains in the multi-class classification of cyberbullying utilizing multi-modal data such as memes and this research aims to bridge this gap by accurately classifying cyberbullying across multiple data modalities through a hybrid deep-learning model that combines RoBERTa for text extraction and Vision Transformer (ViT) for images extraction as hybrid(RoBERTa+ViT) model using late fusion module process. Two datasets were utilized: a Private-dataset was collected from comments on social media videos and a Public-dataset that was downloaded from existing research. The hybrid model was trained on these two datasets and the model demonstrated notable performance, achieving an an accuracy of 99.24% on the Public-dataset and 96.1% on the Private-dataset, with F1-scores of 0.9924 and 0.9599, respectively.

Keywords: cyberbullying; multi-modal Data; multi-class classification; deep-learning; hybrid (RoBERTa+ViT) model; late fusion; social media

1. Introduction

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 [1,2]. 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 [3].

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 [4]. 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

[1] https://www.facebook.com/
[2] https://twitter.com/
[3] https://www.instagram.com/

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. In the study of Collantes et al. [5], as SM continues to grow, it's important to be aware of these dangers and work towards creating a safer, kinder online space for everyone.

According to the 2014 *EU-Kids Online Report*, 20% of kids between the ages of 11 and 16 have experienced cyberbullying [6]. According to the quantitative research of Tokumaga et al. [7], youths experience cyber-victimization at a rate of 20% to 40%. These all highlight how critical it is to identify a strong and all-encompassing solution to this pervasive issue. Qiu et al. [8] claimed that 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. Automated cyberbullying identification and prevention can effectively address this issue. In the research of chen et al. [9], and van et al. [10], there are some approaches available to identify bullying incidents and way to support victims. Teenagers often use online platforms with safety centers, such as YouTube's Safety Centre [4] and Twitter's Safety and Security [5].

Deep-Learning (DL) technologies have produced promising results in various domains including defect detection [11], health diagnosis [12], disaster management [13], stock prediction [14] and also for classifying cyberbullying by processing and learning from large, complex datasets. Current cyberbullying identification and classification methods, which are primarily focused on binary-based multi-modal data [15–18] or multi-class and multi-label textual data [19–21]. However, there is no research has thoroughly investigated multi-class classification of cyberbullying for multi-modal data. This research aims to bridge the gap by employing advanced DL models to classify cyberbullying in a multi-modal context. The study will use both Public-dataset and uniquely collected Private-dataset from SM platforms, and with a focus on comments associated with short videos. This study aims to improve the accuracy of cyberbullying multi-class classification and contribute to safer online environments by creating and testing multiple DL models. Hence, in this work, DL models especially transformer architectures will be used to classify multi-class classification of cyberbullying based on multi-modal data on SM.

In this research, we try to answer the following research questions:

1. Which DL models are best suitable for multi-class classification of cyberbullying using multi-modal data?
2. Does DL perform better than the state-of-the-art algorithms?

The rest of the paper is organized as follows. Section 2 provides the existing literature in the multi-classification of cyberbullying from SM. The considered research methodology, including data acquisition and pre-processing process, use of different classifiers' network architecture, and the descriptions of the publicly available dataset and the acquired dataset are presented in Section 3. Section 4 presents the experimental results obtained using two datasets, and discussed them in Section 5. Finally, a conclusion and future research developments are given in Section 6.

2. Related Work

There is significant research going on in the areas of developing DL algorithms for cyberbullying classification. In this section, we present existing research on applying DL to cyberbullying multi-class classification. We have tried to figure out which models used, and how the data was collected and labeled so that they worked best for multi-modal data.

[4] https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#staying-safe
[5] https://help.twitter.com/en/safety-and-security

The authors in [22] implemented a DL model named bengali BERT to classify multi-classes of the cyberbullying on bengali language data. The researchers used YouTube textual comments dataset, which was the same dataset as previous studies in [23,24]. This dataset contained several classes such as religious, sexual, linguistic, political, personal, and crime-related content. In addition, the dataset contained data related to offensive text, including personal, geographical, religious, and crime-related offensive content, as well as content related to entertainment, sports, memes, and TikTok. The results of developed bengali BERT model demonstrated the best level of accuracy, reaching 0.706, and a weighted F1-score of 0.705.

The authors in [25] performed a classification analysis on Bengali SM comments on bengali language data. In this study, the authors focused exclusively on textual data obtained from comments on Facebook and collected about 42, 036 comments. The study employed the DL models CNN and LSTM for the purpose of multi-classification. The authors categorized the data into several classes such as Political, Religious, Sexual, Acceptable, and Combined. The performance of CNN-based LSTM network, named as: CLSTM architecture exhibits an accuracy rate of 85.8% and an F1 score of 0.86.

The authors in [21] developed a multitask DL framework for the identification of cyberbullying, such as sentiment, sarcasm and emotion aware cyberbullying from multi-modal memes. In their study, the authors collected images and memes from *Twitter* and *Reddit* social site's memes. To scrape images, they used hashtags like MeToo, KathuaRapeCase, Nirbhya, Rendi, Chuthiya, and Kamini on Twitter and subreddits like Desimemes, HindiMemes, and Bakchodi on Reddit, resulting in around 5,854 images or memes. Various DL models such as BERT, ResNET-Feedback and CLIP-CentralNet were developed and trained using textual and visual data. The task of the sentiment-emotion-sarcasm-aware multi-modal cyberbully identification in a code-mixed scenario was introduced for the first time in their paper. To tackle this challenge, they developed a novel multi-modal memes dataset called MultiBully, annotated with labels for bullies, attitude, emotion, and sarcasm. The purpose of this annotation was to determine if this information could aid in more accurate cyberbullying identification. An attention-based multi-task multi-modal framework, CLIP-CentralNet, was developed as a new architecture for sentiment, emotion, and sarcasm-assisted cyberbullying identification. Their suggested model included ResNet, mBERT, and CLIP for effective representations of many modalities and support in learning generic features across several tasks. The newly created CLIP-CentralNet framework performed noticeably better than any single task and uni-modal models in their task. For the purpose of identifyinging cyberbullying, they achieved accuracy of 61.14% for textual data using BERT, GRU, and a fully connected layer, and 63.36% for image data using ResNet and a fully connected layer.

The authors in [26] proposed a model that employed a Convolutional Neural Network (CNN) and Binary Particle Swarm Optimization (BPSO) to classify SM posts from platforms like Facebook, Twitter, and Instagram. The model categorized posts containing both images and written comments into three classes: non-aggressive, medium-aggressive, and high-aggressive. A dataset comprising symbolic images and their corresponding textual comments was created to validate the proposed model. The system employed a pre-trained VGG-16 model to extract the visual features of the image, while also utilizing a three-layered CNN to extract the textual data. The hybrid feature set, consisting of both picture and text features, was optimized using the BPSO algorithm to extract the most pertinent characteristics. The enhanced model, incorporating advanced features and utilizing the Random Forest classifier, achieved a weighted F1-Score of 0.74.

The authors in [27] identified cyber-trolling from SM. The dataset was gathered from various sources, including YouTube API, Twitter API, web scraping, and government sources. Their main goal was to apply the model to both text and video datasets. They developed various machine learning and DL techniques, including multi-modal approaches such as logistic regression, multinomial-NB, perception, random forest, bidirectional-LSTM model. The dataset was divided into topic-specific categories such as misogyny, sexism, racism, xenophobia, and homophobia. Their obtained experimental results showed that the Random Forest model provided highest accuracy with 96.50% than other models, including Bidirectional LSTM model.

The authors in [28] identified multi-label hate speech in their research. They presented "ETHOS" (multi-labEl haTe speecH detection dataset), a textual dataset based on comments from Reddit and YouTube that was validated through the use of the Figure-Eight crowd sourcing platform. It comes in two variants: binary and multi-label. For binary classification, they have got 80.36% accuracy from DistilBERT model as a highest accuracy, and for the accuracies of multi-label classification using BiLSTM were as follows: Violence: 50.86%, Directed versus Generalized: 55.28%, Gender: 70.34%, Race: 75.97%, National Origin: 67.88%, Disability Rate: 69.64%, Religion: 71.65%, Sexual orientation: 89.83%.

The authors in [20], proposed DL based approaches for identifying cyberbullies on SM. They presented an annotated dataset containing 99,991 tweets. They showed result for both binary classification and multi-class classification. For binary classification, the classes were: cyberbully and non-cyberbully classes, and for multi-class classification, the classes were: non-cyberbullying, religion, ethnicity/race, and gender/sexual class respectively. They showed 99.80% accuracy for multi-class classification using RoBERTa model.

From Table 1, we can see that, some research such as [20,22,25–28] has focused on multi-classification, but only on textual data. Where as researcher Maity et al. [21] explored multi-label classification on multi-modal data, but the study did not achieve satisfactory outcomes. While Hossain et al. [29] presented results on multi-classification using multi-modal data, the study was limited to the levels of aggression.

Table 1. Summary of Existing Literature on Social Media Cyberbullying Classification on Multi-classification

Literature Review	Dataset Collection Sources	Model Name	Accuracy	Label Type	Limitation
[21]	Twitter and Reddit memes	BERT, ResNET, GRU	Text accuracy: 61.14% and Image accuracy 63.36%	multi-label	Performance outcome, accuracy rate is low
[22]	YouTube comments	Bengali BERT	Accuracy: 70.6%, F1 score: 0.705	multi-class	Textual data only
[26]	Facebook, Twitter, Instagram	CNN, BPSO	F1-Score of 0.74	multi-class	Focused on Aggression's level
[29]	MultiOFF and TamilMemes	VGG19 and m-distilBERT	Weighted F1-scores of 66.73% and 58.59%	multi-class	Focused on Aggression's level
[27]	YouTube, tiktok, twitter and other social site	Random Forest	accuracy 96.50%	multi-class	Focused only textual data
[28]	Reddit and YouTube	BiLSTM	accuracy 80.36%	multi-label	Focused only textual data
[20]	Twitter	RoBERTa	99.80% accuracy for multi-class classification	multi-class	Focused only textual data

Although various researchers have worked on classifying cyberbullying on SM using multi-modal data, the majority of studies have focused on either identification or binary classification. Several types of cyberbullying have been identified by Van et al. [30,31], but there has been no multi-class classification of cyberbullying using multi-modal data based on these types. Therefore, in this paper, we will explore the use of transformer architectures of DL models for multi-class classification of SM cyberbullying using multi-modal data.

3. Materials and Methods

This section provides an overview of the proposed solution, explains the dataset collection process, describes the pre-processing methods applied, and outlines the network architectures used in the experiments. Additionally, it discusses the evaluation metrics used to assess the model's performance.

3.1. Proposed Solution

For classifying different cyberbullying types, we proposed a generic solution, illustrated in Figure 1. The first step in this solution involved multi-modal data collection, where two datasets were collected. This was followed by a comprehensive workflow that included data preprocessing which was done on both the datasets. Then, splitting the data into train (80%), test (10%) and validation (10%) datasets.

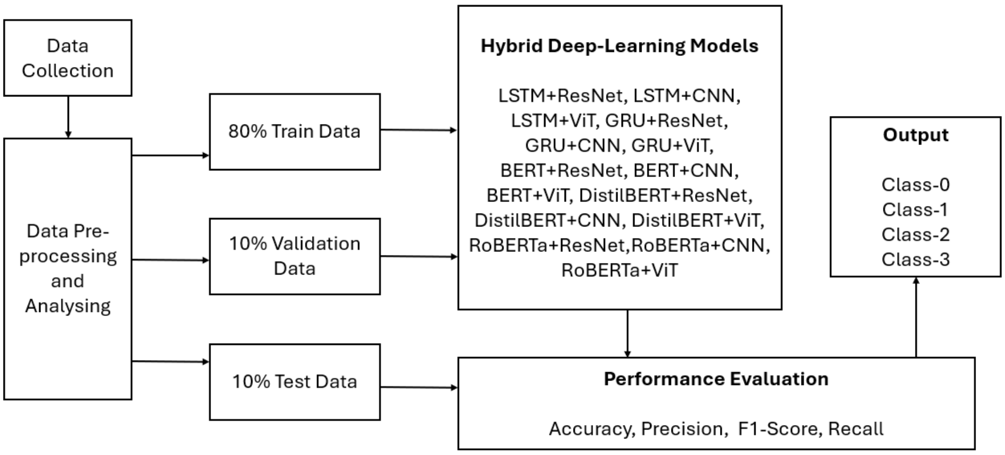

Data Collection
Data Pre-processing and Analysing
80% Train Data
10% Validation Data
10% Test Data
Hybrid Deep-Learning Models
LSTM+ResNet, LSTM+CNN, LSTM+ViT, GRU+ResNet, GRU+CNN, GRU+ViT, BERT+ResNet, BERT+CNN, BERT+ViT, DistilBERT+ResNet, DistilBERT+CNN, DistilBERT+ViT, RoBERTa+ResNet,RoBERTa+CNN, RoBERTa+ViT
Output
Class-0
Class-1
Class-2
Class-3
Performance Evaluation
Accuracy, Precision, F1-Score, Recall

Figure 1. The Proposed Solution

Subsequently, the developed models such as LSTM+ResNet, LSTM+CNN, LSTM+ViT, GRU+ResNet, GRU+CNN, GRU+ViT, BERT+ResNet, BERT+CNN, BERT+ViT, DistilBERT+ ResNet, DistilBERT+CNN, DistilBERT+ViT, RoBERTa+ResNet, RoBERTa+ CNN, and RoBERTa + ViT were trained on the training dataset and evaluated on the testing dataset using various model evaluation metrics such as accuracy, F1-score, precision and recall. More details about each step are explained in the sections below.

3.2. Datasets Collection

In the proposed solution, the first step was data collection (see Figure 1). In this research, two datasets were collected. The first dataset was *Public-dataset*, and the second dataset was *Private-Dataset* (or Self-collected Dataset).

The Public-dataset was compiled by combining datasets from existing studies on cyberbullying classification by Hamza et al. [32] referred as Dataset-1, and Maity et al. [21] referred as Dataset-2. The Dataset-1 was downloaded from existing research on cyberbullying classification, termed hateful memes dataset, which contains 2000 data by Hamza et al. [32]. Maity et al. [21] created a multi-modal sarcasm identification dataset which we referred as Dataset-2, for our experiments. This dataset includes 5,854 samples collected from open-source platforms, Twitter, and Reddit. The two datasets (Dateset-1 and Dataset-2) were merged into a single dataset entitled Public-dataset to enlarge the size of data.

Whereas, Private-dataset contains one thousand multi-modal data, which were self-collected in this research. The dataset was collected from comments on short videos posted on Facebook[6],

[6] https://www.facebook.com/

Instagram [7], YouTube[8], and TikTok[9]. The dataset is available in github [10]. Comments from the aforementioned platforms were extracted using various tools: APIFY[11] for Facebook and YouTube short video's comments, TKCommentExport[12] tool for TikTok's comments, IGCommentExporter[13] for Instagram reels comments.

3.3. Data Pre-Processing and Analysis

After collecting the dataset, we pre-processed the data. Figure 2 shows the data pre-processing pipeline. In this pipeline, the first step was extracting text and images from memes and then analyzing the extracted text and image data.

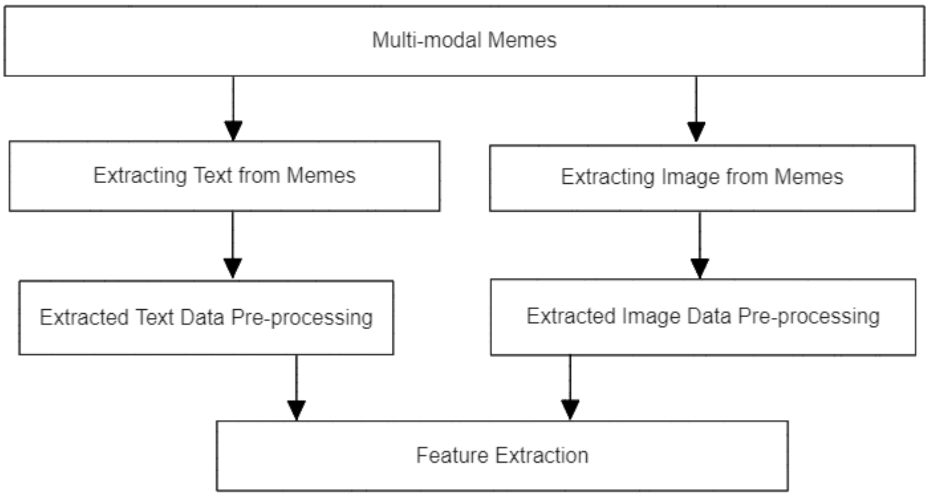

Multi-modal Memes
Extracting Text from Memes
Extracting Image from Memes
Extracted Text Data Pre-processing
Extracted Image Data Pre-processing
Feature Extraction

Figure 2. The Data Pre-processing Pipeline.

3.3.1. Extracting Text and Images from Memes

To extract text and images from memes, we used the tool Optical Character Recognition (OCR)[14], which reads text from memes, and integrated it into Python[15] using the *Pytesseract*[16] package. Before extracting the text, we improved the image quality of the memes using the *OpenCV*[17] library. To enhance memes for text extraction, we used three key techniques: *bilateral filtering*[18] to minimize noise while preserving edges, *grayscale conversion* to convert image format into grayscale, and *thresholding* to make the text more distinct from the background. Once the text was extracted, it was saved in a structured format in a folder named "text_data", separated from the original memes in the Private-dataset.

For extracting images from memes, we resized and standardized the images using techniques from the OpenCV[19] library such as *cv2.imread()*, *cv2.resize()*, *cv2.cvtColor()*, and *cv2.bilateralFilter()*. After the images were processed, they were saved in a folder called "image" for classification.

[7] https://www.instagram.com/
[8] https://www.youtube.com/
[9] https://www.tiktok.com/
[10] https://github.com/israt-tabassum/cyberbullying-classification-private-data
[11] https://apify.com/
[12] https://tkcommentexport.extensionsbox.com/
[13] https://chromewebstore.google.com/detail/igcommentexporter-export/
[14] https://en.wikipedia.org/wiki/Optical_character_recognition
[15] https://www.python.org/
[16] https://pypi.org/project/pytesseract/
[17] https://opencv.org/
[18] Wikipedia for Bilateral filter
[19] https://opencv.org/

3.3.2. Data Categorization

The extracted text data from *Public* and *Private-dataset* were categorized into four cyberbullying categories, based on knowledge acquired from prior studies i.e., [30,31,33]. The identified categories were:

- Non-Bullying (class-0): Data that contains no insulting, defamatory, offensive, or aggressive language based on research [33].
- Defaming Cyberbullying (class-1): Involves behavior where individuals insult or damage another person's reputation and self-worth based on research [30,31].
- Offensive Language Cyberbullying (class-2): Refers to situations where derogatory language was directed at someone based on research [30,31].
- Aggressive Cyberbullying (class-3): Includes threats, violent behavior, and abusive actions aimed at someone based on research [30,34].

Similarly, based on prior research [35–37], images from both Public-dataset and Private-dataset were categorized as follows:

- Non-Bullying (class-0): Images that do not feature defaming, sexual, offensive, or aggressive content.
- Defaming (class-1): Images that contain sexual or nudity-related content.
- Offensive (class-2): In the *Private-dataset*, this includes images such as those showing a middle finger or combining human faces with animal faces. As the *Public-dataset* lacks this specific type, class-2 is reserved for content showing a middle finger.
- Aggressive (class-3): Depictions of violence, such as beating someone or brandishing weapons, fall into this category and class-3.

3.3.3. Data Cleaning

For the extracted text data of multi-modal data, we first converted all text to lowercase, removed leading and trailing spaces, and replaced newline characters with spaces. Next, we eliminated non-alphabetic and non-ASCII characters, fllowed by the removal of URLs. A regular expression tokenizer was then used to break the text into individual words. Common stopwords were then removed and a custom list was created to exclude specific words from the default English stopword list, while eliminating single-character words. The remaining words were rejoined into a single string, and each word was lemmatized to its root form. Furthermore, we addressed numerical values and rectifying duplicate, missing, noise, irreverent data resulted in a cleaner and more comprehensive dataset.

For the extracted image data of multi-modal data, initially we applied *bilateral filtering* to minimize noise and improve clarity. In the next step, *scaling* and *normalization* were used to ensure that all images have uniform dimensions and pixel intensity values, thereby reducing computational complexity and enhancing training convergence. All images were resized to 224x224 pixels via a transformation pipeline, which preserved consistency while standardizing the pixel data. We also handled color channels by converting images to RGB, and then converted them to grayscale to simplify processsing. The *thresholding* was applied to highlight key features, and duplicates were removed along with irrelevant images were manually excluded.

3.3.4. Data Augmentation and Sampling

For the extracted text data, we created different versions of the text using synonym replacement and paraphrasing. To ensure balanced class distribution across the four categories, we adjusted the number of examples for each class. Furthermore, for the extracted image data, we applied techniques such as rotation, flipping, and cropping to introduce variability. Rotation changed the angle, flipping altered the orientation, and cropping focused on specific areas of the images. We also ensured that the number of images from each class was balanced to avoid model bias toward any particular class.

3.4. Feature Extraction of Multi-modal Data

To extract features from the multi-modal data, we used tokenization and word embedding[20] (Word2Vec) for both Public-dataset and Private-dataset. We picked the Word2Vec technique because it captures semantic links between words, which improves the performance of the hybrid DL models on extracted text data. For extracted image data, we used multiple feature extraction approaches tailored to each model such as, *CNN features*, *DL pre-trained models* using transfer learning which allows us to use knowledge from previously trained models to improve categorization, and *patch embedding* approach, which divides images into smaller patches and processes them sequentially.

3.5. Splitting Data into Training, Validation and Testing Datasets

To build a fair and effective cyberbullying classification model, we carefully separated our datasets into different parts for training, validation, and testing. We used a common approach where 80% of the data was set aside for training. This allows the model to learn and identify patterns in comments related to cyberbullying. The remaining 20% of the data was split equally, with 10% used for validation to adjust the model's settings and prevent it from becoming too specialized, and 10% reserved for testing, which helps us evaluate how well the model performs without bias.

We also made sure to pre-process both the *Public* and *Private-datasets*, to maintain consistency and quality. The splitting was done randomly to ensure that all sets were diverse and to prevent any bias that could affect the model's learning process.

3.6. Network Architecture

Figure 3 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.

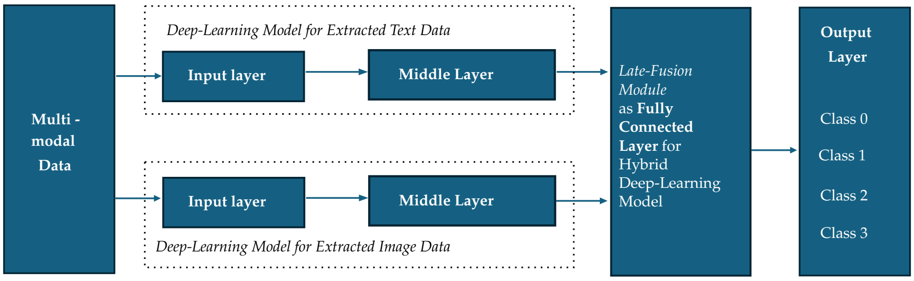

Multi - modal Data
Deep-Learning Model for Extracted Text Data
Input layer
Middle Layer
Input layer
Middle Layer
Deep-Learning Model for Extracted Image Data
Late-Fusion Module as Fully Connected Layer for Hybrid Deep-Learning Model
Output Layer
Class 0
Class 1
Class 2
Class 3

Figure 3. Network Architecture for Classification of Multi-modal Data.

In the DL module, we developed LSTM, GRU, BERT, DistilBERT, and RoBERTa models to extract text data from Public-dataset, and only the RoBERTa model was used to extract the text data from Private-dataset. Table 2 details the employed model architectures for text data. For these DL models, we begin with text tokens as input. Meanwhile, the ResNet, ViT, CNN model were used for extracting image data for Public-dataset, and only ViT model for Private-dataset was used to classify cyberbullying categories. For extracted images, the input typically comprises of RGB images with dimensions of 224x224 pixels. Middle layer was varied accordingly DL models used layer. Later, *Late Fusion Module* [38] technique was utilized to combine the developed DL models for extracted text and image data together as LSTM+ResNet, LSTM+CNN, LSTM+ViT, GRU+ResNet, GRU+CNN, GRU+ViT, BERT+ResNet, BERT+CNN, BERT+ViT, DistilBERT+ResNet, DistilBERT+CNN, DistilBERT+ViT, RoBERTa+ResNet, RoBERTa+CNN, and RoBERTa+ViT models for classifying multi-modal data. Each model concludes with a dense layer and a softmax function that classifies input features into four classes.

[20] https://www.tensorflow.org/text/tutorials/word2vec

Table 2. Developed Hybrid Deep-Learning Model Architectures for Classifying Multi-modal Data.

Model type	Input Layer	Middle Layers	Fully-Connected Layer	Output Layer
LSTM for extracted text data + ResNet for extracted image data	Extracted text tokens as embeddings + Extracted Images (e.g., 224x224x3 RGB)	2 LSTM layers + 50 Convolutional layers (with residual connections)	Combining LSTM and ResNet features as hybrid (LSTM+ResNet) model	Softmax dense layer for four output classes
LSTM for extracted text data + CNN for extracted image data	Extracted text tokens as embeddings + Extracted Images (e.g., 224x224x3 RGB)	2 LSTM layers + 1 Convolutional layer + 1 Activation layer + Additional layers (e.g., Pooling)	Combining LSTM and CNN features as hybrid (LSTM+CNN) model	Softmax dense layer for four output classes
GRU for extracted text data + ResNet for extracted image data	Extracted text tokens as embeddings + Extracted Images (e.g., 224x224x3 RGB)	2 GRU layers + 50 Convolutional layers (with residual connections)	Combining GRU and ResNet features as hybrid (GRU+ResNet) model	Softmax dense layer for four output classes
GRU for extracted text data + CNN for extracted image data	Extracted text tokens as embeddings + Extracted Images (e.g., 224x224x3 RGB)	2 GRU layers + 1 Convolutional layer + 1 Activation layer + Additional layers (e.g., Pooling)	Combining GRU and CNN features as hybrid (GRU+CNN) model	Softmax dense layer for four output classes
BERT for extracted text data + ResNet for extracted image data	Extracted text tokens with positional embeddings + Extracted images (e.g., 224x224x3 RGB)	12 Transformer layers + 50 Convolutional layers (with residual connections)	Combining BERT and ResNet features as hybrid (BERT+ ResNet) model	Softmax dense layer for four output classes
BERT for extracted text data + CNN for extracted image data	Extracted text tokens with positional embeddings + Extracted images (e.g., 224x224x3 RGB)	12 Transformer layers + 1 Convolutional layer + 1 Activation layer + Additional layers (e.g., Pooling)	Combining BERT and CNN features as hybrid (BERT+CNN) model	Softmax dense layer for four output classes
LSTM for extracted text data + ViT for extracted image data	Extracted text tokens as embeddings + Extracted images divided into patches (e.g., 16x16)	2 LSTM layers + 12 Transformer layers	Combining LSTM and ViT features as hybrid (LSTM+ViT) model	Softmax dense layer for four output classes
GRU for extracted text data + ViT for extracted image data	Extracted text tokens as embeddings + Extracted images divided into patches (e.g., 16x16)	2 GRU layers + 12 Transformer layers	Combining GRU and ViT features as hybrid (GRU+ViT) model	Softmax dense layer for four output classes
BERT for extracted text data + ViT for extracted image data	Extracted text tokens with positional embeddings + Extracted images divided into patches (e.g., 16x16)	12 Transformer layers + 12 Transformer layers	Combining BERT and ViT features as hybrid (BERT+ViT) model	Softmax dense layer for four output classes
DistilBERT for extracted text data + ResNet for extracted image data	Extracted text tokens with positional embeddings + Extracted images (e.g., 224x224x3 RGB)	6 Transformer layers + 50 Convolutional layers (with residual connections)	Combining DistilBERT and ResNet features as hybrid (DistilBERT+ ResNet) model	Softmax dense layer for four output classes

Table 2. *Cont.*

Model type	Input Layer	Middle Layers	Fully-Connected Layer	Output Layer
DistilBERT for extracted text data + CNN for extracted image data	Extracted text tokens with positional embeddings + Extracted images (e.g., 224x224x3 RGB)	6 Transformer layers + 1 Convolutional layer + 1 Activation layer + Additional layers (e.g., Pooling)	Combining DistilBERT and CNN features as hybrid (DistilBERT+CNN) model	Softmax dense layer for four output classes
DistilBERT for extracted text data + ViT for extracted image data	Extracted text tokens with positional embeddings + Extracted images divided into patches (e.g., 16x16)	6 Transformer layers + 12 Transformer layers	Combining DistilBERT and ViT features as hybrid (DistilBERT+ViT) model	Softmax dense layer for four output classes
RoBERTa for extracted text data + ResNet for extracted image data	Extracted text tokens with positional embeddings + Extracted images (e.g., 224x224x3 RGB)	12 Transformer layers + 50 Convolutional layers (with residual connections)	Combining RoBERTa and ResNet features as hybrid (RoBERTa+ ResNet) model	Softmax dense layer for four output classes
RoBERTa for extracted text data + CNN for extracted image data	Extracted text tokens with positional embeddings + Extracted images (e.g., 224x224x3 RGB)	12 Transformer layers + 1 Convolutional layer + 1 Activation layer + Additional layers (e.g., Pooling)	Combining RoBERTa and CNN features as hybrid (RoBERTa+ CNN) model	Softmax dense layer for four output classes
RoBERTa for extracted text data + ViT for extracted image data	Extracted text tokens with positional embeddings + Extracted images divided into patches (e.g., 16x16)	12 Transformer layers + 12 Transformer layers	Combining RoBERTa and ViT features as hybrid (RoBERTa+ ViT) model	Softmax dense layer for four output classes

3.7. Fusion Module

To combine both text and image modalities, a late fusion module approach was used. This approach allows each modality to be handled individually, which is useful for memes because they include both text and image [39]. The key features of late fusion module were as follows:

- *Integration:* Combines results from independent classification of extracted text and extracted image data.
- *Decision-Making:* Extracted text and image analyses are both taken into consideration when making decisions about the final output. First, it determines whether extracted text and image classifications are available, indicating multi-modal data. If both agree that there is no cyberbullying, a message appears indicating that the input is no cyberbullying. If both agree on the same type of cyberbullying, a specific class is defined. If the text and image classifications disagree, it acknowledges the existence of cyberbullying but labels it differently. If either classification is missing, the system concludes that the input is not multi-modal data and displays an appropriate message. The process concludes by determining the final message based on these conditions which is presented in Algorithm 1.

Algorithm 1 Fusion Logic Decision Making Logic for Multi-modal Data

```
Begin
if $text\_class$ is not None and $image\_class$ is not None then
    if $text\_class == image\_class$ then
        if $text\_class == 0$ then
            $fusion\_message \leftarrow$ “Input does not contain any Cyber-bullying.”
        else
            $fusion\_message \leftarrow$ “Input contains this class {$text\_class$} of cyberbullying.”
        end if
    else
        $fusion\_message \leftarrow$ “Input contains cyberbullying. Text label is: {$text\_label\_number$} and Image label is: {$image\_label\_number$}”
    end if
else
    $fusion\_message \leftarrow$ “This is not Multi-Modal Data!”
end if
End
```

- *Output Synthesis:* It creates a cohesive response that is presented to the user, effectively communicating the findings of the cyberbullying classification result.

3.8. Model Performance Metrics

We evaluated the efficacy of the developed models using four widely used performance metrics ie., Accuracy, Precision, Recall, and F1-score which was followed by [40] as follows:

1. Accuracy: It measures how often the model makes correct predictions. It is calculated as follows:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}} \tag{1}$$

2. Precision: It shows the quality of positive predictions. It is calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

3. Recall: It measures how well the model finds actual positive cases. It is calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

4. F1 Score: It combines precision and recall into one metric. It is calculated as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

We also evaluated the performance by confusion matix. A confusion matrix is a table that assesses the performance of a classification model. It compares the goal values (true labels) to the model's predictions. The matrix demonstrates how well the model distinguishes between distinct classes by categorizing the findings into four groups, such as, TP (True Positives), TN (True Negatives), FP (False Positives) and FN (False Negatives). TP occur when the model properly predicted the positive class, TN occur when the model predicted the negative class, FP occur when the model wrongly predicts the positive class, and FN occur when the model wrongly predicts the negative class [41]. It aids in computing key performance indicators like as accuracy, precision, recall, and F1-score.

3.9. Models Deployment

After developing the DL models, these models were deployed on a graphical user-interface (GUI) using Flask Python Web Framework [21]. GUI receives inputs in the form of multi-modal data. After deployment, the GUI enables the users to choose between Public-dataset and Private-dataset types. Algorithm 2 describes the full procedure of model deployment, and how the models are responding to multi-modal data. The GUI was developed using tools such as HTML, CSS, and JavaScript. The Flask application, which was the foundation for managing incoming requests and coordinating model predictions using python program. The program carefully pre-processes the data after receiving user inputs to make sure that inputs were formatted correctly for the model to use. Next, hybrid(RoBERTa+ViT) model was used to classify the multi-modal data. After generating classification result, the program compiled the findings and showed them to the user on a results page.

Algorithm 2 Process Input through GUI

```
Start
Receive an input through the GUI (Memes)
Apply the Text Extracting procedure to extract text.
Apply the Image Extracting procedure to extract image.
Use the hybrid DL model to process the extracted text and image.
Generate fusion message based on input data
Render results on the GUI using a template
End
```

4. Results

This section begins with outlining the experimental configuration used to conduct the experiments for classifying the cyberbullying. The details of the obtained results from each experiments are also presented below.

4.1. Experimental Setup and Hyper-Parameter Tuning

To perform multiple experiments, we followed an experimental setup and the hyper-parameter tuning process as follows. To enhance the performance of our models for extracted text data from Public-dataset and Private-dataset and extracted image data on the Public-dataset, we applied a common hyperparameter tuning strategy across all the DL models. We used the Adam[22] optimizer to adjust hyperparameters based on a validation dataset, running for twenty epochs with a batch size

[21] https://flask.palletsprojects.com/
[22] https://keras.io/api/optimizers/adam/

of twenty and a learning rate of 0.00002. To prevent overfitting, we implemented an early stopping mechanism, if there was no improvement after three consecutive epochs. For the textracted text classification, we used the *SparseCategoricalCrossentropy*[23] *loss* function, while for the extracted image classification on the Public-dataset, we applied *CrossEntropyLoss*[24] for multi-class classification using integer labels.

For the Private-dataset's extracted image data, we conducted ten random search trials to fine-tune hyperparameters, testing different batch sizes (8, 16, 20, 32, 64) and learning rates (0.00001, 0.00005, 0.0001, 0.0005, 0.001). The best model configuration was chosen based on the validation loss and accuracy results. We used nn.CrossEntropyLoss along with early stopping (with a patience of three epochs) to avoid overfitting. The final model's performance was evaluated using accuracy, F1-score, precision, and recall metrics.

4.2. Experimental Results on Public-Dataset

In Table 3, the obtained results from a series of experiments conducted on Public-dataset are presented. From the Table, it can be observed that the results are varied and this might be based on the combination of models used. If we see the result obtained from the first hybrid(LSTM+ResNet) model, it achieved an accuracy of 0.7085, with a recall of 0.71. Both the F1-score and precision were 0.665. The next hybrid(LSTM+CNN) model performed slightly better than hybrid (LSTM+ResNet), achieving an accuracy of 0.7285, with a recall of 0.73 and an F1-score and precision of 0.685. The hybrid(LSTM+ViT) model showed further improvement, with an accuracy of 0.7335 and an F1-score of 0.69.

Table 3. Performance Evaluation for Public-Dataset.

Hybrid Deep-Learning Model	Test Accuracy	Recall	F1-Score	Precision
LSTM+ResNet	0.7085	0.71	0.665	0.665
LSTM+CNN	0.7285	0.73	0.685	0.685
LSTM+ViT	0.7335	0.735	0.69	0.69
GRU+ResNet	0.723	0.715	0.655	0.63
GRU+CNN	0.743	0.735	0.675	0.65
GRU+ViT	0.748	0.74	0.68	0.655
BERT+ResNet	0.9585	0.9585	0.9585	0.9585
BERT+CNN	0.9785	0.9785	0.9785	0.9785
BERT+ViT	0.9835	0.9835	0.9835	0.9835
DistilBERT+ResNet	0.9655	0.9655	0.9655	0.9655
DistilBERT+CNN	0.9855	0.9855	0.9855	0.9855
DistilBERT+ViT	0.9905	0.9905	0.9905	0.9905
RoBERTa+ResNet	0.966	0.966	0.966	0.966
RoBERTa+CNN	0.986	0.986	0.986	0.986
RoBERTa+ViT	**0.9924**	**0.9924**	**0.9924**	**0.9924**

For the hybrid(GRU+ResNet) model, the accuracy was 0.723, with a recall of 0.715, an F1-score of 0.655, and a precision of 0.63. The hybrid(GRU+CNN) model performed better, with an accuracy of 0.743, recall of 0.735, F1-score of 0.675, and precision of 0.65. The hybrid(GRU+ViT) model did even better, reaching an accuracy of 0.748, with a recall of 0.74, an F1-score of 0.68, and precision of 0.655.The hybrid(BERT+ResNet) model achieved an accuracy of 0.9585, with recall, F1-score, and precision all at 0.9585. The hybrid(BERT+CNN) model did even better, reaching an accuracy of 0.9785, with matching recall, F1-score, and precision at 0.9785. The hybrid(BERT+ViT) model pushed performance even further, with an accuracy of 0.9835 and all other scores at 0.9835.

[23] https://www.tensorflow.org/api_docs/python/tf/keras/losses/SparseCategoricalCrossentropy
[24] https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

The hybrid(DistilBERT+ResNet) model had an accuracy of 0.9655, with recall, F1-score, and precision all matching at 0.9655. The hybrid(DistilBERT+CNN) model improved this, with an accuracy of 0.9855 and strong recall, F1-score, and precision. The hybrid(DistilBERT+ViT) model achieved the highest score so far, with an accuracy of 0.9905, recall, F1-score, and precision all at 0.9905. The hybrid(RoBERTa+ResNet) model reached an accuracy of 0.966, with all other scores matching at 0.966. The hybrid(RoBERTa+CNN) model did even better, reaching an accuracy of 0.986. Finally, the hybrid(RoBERTa+ViT) model achieved the best performance overall, with an accuracy of 0.9924 and all metrics—recall, F1-score, and precision—matching at 0.9924. Overall, the combination of hybrid(RoBERTa+ViT) for text and image data gives the most accurate and balanced results across all metrics, making it the top-performing model in this table. So we chose this hybrid (RoBERTA+ViT) model to classify the multi-modal data.

Figure 4 depicts the confusion matrix for hybrid(RoBERTA+ViT) model which shows the performance for each classes. Class-0 had the most correct predictions 308 data. Class-1 showed high accuracy, with 332 correct predictions and few errors—only two instances were misclassified as class-0. Class-2 and class-3 both demonstrate flawless predictive accuracy, with 319 and 321 correct predictions, respectively, and no instances incorrectly classified into any other class. This matrix highlights the model's robust ability to accurately identify class-2 and class-3, as well as its performance with class-1.

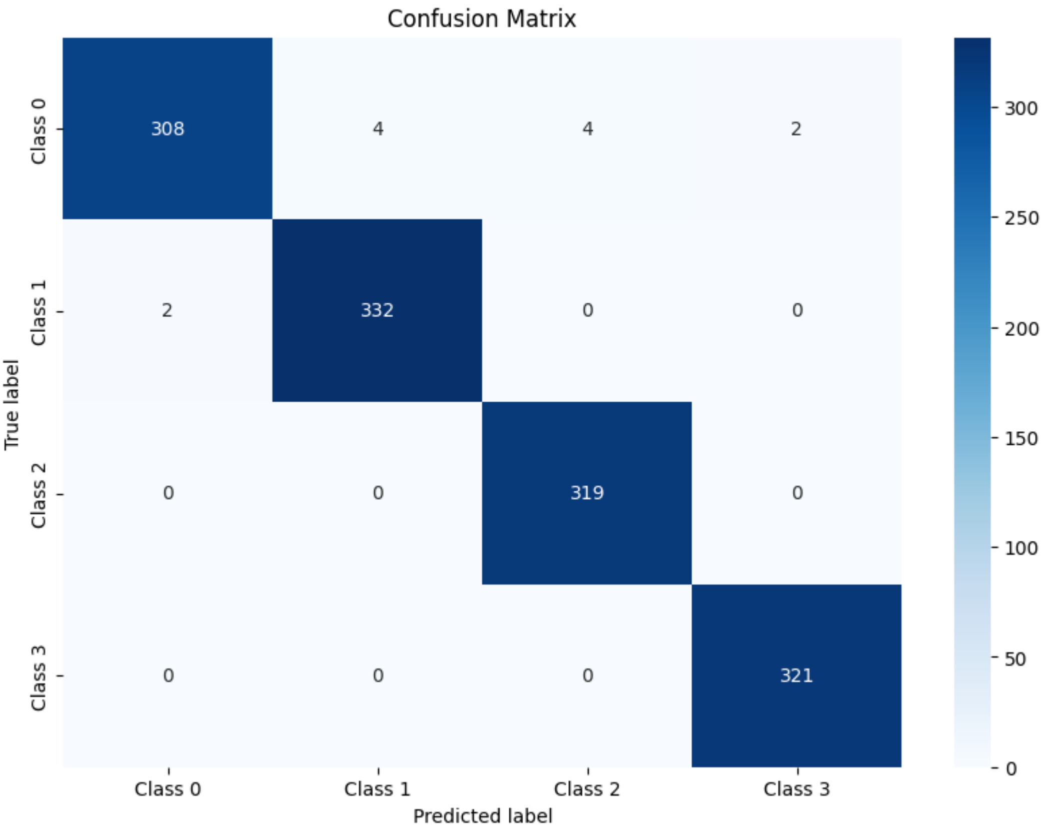


Figure 4. Confusion Matrix of Hybrid(RoBERTa+ViT) Model for Public-Dataset

4.3. Experimental Results on Private-Dataset

To experiment the multi-modal data, we applied hybrid (RoBERTA+ViT) model to classify multi-modal data in our Private-dataset. Based on the experiment with the Public-dataset in Table 3, we can see that the hybrid (RoBERTa+ViT) model performs more accurately than other hybrid DL models. Therefore, we selected the hybrid (RoBERTa+ViT) model to classify multi-modal data in our private dataset. We combined the accuracy of the RoBERTa and ViT models and calculated the average performance of two models to obtain the hybrid (RoBERTA+ViT) model's accuracy using the late fusion module in classifying multi-modal data into multiple classes. Table 4 shows the performance outcome of the hybrid (RoBERTA+ViT) model for Private-dataset, with accuracy of 96.1%, and f1-score of 95.99% and ROC-AUC value of 0.99.

Table 4. Performance Evaluation for Private-Dataset.

Hybrid Deep-Learning Model	Test Accuracy	Recall	F1-Score	Precision	ROC-AUC value
RoBERTa+ViT	0.961	0.9599	0.9599	0.960	0.99

Figure 5 depicts the confusion matrix of hybrid (RoBERTA+ViT) model, which shows the performance for each classes. The model correctly classified 398 instance for class-0. Similarly, in class-1, it mostly got it right with 386 correct predictions, but it made a few mistakes, mislabeling 6 data as class-0, 2 data as class-2, and 4 data as class-3. Class-2 and class-3 had very high correct predictions (408 and 410 data, respectively), with few items mislabeled, this is beacuse, when classe-1, class-2, and class-3 in the confusion matrix seem to be very similar, misclassifications happen, making it challenging to classify the data accurately and possibly leading to mistakes.

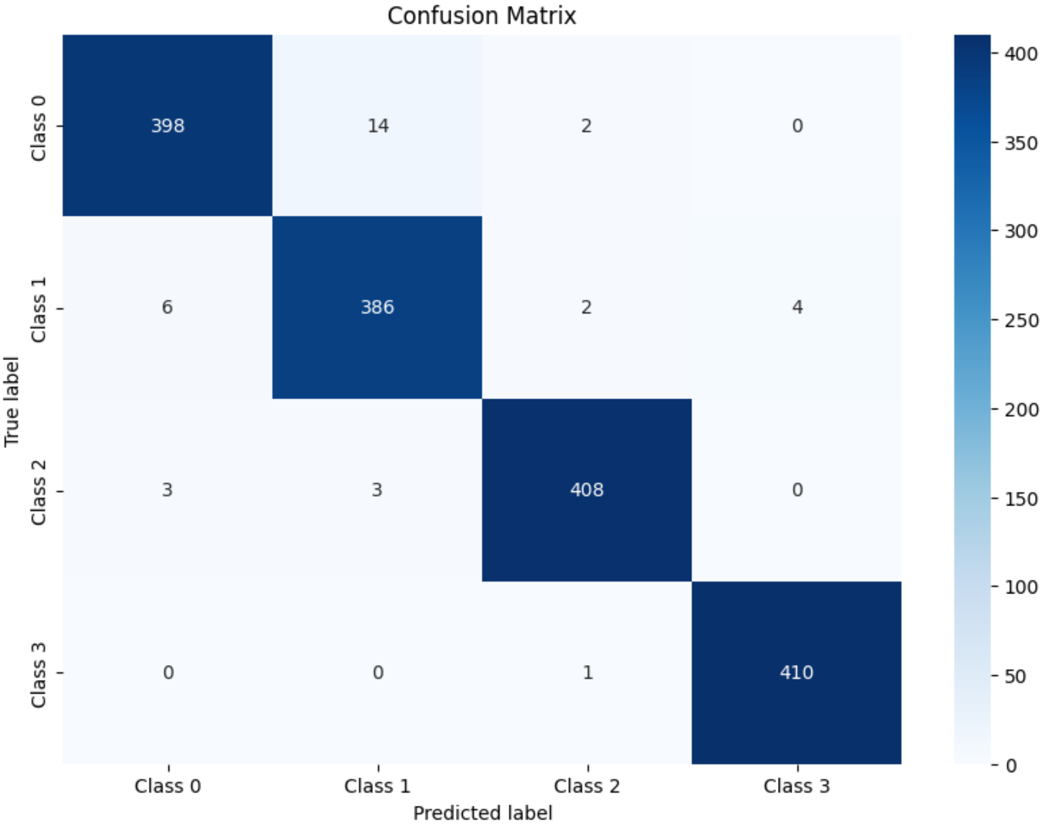


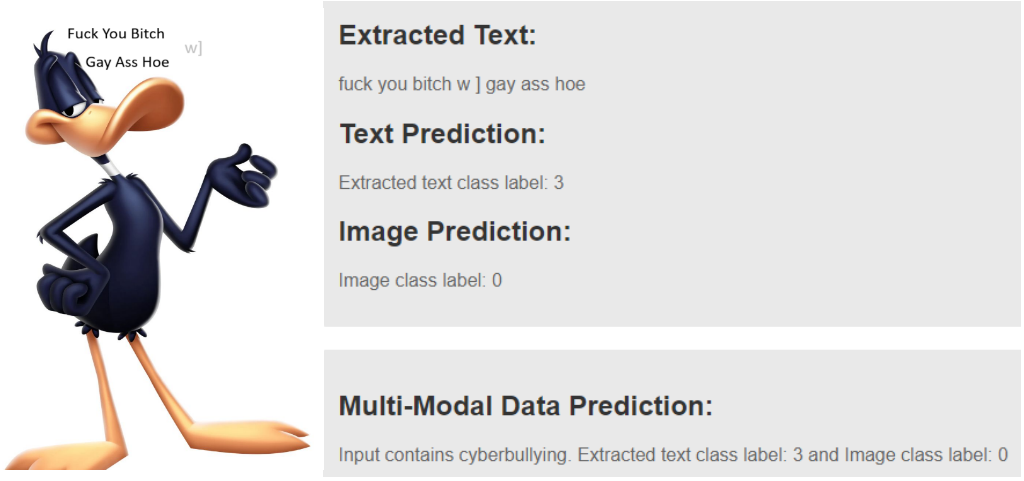
Figure 5. Confusion Matrix of Hybrid(RoBERTA+ViT) Model for Private-Dataset

4.4. Results of Model Deployment

We have deployed the developed model for showing our result on the graphical user interface (GUI). The processing of generating the graphical user interface described in the Section 3.9. In this section, we showed experiment result of model deployment. The result of inputs such as class label 0, 1, 2, 3's explanation was described in Section 3.3.2.

The multi-modal data in Figure 6a was tested to determine if the meme constituted cyberbullying. If it did, a corresponding label would be displayed. The result, shown in Figure 6b, indicates that while the image label is 0 (non-cyberbullying content), the extracted text was classified as aggressive (label

3). Since the text contained cyberbullying elements, the meme is categorized as cyberbullying-related data.



Extracted Text:

fuck you bitch w] gay ass hoe

Text Prediction:

Extracted text class label: 3

Image Prediction:

Image class label: 0

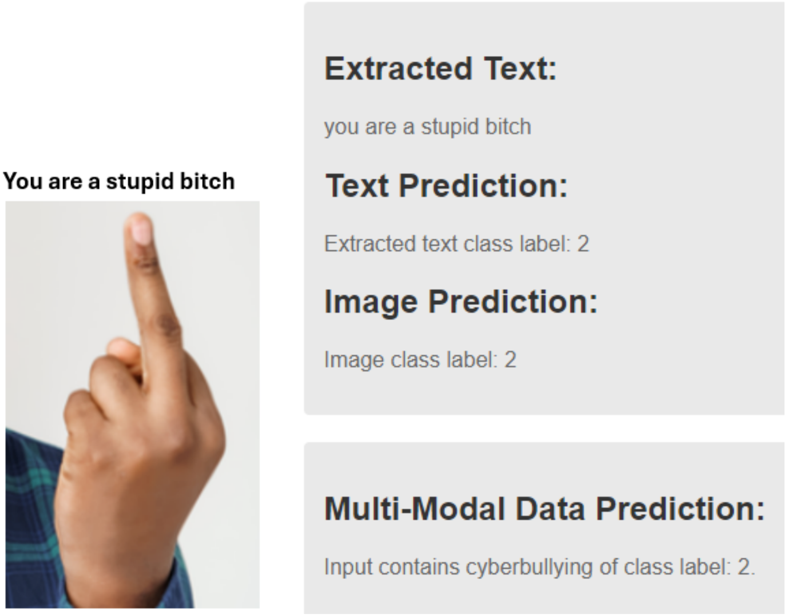
Multi-Modal Data Prediction:

Input contains cyberbullying. Extracted text class label: 3 and Image class label: 0

(a) (b)

Figure 6. Multi-modal Data Testing in GUI.

Figure 7a was used to test the multi-modal data. Figure 7b displayed the result of Figure 7a. The extracted image label was 2 because there was showing middle finger which was evidence of cyberbullying in it, and since the extracted text was offensive word in nature, class label: 2 was the extracted text label.



Extracted Text:

you are a stupid bitch

Text Prediction:

Extracted text class label: 2

Image Prediction:

Image class label: 2

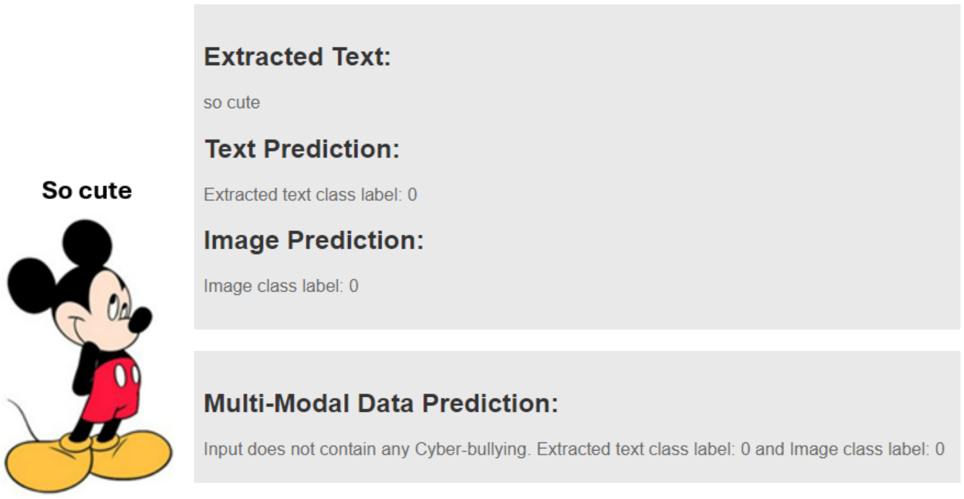
Multi-Modal Data Prediction:

Input contains cyberbullying of class label: 2.

(a) (b)

Figure 7. Offensive Based Multi-modal Data.

Figure 8a was tested as multi-modal data in our GUI, displaying an image without cyberbullying content. Consequently, the result shown in Figure 8b in the GUI indicates non-bullying text data extracted from the image, along with a non-bullying image label. Therefore, the class label for the given multi-modal data is 0.



Extracted Text:

so cute

Text Prediction:

Extracted text class label: 0

Image Prediction:

Image class label: 0

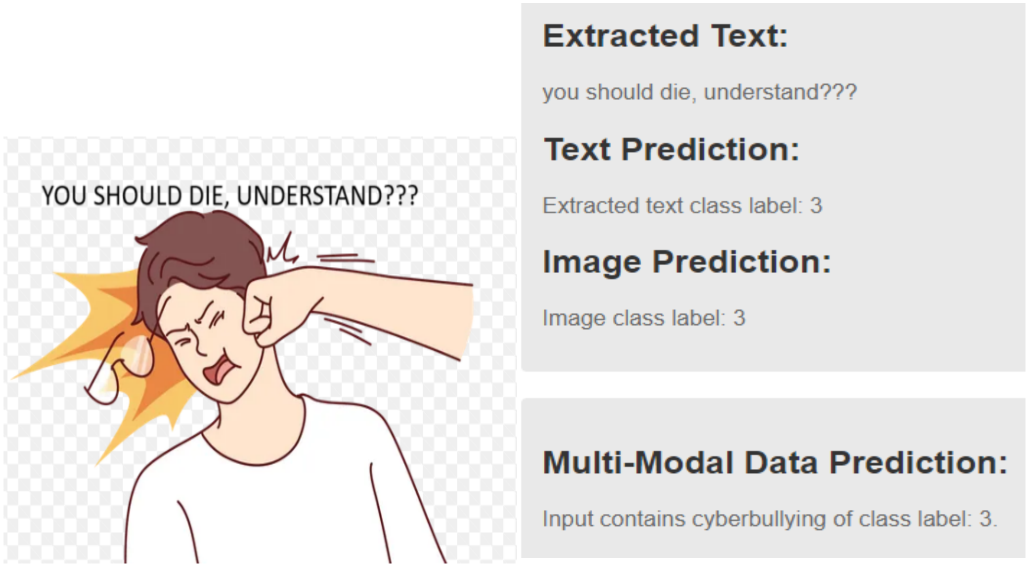
Multi-Modal Data Prediction:

Input does not contain any Cyber-bullying. Extracted text class label: 0 and Image class label: 0

(a) (b)

Figure 8. Non-bullying Based Multi-modal Data.

The Figure 9 showing the aggression based multi-modal data's experiment in our GUI. Figure 9a was displayed that someone was punching to a man's face, which was containing aggressive based cyberbulying, hence the extracted image label was 3, and as the extracted text contains aggressive based bullying language, the text label was class label: 3 which was showing in Figure 9b.



Extracted Text:

you should die, understand???

Text Prediction:

Extracted text class label: 3

Image Prediction:

Image class label: 3

Multi-Modal Data Prediction:

Input contains cyberbullying of class label: 3.

(a) (b)

Figure 9. Aggression Based Multi-modal Data.

This is how the GUI performed to classify multi-class cyberbullying for muti-modal data. Testing has been done on both Public-dataset and Private-dataset in order to confirm the results. The fusion module produced a decision for the multi-modal data and displayed the results for the extracted text and images.

5. Discussion

Table 5 compares our results with previous studies in classifying cyberbullying using various DL models and datasets. In the study by Hamza et al. [32], the authors used a model called RexNeXT-152 combined with Masked R-CNN and BERT on Dataset-1. Their model focused on identifying whether content was hateful or non-hateful and achieved an accuracy of 70.60%. In contrast, Maity et al. [21]

worked with Dataset-2, using a BERT-GRU model for text and ResNet for images. They looked at multiple tasks, including sarcasm identification, sentiment analysis, and emotion recognition. Their results showed accuracies of 59.72% for text and 59.39% for images. Similarly, Yue et al. [42] also worked on Dataset-2 with BERT and ResNet, combining text and image data for their tasks, and achieved a combined accuracy of 64.35%. In the study of Aggarwal et al. [43], the authors used a cross-lingual language model for text and ConvNet with attention for images, focusing on sarcasm identification, with accuracies of 63.83% for text and 62.91% for images.

Table 5. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

Literature Review	Dataset	Label	Accuracy	Used Model
[32]	Dataset-1	Hateful, Non-hateful	70.60%	RexNeXT-152-based Masked R-CNN, BERT
[21]	Dataset-2	Bullies, Attitude, Emotion, and Sarcasm Recognition	Text: 59.72%, Image: 59.39%	Text: BERT-GRU, Image: ResNet
[42]	Dataset-2	Bullies, Attitude, Emotion, and Sarcasm Recognition	64.35%	BERT, ResNet
[43]	Dataset-2	Sarcasm Identification	Text: 63.83%, Image: 62.91%	Text: Cross-lingual Language Model, Image: Self-regulated ConvNet + Lightweight Attention
Our Result	**Public-dataset (Dataset-1 + Dataset-2)**	Non-bullying, Defaming, Offensive, Aggressive	**99.24%**	**Hybrid (RoBERTa + ViT)**
	Private-dataset	Non-bullying, Defaming, Offensive, Aggressive	**96.1%**	**Hybrid (RoBERTa + ViT)**

In this study, we combined Dataset-1 and Dataset-2 to increase the dataset size and to improve the ability of the model to handle multiple classes (non-bullying, defaming, offensive, and aggressive). Using a hybrid (RoBERTa+ViT) model, we achieved a significantly higher accuracy of 99.24% on the Public-dataset. This shows a notable improvement compared to the other studies. By enlarging the dataset and using a more advanced hybrid approach, our results outperformed previous works in Public-dataset.

Furthermore, when comparing the results of the Public-dataset with the Private-dataset, it is observed that the accuracy on the Private-dataset is slightly lower, at 96.1%. This difference may be due to the specific characteristics of the Private-dataset, which could have different features or a smaller amount of data compared to the combined Public-dataset. However, despite this small drop, our hybrid model still outperformed the other approaches mentioned, highlighting its robustness and adaptability to different datasets, which is still much higher than the results achieved by other authors.

6. Conclusions

Cyberbullying on social media comes with serious consequences. It can lead to emotional distress, anxiety, depression, and even self-harm or suicide for victims. The anonymous nature of social platforms often makes it easier for bullies to target individuals, and the public nature of social media amplifies the impact. This makes cyberbullying a significant issue that needs to be addressed. DL models can help reduce cyberbullying by identifying harmful content across platforms. These models can analyze large amounts of data, including text, images, and memes, to spot abusive behavior of cyberbullying. By identifying different types of cyberbullying, these models can help social media companies take quick action, removing harmful posts or alerting moderators before serious damage is done. However, there are still some challenges in effectively identifying cyberbullying. One major issue is the lack of diverse and well-labeled datasets for multi-class classification for training these models. In this research, we utilized various DL models to classify cyberbullying into different types such as harassment, threats, or defamation, using both public and private text and visual data collected from multiple social media platforms.

The experimental results show that the hybrid DL models, LSTM+ResNet, LSTM+CNN, LSTM+ViT, GRU+ResNet, GRU+CNN, GRU+ViT, BERT+ResNet, BERT+CNN, BERT+ViT, DistiBERT+ResNet, DistiBERT+CNN, DistiBERT+ViT RoBERTa+ResNet, RoBERTa+CNN, RoBERTa+ViT achieved accuracies of 70.85%, 72.85%, 73.35%, 72.30%, 74.30%, 74.80%, 95.85%, 97.85%, 98.35%, 96.55%, 98.55%, 99.05%, 96.60%, 98.60%, and 99.24% in classifying cyberbullying using Public-dataset respectively. On the Private-dataset, the hybrid(RoBERTa+ViT) model scored an F1-score of 0.961 and achieved an accuracy of 96.1%.

Based on our results, we believe that, DL models such as hybrid(RoBERTa+ViT) models are well effective at classifying multiple types of cyberbullying for multi-modal data. In hybrid(RoBERTa+ViT) model, RoBERTa works well with text, producing nearly flawless results, whereas ViT excels at handling images. Furthermore, when these models are combined into a hybrid(RoBERTa+ViT) model, they perform even better at multi-class classifying cyberbullying in multi-modal content, such as memes.

6.1. Future Work

As directions for future work, following research initiatives will be taken into consideration:

- Use multi-label classification for representing the work more realistic. As a result, if a comment contains aggressive content with bullying, the result will be displayed for both aggressive and bullying classification type.
- In this study, we have focused only on English language data. Hence, as a future work, we will collect data on multi-languages for multi-class classification on multi-modal data, so that we can classify cyberbullying from multiple language, such as Bengali, Hindi, Urdu, and Norwegian.
- Use stickers and GIF data along with our existing data to classify multi-class cyberbullying. As a result, we will be able to handle all kinds of data from the comments sections of social media short videos.
- Develop different models such as Swin Transformers (Shifted Window Transformer), Multi-scale Vision Transformers from, and BLIP-V2 (Bootstrapping Language-Image Pre-training Version 2).

Author Contributions: Conceptualization, I.T. and V.N.; Methodology, I.T.; Software, I.T.; Validation, I.T.; Formal analysis, I.T. and V.N.; Investigation, I.T.; Resources, I.T.; Data curation, I.T.; Writing—original draft preparation, I.T.; Writing—review and editing, I.T. and V.N.; Visualization, I.T.; Supervision, V.N.; Project administration, I.T. and V.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available online at https://drive.google.com/drive/folders/1Rp5t0T49u_6h9Nxu5KPz_NP8N9t4mVYK?usp=drive_link.

Acknowledgments: This work is the result of a 60-ECTS master's thesis at Department of Science and Industry Systems, Faculty of Technology, Natural Sciences and Maritime Sciences, University of South-Eastern Norway, Kongsberg, Norway.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SM	Social Media
DL	Deep-Learning
CNN	Convolutional Neural Network
LSTM	Long Short-Time Memory
GRU	Gated Recurrent Unit
RoBERTa	Robustly Optimized BERT Pre-training Approach
ViT	Vision Transformer
ResNet	Residual Network
BERT	Bidirectional Encoder Representations from Transformers
DistilBERT	Distilled BERT
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
OCR	Optical Character Recognition
GUI	Graphical User Interface

References

1. Mayfield, A. What is social media **2008**.
2. 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.
3. Kaye, D.B.V.; Zeng, J.; Wikstrom, P. *TikTok: Creativity and culture in short video*; John Wiley & Sons, 2022.
4. 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.
5. 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.
6. 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**.
7. 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.
8. 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.
9. 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.
10. 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.
11. Dingming Yang, Yanrong Cui, Z.Y.; Yuan, H. Deep Learning Based Steel Pipe Weld Defect Detection. *Applied Artificial Intelligence* **2021**, *35*, 1237–1249. https://doi.org/10.1080/08839514.2021.1975391.
12. Mabrook S. Al-Rakhami, S.A.A.; Alawwad, A. Effective Skin Cancer Diagnosis Through Federated Learning and Deep Convolutional Neural Networks. *Applied Artificial Intelligence* **2024**, *38*, 2364145. https://doi.org/10.1080/08839514.2024.2364145.
13. Nunavath, V.; Goodwin, M. The Use of Artificial Intelligence in Disaster Management - A Systematic Literature Review. In Proceedings of the 2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), 2019, pp. 1–8.
14. Sang, S.; Li, L. A Stock Prediction Method Based on Heterogeneous Bidirectional LSTM. *Applied Sciences* **2024**, *14*, 9158.

15. Chandrasekaran, S.; Singh Pundir, A.K.; Lingaiah, T.B.; et al. Deep learning approaches for cyberbullying detection and classification on social media. *Computational Intelligence and Neuroscience* **2022**, *2022*.
16. Dadvar, M.; Eckert, K. Cyberbullying detection in social networks using deep learning based models. In Proceedings of the Big Data Analytics and Knowledge Discovery: 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 22. Springer, 2020, pp. 245–255.
17. Singh, N.K.; Singh, P.; Chand, S. Deep Learning based Methods for Cyberbullying Detection on Social Media. In Proceedings of the 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). IEEE, 2022, pp. 521–525.
18. Alotaibi, M.; Alotaibi, B.; Razaque, A. A multichannel deep learning framework for cyberbullying detection on social media. *Electronics* **2021**, *10*, 2664.
19. Faraj, A.; Utku, S. Comparative Analysis of Word Embeddings for Multiclass Cyberbullying Detection. *UHD Journal of Science and Technology* **2024**, *8*, 55–63.
20. Ahmadinejad, M.; Shahriar, N.; Fan, L. Self-Training for Cyberbully Detection: Achieving High Accuracy with a Balanced Multi-Class Dataset. PhD thesis, Faculty of Graduate Studies and Research, University of Regina, 2023.
21. Maity, K.; Jha, P.; Saha, S.; Bhattacharyya, P. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In Proceedings of the Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1739–1749.
22. Titli, S.R.; Paul, S. Automated Bengali abusive text classification: Using Deep Learning Techniques. In Proceedings of the 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS). IEEE, 2023, pp. 1–6.
23. Romim, N.; Ahmed, M.; Talukder, H.; Saiful Islam, M. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In Proceedings of the Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020. Springer, 2021, pp. 457–468.
24. Karim, M.R.; Dey, S.K.; Islam, T.; Shajalal, M.; Chakravarthi, B.R. Multimodal hate speech detection from bengali memes and texts. In Proceedings of the International Conference on Speech and Language Technologies for Low-resource Languages. Springer, 2022, pp. 293–308.
25. Haque, R.; Islam, N.; Tasneem, M.; Das, A.K. Multi-class sentiment classification on Bengali social media comments using machine learning. *International Journal of Cognitive Computing in Engineering* **2023**, *4*, 21–35.
26. Kumari, K.; Singh, J.P.; Dwivedi, Y.K.; Rana, N.P. Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization. *Future Generation Computer Systems* **2021**, *118*, 187–197.
27. Barse, S.; Bhagat, D.; Dhawale, K.; Solanke, Y.; Kurve, D. Cyber-Trolling Detection System. *Available at SSRN 4340372* **2023**.
28. Mollas, I.; Chrysopoulou, Z.; Karlos, S.; Tsoumakas, G. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems* **2022**, *8*, 4663–4678.
29. Hossain, E.; Sharif, O.; Hoque, M.M.; Dewan, M.A.A.; Siddique, N.; Hossain, M.A. Identification of Multilingual Offense and Troll from Social Media Memes Using Weighted Ensemble of Multimodal Features. *Journal of King Saud University-Computer and Information Sciences* **2022**, *34*, 6605–6623.
30. Van Hee, C.; Lefever, E.; Verhoeven, B.; Mennes, J.; Desmet, B.; De Pauw, G.; Daelemans, W.; Hoste, V. Detection and fine-grained classification of cyberbullying events. In Proceedings of the Proceedings of the international conference recent advances in natural language processing, 2015, pp. 672–680.
31. Van Hee, C.; Jacobs, G.; Emmery, C.; Desmet, B.; Lefever, E.; Verhoeven, B.; De Pauw, G.; Daelemans, W.; Hoste, V. Automatic detection of cyberbullying in social media text. *PloS one* **2018**, *13*, e0203794.
32. Hamza, A.; Javed, A.R.; Iqbal, F.; Yasin, A.; Srivastava, G.; Połap, D.; Gadekallu, T.R.; Jalil, Z. Multimodal Religiously Hateful Social Media Memes Classification based on Textual and Image Data. *ACM Transactions on Asian and Low-Resource Language Information Processing* **2023**.
33. Hasan, M.T.; Hossain, M.A.E.; Mukta, M.S.H.; Akter, A.; Ahmed, M.; Islam, S. A review on deep-learning-based cyberbullying detection. *Future Internet* **2023**, *15*, 179.
34. Dewani, A.; Memon, M.A.; Bhatti, S. Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data. *Journal of big data* **2021**, *8*, 160.

35. Ahsan, S.; Hossain, E.; Sharif, O.; Das, A.; Hoque, M.M.; Dewan, M. A Multimodal Framework to Detect Target Aware Aggression in Memes. In Proceedings of the Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 2487–2500.
36. Paciello, M.; D'Errico, F.; Saleri, G.; Lamponi, E. Online sexist meme and its effects on moral and emotional processes in social media. *Computers in human behavior* **2021**, *116*, 106655.
37. Sharma, S.; Alam, F.; Akhtar, M.S.; Dimitrov, D.; Martino, G.D.S.; Firooz, H.; Halevy, A.; Silvestri, F.; Nakov, P.; Chakraborty, T. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274* **2022**.
38. Pandeya, Y.R.; Lee, J. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications* **2021**, *80*, 2887–2905.
39. Gupta, P.; Gupta, H.; Sinha, A. Dsc iit-ism at semeval-2020 task 8: Bi-fusion techniques for deep meme emotion analysis. *arXiv preprint arXiv:2008.00825* **2020**.
40. Cai, Y.; Cai, H.; Wan, X. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In Proceedings of the Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 2506–2515.
41. Room, C. Confusion matrix. *Mach. Learn* **2019**, *6*, 27.
42. Yue, T.; Mao, R.; Wang, H.; Hu, Z.; Cambria, E. KnowleNet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion* **2023**, *100*, 101921.
43. Aggarwal, S.; Pandey, A.; Vishwakarma, D.K. Modelling Visual Semantics via Image Captioning to extract Enhanced Multi-Level Cross-Modal Semantic Incongruity Representation with Attention for Multimodal Sarcasm Detection. *arXiv preprint arXiv:2408.02595* **2024**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.