EasyChair Preprint

№ 15638

A Deep-Learning Based Approach for Multi-Class Cyberbullying Classification Using Social Media Text and Image Data

Israt Tabassum and Vimala Nunavath

January 6, 2025

A Deep-Learning Based Approach for Multi-class Cyberbullying Classification Using Social Media Text and Image Data

Israt Tabassum* and Vimala Nunavath

Department of Science and Industry Systems, University of South-Eastern Norway, Kongsberg, Norway
israt.tabassum34@gmail.com and vimala.nunavath@usn.no

Abstract. Social media sites like Facebook, Instagram, Twitter, LinkedIn, have become crucial for content creation and distribution, influencing business, politics, and personal relationships. Users often share their daily activities through pictures, posts, and videos, making short videos particularly popular due to their engaging format. However, social media posts frequently attract mixed comments, both positive and negative, and the negative comments can in some cases take the form of cyberbullying. To identify cyberbullying, a deep-learning approach was employed using two datasets: one self-collected and another public dataset. Nine deep-learning models were trained: ResNet-50, CNN and ViT for image data, and LSTM-2, GRU, RoBERTa, BERT, DistilBERT, and Hybrid (CNN+LSTM) model for textual data. The experimental results showed that the ViT model excelled in multi-class classification on public image data, achieving 99.5% accuracy and a F1-score of 0.995, while RoBERTa model outperformed other models on public textual data, with 99.2% accuracy and a F1-score of 0.992. For the private dataset, the RoBERTa model for text and ViT model for images were developed, with RoBERTa achieving a F1-score of 0.986 and 98.6% accuracy, and ViT obtaining an F1-score of 0.9319 and 93.20% accuracy. These results demonstrate the effectiveness of RoBERTa for text and Vision Transformer (ViT) for images in classifying cyberbullying, with RoBERTa delivering nearly perfect text classification and ViT excelling in image classification.

Keywords: Cyberbullying · Deep-learning · RoBERTa · ViT · DistilBERT · BERT · CNN · LSTM-2 · GRU · ResNet-50 · Multi-class classification · Social Media · Text data · Image data.

1 Introduction

Social media platforms, such as Facebook [1], Twitter [2], Instagram [3], and many others, have completely changed how individuals create, share, and interact with

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

one another in online communities [18]. A large number of comments are typically posted in the comment sections of social media platforms and these comments come in a variety of formats, including text, photos, audio, and video. Some people post negative comments or aggressive content on social media to insult others, which is called cyberbullying [27]. Constantly receiving negative comments can lead to severe psychological effects such as depression or suicide. It can also have a significant negative impact on an individual's physical and mental health by eroding their self-confidence [5]. The 2014 EU-Kids Online Report [15] shows that 20% of children aged 11 to 16 have experienced cyberbullying. Another quantitative research [27] claims that youths experience cyberbullying at a rate of 20-40%. As a result, it is vital to keep social media platforms secure and free of unpleasant interactions continue to attract millions of viewers globally [23].

There have been many attempts made to classify cyberbullying using deep-learning methods [9], largely because, deep-learning (DL) has made substantial contributions in various fields, such as healthcare [21] [20], the automotive industry [4], and retail [11]. Furthermore, several studies [9] have used deep-learning models to classify cyberbullying based on text or images. However, these classifications are often limited to binary classification [7] [12], and [14], [19], using traditional deep-learning models [1] [13], or are limited to age-sex-race based cyberbullying [19] instead of focusing actual types of cyberbullying such as aggression, harassment, and offensive [32][29][28][30]. More advanced techniques, particularly improved natural language processing capabilities, are needed, as current models often fail to capture subtleties of language, such as distinct types of bullying [9].

In this study, we explore the use of various deep-learning models especially Long Short-Term Memory with 2 layers (LSTM-2), Gated Recurrent Unit(GRU), Robustly Optimized BERT Pretraining Approach(RoBERTa), Bidirectional Encoder Representations from Transformers(BERT), Distilled Bidirectional Encoder Representations from Transformers(DistilBERT), Residual Network with 50 layers(ResNet-50), Convolutional Neural Network(CNN), the Vision Transformer (ViT), and Hybrid (CNN+LSTM) models to classify the multi-classes of cyberbullying based on textual and image data collected from various social media sites, as well as using publicly available dataset.

The rest of the paper is organized as follows. The related work is presented in Section 2. Section 3 outlines the system architecture that we proposed to carry out the experiments, while the data preprocessing of both the publicly available dataset and the acquired dataset is presented in Section 4. Section 5 presents and discusses the experimental results obtained using the two datasets. Finally, the conclusion and future research directions are provided in Section 6.

2 Related Work

Researchers have proposed and applied various deep-learning models for binary classification of cyberbullying using text or image data [7] [12], and [14]. However,

in this section, we present the state-of-the-art related to applying deep-learning for classifying multi-classes of cyberbullying using text and image data.

In [19], the researchers goal was to identify the multi-label hate speech using deep-learning models. To achieve the goal, the researchers used "ETHOS" (multi-label hate speech detection dataset) which includes text data from YouTube and Reddit comments. For multi-label classification, the researchers employed BiLSTM model and achieved an accuracy of gender (70.34%), race (75.97%), national origin (67.88%), disability (69.64%), religion (71.65%), sexual orientation (89.83%), violence (50.86%), and directed vs. generalized (55.28%).

In [7], the researchers classified multi-modal data consisted of religiously abusive memes using deep leraning models. The used dataset contained textual and image data of approximately 2000 meme images from social media platforms including several social media platforms including Twitter, Instagram, Facebook, and Reddit called *religiously hateful memes dataset*. TThe study uses ResNeXT-152, a type of Convolutional Neural Network (CNN), to extract visual features from masked image regions. For the text part, it uses BERT (a model that reads text) to encode the words. These visual and text features are combined early in the model using an early fusion module. From the experiments, the model ResNeXT-152 + BERT (uncased) early fusion model for processing both image and text data together obtained an accuracy rate of 70.60%.

In [17], the authors developed a multitask deep-learning framework for the identification of cyberbullying, such as sentiment, sarcasm and emotion aware cyberbullying referred as "MultiBully" from multi-modal memes. In their study, the authors collected images and memes from *Twitter* and *Reddit* social site's memes, resulting in around 5854 data. For the purpose of identifying cyberbullying, they achieved accuracy of 59.72% for textual data using BERT-GRU, and a fully connected layer, and 59.39% for image data using ResNet-50.

In [1], the researchers worked on the dataset that was used in Maity *et al.* [17] for multi-modal sarcasm identification. In order to extract an improved multi-level cross-modal semantic incongruity representation with consideration for multi-modal sarcasm identification, their research focuses on modeling visual semantics through image captioning. For the textual data, they have got 63.83% accuracy by using Cross-lingual language model, and using the Self-regulated ConvNet + Lightweight Attention model for image data, they have achieved 62.91% accuracy.

In [13], the authors presented a model to categorize posts from social media platforms such as Facebook, Twitter, and Instagram using a Convolutional Neural Network (CNN) in conjunction with Binary Particle Swarm Optimization (BPSO). Posts with both text and images were divided into three categories by the model: non-aggressive, medium-aggressive, and high-aggressive. They used a three-layered CNN to extract textual data and a pre-trained VGG-16 model to extract visual features from the photos. The BPSO algorithm was used to optimize the hybrid feature set, which combined text and image information, and the improved model obtained a weighted F1-Score of 0.74.

Although various scholars have worked on cyberbullying classification using various deep-learning algorithms [13] [19] [26] [8] [3] [17] [10], none of the above studies have addressed the multi-class classification of cyberbullying (i.e., Non-bully, Defaming, Offensive, and Aggressive) using text and image data. Therefore, in this research, we will explore the effectiveness of various deep-learning models for multi-class classification of cyberbullying.

3 System Architecture

For multi-class cyberbullying classification, the overall system architecture for both public and private datasets is proposed and depicted in Figure 1. The system begins by taking input i.e., text, and image data. In the proposed system, all steps are common except the used deep-learning (DL) models and type of data used.

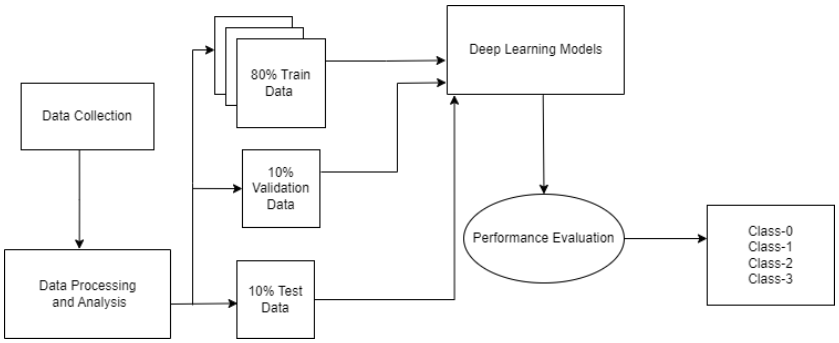

Fig. 1. The proposed System Architecture[24].

The second step was data preprocessing, and the detailed steps can be found in subsection 4.2. After performing data preprocessing and feature extraction, the next step was to train and test the DL models using both public and private datasets. To train the DL models, 80% of the data was used for training on both textual and image data. The remaining 20% of the data was split equally: 10% for validation, which was used to fine-tune model hyperparameters and prevent over-fitting, and 10% for testing, providing an unbiased evaluation of the model's performance.

For the text data of the public dataset, six different deep-learning models were developed: *LSTM-2, Hybrid model(CNN+LSTM), GRU, BERT, DistilBERT, and RoBERTa.* Additionally, for image data from the public dataset, DL models such as ResNet-50, CNN, and ViT, as depicted in subsection 4.4, were employed. For the private dataset, a RoBERTa model was employed for text data, while ViT model was developed for image data. The final step was to evaluate the performance of the employed models. The performance metrics used were [16]: Accuracy, F1-score, Precision, and Recall.

4 Datasets and Data Pre-processing

This section describes the used datasets, different steps considered for data pre-processing, feature extraction techniques, employed DL models, and hyperparameter tuning setup.

4.1 Public and Private Datasets

Two datasets were used in this research. The first dataset was public dataset, and the second dataset was private dataset (i.e., self-collected dataset).

In the literature, two researchers [7] and [17] independently collected cyberbullying data from various open social media sources. These datasets are called as the *Religious hateful memes* dataset (i.e, dataset-1) and the *multi-modal sarcasm detection* dataset (i.e, dataset-2). The first dataset contained 2000 data of both text and image data, while the second dataset contained 5,854 samples, both consisting of text and image data. In this research, the two datasets were downloaded and combined into one to create a larger dataset, and referred to as the *Public Dataset*.

In addition to the public data, we used another dataset called the *Private dataset*, which was collected by the author of this research. This dataset comprised approximately twelve thousand textual samples and around one thousand image samples, including images and memes. The dataset was sourced from platforms such as Facebook, Instagram, YouTube[4], and TikTok[5], specifically from comments on short videos. The dataset is made available on GitHub[6]. Comments from the aforementioned platforms were extracted using various tools: APIFY[7] for Facebook and YouTube short video comments, TKCommentExport[8] for TikTok's comments, IGCommentExporter[9] tool for Instagram reels comments. The dataset includes text, images, and memes related to cyberbullying.

4.2 Data Pre-processing and Analysis

After collecting the dataset, then next step was to process and analyse the data. The techniques used for pre-processing were as follows:

Step 1 - Extracting Text and Images from Memes: Although the public dataset has both text and image data separately, the private dataset includes memes data along with text and images data. In this research, our goal was to classify different classes using only the text and image data. Therefore, we

[4] https://www.youtube.com/
[5] https://www.tiktok.com/
[6] https://github.com/israt-tabassum/cyberbullying-classification-private-data
[7] https://apify.com/
[8] https://tkcommentexport.extensionsbox.com/
[9] https://chromewebstore.google.com/detail/igcommentexporter-export/ehaaocefdhppmemaaeedemaokjooldgm

extracted the text and image data from the available memes. To achieve this, we used a tool called *Tesseract-OCR*, which helps reading text from memes. To integrate this in Python, we installed a package called *Pytesseract*[10]. Before extracting the text, we improved the quality of the meme images using the *OpenCV*[11] library.

Additionally, we employed three main methods to enhance the quality of the meme images for text extraction. The used methods were: *bilateral filtering* to reduce noise while keeping edges clear, *converting the image to grayscale*, and *applying thresholding* to make the text stand out from the background. After extracting the text, we stored it in a structured format in a folder called "text_data", separate from the original image data in the private dataset.

For extracting only images from memes, we read, and resized the images to a consistent size while enhancing their features using various techniques available in the OpenCV library, such as *cv2.imread(), cv2.resize(), cv2.cvtColor(), and cv2.bilateralFilter()*. Once the images were prepared, we stored them in a separate folder called "image_data," for further analysis and classification tasks.

Step 2 - Data Categorization: As mentioned above, both public and private datasets contain text and image data. The text data from these two datasets has been classified into the following four categories for cyberbullying. The categorization below was based on the studies mentioned in [9] [28] [29].

- Non-Bullying (class-0): This category includes data that does not include any content that is insulting, defamatory, offensive, or that uses threatening or aggressive language [9].
- Defaming Cyberbullying (class-1): Defaming cyberbullying refers to behaviors in which individuals insult or defame another person. This type of cyberbullying is specifically targeted at damaging an individual's reputation and self-worth [28] [29].
- Offensive Language Cyberbullying (class-2): This category includes situations where individuals use derogatory language to target someone [28] [29].
- Aggressive Cyberbullying (class-3): This refers to the act of making direct threats, displaying violent behaviour, and engaging in abusive conduct towards an individual [6] [29].

Similarly, according to the previous studies [2] [22], and [25], the images from the public and private datasets were categorized into the following four different categories.

- Non-Bullying (class-0): The image contains normal content, which does not contains any defaming, sexual, offensive, or aggressive content.
- Defaming (class-1): The image contains sexual or nudity content.

[10] `https://pypi.org/project/pytesseract/`
[11] `https://opencv.org/`

- Offensive (class-2): For the private dataset, showing a middle finger, and mixing other creatures' faces into people's faces have been categorized as class-2. Due to the absence of content that mixes other creatures faces into people's faces in the public dataset, we refer to only *showing a middle finger* in class-2 as Offensive content.
- Aggressive (class-3): Beating someone or showing weapon to someone was considered as class-3.

Step 3 - Data Cleaning: For the text data, we used *regular expressions (regex)* to remove noise by eliminating unwanted characters and special symbols. We applied the *lower()* function to convert all text to lowercase for consistency. To enhance meaningful analysis, we removed *stop words* that do not contribute significantly to the text's meaning. We utilized *stemming* to reduce words to their base forms, treating variations like "running" and "run" as equivalent. Additionally, we applied *lemmatization* to further reduce words to their dictionary forms, such as changing "better" to "good." Finally, we employed *dropna()* to remove null values and *drop_ duplicates()* to eliminate duplicate text entries.

For the image data, we applied *bilateral filtering* to reduce visual noise and enhance clarity. We converted images to *grayscale* to simplify data processing by focusing on essential features. We used *thresholding* to binarize the images, highlighting important features against the background. Additionally, we removed duplicate entries with *drop_ duplicates()* and manually eliminated irrelevant image data.

Step 4 - Data Augmentation and Sampling: For text data, we used synonym replacement and text paraphrasing to create different versions of the text. We also ensured that each class had the same number of examples to maintain a balanced class distribution, particularly given the four different classes.

For the image data, we applied rotation, flipping, and cropping to increase images variability. Rotation changes the angle of the images, flipping alters their orientation i.e., horizontally or vertically, and cropping focuses on different parts of the images. Additionally, we ensured that the number of images from each class was equal to maintain balance and prevent the model from favoring one class over another.

4.3 Feature Extraction

For text data, we utilized tokenization and word embeddings (specifically, *Word2Vec*) as feature extraction techniques for both public and private datasets. We chose Word2Vec technique because it captures semantic relationships between words, enhancing the performance of sequential models like hybrid (CNN+LSTM), LSTM-2 and GRU, as well as transformer-based models like BERT, DistilBERT and RoBERTa.

For image data, we employed different feature extraction techniques suited for each model. For Convolutional Neural Networks (CNN), we applied *CNN*

features, allowing the model to automatically learn patterns such as edges and shapes from images. For LSTM-2, we utilized *deep-learning pre-trained models* through transfer learning, which enables us to leverage knowledge from previously trained models to improve classification. For the Vision Transformer (ViT), we employed the *patch embedding* technique, where images are divided into smaller patches and processed as sequences.

4.4 Employed deep-learning Models

Table 1 shows the model architectures we used for working with text data, including details for models such as Hybrid (CNN + LSTM), LSTM-2, GRU, BERT, DistilBERT, and RoBERTa. For these text models, we start with text tokens as the input. The middle layers vary according to the specific deep-learning model. All models end with a dense layer that used softmax to classify the text into four different classes.

Table 1. Employed Model Architectures for Text Data

Model	**Input Layer**	**Middle Layers**	**Output Layer**
Hybrid (CNN + LSTM)	Text tokens as embeddings	CNN layers + LSTM layers (e.g., 1-2 layers each)	Dense layer with softmax (4 classes)
LSTM-2	Text tokens as embeddings	2 LSTM layers	Dense layer with softmax (4 classes)
GRU	Text tokens as embeddings	2 GRU layers	Dense layer with softmax (4 classes)
BERT	Text tokens with positional embeddings	12 Transformer layers (base version)	Dense layer with softmax (4 classes)
DistilBERT	Text tokens with positional embeddings	6 Transformer layers	Dense layer with softmax (4 classes)
RoBERTa	Text tokens with positional embeddings	12 Transformer layers	Dense layer with softmax (4 classes)

In contrast, Table 2 describes the model architectures employed for image data, which include ResNet-50, CNN, and ViT deep-learning models. For images, the input typically consists of RGB images sized 224x224 pixels. In the middle layers, ResNet-50 utilizes many convolutional layers with special connections, CNN employs a mix of convolutional and activation layers, and ViT incorporates twelve Transformer layers with image patches. Each model ends with a dense layer and softmax function to categorize images into 4 classes. We used the same model architectures for both our public and private datasets.

Table 2. Employed Model Architectures for Image Data

Model	Input Layer	Middle Layers	Output Layer
ResNet-50	Images (e.g., 224x224x3 RGB)	50 Convolutional layers (with residual connections)	Dense layer with softmax (4 classes)
CNN	Images (e.g., 224x224x3 RGB)	1 Convolutional layer + 1 Activation layer + Additional layers (e.g., Pooling, Convolutional)	Dense layer with softmax (4 classes)
ViT (Vision Transformer)	Images divided into patches (e.g., 16x16)	12 layers Transformer layers	Dense layer with softmax (4 classes)

4.5 Hyperparameter Tuning

To improve the model performance for text classification on both datasets and image classification on public dataset, we used a common hyperparameter tuning strategy for all deep-learning models. The *Adam* optimizer was used to tune hyperparameters based on a validation dataset with 20 epochs, batch size of 20 and learning rate of *0.00002*. To avoid overfitting, we implemented an early stopping strategy with a patience of three epochs. We utilized *SparseCategoricalCrossentropy* as the loss function for text dataset and *CrossEntropyLoss* as the loss function for image data of public dataset for multi-class classification using integer labels.

For the private dataset's image data, we conducted ten trials of random search hyperparameter tuning with varying batch sizes (8, 16, 20, 32, 64) and learning rates (0.00001, 0.00005, 0.0001, 0.0005, 0.001). The best model configuration, selected based on validation loss and accuracy, was evaluated using *nn.CrossEntropyLoss* and early stopping with a patience of three epochs to avoid overfitting. We used accuracy, F1-score, precision, and recall metrics to evaluate the models' performance.

5 Experimental Results and Discussion

In this section, we present the obtained results for both text and image data. The models were built using the Keras deep-learning library in Python. TensorFlow was used to train text data, while PyTorch was used for image data due to its flexibility and performance.

5.1 Experimental Results on The Public Dataset

To classify cyberbullying using the text data from the public dataset, six different deep-learning models were employed (see Table 1). For the image data from public data, three different deep-learning models were utilized (see Table 2). The results obtained from these models are presented in Table 3.

Table 3. The Obtained Results on Public Textual Data.

Model Name	Test Accuracy	Recall	F1-Score	Precision
Hybrid (CNN+LSTM)	0.490	0.492	0.363	0.316
LSTM-2	0.477	0.48	0.39	0.39
GRU	0.506	0.49	0.37	0.32
BERT	0.977	0.977	0.977	0.977
DistilBERT	0.991	0.991	0.991	0.991
RoBERTa	**0.992**	**0.992**	**0.992**	**0.992**

From Table 3, it is evident that the RoBERTa model performed best for the text data classification, achieving a high accuracy of 0.992 with an F1-score of 0.992, outperforming the other models. This superior performance of the RoBERTa model can be attributed to its extensive pre-training on large text corpora, which enables it to deeply understand linguistic patterns associated with cyberbullying.

For the multi-class classification of cyberbullying using image data, as shown in Table:4, the Vision Transformer (ViT) model achieved the highest accuracy of 0.995 and an F1-score of 0.995, outperforming the other models. This strong performance can be explained by ViT's ability to capture intricate visual features, making it particularly effective in recognizing offensive imagery.

Table 4. The Obtained Results on Public Image Data.

Model Name	Test Accuracy	Precision	Recall	F1-Score
ResNet-50	0.94	0.94	0.94	0.94
CNN	0.98	0.98	0.98	0.98
ViT	**0.995**	**0.995**	**0.995**	**0.995**

5.2 Experimental Results on The Private Dataset

To classify cyberbullying using private data, we used the RoBERTa model for text data and the ViT model for image data. The RoBERTa model was selected based on its superior performance on the public dataset (as shown in Table 3), where it outperformed other models in capturing linguistic nuances of cyberbullying. Similarly, ViT was chosen for image data because it achieved better results than CNN and ResNet-50 in the public dataset (as seen in Table 4), showcasing its strength in extracting visual features.

When these models were employed to the private dataset, RoBERTa achieved 98.2% accuracy with an F1-score of 0.982 for text data, while ViT obtained 93.2% accuracy and an F1-score of 0.932 for image data, as shown in Table 5. The high accuracy of RoBERTa on the private dataset suggests that the model's extensive pre-training on large text corpora allows it to generalize well across

Table 5. The Obtained Results Using Private Text and Image Data

Model Name	Accuracy	Recall	F1-Score	Precision
RoBERTa for Text Data	0.982	0.982	0.982	0.982
ViT for Image Data	0.932	0.932	0.932	0.933

different datasets, effectively identifying complex language patterns related to cyberbullying. Although, ViT achieved slightly lower accuracy on the private dataset compared to the public dataset, its performance is still impressive given that private datasets typically have more variability and noise. ViT's ability to handle diverse and potentially less structured visual data underscores its robustness in real-world scenarios.

Table 6. The Comparison of Existing Literature with Our Obtained Results

LR	Dataset	Used DL Models	Categories	Accuracy
[7]	Dataset-1	RexNeXT-152-based Masked R-CNN, BERT	Hateful, non-hateful	70.60%
[17]	Dataset-2	Text: BERT-GRU, Image: ResNet-50	Sarcasm detection, sentiment analysis, recognition of emotions	Text: 59.72% and Image: 59.39%
[31]	Dataset-2	BERT, ResNet-50	Sarcasm detection, sentiment analysis, recognition of emotions	Text and Image together 64.35%
[1]	Dataset-2	Text: Cross-lingual language model, Image: Self-regulated ConvNet + Lightweight Attention	Sarcasm detection	Text: 63.83% and Image: 62.91%
Our Result	**Public Dataset**	**Text: RoBERTa, Image: ViT**	Non-bullying, defaming, offensive, aggressive	**Text: 99.2%, Image: 99.5%**
	Private Dataset	**Text: RoBERTa, Image: ViT**	Non-bullying, defaming, offensive, aggressive	**Text: 98.2%, Image: 93.2%**

If we compare our results (see Table 6) with the existing literature (LR), we observe that the RoBERTa model achieved 99.2% accuracy on the text data from the public dataset, and the ViT model obtained 99.5% accuracy on the image data from the same dataset. These results significantly outperform the

models reported in the literature, which used the same public datasets. Furthermore, our approach not only surpasses existing works on public datasets but also shows strong performance on private dataset. Specifically, the RoBERTa model achieved 98.2% accuracy using text data, while the ViT model attained 93.2% accuracy using image data in classifying cyberbullying multi-classes, including non-bullying, defaming, offensive, and aggressive types.

These results not only exceeds the benchmarks reported in previous research but also underscore the growing effectiveness of transformer-based models like RoBERTa and ViT in cyberbullying classification. Our findings align with recent trends in deep-learning, demonstrating that transformers excel both text and image classification tasks.

6 Conclusion

Social media platforms like Facebook, Instagram, and Twitter have become central to content creation and interaction, but they also present significant challenges such as rise of the cyberbullying. To address this issue, deep-learning models have been employed to classify cyberbullying content in both text and image data. However, most existing work focuses on binary, multi-task, or multi-label classification, with limited emphasis on multi-class classification of cyberbullying. In this research, we employed several deep-learning models for multi-class classification of cyberbullying using both public and private textual and image data collected from various social media sources. For classifying cyberbullying using public text data, deep-learning models such as Hybrid (CNN+LSTM), LSTM-2, GRU, BERT, DistilBERT, and RoBERTa were evaluated, achieving accuracies of 49%, 47%, 50%, 97.7%, 99.1% and 99.2% respectively. For public image data, CNN, ResNet-50 and Vision Transformer (ViT) models were employed, attaining accuracies of 94%, 98%, and 99.5% respectively. On the private dataset, the RoBERTa model achieved an F1-score of 0.982 and an accuracy of 98.2% for text data, while the Vision Transformer (ViT) model obtained an F1-Score of 0.932 and an accuracy of 93.2% for image data. When compared to existing literature (see Table 6), our models, particularly RoBERTa and ViT, demonstrate superior performance on public datasets, outperforming previously reported results in studies [7], [17], [31], and [1].

As a future work, we plan to carry out the following research: collecting additional data such as multi-modal (memes) data, GIFs, video and audio for the multi-class classification of cyberbullying. Additionally, we aim to try different models such as Swin Transformers, Multiscale Vision Transformers, and BLIP-V2 models. Also, plan to collect text from different languages, including Bengali, Hindi, Urdu, and Norwegian, for the multi-class classification of cyberbullying.

Acknowledgments. This work is the result of a 60-ECTS master's thesis in Computer Science at IRI, USN, Kongsberg.

References

1. 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)
2. 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 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2487–2500 (2024)
3. Barse, S., Bhagat, D., Dhawale, K., Solanke, Y., Kurve, D.: Cyber-trolling detection system. Available at SSRN 4340372 (2023)
4. Chen, C.: Deep learning for automobile predictive maintenance under Industry 4.0. Ph.D. thesis, Cardiff University (2020)
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: 2020 4th International Conference on Vocational Education and Training (ICOVET). pp. 30–35. IEEE (2020)
6. Dewani, A., Memon, M.A., Bhatti, S.: Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for roman urdu data. Journal of big data **8**(1), 160 (2021)
7. 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)
8. 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 **4**, 21–35 (2023)
9. 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 **15**(5), 179 (2023)
10. 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 **34**(9), 6605–6623 (2022)
11. Kaneko, Y., Yada, K.: A deep learning approach for the prediction of retail store sales. In: 2016 IEEE 16th International conference on data mining workshops (ICDMW). pp. 531–537. IEEE (2016)
12. Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Benchmarking aggression identification in social media. In: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018). pp. 1–11 (2018)
13. 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 **118**, 187–197 (2021)
14. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
15. 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)

16. Magboo, V.P.C., Magboo, M.S.A.: Machine learning classifiers on breast cancer recurrences. Procedia Computer Science **192**, 2742–2752 (2021)
17. 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 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1739–1749 (2022)
18. Mayfield, A.: What is social media (2008)
19. Mollas, I., Chrysopoulou, Z., Karlos, S., Tsoumakas, G.: Ethos: a multi-label hate speech detection dataset. Complex & Intelligent Systems **8**(6), 4663–4678 (2022)
20. Nunavath, V., Goodwin, M., Fidje, J.T., Moe, C.E.: Deep neural networks for prediction of exacerbations of patients with chronic obstructive pulmonary disease. In: Engineering Applications of Neural Networks: 19th International Conference, EANN 2018, Bristol, UK, September 3-5, 2018, Proceedings 19. pp. 217–228. Springer (2018)
21. Nunavath, V., Johansen, S., Johannessen, T.S., Jiao, L., Hansen, B.H., Berntsen, S., Goodwin, M.: Deep learning for classifying physical activities from accelerometer data. Sensors **21**(16), 5564 (2021)
22. 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 **116**, 106655 (2021)
23. Qiu, J., Moh, M., Moh, T.S.: Multi-modal detection of cyberbullying on twitter. In: Proceedings of the 2022 ACM Southeast Conference. pp. 9–16 (2022)
24. Reichert, J.R., Kristensen, K.L., Mukkamala, R.R., Vatrapu, R.: A supervised machine learning study of online discussion forums about type-2 diabetes. In: 2017 IEEE 19Th International Conference on E-health Networking, Applications and Services (Healthcom). pp. 1–7. IEEE (2017)
25. 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)
26. Titli, S.R., Paul, S.: Automated bengali abusive text classification: Using deep learning techniques. In: 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS). pp. 1–6. IEEE (2023)
27. Tokunaga, R.S.: Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in human behavior **26**(3), 277–287 (2010)
28. 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 **13**(10), e0203794 (2018)
29. 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 international conference recent advances in natural language processing. pp. 672–680 (2015)
30. Van Houdt, G., Mosquera, C., Nápoles, G.: A review on the long short-term memory model. Artificial Intelligence Review **53**(8), 5929–5955 (2020)
31. Yue, T., Mao, R., Wang, H., Hu, Z., Cambria, E.: Knowlenet: Knowledge fusion network for multimodal sarcasm detection. Information Fusion **100**, 101921 (2023)
32. Van der Zwaan, J., Dignum, V., Jonker, C.: Simulating peer support for victims of cyberbullying. In: Proceedings of the 22st Benelux conference on artificial intelligence (BNAIC 2010) (2010)