International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 10 | Oct - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

AI Powered Cyber Bullying Detection: Enhancing Safety in Digital Communities

Srinidhi Pudipeddi[1], Shriya Aishani Rachakonda[2], Dr. Alagiriswamy A A[3]

[1] *Department of Computing Technologies and SRM Institute of Science and Technology*
[2] *Department of Computational Intelligence and SRM Institute of Science and Technology*
[3] *Department of Physics & Nanotechnology and SRM Institute of Science and Technology*

Abstract - Cyberbullying has a profound effect on its victims, and each person's reaction is unique. This variation makes it difficult to find trustworthy cyberbullying content. There is less emphasis on image-based cyberbullying identification than there is on text-based methods in certain research. The goal of this effort is to improve a model that stops hazardous image-based content from being shared on social media. We provide an automated method that recognizes dangerous cyber-symbol images from social media networks using switch learning. Images used in cyberbullying are analyzed using transfer learning to find contextual information. Two photo datasets—one with and one without photos of cyberbullying—are used in our testing. These datasets are beneficial for additional study in this field. Since finding a useful model to identify bullying photos is difficult, we investigate deep learning (DL) as well as transfer learning.

Key Words: Cyberbullying, Data Processing, Data Collection, MoblieNet V2 Algorithm, MobilenetV2

1.INTRODUCTION

Internet 2.0 changed the way people communicated with each other, changing friendships and relationships and putting users at danger of cyberbullying. To address this problem, a multimodal strategy that incorporates automated detection and prevention methods is necessary. Cyberbullying, which mostly targets teenagers and manifests in a variety of destructive forms, is nevertheless an ongoing problem despite substantial studies. Pew Research Centre (2014) found that 40% of internet users had encountered online harassment. The goal of this thesis is to create a prediction algorithm that will warn viewers to stay away from bullyable photographs, which are edited or receive offensive remarks. By exploiting the popularity and high quality of images on Instagram, we use machine learning to identify patterns in manually labelled photos, allowing us to research trends in teen cyberbullying.

Cyberbullying is a type of digital abuse in which people are threatened, harassed, or singled out via technology. Doxing, disseminating objectionable content, and online harassment are examples of common strategies. Because of their anonymity, perpetrators are frequently bolder, and victims' psychological effects are worsened, increasing feelings of fear, despair, and even suicide thoughts. Particularly at risk are marginalized groups. Strict platform policies and educational initiatives are crucial in the fight against cyberbullying. The goal of this thesis is to create a machine learning predictive approach that can recognize and recommend not to post photographs that could be bullied based on textual comments. Through the analysis of Instagram data, the project aims to improve online safety—especially for teenagers.

Adolescents' mental health is seriously threatened by cyberbullying, which can result in self-harm, sadness, and anxiety. The detection and prevention techniques used today are insufficient. The goal of this project is to create a predictive algorithm that can recognize "bullyable" photos on the teen-focused Instagram platform. In order to forecast probable cyberbullying, the program examines the metadata and content of images. The algorithm increases kids' online safety without restricting their freedom of speech by assisting users in identifying images that pose a risk. This allows teens to make more educated judgements about what they post. This project uses artificial intelligence (AI) to make the internet a safer place and may spread to other social media sites in order to completely eradicate cyberbullying.

The Cyberbullying Prevention Project uses Instagram data and machine learning algorithms to combat cyberbullying on social media, especially among young people. The goal is to create a predictive method that can recognise photographs that are susceptible to damage based on the wording of comments. This would allow for interventions such as warnings or suggestions to stop damaging posts. Rigid analysis guarantees accuracy even in the face of obstacles like changing cyberbullying methods and poor data quality. By fostering an informed social media environment, the project aims to provide users the ability to make safer judgements. Through the utilisation of prominent social media platforms and continuous investigation, the aim is to significantly reduce cyberbullying.

2. THEORETICAL FRAMEWORK

With an emphasis on Instagram, the Cyberbullying Prevention Project uses a comprehensive strategy to end image-based cyberbullying on social media. The system processes and analyses data using machine learning, namely convolutional neural networks and transfer learning models, to find patterns in cyberbullying. The Flask framework is used to streamline the process, which entails gathering data, preprocessing, training the model, and evaluation. A use case diagram is included in the architecture to show how users interact with the system and how it functions. In order to guarantee accuracy and dependability, every step of the process—from data preparation to prediction—is optimized with the goal of eliminating damaging posts before they happen and fostering a safer social media environment.

During the model implementation phase, the system is trained on the annotated dataset using convolutional neural networks (CNNs) and transfer learning models. Convolutional neural networks (CNNs) excel in image analysis because they possess the capability to automatically extract and acquire hierarchical features from visual data. Transfer learning improves model performance by utilizing pre-trained networks, enabling the system to utilize existing knowledge and apply it to the unique goal of identifying patterns that indicate cyberbullying.

The training method entails refining these models using the dataset, which comprises photos with annotations that emphasize traits associated with cyberbullying. Through the use of transfer learning, the system is able to swiftly adjust to the subtleties of the fresh dataset, hence enhancing its precision and effectiveness in recognizing pertinent patterns.

3. METHODOLOGY

3.1 Objective

The development of a model that may effectively combat image-based cyberbullying on social media platforms is the primary objective of this research. In the first method, the model is constructed through the utilization of a convolutional neural network (CNN) that is founded on deep learning capabilities. When it comes to identifying patterns that are connected with cyberbullying in photographs, this basic model offers a solid beginning point. After that, the research integrates transfer learning models, which make use of pre-trained networks in order to improve performance.

As a result of the experimental findings that indicate higher accuracy across a variety of hyper-parameter settings, transfer learning is shown to be the preferable alternative for solving this problem. In the best-case scenario, the suggested model has an accuracy that is sufficiently high, which indicates that it is able to properly recognize and categorize the majority of posts that involve cyberbullying. This new development highlights the model's potential for considerably reducing instances of cyberbullying on social media platforms, hence helping to the creation of safer online environments through the use of improved processes for image classification.

3.2 Design

A multi-pronged strategy is utilized in the creation of the system for the classification of cyberbullying images. This strategy is intended to identify and address harmful content distributed on social media platforms. Beginning with the collection and preprocessing of a broad dataset of annotated photos, which includes scaling, normalization, and data augmentation in order to improve the resilience of the model, the process begins. Deep learning approaches are used to construct the core model. These techniques begin with Convolutional Neural Networks (CNNs), and then proceed to apply transfer learning with pre-trained networks such as VGG16 or ResNet.

During training, hyper-parameters are adjusted to optimize performance, and a separate test dataset is used to validate the model and ensure accurate identification of cyberbullying. After being validated, the model is implemented in a server or cloud environment, where it interacts with social media networks to perform real-time image analysis. A user interface (UI) is created to oversee and evaluate identified occurrences, offering administrators with practical and useful information.

A feedback loop enables the ongoing improvement of the model by incorporating new data and user feedback, guaranteeing adaptability to evolving patterns. The user interface presents organized photographs and in-depth analysis, simplifying the review procedure and facilitating prompt intervention for effective oversight, hence fostering a more secure online environment.

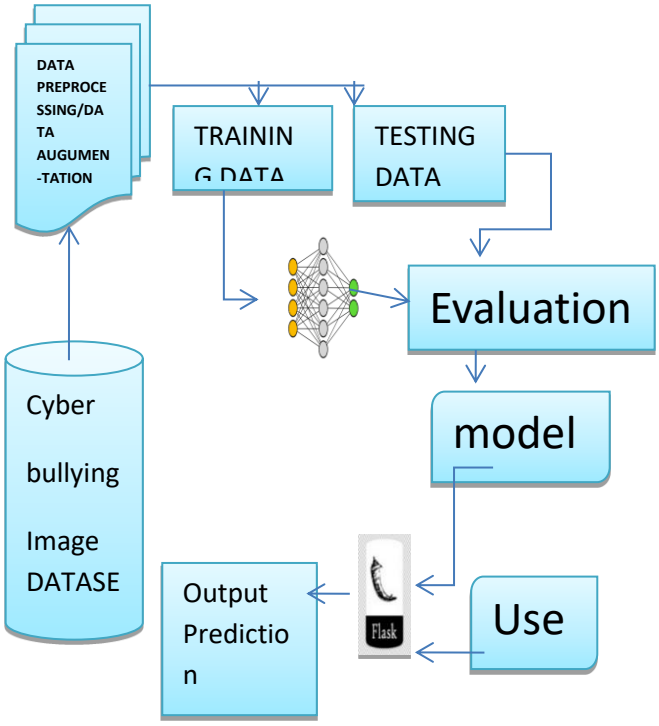

Figure 1 – System Design Related to Cyber Bullying

3.3 Data Preprocessing

Data preprocessing is an essential step in preparing the dataset for training an action recognition model that aims to identify human actions from photos. At first, the photos undergo resizing and standardization to a consistent resolution, such as 224x224 pixels. This process guarantees uniformity and avoids any distortions that may occur during the training of the model. Standardization is crucial for ensuring consistent scale and aspect ratio throughout the collection. The pixel values are subsequently normalized to a range of 0 to 1. This normalization process enhances the convergence of the model and improves training efficiency by reducing the impact of differences in brightness and contrast. In order to improve the dataset, data augmentation techniques like as rotation, translation, scaling, and flipping are utilized.

These enhancements enhance the heterogeneity of the dataset, mitigate overfitting, and enhance the model's capacity to generalize to new data. Ultimately, the preprocessed dataset is partitioned into training, validation, and test sets. The training set is utilized to construct the model, the validation set is employed for hyperparameter tuning and progress monitoring, and the test set is used to evaluate the performance on data that has not been previously observed. The preprocessing processes cumulatively improve the dataset, increasing the accuracy and robustness of the model in recognizing human actions from photos

3.4 Data Acquisition and Comprehension

We compile a heterogeneous dataset comprising a vast array of photographs that document different human acts. These images are obtained from publicly accessible databases like ImageNet, as well as from private collections. The selection of these photos is meticulously done to encompass crucial properties for efficient action identification. The wide range of activities and scenarios in the dataset ensures that it encompasses a diverse spectrum of human actions, which in turn provides a strong basis for training and testing the action recognition model. By integrating data from both public and private sources, the dataset is enhanced with diverse examples, hence improving the model's capacity to effectively detect and categorize various activities in real-life situations.

3.5 Image resizing and Rescaling

Resizing and rescaling photos is an essential preprocessing step to guarantee consistency throughout the dataset. To reduce variances in image dimensions, we employ a process of standardizing the image resolution and adapting all photos to a consistent size. The homogeneity of the image processing pipeline enhances its efficiency for the machine learning model. Utilizing standardized image sizes guarantees that the model is provided with input in a uniform format, hence minimizing the possibility of distortions and enhancing the precision and efficiency of the training procedure. Performing this preprocessing step is crucial for attaining dependable and resilient performance in action recognition tasks.

3.6 Noise Reduction and Contrast Enhancement

The application of noise reduction and contrast enhancement techniques enhances the quality and clarity of the photos. These tactics aid in minimizing unwanted distortions and enhancing crucial visual characteristics, so ensuring that the machine learning model receives unambiguous and informative input.

3.7 Data Normalization

Data normalization is an essential preprocessing technique that is employed to standardize the intensity values of pixels across all photographs in a collection. This procedure entails normalizing pixel values to a standardized scale, usually ranging from 0 to 1, by dividing the intensity of each pixel by the maximum attainable value. Standardization guarantees that all photos have a uniform distribution of pixel values, which is crucial for efficient model training. Normalization corrects discrepancies in image luminosity and contrast that could potentially distort the learning process of the model. Normalization is a process that brings pixel values to a consistent scale. This is beneficial for machine learning models because it allows them to learn more efficiently. By normalizing the pixel values, the model can concentrate on extracting important features without being affected by variations in pixel intensity among different images. The model's consistency allows it to more effectively extrapolate patterns from the training data, resulting in enhanced performance in the identification and categorization of objects or actions.

3.8 Feature Extraction

Feature extraction is a crucial procedure in creating pre-processed images for machine learning models, especially for the purpose of detecting human actions. This procedure entails the identification and isolation of prominent visual attributes from the images in order to improve the model's capacity to detect pertinent patterns. Methods like as edge detection, texture analysis, and pattern recognition are used to emphasize particular characteristics that are crucial for recognizing actions. Edge detection is a useful technique for identifying the boundaries and forms present in an image. This information is important for accurately comprehending the contours and movements of objects.

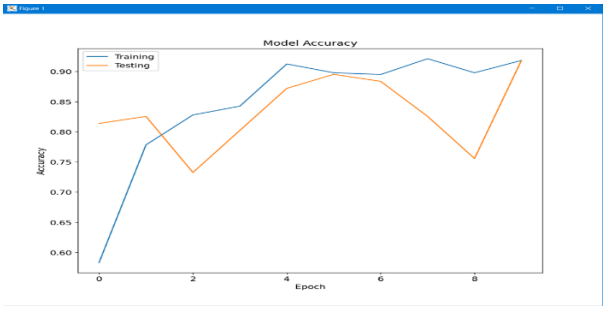

Figure 2 – Graph of Model Accuracy (gain)

The image displays a line graph called "Model Accuracy" which illustrates the performance of a machine learning model throughout multiple epochs. The accuracy scores for the "Training" and "Testing" datasets are represented by distinct lines. The y-axis demonstrates the precision, ranging from approximately 0.60 to 0.90, while the x-axis represents the epochs, ranging from 0 to 10. The graph illustrates the performance of the 'Training' dataset, represented by the orange line. It demonstrates an initial accuracy of roughly 0.85, which experiences a slight decline at epoch 2, followed by a constant increase until epoch 5. Subsequent little oscillations occur, exhibiting a slight downward trend towards epoch 10. Generally, as the number of epochs increases, both the training and testing accuracies exhibit an upward trend, indicating that the model's performance has improved over time. However, there is a noticeable variation in performance, particularly in the testing dataset, suggesting the presence of potential overfitting or generalization issues that require more investigation or enhancement of the model development.

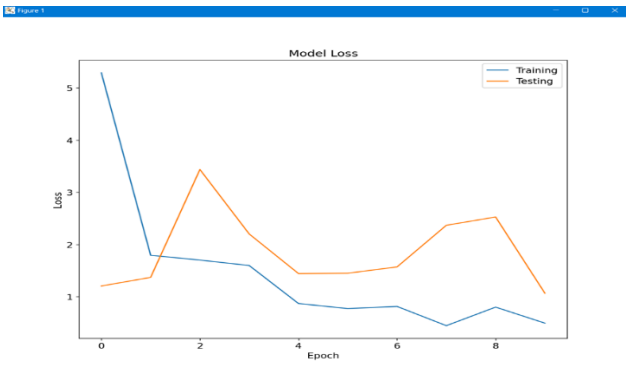

Figure 3 – Graph of Model Accuracy (loss)

The image above is a line graph titled "Model Loss," illustrating the testing and training losses of the model throughout multiple epochs. The x-axis represents the epochs, which are complete iterations of the training dataset, ranging from 0 to 8. Meanwhile, the y-axis represents the loss, which is a measurement of the model's performance on a scale of 0 to 5. The training loss, which represents the model's ability to learn from the training dataset, is illustrated by the blue line. In contrast, the orange line represents the testing loss, indicating the model's ability to apply its acquired knowledge to fresh datasets. The training and testing losses exhibit a consistent downward trajectory over the epochs depicted in the graph, suggesting that the model is progressively improving its accuracy and acquiring knowledge from the data. A noteworthy observation is that the training loss consistently remains lower than the testing loss.

This is a regular occurrence as models often perform better on training data compared to test data. An interesting discovery is that the testing loss somewhat rises after approximately six epochs. This behavior may indicate that the model is beginning to exhibit overfitting on the training set. Overfitting refers to the situation when a model performs well on the training data but struggles when presented with unfamiliar data. This realization suggests that further investigation or adjustments to the model may be necessary in order to tackle potential issues of overfitting and enhance the model's ability to generalize.

4. SYSTEM ARCHITECTURE

The architecture of the system for the classification of images of cyberbullying incorporates a number of essential components that are meant to guarantee the trustworthy and efficient identification of harmful content. The architecture is organized into separate layers that, when taken as a whole, make it possible to filter, analyze, and categorize photos in order to identify instances of cyberbullying.

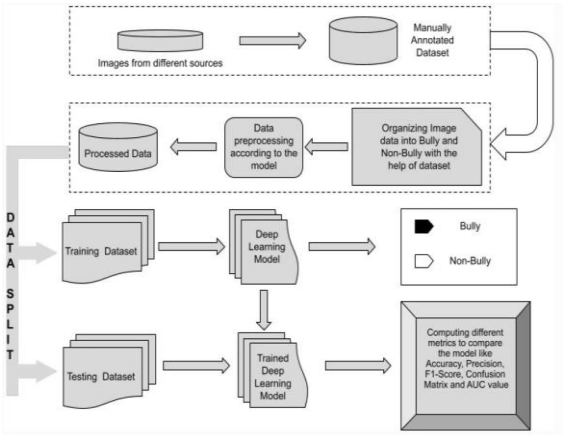

Figure 4 – Architecture Diagram

The MobileNetV2 architecture makes use of an inverted residual structure, in which the residual blocks' input and output are narrow bottleneck layers. It additionally filters features in the expansion layer using lightweight convolutions. Ultimately, non-linearities in the narrow layers are eliminated. The photos are loaded using the data frame by the flow-from-data frame method. The precise location of the photos is specified by the directory argument. The independent and dependent variables in this scenario are the labels and the images, or x_col and y_col. Class_mode=" binary" indicates that there are just two different classes in the data. Using target size = (224,224), a 224 x 224 image will be produced. How many photos are sampled at once is known as the batch size.

In the inverse residual shape used by the MobileNetV2 structure, the input and output residual blocks are thin bottleneck layers. It also uses mild convolutions to spread the functions on the growth layer. Finally, it gets rid of nonlinearities in narrow layers. The normal architecture seems like this:

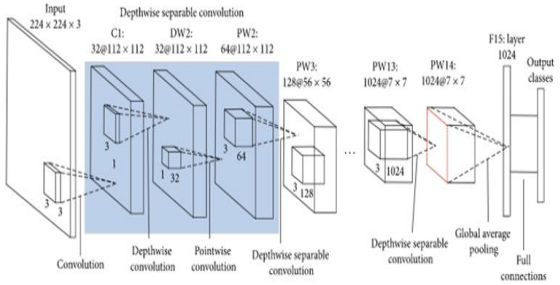

Figure 5 – Architecture Diagram of MobileNetV2

Transfer learning models are valuable for making predictions in several fields and offer the benefit of utilizing state-of-the-art deep learning architectures. This study predicts the occurrence of message-based cyberbullying by using the benefits of pre-structured transfer learning methods. The selected dataset was initially processed using popular switch architectures such as VGG16, MobilenetV2, and others available in the Keras toolkit. Based on the empirical findings of multiple models, it has been determined that Mobilenet-V2 outperforms other models. This individual became one of the most successful architects in the ILSVRC project during the same year. We continued our investigation using Mobilinet-V2, a machine learning technique developed by experts in the field and commonly employed for image recognition tasks.

Due to the pre-processed image records, we have, we will configure the lowest model with the input length of 224 x 224.

The base model could possess equal image dimensions. To omit the highest volumes from the reordered fashions, we can use the parameter include top=False, which is excellent for feature extraction. The provided code will download the pre-processed version and initialize it using the specified settings. It is crucial to verify that the convolution weights are not updated prior to drawing and setting the version.
Transfer learning models are beneficial for predictive tasks across several domains and offer the advantage of utilizing state-of-the-art deep learning architectures. This study predicts the occurrence of message-based cyberbullying by using the benefits of pre-structured transfer learning methods. The selected dataset was initially processed using popular switch architectures available in the Keras toolkit, such as VGG16 and MobilenetV2. Based on the empirical findings of multiple models, it has been determined that Mobilenet-V2 outperforms other models.
This individual became one of the most accomplished architects in the ILSVRC project during the same year. We continued our investigation using Mobilinet-V2, a machine learning technique developed by experts in the field and commonly employed for image recognition tasks.
Due to the pre-processed image records, we have, we will configure the lowest model with the input length of 224 x 224. The base model may possess same image dimensions. To omit the highest volumes from the reordered fashions, we can use the parameter include top=False, which is excellent for feature extraction.

Performance Metrics

True Positive Rate (Recall) The true positive rate, or recall, is defined as:

Recall=True Positives (TP)/ True Positives (TP) + False Negatives (FN)

This metric indicates the proportion of actual positive instances that are correctly identified by the model

True Negative Rate (Specificity) The true negative rate, or specificity, is defined as:

Specificity=True Negatives (TN)/True Negatives (TN) + False Positives (FP)

This metric measures the proportion of actual negative instances that are correctly identified by the model.

False Positive Rate The false positive rate is defined as:

False Positive Rate=True Negatives (TN)/ False Positives (FP) + False Positives (FP)

This metric indicates the proportion of negative instances that are incorrectly classified as positive.

5. RESULT

The dataset consists of 429 photos, divided into two categories: Bullying (215 images) and Non-Bullying (214 images). The dataset is divided into separate subsets for training and testing purposes. After completing the training phase of the model, the testing phase requires the utilization of a minimum of 100 sample images from the dataset to assess the model's performance. During the testing phase, screenshots are taken to record the model's predictions. The model analyzes each input image, containing examples of individuals who have encountered bullying, to ascertain the presence of bullying. The algorithm categorizes the image as "Bullying" if it detects indications of bullying; otherwise, it labels it as "No Bullying." This methodology guarantees a thorough evaluation of the model's precision in identifying instances of bullying from the photos.

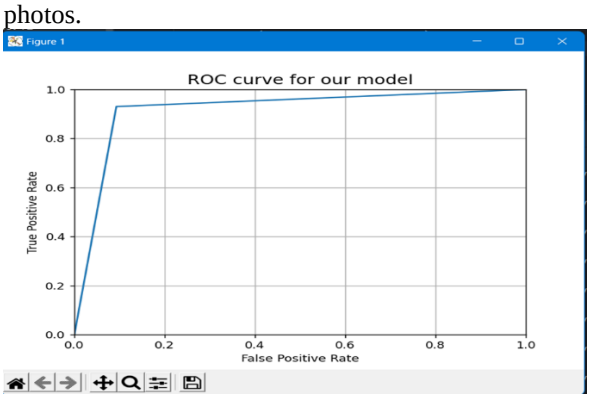

Figure 6 – ROC Curve Illustrating Different Threshold Settings

The curve's shape indicates if the model can effectively distinguish between the classes. When the curve is positioned in close proximity to the upper left-hand corner, the performance is enhanced. The AUC measure provides a concise summary of the entire performance, with a greater value indicating a superior discriminating performance. The threshold value of the output score determines whether it should be considered as positive or negative.
Typically, in classification models, the probability of an instance belonging to a specific class is represented as a value ranging from 0 to 1. This graph indicates a reliable indicator of distinguishability since the curve appears to be in close proximity to the upper-left corner of the ROC space. This indicates that the model has the ability to accurately differentiate between positive and negative classes, which is a desirable trait in classification tasks. An AUC (Area under curve) value of 1 signifies that there is a complete distinction between positive and negative cases by a flawless classifier. On the other hand, a perfectly random classifier would have an AUC of 0.5. Analysts can assess the classifier's overall effectiveness by computing its AUC, where higher values indicate superior model performance in distinguishing between classes.

6. CONCLUSION

To summarize, the action recognition project has shown that convolutional neural network (CNN) architecture, MobileNet, are effective in properly identifying human activities from photos. By conducting thorough experimentation and evaluation, we have acquired useful insights into the performance of various designs and their appropriateness for the task of action recognition.
Throughout the investigation, we acquired an extensive dataset comprising photographs that depict a diverse range of human actions. This was done to ensure that the trained models

possessed robustness and exhibited a significant level of generalizability. We utilized rigorous data preparation techniques, including resizing, rescaling, noise reduction, and contrast enhancement, to enhance the quality and clarity of the images. This strategy improves the effectiveness of model training.

The training and validation phases adhered to traditional machine learning techniques, wherein measures such as accuracy, precision, and loss were consistently monitored to assess the model's performance and guide optimization endeavors. The evaluation resulted in positive outcomes, as each design shown outstanding accuracy and competency in detecting specific human activities.

Furthermore, the examination of performance metrics across various designs revealed noteworthy insights regarding their distinct strengths and weaknesses. MobileNet exhibited outstanding computational efficiency while maintaining consistent performance.

However, it is essential to acknowledge the limitations and challenges encountered throughout the process. Notable achievements have been made, although the performance and capability of the models may have been affected by factors such as the size of the dataset, imbalanced class distribution, and optimization of hyperparameters. Moreover, it is imperative to do further research and experimentation in order to explore novel methodologies, improve model architectures, and tackle existing constraints.on

ACKNOWLEDGEMENT

We would like to extend our heartfelt gratitude to the College of Engineering and Technology at SRM Institute of Science and Technology, Kattankulathur, for their invaluable support and resources throughout the duration of this project. The expertise and encouragement provided by the faculty and staff have been crucial in guiding and shaping our research. We are deeply appreciative of the faculty and staff who provided insights and assistance, making this endeavor both educational and rewarding.

REFERENCES

[1] A. López-Martínez, J. A. García-Díaz, R. Valencia-García, and Ruiz-Martínez, ‘‘CyberDect. A novel approach for cyberbullying detection on Twitter,’’ in Proc. Int. Conf. Technol. Innov., Guayaquil, Ecuador: Springer, 2019, pp. 109–121, doi: 10.1007/978-3-030-34989- 9_9.

[2] R. M. Kowalski and S. P. Limber, ‘‘Psychological, physical, and academic correlates of cyberbullying and traditional bullying,’’J. Adolescent Health, vol. 53, no. 1, pp. S13–S20, Jul. 2013, doi: 10.1016/j.jadohealth.2012.09.018.

[3] Anwar, D. M. H. Kee, and A. Ahmed, ‘‘Workplace cyberbullying and interpersonal deviance: Understanding the mediating effect of silence and emotional exhaustion,’’ Cyberpsychol., Behav., Social Netw., vol. 23, no. 5, pp. 290–296, May 2020, doi: 10.1089/cyber.2019.0407.

[4] D. M. H. Kee, M. A. L. Al-Anesi, and S. A. L. Al-Anesi, ‘‘Cyberbullying on social media under the influence of COVID-19,’’ Global Bus. Organizational Excellence, vol. 41, no. 6, pp. 11–22, Sep. 2022, doi: 10.1002/joe.22175.

[5] Kwan, K. Dickson, M. Richardson, W. MacDowall, H. Burchett, C. Stansfield, G. Brunton, K. Sutcliffe, and J. Thomas, ‘‘Cyberbullying and children and young people’s mental health: A systematic map of systematic reviews,’’ Cyberpsychol., Behav., Social Netw., vol. 23, no. 2, pp. 72–82, Feb. 2020, doi: 10.1089/cyber.2019.0370.

[6] R. Garett, L. R. Lord, and S. D. Young, ‘‘Associations between social media and cyberbullying: A review of the literature,’’ mHealth, vol. 2, p. 46, Dec. 2016, doi: 10.21037/mhealth.2016.12.01.

[7] M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, and K. Araki, ‘‘Automatic extraction of harmful sentence patterns with application in cyberbullying detection,’’ in Proc. Lang. Technol. Conf. Poznań, Poland: Springer, 2015, pp. 349–362, doi: 10.1007/978-3-319-93782-3_25.

[8] M. Ptaszynski, P. Lempa, F. Masui, Y. Kimura, R. Rzepka, K. Araki, M. Wroczynski, and G. Leliwa, ‘‘"Brute-force sentence pattern extortion from harmful messages for cyberbullying detection,’’’ J. Assoc. Inf. Syst., vol. 20, no. 8, pp. 1075–1127, 2019.

[9] M. O. Raza, M. Memon, S. Bhatti, and R. Bux, ‘‘Detecting cyber- bullying in social commentary using supervised machine learning,’’ in Proc. Future Inf. Commun. Conf. Cham, Switzerland: Springer, 2020,pp. 621–630.

[10] D. Nguyen, M. Liakata, S. Dedeo, J. Eisenstein, D. Mimno, R. Tromble, and J. Winters, ‘‘How we do things with words: Analyzing text as social and cultural data,’’ Frontiers Artif. Intell., vol. 3, p. 62, Aug. 2020, doi: 10.3389/frai.2020.00062.

[11] Cai, J. Li, W. Li, and J. Wang, ‘‘Deep Learning model used in text classification,’’ in Proc. 15th Int. Compute. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP), Dec. 2018, doi: 10.1109/ICCWAMTIP.2018.8632592.