International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Cyberbullying Image ClassificationUsing Transfer Learning Model

A. Koteswaramma[1], E. Pravallika[2], G. Rasi[3], B. Puneeth[4], J. Yuvraj[5]

[1]Assistant Professor, Department of IT, Seshadri Rao Gudlavalleru Engineering College (SRGEC), Krishna-521356, A.P., India
[2,3,4,5]Department of IT, Seshadri Rao Gudlavalleru Engineering College (SRGEC), Krishna-521356, A.P., India

Abstract:
The impact of cyberbullying is immeasurable on the lives of victims as it is very subjective to how the person would tackle this. The message may be a bully for victims, but it may be normal for others. The ambiguities in cyberbullying messages create a big challenge to find the bully content. Some research has been reported to address this issue with textual posts. However, image-based cyberbullying detection has received less attention. This Project aims to develop a model that helps to prevent image- based cyberbullying issues on social platform posts. We proposed a transfer learning-based automated model to detect image-based cyberbullying posts from the social platform. The transfer learning models are capable of extracting hidden contextual features from cyberbullying posts. Our experiment consists of two sets of datasets (i.e.) images consisting of cyberbullying and non cyberbullying images. The datasets can be useful for future researchers to extend the research. Finding the best-suited model to detect the bully images is a challenging task, hence experimented with both DL and transfer learning models to find the best model. The experimental outcomes confirmed that the transfer learning models are the better choice for predicting image-based cyberbullying posts.

Keywords: cyberbullying, transfer learning, MobilenetV2, image- based threats.

1. INTRODUCTION

The emergence of Web 2.0 has had a significant impact on social communication, redefining friendships and relationships in new ways. Teenagers spend a lot of time online and on various social platforms. While their online presence may provide them with many benefits, it also exposes them to threats and social misbehaviors like cyberbullying.

Digital harassing should be perceived and tended to according to alternate points of view. The automatic detection and prevention of these incidents can significantly contribute to the solution to this issue. Programs that attempt to offer victims of bullying support and tools that can identify bullying have already been developed. In addition, the majority of the online platforms that teenagers frequently utilize have safety centers, such as the YouTube Safety Centre and Twitter Safety and Security, that support users and monitor communications. There has also been a lot of research done on automatic detection and prevention of cyberbullying, which we will go into more detail about in the next section. However, this issue is still a long way from being solved, and there needs to be more work done to find a real solution. To identify instances of cyberbullying, the majority of existing studies have utilized

conventional Machine Learning (ML) models. In recent times, models based on deep neural networks (DNNs) have also been used to detect cyberbullying.

2. LITERATURE REVIEW

The issue of cyberbullying has recently received a lot of attention from researchers. The relevant contribution of research to cyberbullying detection is discussed in this section. Dadvar Maral. 2018. There are frequently very few posts in the datasets used to detect cyberbullying. This unevenness issue can be somewhat covered by oversampling the tormenting posts. However, more research is needed to determine how this prevalence affects models' performance. The DNN models' adaptability and transferability to the new dataset are demonstrated by our investigation. When it came to the detection of cyberbullying in this YouTube dataset using machine learning models, DNN-based models and transfer learning performed better than any of the previous results. The authors trained the ML models with context- based features like user profile information and personal demographics, which led to an F1-score of 0.64. By incorporating expert knowledge, the detection methods' discrimination capacity was raised to 0.76.

H. H. Aldhyani theyazn. 2022. The efficacy of the classifiers like CNN-BiLSTM which is a amalgamated deep learning technique and BiLSTM was compared to organize the posts which are posted on the media into varied bullying patterns. When compared the BiLSTM classifier had a detection rate of 94% than the CNN-BiLSTM with the binary classification dataset (aggressive or non-aggressive bullying). As the CNN-BiLSTM classifier consists of Multiclass dataset, it is combined with the BiLSTM to achieve 100% results.

Bozyigit and others a Turkish dataset for identifying cyberbullying was created using data from. In their experiments, the authors used a variety of techniques for neural networks. Before the models were trained, the data obtained was used for identifying the importance and ranking the features with which to minimize space dimensions and discard redundant words. This method using artificial neural networks had an F-measure of 91 percent. Wang and co. represented a GCN (Graph Convolutional Neural Network) method for detecting multiple classes of cyberbullying using 40,000 tweets. The authors also compared XGBoost, Nave Bays (NB), SVM, MLP, and KNN, among other machine- learning methods. The GCN model had the highest F1-score, 92%, according to their experiments. A different study carried out by Bozyi et al. in which used several machine learning methods like SVM, logistic regression, NB, random forest, and AdaBoost has been compared. Amongst them, the AdaBoost algorithm delivered the best results in the criteria of cyberbullying.

Chen and Co. proposed a CNN-based text classification model for the dataset on de facto verbal aggression in which the researchers have used Facebook comments and Tweets manually, Except the stickers and emojis that have been sent to the post. They used the "sentiment140 corpus" to collect social network comment data in addition to the comments that were manually labeled. After altering, the tweets were classified as aggressive or passive. They performed lowercasing during preprocessing, removed usernames, hashed topics with stickers, and proceeded at the rate. The method extracted features. With an efficacy of 0.92 and an AUC value of 0.98, the DL-based CNN model produced the best outcome.

In terms of both accuracy and F1-score, as well as the number of studies, the existing study on image-based cyber bullying detection lags significantly behind text-based detection. The algorithm created to detect text-based cyberbullying has an accuracy value of more than 90% and a F1 score of 0.90. Image-

based cyber bullying detection received less attention than textual cyber bullying. However, at the moment, posts can also be made with images or a combination of images and text. Therefore, to guarantee a cyberbullying-free network, it is necessary to capture the image-based bullied post which will be published on the media platform. This study zeroed in on fostering a robotized model with a profound exchange learning way to deal with distinguish picture put together digital harassing presents with respect on friendly stages to fill this examination hole.

3. METHODOLOGY

Transfer learning models outperform the current deep learning architecture for the prediction task and work well across a variety of industries. As a result, image-based cyberbullying posts can be predicted using the advantages of pre-trained transfer learning models in this study. The selected dataset was initially utilized with models like VGG16, MobilenetV2, and others from the Keras library. When compared the Mobilenet-V2 was best than the other models in terms of performance. In the same year, it finished as one of the best architectures in the ILSVRC challenge. As a result, we have carried on our research by employing Mobilenet-V2 techniques.

A. Data Collection

The aim of the data collection is to give the model the ability to understand both the typical case and the cyberbullying case. For instance, the model might perceive the typical human face as bullying if she sees offensive images of human faces.

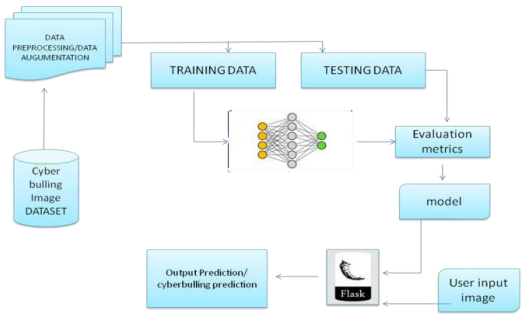

Fig1 Block Diagram

B. Preprocessing

- With sufficient data, Convolutional Neural Networks (CNNs) can accomplish incredible feats. However, it is challenging to determine the appropriate amount of training data for each feature that must be trained. The network may overfit the training data if the user does not provide sufficient information. Sizes, poses, zoom, lighting, noise, and other characteristics are all present in realistic images.
- Data Augmentation technique is utilized in order to strengthen the network's resistance to these typical threats. During training, the network will encounter these phenomena by rotating input images to differentangles, flipping images along different axes, or translating or cropping the images.
- "ImageDataGenerator" contains a handy set of arguments that can be used with Keras to enhance images. We can completely rearrange the pixels of the image by flipping it horizontally or vertically, butthe features remain unchanged. Numpy can be used toaccomplish this.
- If there is an urgent need for more data, rotation is an enhancement that can be applied at angles smaller than 90 degrees. To blend with the image when it is rotated, the background color is typically fixed for rotation. Orelse, the model will assume that the background change is an unique feature. This functions most effectively when the backgrounds of all rotated images match.

- The width_shift_range and height_shift_range main arguments of the Keras ImageDataGenerator class determine whether the image's pixels will move left- to-right or top-to-bottom.
- At the time of learning, this simply consists of learning one binary classifier or regressor for each class. To accomplish this, multi-class labels must be converted to binary labels (either belong to the class or not). With the transform method, LabelBinarizer makes this process simple.

C. Model Training

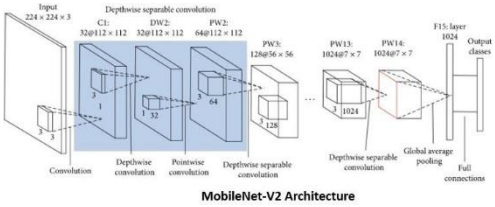

Fig 2 Pictorial representation of MobileNetV2 architecture

MobileNetV2 architecture makes use of an inverted residual structure, with thin bottleneck layers acting as the residual blocks' input and output.

It additionally utilizes lightweight convolutions to channel highlights in the development layer. Finally, non- linearities in the narrow layers are eliminated. The architecture as a whole looks something like this:

We will divide the training data into two distinct datasets: a set to train the model and another set to evaluate themodel's performance.

The data frame is used to load the images by the flow_from_dataframe method. The precise location of the images is specified by the directory parameter. The images and labels in this case are the independent and dependent variables, respectively, represented by x_col and y_col. class_mode="binary" indicates that there are only two distinct classes in the data. An image with the size 224 x 224 will be produced if target_size=(224,224). Clump size is the quantity of pictures examined on the double

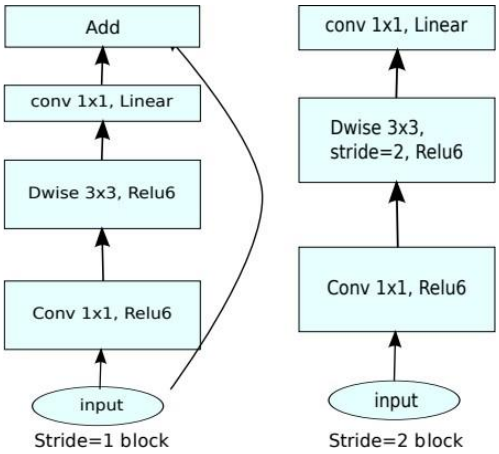

Fig 3 Sample Model

In addition, we will set the base model's input size to match the 224 x 224 preprocessed image data we have. The same image net weights will be used in the base model. By setting include_top=False, which is ideal for feature extraction, we will exclude the top layers of the pre-trained model.The pre-trained model will be downloaded and initialized using the parameters specified in the preceding code block. Additionally, we should prevent the convolution's weights from being updated before the model is assembled and trained. We set the trainable attribute to false to accomplish this.Accuracy of Classification: When we talk about accuracy, we typically mean classification accuracy. It is the

proportionbetween the number of samples used as input and the numberof accurate predictions.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

It only works well when there are equally many samplesfrom each class.

Loss in logarithms: The way that Log Loss, also called log loss, operates is by penalizing incorrect classifications. Multiple classes can be classified using it effectively. When using this algorithm, the classifier should assign probabilities to each class for every sample. The following formula can be used

$$LogarithmicLoss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

to calculate the Log Loss if there are N samples from M classes:

Where y_ij denotes whether sample i belongs to class j or not and p_ij denotes the likelihood that sample i belongs to classj.

On the range [0, ∞], It has no upper bound. When the valueis closer to zero it is considered as higher accuracy, while a log loss that is further from zero indicates lower accuracy. Ingeneral, reducing Log Loss increases the classifier's accuracy.

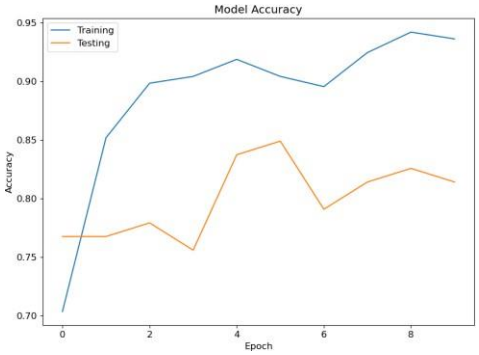

Fig 4 Graphical representation of Log Loss

Area Under the Curve (AUC) is the widely used metrics for evaluation which is utilized for binary classification-related issues. A classifier's AUC is the probability that it will place a positive example chosen at random higher than a negative example chosen at random. Before characterizing it, let us figure out two essential terms:

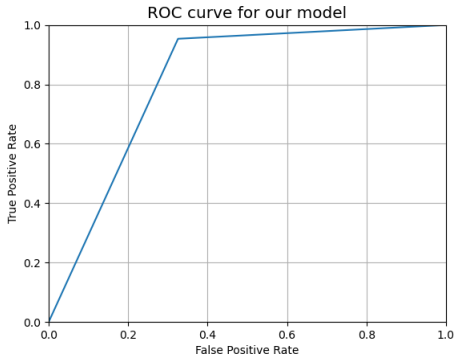

Fig 5 ROC curve

- **True Positive Rate (Sensitivity)** is calculated as TP/(FN+TP). This rate is the portion of positive data points among all positive data points that are correctly classified as positive.

$$TruePositiveRate = \frac{TruePositive}{FalseNegative + TruePositive}$$

- **True Negative Rate (Specificity)** is obtained by dividing (FP+TN), which is equal to TN. The False Positive Rate is the percentage of negative data points that are correctly classified as negative

relative to all negative data points.

$$TrueNegativeRate = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

- **False Positive Rate** is Divided by (FP+TN), which is denoted by FP. It is the percentage of negative data points that are incorrectly interpreted as positive when compared to all negative data points.

$$FalsePositiveRate = \frac{FalsePositive}{TrueNegative + FalsePositive}$$

Confusion Matrix: As the name suggests, the Confusion Matrix outputs, the matrix which describes the model's entireperformance.

Confusion Matrix with Normalized ValuesPrecision recall f1-score support

0	0.94	0.67	0.78	43
1	0.75	0.95	0.84	43
Accuracy			0.81	86
Macro average	0.84	0.81	0.81	86
Weighted average	0.84	0.81	0.81	86

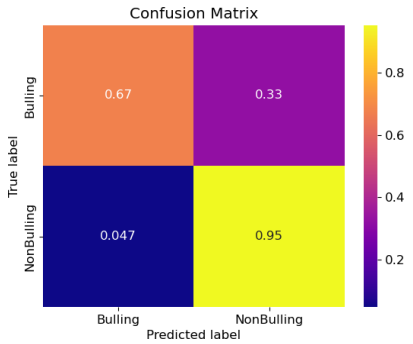

Fig 6 Confusion Matrix model

D. *Flask Framework:*

The part of a website that the user interacts with right away is called the front end. It includes everything that users interact with and see: the styles and colors of the text, images and videos, graphs and tables, the colors of the buttons, and the navigation menu. The front end is built using JavaScript, HTML, and CSS. Flask is used to create Python-based web applications. Importing the Flask class was our first step. After that, we create this class's instance. The application's module or package's name is passed as the " name_" argument. This is necessary for Flask to know where to look for resources like static files and templates. Flask is then informed by the route () decorator which URL should activate our method. The message that should be displayed in the user's browser is returned by this method.

Fig 7 Snippet 1

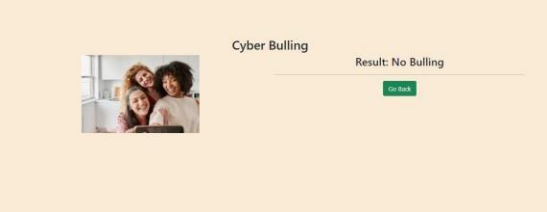

Fig 8 Snippet 2

Fig 9 Snippet 3

CONCLUSION

The typical system makes it difficult to trace complex issues like cyberbullying, which are comprised of numerous issues. Post-detection image-based social cyberbullying is especially difficult. In order to determine the most suitable model for predicting image-based posts on social platforms, this study investigated transfer learning and deep learning frameworks. The mobile net transfer learning models had better predictions. As a result, it can be concluded that the majority of image-based cyberbullying posts are detected by the proposed system.

The following are some of the model's limitations:

i) A post with only text is not included in this study because it is not considered textual cyberbullying detection; likewise, images and text have been found in posts about cyberbullying. This study, on the other hand, only looks at image-based cyberbullying detection. Because of its numerous subproblems, the scope of this research in the future is always up for debate. The text-based part can be considered alongside the picture to get more digital harassing-related posts on friendly stages.

REFERENCES

1. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251–1258
2. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778
3. Roy PK, Ahmad Z, Singh JP, Alryalat MAA, Rana NP, Dwivedi YK (2018) Finding and ranking high-quality answers in community question answering sites. Global J Flex Syst Manag 19(1):53–68
4. Bhat S, Koundal D (2021) Multi-focus image fusion using neutrosophic based wavelet transform. Appl Soft Comput 106:107307

5. Kumari K, Singh JP, Dwivedi YK, Rana NP (2021) Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization. Future Gener Comput Syst 118:187–197.
6. Aggarwal S, Gupta S, Alhudhaif A, Koundal D, Gupta R, Polat K (2021) Automated COVID-19 detection in chest x-ray images using fine-tuned deep learning architectures. Expert Syst 39:e12749
7. Modha, S.; Majumder, P.; Mandl, T.; Mandalia, C. Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. Expert Syst. Appl. 2020, 161, 113725. [CrossRef]
8. Dinakar, K.; Reichart, R.; Lieberman, H. Modeling the detection of textual cyberbullying. In Proceedings of Proceedings of the International AAAI Conference on Web and Social Media, Atlanta, GA, USA, 6–9 June 2022.
9. Dadvar, M.; Jong, F.D.; Ordelman, R.; Trieschnigg,
10. Improved cyberbullying detection using gender information. In Proceedings of the Title of host publicationProceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), Ghent, Belgium, 24 February 2012.
11. Kontostathis, A.; Reynolds, K.; Garron, A.; Edwards,
12. L. Detecting cyberbullying: Query terms and techniques. In Proceedings of the 5th Annual Acm, Web Science Conference, online, 2 May 2013; pp. 195–204.