Volume: 11 Issue: IV Month of publication: April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50438

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

A Review of Machine Learning and AI-Based Approaches to Detecting Cyberbullying on Social Media

Harish D[1], Manimaran M[2], Jayashakthi Vishnu P[3]
[1, 2, 3]Department of Information Technology, Kumaraguru College of Technology

Abstract: With rapid rise in the use of social media, cyber-bullying has become a significant concern, affecting individuals of all ages and backgrounds. Finding cyberbullying on social media is essential to building a secure and welcoming online community. The various approaches and techniques used to identify cyberbullying on social media are examined in this research, ranging from traditional keyword-based strategies to more sophisticated machine learning and NLP-based algorithms. We also draw attention to the difficulties and restrictions posed by the use of current techniques, such as their poor accuracy, lack of standardization, and privacy issues. We also go over the significance of addressing social and cultural disparities as well as the necessity of ethical concerns when identifying cyberbullying. Our review provides a comprehensive understanding of the current state of cyberbullying detection on social media, along with recommendations for future research. Ultimately, combating cyberbullying and fostering a more supportive and positive online environment depend on the development of more precise and reliable detection techniques.

Keywords: Cyber bullying, Machine learning, deep learning, review, NLP, social media, survey.

I. INTRODUCTION

With more than 4.7 billion active users globally [23], social media has integrated itself into every aspect of our daily lives. Cyberbullying is a growing concern in today's digital world, particularly on social networking platforms. Cyberbullying [22] is utilising technology to harass, humiliate, or threaten people, and can take numerous forms, such as derogating, spilling rumours, or spreading embarrassing images or recordings. The anonymity and breadth of social media platforms can make it easier for cyberbullies to target and victimise others, frequently with fatal effects.

The detection of cyberbullying is vital to avoiding and treating its detrimental effects. Early detection can help identify victims and provide them with support and resources to cope with the bullying. It can also help identify and intervene with cyberbullies, thereby preventing them from participating in additional destructive acts. Consequently, it is crucial to provide efficient and precise techniques for identifying cyberbullying on social media sites.

Given that cyberbullying has societal repercussions in addition to affecting individuals and communities, this issue is significant and relevant on a larger scale. For example, cyberbullying can damage academic and workplace productivity, contribute to mental health difficulties, and even lead to legal or criminal penalties in some situations. Therefore, creating effective and precise ways for identifying cyberbullying on social media platforms is vital for maintaining safe and healthy online environments.

II. LITERATURE REVIEW

In a solution put up by Fan et.al [1] (2021), to classify the gathered tweets, a labelled harmful comment dataset was used to fine-tune the BERT pre-trained model and its three versions. Punctuation, links, and non-English terms were eliminated during the pre-processing stage. A language model was created using transformer-based models. The Kaggle competition's document retrieval from Wikipedia was used to train and test 1,59,571 training data and 1,53,165 testing data for the BERT-base model. With 1,59,571 training data and 1,53,165 testing data, the dataset consists of 16,225 instances with labels. By introducing a dropout and segregation layer, with an at most sequence-size of 169 tokens, a batch-size of 16, and a learning-rate of 2105, BERT-base is fine-tuned for a multi-label text classification task. The methodology was used to a Twitter data flow that was gathered and separated into two time periods. Search phrases and hashtags like Brexit, #Brexit, #BrexitBetrayal, and #StopBrexit were used to gather data. Emojis, numerals, punctuation, and URL links were all eliminated since they were superfluous.

Datasets 1 and 2 revealed that the majority of tweets were labelled as poisonous, profane, insulting, or threatening. The majority of tweets are classified by the multilingual BERT, RoBERTa, and DistilBERT models as insult, threat, and toxic, with severe toxic being the most classified.

This paper by Beddiar et.al [2] (2021), suggests the use of back-translation and paraphrase methods in conjunction for data augmentation. Later, the enhanced data is utilised to identify hate-speech and cyber-bullying materials on social media platforms. A larger dataset is first created by concatenating the original data with synthetic data that was previously generated through back translation. Results are examined to gauge how well the back translation produces material that is semantically similar. The text is cleaned and normalised using pre-processing methods once the datasets have been extended. The computation and use of FastText embeddings as a feature set. The larger datasets are then used to train a CNN/LSTM to categorise articles containing hate speech and cyberbullying.

The suggested approach [2] functions for both baseline and neural-network models, demonstrating the viability of the suggested approaches. When back-translation and paraphrasing were used with the CNN-classifier for four extended datasets, the accuracy and F1-score of the CNN and LSTM classifications improved by 4–10%. For miniscule to huge collections of data, data augmentation demonstrated improved accuracy and F1-score of CNN using FastText-classification representation. In terms of performance, the (CNN and LSTM)-classifiers produced good results for the classification of hate-speech. CNN model had 1.8% greater accuracy and 42% higher F1 scores than PCNN model after being trained on enlarged Formspring datasets using back-translation, indicating it might be applied to additional NLP difficulties.

Rui Zhao et.al [3] (2016), suggested representation-learning approach for detecting cyber-bullying, which is divided into three sections: Latent semantic features, bullying attributes, and bag-of-words attributes based on word embeddings. Latent Semantic Analysis(LSA) is used to juice out latent semantic attributes, whereas unigram and bigram are used to create bag-of-words features. Based on word embeddings, pre-defined bullying traits are expanded. Word embeddings utilize real-valued, low-dimensional vectors to express word semantics, and they can also represent semantic similarity via their cosine similarity. The top ten terms that are related to the insulting word are concatenated to create the final bullying characteristics after we apply an additive model to produce a corresponding embedding for a bigram. In all three-evaluation metrics, the proposed EBoW model performs better than other compared approaches. By using word embeddings, it expands the pre-defined insulting terms and gives these expanded bullying traits various weights based on the cosine-similarity among word embeddings. Additionally, it outperforms LSA and LDA, two state-of-the-art techniques for learning text representation. The standard of the learned-attribute space is largely dependent on the decreased dimension of the latent space.

Raj et.al [4] (2022), suggested a model, which is a working prototype for a system, that can be used to identify cyberbullying on social media sites. Before being fed into CNN-BiLSTM deep learning models, the training data is cleaned and prepared. The model is then utilized on the website, and administrators have access to see the progress of the content. Data-cleaning, data-integration, data-transformation, data-reduction, and data-discretization are all steps in the study of a dataset.

The most crucial information is that a 0-1 classifier is used to determine whether the text contains content related to cyberbullying, that data is integrated into a single csv file, that the sentence is divided into words using data discretization, that urls, special characters, '@', and stopped words are removed using data reduction, that a word is stemmed to its root form, and that words are lemmatized to their simplest form using data reduction. In order to assess the suggested CNN-BiLSTM model's accuracy, an ensemble ML model is contrasted with it.

The best results were obtained when Sigmoid activation and Adam optimizer were compared to other combinations of activations and optimizers for a baseline LSTM model on accuracy. Due to ReLU's non-linear properties and quicker modelling creation time, it is chosen over Sigmoid for the activation-layer of the CNN-BiLSTM model's hidden-layer. The CNN-BiLSTM model, according to the results, performs the best of all the models examined.

According to Nayel et.al [5] (2020) suggested method, transforms tweets into vectors using a TF/IDF vector space model, then uses the vector space as input for a linear classifier. Pre-processing, abbreviation removal, punctuation and digit elimination, elongation elimination, and feature extraction make up the overall structure. English abbreviations, punctuation, and numbers are removed by pre-processing, whilst redundancy is eliminated, and the feature space is shrunk using elongation elimination. The feature extraction stage of the pipeline's second stage used TF/IDF with an n-gram range. The dataset was split into training, development, and test sets, and various classifiers were utilised to train the classifier. The suggested models make use of the Stochastic Gradient Descent (SGD) optimisation technique, the Hinge loss function, linear kernels for the SVM and MLP classifiers, logistic function for the MLP classifiers, and hard voting strategy for ensembles.

Linear Classifier-[0.8421,0.8182], SVM-[0.8115,0.8043], MLP (n = 60)-[0.8033,0.7831], and Voting-[0.8265,0.8129] are the recommended classifiers that perform the best on development and test sets. The performance of the suggested models is constrained by the local context representation of comments, TF/IDF, and traditional classification techniques. When applied to various NLP tasks, deep learning models—which rely on word embeddings—show improvement.

This review by, Aljabri et.al [6] (2023), looks at the research that has been published between 2015 and 2022 in the area of bot identification and classification using ML approaches on different social media platforms. A taxonomy was developed based on ML-based methodologies, kind of social media platform, and type of social media bot, with a total of 105 publications reviewed in total. The taxonomy's goals are to identify the social media platforms that are most affected by bots, the class of bots that are most prevalent on those platforms, to highlight the social media platforms that have received the most research, and to examine the research gaps on other platforms.

In the research work of A. Shekhar et.al [7] (2018), the study of sentiment requires the preparation and collecting of datasets. To do sentiment analysis on Twitter data, key processes include tweet preparation and tokenization. Two steps are involved in automatically detecting cyberbullying using machine learning: Classification and Representation Learning for Tweets. An exciting result of deep learning is word embeddings. One-hot encoded words are mapped to words using word embeddings. There are two categories of word representations used in NLP: word embeddings. Prediction-based embeddings are constructed using a 2-layer neural network, while frequency-based embeddings use a deterministic methodology. Unsupervised learning techniques like clustering and classification are used to categorise related instances based on features. Text classification has made substantial use of both Naive Bayes Classifier and SVM classifiers, the latter of which is based on the idea of a hyperplane and the former being probabilistic. Topics and words were categorised using unsupervised machine learning (LDA) and K-Means clustering (K-Means). Manually tagged datasets were categorised using supervised machine learning (SML). The analysis of 2500 tweets on Jason Colins included the selection of 100 at-random tweets, which were then categorised as toxic, non-toxic, and neutral. The analysis used uni-gram + bi-gram attributes, although manual labelling encountered issues. Over several iterations, the SVM-classifier with a (linear)-kernel produced accuracy between (55 and 57) %, whereas the Multinomial-Naive Bayes(NB) with ternary segregation produced accuracy of 0.57. The overall accuracy of supervised-ML on a marked dataset was 97%. Using the Soundex Algorithm, unigram characteristics based on word pronunciation were retrieved.

Kumari et.al [8] (2021), suggested a hybrid model made up of four parts: fine-tuned VGG-16, CNN, Genetic Algorithm, and Classifier for Non-bullying and Bullying. A convolutional and max-pooling model is the second component, whereas the first is a pretrained model for pictures. The Sigmoid activation-function with categorical-cross-entropy was utilized, along with ReLU activation functions, GloVe42 embeddings, a 100-dimensional pre-trained GloVe42 embedding, a max-pooling layer, a flatten-layer, double fully connected-layers, and the CNN and VGG-16 networks. In order to create a combined feature, set of 384, the number of epochs was fixed at the aforementioned values. The highest performing features were chosen using GA, and then they were categorised using a typical machine-learning classifier. The main purpose of the study was to extract attributes from multi-modal data by simultaneously training the CNN for text and the VGG-16 for images. While CNN outperformed LSTM, VGG-16 performed better than the other three models. The attributes from the second final dense-layer of (CNN and VGG-16) were stored in order to merge the attributes from the text and the image. To establish the ideal size for image and text components, experiments were carried out. The system performs best when the size is 128 for the picture and 256 for the text, according to our testing of various combinations of image and text attribute sizes. The optimal features vector size was discovered using the random forest classifier. In order to acquire the (weighted) F1-score in the range of (76 to 81)%, they additionally ran the system 20 times.

In a research work put up by, T. Saha et.al [9] (2019), a potential NLP strategy called deep learning(DL) incorporates feature creation and classification into the learning process. Various NLP-tasks, including automatic-speech recognition, sentiment-analysis, dialogue-systems, and machine-translation, have been successfully completed by it. In this study, we propose a linear SVM architecture and two parallel 1-D convolutional layers as a standby to the softmax-function for tweet act segregation. In multinomial classification, (SVM) is used to identify the best hyperplane separating two-classes in a dataset. To make the CNN-based model more robust, other elements were added, such as Twitter-specific characters, abbreviations, punctuation, and emoticons. The created classification system makes use of binary features to identify tweet acts that are indicated by sentiment, opinion words, vulgar terms, and n-grams. In contrast to the CNN-Softmax and CNN-SVM models, the (CNN-SVM) model performed well. In terms of accuracy and F1-measure, the combined attribute-based classifier fared better than both the syntactic and semantic attributes. The majority of tags in the dataset have skewed representations, according to error analysis, with the top performing tags having the highest percentages of "EXP" and "STM" tags.

The proposed best performing model surpassed both state-of-the-art techniques in terms of accuracy and F1-score, and the F1-scores of the individual classes were also higher in the generated classifier as compared to the state of the art. These are the most crucial points in this article. This is because there isn't just one dataset or set of tags that can be used for tweet act classification.

Nabi Rezvani et.al [10] (2021), in order to do Cyber-bullying detection, they suggest a method for collecting attributes, merging them, and classifying social media target content. Maximizing context use within text, using context from sources outside of text, and a novel architecture for merging characteristics are the three primary contributions. The suggested method extracts three characteristics—average reactions, replies, and frequent mentions—that are representative of a user's most popular categories. Visual features are retrieved from articles' associated photos, and enrichment features are taken from a list of common profane words provided by Google. To get the appropriate class label, context features are integrated using a fully connected neural network or a boosting model with soft voting. They contrasted the two suggested models with the other three predictors. The third model makes use of an LSTM network, while the first two models employ a traditional Neural Network (NN) architecture. The results indicate accuracy, precision, recall, and f-score for the five models, which incorporate textual and contextual information. By using contextual variables, the suggested contextualised LSTM model has greatly outperformed the baseline model on most criteria. The boosting combiner outperformed the LSTM in terms of recall, while the NN combiner fared better for all other measures. With the exception of the boosting combiner, which has not performed as well, the findings of the Twitter dataset are comparable to those of the Instagram dataset on all measures.

Paul et.al [11] (2022), in their paper came up with a solution suggesting, bi-directional language models are used by BERT, a multi-layer bidirectional Transformer encoder, to learn broad language representations. A single sentence or two sentences can be represented in one token sequence using the BERT input representation. Every sequence begins with a specific categorization token, denoted by [CLS]. We employed knowledge distillation, a more straightforward variant of BERT that condenses the data learned by a large model into a comparable little model, to reduce the cost of the network. Using Scikit-Learn-0.22.2 and PyTorch-1.4.0, the BERT model is a text categorization model based on machine learning that is trained on the (tf-idf) vectors of the corpus. It is contrasted against three conventional machine learning-based text categorization models, (CNN, LSTM, BiLSTM) with attention-layer, and these four other models. By applying a soft target distribution and training the distilled model on a transfer set, knowledge is transferred to the model. The mean F1 scores across five runs, the inference times on the validation sets, and the comparison of the prediction standard on the validation set and test set are the most crucial information in this book. The experiments were carried out five times, and the results are statistically significant. According to the data description section, there are much less cases of the bully class than there are overall non-bully instances. When compared to other state-of-the-art models for various data, the suggested model performs statistically significantly better. The proposed models are unable to categorise specific posts or comments that contain comments that are likely to occur in other categories, according to error analysis. The use of offensive or profane language in the postings or comments as well as non-standard English terminology are factors in this.

In the research work of Roy et.al [12] (2022), cyber-bullying detection has been carried out using Convolutional Neural Network (CNN) frameworks, and a two-dimensional CNN (2DCNN) operates in three stages: feature extraction, feature selection, and feature flattening. A 2D matrix that has been calculated using element-wise multiplication and taking the sum is the output matrix. Convolutional feature extraction is used by the CNN to extract features, which are then pooled before being passed on to a fully connected-layer at the very last. ReLU serves as the activation function, while the dropout prevents overfitting. Image-based cyber-bullying posts are predicted using transfer-learning models, with VGG-16 and InceptionV3 outperforming other models. The majority of the information for data garnering and annotation of image-data for cyber-bullying came from MMHS150K dataset and Google image searches. Bully and non-bully labels were applied to images that had been converted to.jpg format. Images were converted to the same target size for each model by data pre-processing. The three stages of system design include data collection, pre-processing of the data, and training and testing. Models VGG16 and InceptionV3 were chosen for additional experimentation. 50 epochs of experimentation with different 2DCNN, VGG16, and InceptionV3 models have been conducted with an early-stopping setting.

The suggested system [12] is a supervised model that has been assessed for accuracy, precision, recall, F1-score, confusion matrix, and area under the ROC-curve. The tests revealed that a CNN model with a convolution-layer and a learning-rate of (0.001) without dropout outperformed a model with a dropout layer in terms of performance. With varying dropout levels and learning rates, transfer-learning-based models (VGG-16 and InceptionV3) performed finer than VGG16 and had a reduced rate of misclassification. While the transfer learning model with 3000 samples achieved precision, recall, and F1-score of 0.66, 0.72, and 0.69, the 2DCNN model with a single convolution layer achieved accuracy and an AUC-value of 0.51 instead.

With more data samples, the transfer-learning models VGG-16 and InceptionV3 performed better. VGG-16 had the best results with a dropout value of 0.5 and InceptionV3 had the peak results with an F1-score of 0.89 for the bully class. The proposed transfer learning model surpassed previous work in image-based cyberbullying prediction with optimised hyperparameter settings.

Data collection, data removal, and data annotation are the three processes that make up the data preparation procedure for the study of Alican Bozyiğit et.al [13] (2021). A balanced dataset of 5,000 Turkish labels containing a variety of social media properties was created and made available to the public in Comma Separated Value format. Cyberbullying To obtain and return models from a relational database, context was created. To decrease the amount of data to be analysed and the size of the dataset, data elimination was used. A strategy that relied on crowdsourcing was used for data annotation. Users can annotate random tweets from the Cyberbullying Context and flag them as having objectionable content using the built web application. As users/annotators, three data scientists with master's degrees from departments with ties to computer science were registered. A dataset of 5000 tweets was chosen. This study investigates which social media characteristics are associated with instances of cyberbullying and relevant to machine learning techniques. The association between the class of the samples and their characteristics was examined using a Chi-square test. Discretization was used to visually analyse the dataset's relationship between social media elements and cyberbullying, demonstrating how heavily dependent the samples' class is on social media features. The analysis's findings indicate that people with more followers are less likely to upload content that promotes cyberbullying, but those with less followers are more likely to do so. The dataset underwent pre-processing and normalisation procedures to enhance machine learning outcomes. Numerical social media features were subjected to min-max normalization, characters, punctuation, and weblinks were eliminated, and text pre-processing was used to fix misspelt terms related to online bullying. To lessen the noise in the dataset, this study used numerical normalization, text cleaning, lowercase conservation, and correction for misspelt slang words. The bag of words method was used for feature extraction, and recursive feature selection was employed to choose pertinent textual features. Supervised techniques for classification and regression include SVM, LR, and KNN. High-dimensional spaces benefit from SVM, training using LR is effective, and KNN is a slow learner. The ensemble learning algorithms NBM, AdaBoost, and RF are untrained and robust to noisy data.

Viktor Golem et.al [14] (2018), in their research, they tested models using 5-fold cross-validation and five trained models, holding out test data from the official development set. They used characteristics including problematic word occurrences, POS tags, text length, and capitalization features with three basic models: logistic regression, CNN, and BiLSTM. The emotion of the text was extracted and predicted using deep learning models, numerical tokens, named entities, sentiment polarity, and other techniques. Adam was used to train BiLSTM and CNN while hard predictions and an SVM with an RBF kernel were used to test the voting combination of seven models. On two data sets, the task creators examined three models, with the best performances coming from BiLSTM with Common Crawl embeddings, logistic regression with bigrams, and SVM. Two models were among the top 5 on both datasets, according to error analysis. Despite having a tendency to predict covert aggressiveness when the sentence should have been classified as non-aggressive, their model was able to distinguish between open and covert hostility. No model accurately predicted non-aggression, leading to misclassifications of overt and covert aggressiveness. The same choice would have been made by human annotators.

This study of N. Hettiarachchi et.al [15] (2020), seeks to separate hate speech from regular pornography by identifying it in online forum comments. The steps involved in data pre-treatment and feature extraction, as well as the layout of the experimental setup and the employed methods, are discussed. The data collection consists of various articles with user-authored comments from social media, all of which are written in romanized Sinhala. If a hate tale exists, the data set will be manually annotated. 1400 records from the collection and 1100 comments classified as hate speech were manually annotated. The sklearn toolkit in Python was used to implement feature extraction, and the supervision learning technique was employed to detect hate speech in Romanized Sinhala. Pre-processing was utilised to reduce noise.

The best algorithms for identifying hate Speech, according to the results, were Logistic Regression(LR), Multinomial-Naive Bayes(NB) Classifier, Linear SVM, and Random Forest(RF) Classifier. comparing the accuracy, precision, recall, and F1-score of four classifiers and feature sets. The data was split into training and testing datasets, and count-vectorizer and Tf-idf Vectorizer were used to extract features.

We analysed the accuracy values for logistic regression, SVM, and random forest. The Multinomial-Naive Bayes(NB) Classifier, which has the best accuracy, precision, recall, and f1-score values, is the best classification model. Tf-Idf Vectorizer and Multinomial-Naive Bayes(NB) Classifier with bi-gram and min_Df value 3 are chosen as the training models because they outperform other ML-classification models with bi-gram and min_Df value 3 in this study.

Wessel Stoop et.al [16] (2019), in their work, a framework HaRe (Harassment Recognizer) was created to track ongoing dialogues and assess their effectiveness. By concatenating all utterances for each speaker, separated by [NEW UTTERANCE] tags, then segregating the resulting text, it keeps track of toxicity estimations for each participant independently. HaRe's source code and application are accessible at [21]. 1000 talks from the dataset were used for evaluation, while 4000 were used for training. Concatenating all player utterances produced training texts that resemble the dialogues that are presented to the classifier during the classification phase. When comparing the training and classification phases, it is important to note that only 10.3% of the texts were classified as harmful during the training phase, but the training phase texts were down-sized to have an equal (50–50)% distribution of toxic and harmless textual data. While the talks during the categorization phase were frequently not yet finished, training was conducted on fully finished dialogues. The level of confidence over which a player is regarded as toxic has a significant impact on the rate of recall growth and precision decline. As a result, the ideal threshold—i.e, the threshold that yields the highest F-score—can fluctuate over a conversation. The size of the training set determines how quickly this sliding threshold should be raised; the bigger the train data, the slower the threshold can be raised.

Elizaveta Zinovyeva et.al [17] (2020), in their research to automatically train abstract data representations and extract features, TML and DL models are utilised. Sequential data may be handled by RNNs, whereas long-term dependencies can be handled by LSTM and GRU. Higher parallelization and fine-tuning on certain datasets are made possible by transformers. When building a representation of the document, the HAN and psHAN models employ attention mechanisms at several levels to give different content areas more or less attention. PsHAN is a network that streamlines and accelerates pre-processing by dividing incoming documents into pseudo-sentences of arbitrary length. The length of a sentence and the number of sentences in an input document are two new hyper-parameters introduced by psHAN, which enables data rearranging without obliterating any information. For shorter sentences, psHAN adds additional zeros, while for longer phrases, it drops certain words. Pseudo-sentences are used in place of actual sentences in the same architecture as HAN. Pre-trained language models produce the greatest results, while CNNs and DL approaches may also outperform TML. Bidirectionality does not always boost performance. Since the effective deployment of DS systems depends on the interpretability of DL models, Ribeiro et al.'s LIME approach offers explanations to support decision-making and foster confidence in the detection model. LIME can show unintentional bias in AOB detection models, which can be eliminated by enhancing the DS system with explainability.

This study by M. T. Ahmed et.al [18] (2021), uses NLP approaches, ML, and DL-algorithms to find cyber-bullying texts in comments on YouTube videos. Datasets were manually chosen, pre-processed, feature extracted, and divided into bullying and not bullying classes. Multinomial-Nave Bayes(NB), SVM, Logistic Regression(LR), and XG-Boost were the four models they employed. The probabilities for each category are calculated by Nave Bayes using the Bayes theorem, the largest margin is selected by SVM, predicted values are converted to probabilities by Logistic Regression using a logistic sigmoid function, and DL-algorithms including (CNN, LSTM, BLSTM, and GRU) are utilized. Four layers make up the CNN architecture: the Dropout Layer, the Convolutional-Layer, the Max-Pooling Layer, the Flatten-Layer, and the Dense-Layer. For classification, a layer is used for each. LSTM is an RNN enhancement that lessens the vanishing gradient problem and remembers dependencies over wide gaps. Two LSTMs are trained on the input sequence in the BLSTM, an extended form of LSTM. An update and reset gate is used by GRU, an upgraded version of RNN, to address the vanishing gradient issue. The most accurate algorithm for Dataset 1 was CNN, whereas the best algorithms for Dataset 2 were Multinomial Nave Bayes, SVM, Logistic Regression, XGBoost, CNN, LSTM, BLSTM, and GRU. The most effective algorithms in Dataset 3 were Multinomial Nave Bayes, SVM, Logistic Regression, XGBoost, CNN, LSTM, BLSTM, and GRU. All Machine Learning and Deep Learning algorithms performed best on Dataset 1, with CNN classifying 841 occurrences correctly and 159 instances inaccurately. The least amount of correctly identified examples (740 correctly and 260 wrongly) came from Multinomial Naive Bayes. 1180 examples were properly identified by Multinomial Nave Bayes, 1091 by BLSTM, 1784 by XGBoost, and 0.84 by CNN for both the negative and positive classes. Multinomial Nave Bayes displayed the lowest scores. The top algorithm for Dataset 2 was Multinomial-Nave Bayes(NB), with precision-scores of 0.82 and 0.87, recall-scores of 0.89 and 0.79, and f1-scores of 0.86 and 0.83 for the negative and positive classes, respectively.

Mangaonkar et.al [19] (2022), in order to identify cyberbullying behaviour, this research suggests a distributed collaborative detection framework that uses a collaborative method to categorise tweets as either bullying or non-bullying. Different distributed-collaborative setups perform differently depending on parameters in the suggested technique. When conducting experiments for the distributed-collaborative technique, it is crucial to take the network's nodes, the training set connected to each node, and any entry point(s) into account. Collaboration may be used to improve classification performance, however better performing nodes may not be impacted. A distributed-collaborative configuration's total performance will be impacted by the algorithms that the DNs apply.

In a distributed collaborative system, the number of opinions offered by other DNs to the entry point is crucial, and the number of collaborations desired determines when to participate. Using merging approaches, the performance of individual classifiers is enhanced by combining the results of various DNs. To evaluate the effectiveness of various distributed-collaborative arrangements, Twitter data was collected. Based on the training data and the classification method, Weka algorithms were utilised to construct DNs that can categorise tweets as bullying or non-bullying.

In the research work by A. Malte et.al [20] (2019), the pre-processing procedures guarantee the conversion of unclean data to clean text, and the Deep Bi-directional Transformer-architecture offers significant performance gains. It comprises of a Multi-Head Attention-layer, residual connections, a normalisation layer, and a generic feed-forward layer in an encoder-decoder architecture. A semi-supervised generative model is trained using the Bidirectional Encoder Representations from Transformer (BERT) architecture for context-based understanding of sequences. The Next Sentence Prediction challenge is used to train the model after it has already been trained as a Masked Language Model. By including linear layers to the output, fine-tuning is carried out. To increase the model's convergence and lessen the impact of overfitting in neural networks, SLTR and Dropout are used. The (weighted) macro F1-score metric was used to assess the model's performance. The results demonstrated that the model performed at the cutting-edge level on the Hindi Facebook test-data and attained a respectable third place on the English test-data. The model was tested on the unexpected test-data of Twitter comments without being tuned in order to further highlight its adaptability. It was discovered that pre-processing with transliteration considerably increased performance on the Hindi-test set.

III. CHALLENGES AND LIMITATIONS OF EXISTING APPROACHES AND TECHNIQUES

A. Sample Size And Representativeness

Choosing an appropriate sample size that is both significant and representative of the target population is one of the major issues in research on the detection of cyber-bullying. Small sample sizes and non-representative samples both have the potential to limit the generalizability of conclusions. Because cyberbullying can differ depending on aspects like age, gender, and ethnicity, researchers must also take into account the demographics and sample characteristics.

B. Validity And Reliability Of The Measures

To guarantee accurate and dependable results, the validity and reliability of cyber-bullying detection measures are essential. Reliability is the consistency of results across time and in many circumstances, whereas validity is the degree to which a measurement tool evaluates what it is meant to measure. Researchers must make use of legitimate, reliable measurement methods that are targeted at the right demographic.

C. Ethics And Privacy Issues

There are a number of ethical and privacy issues that come up when recognising cyber-bullying on social media, including the possibility of labelling innocent individuals as bullies or victims, invasions of privacy, and aggravation of victimisation. The ethical standards and principles that researchers must follow include gaining informed consent, maintaining participant confidentiality, and minimising potential harm. Researchers must also make sure that detection techniques don't invade users' privacy or result in more victimisation.

D. Methodological Issues

There are several methodological issues and limitations with regard to the detection of cyber-bullying, such as the absence of a widely accepted definition of the phenomenon, the dynamic nature of social media platforms, and the requirement for ongoing updates and improvements to the detection techniques. Additionally, the accuracy and dependability of detection systems may be impacted by the usage of various data sources and formats, such as text-based or multimedia-based data. To guarantee the efficacy and precision of cyberbullying detection systems, researchers must overcome these methodological difficulties and constraints.

IV. CONCLUSIONS

In the current digital era, cyberbullying is a developing issue, especially on social media platforms, and spotting it is essential to preventing and treating its negative impacts. There are now several ways to identify cyberbullying on social media platforms, including hybrid methods, machine learning-based methods, and keyword-based methods. These techniques do have drawbacks, too, include low accuracy rates, ethical issues, and difficulties spotting subtle kinds of cyberbullying.

Sample size and representativeness, measure validity and reliability, ethical considerations, and privacy issues, as well as methodological difficulties and constraints, are difficulties and limitations of existing approaches. Standardizing definitions of cyberbullying, creating new detection tools, investigating the use of artificial intelligence and machine learning techniques, concentrating on certain demographics, carrying out longitudinal studies, and taking ethical considerations into account should be the main goals of future study. For the purpose of promoting safe and healthy online environments, preventing harm to individuals and groups, and addressing the wider societal implications of cyberbullying, it is vital to develop reliable and accurate methods for catching cyber-bullying on social media platforms.

V. RECOMMENDATIONS FOR FUTURE RESEARCH

Recommendations for future research on cyberbullying detection may include:

1) *Standardization of Definitions:* The lack of a uniform definition of cyberbullying can make it difficult to compare and generalise findings across studies. A standard definition of cyber-bullying that takes into account various types and forms of online harassment should be developed in the future.
2) *Creation of new detection techniques:* To keep up with the changes occurring in social media platforms, new detection techniques must be created. Future research should focus on developing unique detection systems that can accurately identify and characterise cyberbullying in real-time.
3) *Utilization Of Artificial Intelligence And Machine Learning:* The detection of cyberbullying on social media platforms has showed promise when using artificial intelligence and machine learning approaches. Future studies should examine the application of these methods to create more precise and reliable detection methods.
4) *Focus on Specific Populations*: Cyberbullying can affect different populations differently, and research should focus on establishing detection systems that are relevant to diverse age groups, genders, and cultures.
5) *Studies that are Longitudinal:* Studies that are longitudinal can shed light on the long-term impacts of cyberbullying as well as the effectiveness of prevention strategies over time. To better understand the evolution of cyberbullying and the role that detection plays in reducing its consequences, future study should take into account undertaking longitudinal studies.
6) *Ethical Considerations:* The identification of cyberbullying poses ethical questions involving privacy, confidentiality, and potential harm to participants. The ethical issues should be taken into account in future studies while creating and using detection techniques.

In general, future study should focus on improving the tools for spotting cyberbullying on social media platforms while also taking into account their ethical and practical ramifications.

REFERENCES

[1] Fan, H.; Du, W.; Dahou, A.; Ewees, A.A.; Yousri, D.; Elaziz, M.A.; Elsheikh, A.H.; Abualigah, L.; Al-qaness, M.A.A. Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit. Electronics 2021, 10, 1332. https://doi.org/10.3390/electronics10111332

[2] Beddiar, D.R., Jahan, M.S. and Oussalah, M., 2021. Data expansion using back translation and paraphrasing for hate speech detection. Online Social Networks and Media, 24, p.100153

[3] Rui Zhao, Anna Zhou, and Kezhi Mao. 2016. Automatic detection of cyberbullying on social networks based on bullying features. In Proceedings of the 17th International Conference on Distributed Computing and Networking (ICDCN '16). Association for Computing Machinery, New York, NY, USA, Article 43, 1–6. https://doi.org/10.1145/2833312.2849567

[4] Raj, M., Singh, S., Solanki, K. et al. An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques. SN COMPUT. SCI. 3, 401 (2022). https://doi.org/10.1007/s42979-022-01308-5

[5] Nayel, H.A., 2020. NAYEL at SemEval-2020 task 12: TF/IDF-based approach for automatic offensive language detection in Arabic tweets. arXiv preprint arXiv:2007.13339.

[6] Aljabri, M., Zagrouba, R., Shaahid, A. et al. Machine learning-based social media bot detection: a comprehensive literature review. Soc. Netw. Anal. Min. 13, 20 (2023). https://doi.org/10.1007/s13278-022-01020-5

[7] A. Shekhar and M. Venkatesan, "A Bag-of-Phonetic-Codes Modelfor Cyber-Bullying Detection in Twitter," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, 2018, pp. 1-7, doi: 10.1109/ICCTCT.2018.8550938.

[8] Kumari, K, Singh, JP. Identification of cyberbullying on multi-modal social media posts using genetic algorithm. Trans Emerging Tel Tech. 2021; 32:e3907. https://doi.org/10.1002/ett.3907

[9] T. Saha, S. Saha and P. Bhattacharyya, "Tweet Act Classification : A Deep Learning based Classifier for Recognizing Speech Acts in Twitter," 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019, pp. 1-8, doi: 10.1109/IJCNN.2019.8851805.

[10] Nabi Rezvani, Amin beheshti, and Alireza Tabebordbar. 2021. Linking textual and contextual features for intelligent cyberbullying detection in social media. In Proceedings of the 18th International Conference on Advances in Mobile Computing &amp; Multimedia (MoMM '20). Association for Computing Machinery, New York, NY, USA, 3–10. https://doi.org/10.1145/3428690.3429171

[11] Paul, S., Saha, S. CyberBERT: BERT for cyberbullying identification. Multimedia Systems 28, 1897–1904 (2022). https://doi.org/10.1007/s00530-020-00710-4

[12] Roy, P.K., Mali, F.U. Cyberbullying detection using deep transfer learning. Complex Intell. Syst. 8, 5449–5467 (2022). https://doi.org/10.1007/s40747-022-00772-z

[13] Alican Bozyiğit, Semih Utku, Efendi Nasibov, Cyberbullying detection: Utilizing social media features, Expert Systems with Applications, Volume 179, 2021, 115001, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2021.115001.

[14] Viktor Golem, Mladen Karan, and Jan Šnajder. 2018. Combining Shallow and Deep Learning for Aggressive Text Detection. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pages 188–198, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[15] N. Hettiarachchi, R. Weerasinghe and R. Pushpanda, "Detecting Hate Speech in Social Media Articles in Romanized Sinhala," 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2020, pp. 250-255, doi: 10.1109/ICTer51097.2020.9325465.

[16] Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. 2019. Detecting harassment in real-time as conversations develop. In Proceedings of the Third Workshop on Abusive Language Online, pages 19–24, Florence, Italy. Association for Computational Linguistics.

[17] Elizaveta Zinovyeva, Wolfgang Karl Härdle, Stefan Lessmann, Antisocial online behavior detection using deep learning, Decision Support Systems, Volume 138, 2020, 113362, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2020.113362.

[18] M. T. Ahmed, M. Rahman, S. Nur, A. Islam and D. Das, "Deployment of Machine Learning and Deep Learning Algorithms in Detecting Cyberbullying in Bangla and Romanized Bangla text: A Comparative Study," 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2021, pp. 1-10, doi: 10.1109/ICAECT49130.2021.9392608.

[19] Mangaonkar, A., Pawar, R., Chowdhury, N.S. et al. Enhancing collaborative detection of cyberbullying behavior in Twitter data. Cluster Comput 25, 1263–1277 (2022). https://doi.org/10.1007/s10586-021-03483-1

[20] A. Malte and P. Ratadiya, "Multilingual Cyber Abuse Detection using Advanced Transformer Architecture," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 784-789, doi: 10.1109/TENCON.2019.8929493.

[21] https://github.com/woseseltops/HaRe

[22] Cyberbullying: What is it and how to stop it. (n.d.). Cyberbullying: What Is It and How to Stop It | UNICEF. https://www.unicef.org/end-violence/how-to-stop-cyberbullying

[23] Global Social Media Statistics &mdash; DataReportal – Global Digital Insights. (n.d.). DataReportal – Global Digital Insights. https://datareportal.com/social-media-users