Naif Arab University for Security Sciences
Journal of Information Security & Cybercrimes Research
مجلة بحوث أمن المعلومات والجرائم السيبرانية
https://journals.nauss.edu.sa/index.php/JISCR

JISCR

A Systematic Review of Machine Learning Algorithms in Cyberbullying Detection: Future Directions and Challenges

Muhammad Arif*
Department of Computer Science, College of Computers and Information Systems, Umm Al-Qura University, Mecca, Saudi Arabia.

Received 07. Apr. 2021 ; Accepted 20. May. 2021 ; Available Online 01. Jun. 2021

Abstract

Social media networks are becoming an essential part of life for most of the world's population. Detecting cyberbullying using machine learning and natural language processing algorithms is getting the attention of researchers. There is a growing need for automatic detection and mitigation of cyberbullying events on social media. In this study, research directions and the theoretical foundation in this area are investigated. A systematic review of the current state-of-the-art research in this area is conducted. A framework considering all possible actors in the cyberbullying event must be designed, including various aspects of cyberbullying and its effect on the participating actors. Furthermore, future directions and challenges are also discussed.

I. Introduction

Bullying refers to aggressive behavior, which can be physical, verbal, or social [1]. Bullying is distinguished by three criteria, aggressive motive, repetition, and imbalance of power. It hurts individuals physically, mentally, or emotionally [2]. A bullying culture can develop in any environment where humans interact with each other. It may happen in the family, school, or workplace. Sometimes it is also referred to as peer abuse. If bullying behaviors prevail in society, it may affect the mental health of the underprivileged portion of the community in many ways. People who face bullying in their childhood or at the adolescent age are at higher risk of suffering from anxiety, depression, and low esteem than those who are not bullied [3]. Children or youth bullying others may also face psychological problems in their later years [4]. Even bystanders to bullying also develop mental stress and fear [5].

With the advancement of technology, the use of social media is growing every day. We are living in the internet-enabled world in which reaching other people is just a click away. People can openly share on social media, websites, and community forums. Although digital media offers various resources, many people misuse it in the name of freedom of speech or hatred towards a particular race, social group, or individual. Cyberbullying is a bullying action performed through digital means to embarrass, threaten, or socially exclude others [6]. In cyberbullying, elements of repetition, power differential, and motive are considered essential fac-

Keywords: Cybercrime, Information Security, Cyberbullying, Social media, machine learning, natural language processing.

Production and hosting by NAUSS

* Corresponding Author: Muhammad Arif
Email: mahamid@uqu.edu.sa
doi: 10.26735/GBTV9013

tors [7]. Cyberbullying victims report a higher level of depression and anxiety [8], suicidal trends and attempts [9], academic performance [10], work performance [11], and poorer physical and mental health [1]. The harmful effects of cyberbullying are more severe than traditional bullying because of the broader audience on the internet and the faster spread of messages. The harmful effects of cyberbullying are more stringent than conventional bullying because of a wider audience on the internet. Tommy et al. [12] described triadic reciprocal relationships between perpetrators, victims, and bystanders. This framework consolidated personal factors, environmental events, and behavioral patterns that influence each other in a triadic manner.

Early detection of different social anomalies, including cyberbullying, hate speech [13, 14], trolling [15], fake news [16], rumors [17], counterfeit profile detection [18], misogyny [19], etc., is becoming a trend in recent social media-based research. Abusive text can be detected in the messaging/comments on social media by maintaining a list of offensive words. However, words and phrases can be obfuscated by the users. The human readers can easily understand these confused words but extremely difficult for an automated system [20]. One solution to this problem is by updating the word list continuously. A survey on hate speech detection using natural language processing can be found in [21]. Cyberbullying can take different forms, including flaming (online fights), harassment, denigration, impersonation, outing (sharing secrets of someone to others), trickery, exclusion, and cyberstalking [22]. A good list of 118 causes of cyberbullying and methods are given in [23]. The role of a person's personality is also important to predict whether he/she may engage in the activity of cyberbullying [24]. Personality traits related to the dark triad (psychopathy, Machiavellianism, and narcissism) may influence a person's cyberbullying behavior.

A systematic review is conducted to identify and compare different machine learning methods used to detect cyberbullying. This methodological review is undertaken to answer the following questions:

RQ1: What are the existing machine learning methods applied to detect cyberbullying in social media?

RQ2: What are the possible applications of cyberbullying detection tools?

RQ3: What are the challenges and future perspectives of the cyberbullying detection frameworks?

The systematic review consisted of three steps. In the first step, three major research databases, IEEE explore, ScienceDirect, and Springer was searched through queries and collected as many papers as possible. The search queries are "cyberbullying detection machine learning," "cyberbullying detection," and "Cyberbullying natural language processing." Based on initial exclusion criteria, papers were selected after carefully reading the abstract of the papers in the second step. A final list of papers is prepared after reading the full articles and applying further exclusion criteria.

Exclusion criteria in the first and second steps include:

EC1. Studies must be peer-reviewed articles published in the English and Arabic languages.

EC2. Books, notes, theses, letters, and patents are not included in this review.

EC3. Only papers focusing on applying machine learning methods to the problem of cyberbullying are considered.

Three further exclusion criteria are used in the final step as follows:

EC4. A unique contribution is considered for inclusion, and repeated studies are not included.

EC5. Those articles which do not describe the methodology and result sufficiently are excluded.

EC6. Those articles that do not address the research questions mentioned above are not included. Cyberbullying research on languages other than English and Arabic is excluded.

Fig. 1 depicts the literature review process. Query searches on the three databases produced a total of 1032 articles. Article search is not limited to any specific period. In these articles, some of the articles are repeated. After carefully reading the papers' title and abstract, 803 articles are excluded based on the exclusion criteria EC1, EC2, and EC3.

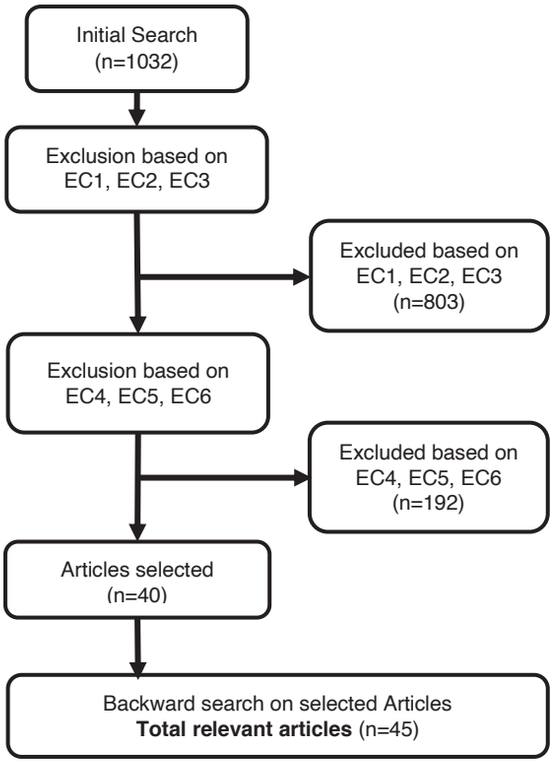

Fig. 1 Systematic literature review process.

After reading the remaining articles' text, 192 articles are excluded based on the exclusion criteria EC4, EC5, and EC6. Finally, 40 articles are identified. Based on the references or bibliographies of these papers, a backward search is performed to identify any additional relevant articles. Including five more articles, a total of 45 articles are shortlisted in the final list for review.

II. Role of Machine Learning in Cyberbullying Research

People connect on social media through various platforms. Every platform has certain limitations on shared content. Social media content can be text, image, video, or infographics. Monitoring cyberbullying on social media requires understanding the content, the social network of the connecting people, user activities, connection behavior on the social media, and users' profile. Fig. 2 describes a generic framework of monitoring and detecting cyberbullying events on social media using natural language processing and machine learning algorithms. Due to the diversity of the data, different algorithms are required to extract and select useful features from the data. Once important features are collected from the data, a classifier must be trained on this feature set to predict the correct events.

Performance analysis can be based on different performance metrics in machine learning and may include classification accuracy, precision, recall, or F-Score. Classifier structure can be fine-tuned to achieve an optimal performance of the cyberbullying detection and monitoring framework. In the following sub-sections, the use of different feature extraction and classification methods in cyberbullying detection literature is described in detail.

A. Feature Extraction

Features for cyberbullying detection can be broadly classified into Content features, Network features, Activity features, User profile features, and sentiment features. Table I summarizes a list of features used in cyberbullying detection in the literature. Word embedding features are the most common features used for cyberbullying detection in the literature.

Further details about different types of features used in cyberbullying research can be found in the following subsections.

1) Content features

Content features may include textual features, emoji-based features, and features extracted from audio, images, or video contents. Many valuable features can be extracted from the text posted by the user based on natural language processing. Some features may depend on the word dictionary, and some features also consider the context of a sentence. Based on the vulgar/profane words dictionary, the vulgarity feature can be calculated by the number of offensive words present in the user's post.

N-gram features: N-gram is the probabilistic model of a sequence of n adjacent items (words, phonemes, syllables, letters, etc.). In the linguistic sense, n-grams are collected from text or speech corpus. In n-gram, n refers to the number of items or words in the text. N-gram features have numerous applications in natural language processing and other areas [51-53]. In this model, conditional

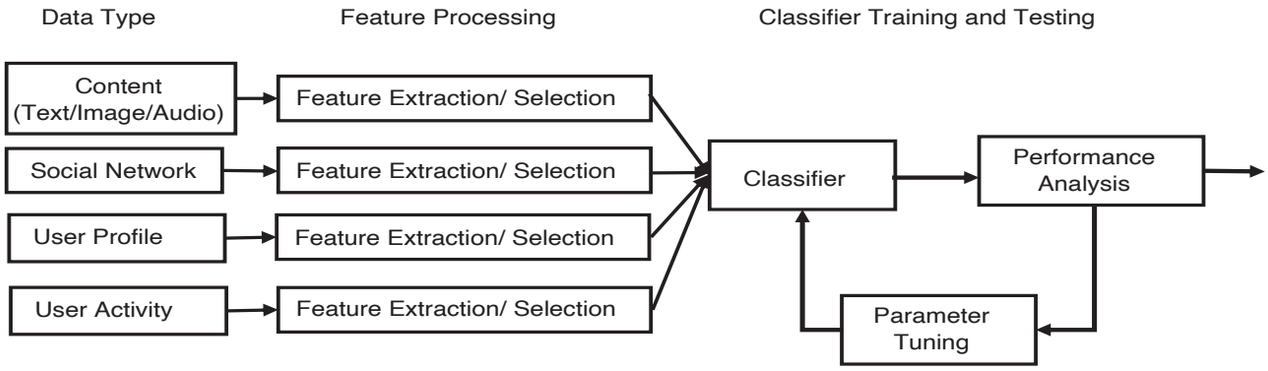

Fig. 2 Machine learning framework for monitoring cyberbullying.

TABLE I
USAGE OF DIFFERENT FEATURES TYPE IN CYBERBULLYING DETECTION

Source of Features	Category	Feature Type	Usage
Content based Features	N-grams based Features		[20], [25], [26], [27], [28], [29], [30]
	Distance Measures	Edit-Distance	[20]
	Word Embedding	Word2Vec	[31], [32], [28], [33], [34], [35], [30], [36], [37], [38], [39]
		Skip-gram	[25]
		CBoW	[25], [32]
		BoW	[40], [31], [33], [34]
		TF-IDF	[26], [27]
		FastText	[25], [36]
		GLoVe	[41], [42], [37]
		LSHWE	[37]
	Vulgarity/Hate Features		[43], [25], [32], [33], [44]
Sentiment	Sentiment Analysis		[27], [45], [32], [33], [41], [46], [30], [39]
User	Profile		[27], [47], [43], [25], [32], [33], [46], [44]
	Personality traits		[45], [24]
	Dark Triad		[24]
Network			[45], [43], [48], [32], [33], [46], [49]
Media Related Features			[31], [41], [49], [50], [44]

probabilities of a word given all the previous words $P(W_i|W_{i-(n-1)},\ldots,W_{i-1})$ are calculated. Probability distribution is further smoothed for unseen n-grams. Through these conditional probabilities, we can predict the next most likely word. Binary features can be presence or absence or n-grams in the text. N-gram frequency profile is the frequency of presence of the n-grams in the text.

Word Embedding Techniques: Word embedding is a vector representation of the text where words with similar meanings have similar representations. Words are represented as real-valued vectors in a vector space. According to the usage of the words, the similarity of the context of the words is learned. By word embedding, the semantic and syntactic similarity of the word in the corpus can be understood. It can also capture the relation of a word with other words. There are different methods to learn this representation from the text corpus. Word2Vec [54, 55] is a statistical method to learn word embedding from the text corpus using neural networks. The Skip-gram model [55] predicts the context words for a target word. It is an unsupervised learning technique that can learn the context of any word. The bag of words (BoW) model extracts features from the text by occurrence frequency of words in the text. A histogram of words can be used as an input to the classifiers [56]. BoW model does not consider the context of the word in the sentence, and hence semantic of the word is lost. Continuous Bag Of Words (CBOW) [55] predicts the current word by using the contexts (surrounding words). Both methods are used to learn the usage context of a word. Global vector space

representations of words can be constructed [57]. Word-word co-occurrence counts are tabulated from the corpus. The probabilities of a word i appearing in the context of a word j are calculated and called co-occurrence probabilities. The ratio of these probabilities can be used to model the context of the words. Distance between the words relates to the semantic similarity of the words. Term Frequency-Inverse Document Frequency (TF-IDF) finds the importance of the word in the document or corpus [58]. Term frequency (TF) measures the frequency of occurrence of a term or word in the document. Whereas inverse document frequency (IDF) measures how important this term or word is? It is calculated by taking the logarithm of the ratio of the total number of documents in the corpus and the number of documents having this term or word.

TF-IDF score is calculated for each word in the corpus by multiplying TF and IDF. Fasttext [59] is an extension of the word2vec model. In this model, every word is represented as an n-gram of characters. A skip-gram is used to learn the embedding. Zhaou et al. [37] proposed a new word embedding method called Locality Sensitive Hashing Word Embedding (LSHWE) to represent obfuscated words in cyberbullying events. The method assumes that deliberately a obfuscated word has a high context similarity with their corresponding actual word. The assumption is fair; otherwise, the victim may not be understanding this word either. Based on the co-occurrence matrix and rare word list, a nearest neighbor matrix is generated through locality-sensitive hashing from the corpus. A LSH based autoencoder model is used to learn every word representation.

Sentiment Analysis: Sentiment is a feeling provoked by text, image, or video. Positive, neutral, or negative scores of the sentiments can be used as features towards cyberbullying detection [60].

Media Related Features: In the case of media (audio/image/video) sharing by the user, features such as a number of likes/dislikes, comments, and sharing, the subjectivity of the media caption are important in the prediction of the cyberbullying associated with it [40]. Further features can be extracted from the comments section of the media, including a percentage of negative comments, profane words in the comments, average polarity/subjectivity of the comments, top topics using Latent Dirichlet Allocation (LDA) from all the comments, etc.

Vulgarity Features: Vulgar words represent hostile or aggressive behavior and can be related to the perpetrator's cyberbullying effort. Based on a dictionary of vulgar or profane words, the vulgarity feature can be a number of vulgar or offensive words present in the post's content.

2) User Network features

There is a strong correlation between cyberbullying and how social a perpetrator is on the social network [61]. Therefore, the social network features of users are essential for the detection of cyberbullying. It includes followers of a person, the number of users a person follows. A following to follower ratio is also an important feature. The popularity of the user plays a vital role in the severity of cyberbullying. Any attack from a popular user hurts the victim more as the perpetrator's popularity creates the power imbalance between the perpetrator and the victim [62].

3) User Profile Features

Various features can be extracted from the user profile and his/her activity on social media. The age and gender of the user can be important of the cyberbullying perpetrator and victims. The age difference between the perpetrator and the victim sometimes plays a vital role in assessing the severity of the bullying. Types of social groups of the user can be predicted from his/her social network [32]. Many researchers have worked on predicting the personality of the user by his/her social activity data. By knowing the character of the person, we can understand human behavior. Cyberbullying sometimes correlates with the aggressive or hostile behavior of the user [63].

Similarly, neurotic people show anger, anxiety, and moodiness and can engage in an activity leading to cyberbullying [64]. Many studies have shown that there is no relationship between gender and cyberbullying. However, few experts suggest that including gender in the machine learning models may

improve cyberbullying detection [47]. Gender information of a user can be predicted by his/her writing style [65, 66]. There may exist some correlation between cyberbullying and age groups of the social media user [67]. Therefore, prediction of the age group from the social media data can be beneficial in cyberbullying classification [68]. Personality traits also correlate with the bullying activity of the person. Big Five model [69] identifies five personality traits: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. Mitsopouloua and Giovazolias [70] found that lower levels of Agreeableness and Conscientiousness and higher levels of Neuroticism and Extraversion are associated with bullying behavior. Researchers have tried to predict these five personality traits from a person's digital footprint on social media that includes posted text, images, videos [68, 71, 72]. Contextual features may consist of the profession, religion, family, and financial factor of the user.

4) User Activity features

Activity features can measure the online communication activity of a user. It may include the number of posts/tweets, the number of posts/tweets liked or disliked, and hashtag activity.

5) Sentiment features

Sentiment features can also help detect aggressive and abusive behaviors [32, 60]. The post's sentiment can be decided by a well-trained classifier or from a dictionary of the words in which sentiments are related to the words [73]. Jain et al. [74] compared various word embedding techniques for hate speech detection. Few works are reported about the use of personality features in cyberbullying detection. Personality prediction from the social media traces can help predict the roles of users in a cyberbullying event. Recent advances in increased internet speed, and users more frequently sharing video clips and other visual content forms. Hence, more research is needed to identify cyberbullying events through images or videos.

B. Classification

A range of classifiers is available for the detection and monitoring of cyberbullying events on social media. The feature extraction section mentions that textual features can be converted into vectors by embedding techniques and other features are also numeric, categorical, or binary values. Hence, many well-known classifiers, already proven in various real-life applications, can be used to classify the presence of cyberbullying events, the severity of the cyberbullying events, and the identification of perpetrators and victims.

Table II summarizes the number of articles using different classifiers in cyberbullying detection. Naïve Bayes classifier, SVM, RF, and RNN are the most commonly used classifiers. In recent years, deep learning architecture have the attention of researchers in natural language processing. Hence,

TABLE II
DIFFERENT CLASSIFICATION METHODS IN CYBERBULLYING DETECTION

Type of Classifier	Classifier	Usage
Regression/ Statistical Classifiers	Logistic Regression	[75], [25], [48], [28], [41], [46], [30], [38]
	KNN	[43], [48], [41]
	Naïve Bayes	[27], [43], [25], [28], [33], [41], [46], [76], [29], [77]
Decision Tree Algorithms	CHAID	[23]
	Exhaustive CHAID	[23]
	QUEST	[23]
	Regression Tree	[30]
	J48	[27], [45], [39]
Support Vector Machines		[26], [27], [43], [25], [48], [28], [41], [46], [76], [29], [78], [39]
Ensemble of Classifiers	Random Forest	[32], [45], [43], [24], [48], [28], [33], [41], [46], [30], [38], [39], [44]
	Bagging/Boosting	[24], [79], [78], [44]
	Voting Ensemble	[33]
Deep Learning	CNN	[80], [81], [28], [42], [36], [78]
	CapsNet + ConvNet	[82]
	RNN	[33], [75], [28], [42], [36], [37], [83], [84], [49], [50]
	BERT	[25], [85]
Reinforcement Learning		[86]

many papers have focused on using deep learning architectures like RNN, CNN, BERT, etc., especially in text-based analysis. A detail of the classifiers is given below in the subsequent subsections.

1) Regression/ Statistical Classifiers

In regression analysis, statistical methods are used to estimate the relationship between dependent and independent variables. Linear and logistic regression methods are used in classification or prediction modeling [87]. The linear regression model assumes a linear relationship between dependent and independent variables. Logistic regression [88] is an extension of the linear regression model used for classification purposes. It models the probabilities of two classes by a logistic function at the output of the linear model. Maximum likelihood estimation is one of the cost functions that can be used to find the coefficients of the logistic regression model. A proper threshold at the output of the logistic function can give binary classification. This classifier is used in the literature on cyberbullying when there are two classes, cyberbullying or no cyberbullying. K-nearest neighbor (KNN) classifier [89] is a nonparametric classifier that classifies based on the majority class of k number of nearest neighbors. Naïve Bayes classifiers [90] are based on Bayes' theorem using conditional probabilities. All the features used in the classification are considered independent from each other. Hence, the presence or absence of any feature does not affect other features. Probability model of every feature is combined with a decision rule to select the class. The decision about the class can be taken by choosing the most probable class (Maximum a posteriori decision rule) or maximum likelihood estimate (MLE decision rule).

2) Decision tree-based Classifiers

Decision trees, more recently called Classification and Regression Tree (CART) [91], are used for classification and regression modeling. Each nonleaf node is labeled with a feature or attribute and arcs from this node are possible values of this attribute in the classification tree. The leaf nodes represent the class or the probability distribution over the classes. In the case of the regression tree, predicted output is a continuous real number. J48 classifier is an open-source implementation of C4.5 decision tree algorithm [92]. This algorithm builds a tree based on the training dataset. For every node in the tree, the algorithm selects the best attribute to split with the highest normalized information gain. More nodes are generated on the remaining attributes. Once all attributes are covered, a single pass pruning is performed on the tree to minimize the overfitting. Chi-square Automatic Interaction Detector (CHAID) algorithm [93] builds the tree to detect the relationship between independent variables. This algorithm can be applied to nominal, ordinal, and continuous data split into categories. Chi-square test of independence (p-value) is applied at each of the stages to find the association between two categorical variables, and F-test is used for continuous variables. Bonferroni corrections are calculated to counter the problem of multiple comparisons/ testing. CHAID algorithm selects predictors having the strongest interaction with the dependent variable. Exhaustive CHAID [94] creates all possible splits for each predictor and groups the categories of each predictor optimally.

Moreover, Bonferroni correction in exhaustive CHAID is revised to counter excessive penalization in the CHAID algorithm. Quick Unbiased Efficient Statistical Tree (QUEST) algorithm [95] is another binary-split decision tree-based algorithm. QUEST uses the ANOVA F-test for non-categorical variables and the chi-square test for categorical variables to select the splitting variables. A Variable having the smallest significance probability is assigned for the split. It has low variable selection bias and computationally simple. One of the advantages of QUEST algorithm is its unbiasedness in variable selection for the split. If there are more than two classes, then classes are grouped into two super classes before the application of quadratic discriminant analysis.

3) Support Vector Machines

Support vector machines [96] are well-known classifiers in the community of machine learning, and it is used successfully in a variety of real-life problems [97]. SVM is a supervised classifier that trains itself to find maximum margin hyperplanes

between classes to minimize generalization errors. Training data is mapped to a higher dimension using kernels to maximize the separation between classes. With this kernel trick, SVM can classify nonlinearly separable classes. Utilizing the support vectors, which are the data points near the decision boundary, the hyperplane's position and orientation are adjusted to maximize the separating margins. A hinge loss function for margin maximization along with a regularization parameter makes the SVM a robust classifier. A good review about variants of SVM can be found in [98].

4) Ensemble of Classifiers

In an ensemble of classifiers, individual decisions of a set of classifiers are combined in a certain way (for example, voting) to predict the class of a new data point. Ensemble of classifiers improves the predictive performance as compared to an individual classifier. The idea is to combine several weak learners to form a strong learner. Recently a lot of research has been done on the ensemble of classifiers [99]. Few of such classifiers are intensively used in cyberbullying detection.

In the bagging classifier [100], several training data subsets are chosen randomly from the training data with replacement. For each subset of data, a decision tree is trained, and hence an ensemble of classifiers is created. This ensemble of classifiers is trained independently, and prediction is based on the aggregate of their outputs. In this way, the variance of the classifier is reduced. Aggregation of the outputs of weak learners can be through hard voting (majority voting) or through soft voting. In soft voting, class probabilities of all the classifiers are averaged, and the class with the highest probability is selected. Bagging can be implemented in parallel. In boosting [101], classifiers learn sequentially. A learner learns the simpler model, and by analyzing the classification errors on this classifier, subsequent classifiers are fitted on the training data. All weak learners are combined by majority voting. There are many variants of boosting classifier including AdaBoost [102], LogitBoost, BrownBoost [103], GentleBoost [104] etc., LogitBoost [105], BrownBoost [103], GentleBoost [104] etc. The voting ensemble method uses voting of multiple classifiers to decide the class of an unknown data point. It uses different classifiers on the same dataset or uses the same base classifier on a different subset of the data. Due to the diversity of the classifier, an ensemble of classifiers performs better than a single classifier.

A random forest classifier was proposed by Ho et al. [106] and later formulated by Breiman et al. [107]. It grows a forest of decision trees that split the feature space with hyperplanes. Training data is projected to a randomly chosen subspace (a randomly selected subset of features) and used to train the decision tree using bagging. Hence, in the random forest classifier, many decision trees are trained in parallel with bootstrapping followed by bagging. The random forest classifier variance is minimized by ensuring that every decision tree in the forest is unique. It exhibits good generalization without overfitting issues. Different base classifiers can be used in the random forest classifier [108, 109]. Random forest classifier is successfully applied in many practical applications [110-112]. Several studies demonstrated the potential of using kernel functions in the random forest classifier [113, 114].

5) Deep Learning Classifiers

Traditionally, features are designed by a human to train the machine learning algorithms, which require a lot of expertise and domain knowledge. Deep learning architectures exploit powerful neural networks containing multiple layers without the burden of feature engineering. Several useful deep learning architectures are successfully used in natural language processing [115, 116], image, and video processing. These deep learning architectures include Convolutional neural networks [113], Deep belief networks [115], Deep recurrent neural networks [113], Deep stacking networks [117], Generative Adversarial networks [118], LSTM [117] and variants of LSTM [119, 120], etc.

Convolutional neural networks were originally designed for image processing [121]. It contains many hidden layers, including convolution layers which perform convolution on the input, pooling layers, fully connected layers, and normalization layers. Convolution layers contain filters to learn features from the input (feature map). Pooling layers reduce the

dimensions of data. Max pooling, sum pooling, and average pooling are the common operations in the pooling layers. Fully connected layers are the same as the multi-layer perceptron. There are different types of normalization layers in deep learning architectures [122]. In batch normalization, input to any layer is normalized to have a pre-defined mean and variance. Another type of normalization is weight normalization, in which weights are normalized to improve convergence [123]. Layer and group normalization normalize the activations in feature direction [124, 125]. Existing popular CNN architectures are AlexNet [126], VGGNet [127], ResNet [128] etc. A good survey can be found in [129].

Recurrent neural networks (RNN) are designed to handle the temporal sequences of the inputs, which is very important in natural language processing. The recurrent neural network allows the previous outputs to be used as inputs by looping the output to the input of the network. Hence, an input of any length can be handled by RNN. RNN usually has short-term memory and cannot tackle long-term dependencies. Vanishing gradients, when backpropagated through time, cannot contribute to the learning. Sometimes long-term dependencies also matter in the correct prediction of the output. These scenarios exist more frequently in natural language processing. Long short-term memory (LSTM) networks are widely used in applications requiring temporal processing of the inputs and outputs [130]. These networks are especially suited to learn the long-term dependencies. LSTM consists of connected subnet call memory blocks, and each block contains one or more self-connected memory units comprising a cell, an input gate, an output gate, and a forget gate. Information flows from the input to the output through cells and removing information or adding new information is done through gates. LTSM stores information in these gated cells and passes the long chains of sequences to make the prediction. Bidirectional LSTM (Bi-LSTM) is an extension of classical LSTM in which two independent RNN structures are used to forward and backward the information at every time step.

Bi-LSTM learns the context in the text better than LSTM [131]. Encoder-Decoder LSTM [132] reads the input sequence and encodes it to a fixed-length vector. The decoder part decodes the vector into the predicted sequence. In the context of machine learning, concentrating relevant concepts in the data and ignoring the irrelevant ones is done by an attention mechanism. Attention network is used between encoder and decoder layers of LSTM by Google neural machine translation. Attention network is a single layer RNN encoder whose weights are adjusted using a fully connected shallow network and a softmax function [85]. The attention mechanism does not consider the order of the sequence and model the relevance between representation pairs.

The transformer model [85, 133] uses the attention to boost the training speed. It transforms one sequence into another sequence by using encoder and decoder layers. Bidirectional Encoder Representations from Transformers (BERT) [134] learns the text representation bidirectionally. By doing so, BERT captures the language context more accurately. BERT uses an attention mechanism called a transformer to understand the contextual relationship between words. It is trained by the text sequence with a certain percentage of the masked tokens, and the network must predict the masked tokens.

Moreover, pairs of sentences as input and target sentences are used in the training process. Few recent works have used more advanced language models like BERT and variants of BERT. Due to the BERT model's high computational cost and larger memory footprint, many lighter versions of BERT are also proposed [135-137]. Hardware acceleration of the BERT model is also an active area of research [138]. Emotion detection using BERT-based models is presented in a recent paper [74], which may help extract the sentiment features for cyberbullying detection.

6) Reinforcement Learning

Reinforcement learning [139] is another paradigm of artificial intelligence in which an agent learns an optimal or near-optimal policy to maximize a reward function. For each action of the agent, there is either a positive or negative reward associated with it. Hence reinforcement learning uses a balance between exploration and exploitation and generates a sequence of actions to maximize the cumulative reward function. Q-learning is an algorithm that does not require a model to learn action. Based on the actions of the agent to enter

a new state, the Q value is updated accordingly using the Bellman equation. Deep reinforcement learning [140] proposed recently will be a new area to build more autonomous systems with a high level of understanding.

III. Text-Based Cyberbullying Detection

Balakrishnan et al. [45] described various feature sets that can be useful in cyberbullying detection. These features include user personality (Big Five model [69]), Twitter features (Text/content features, user features, network features), emotion analysis, and sentiment-based features. After a comparative study of different combinations of the features, the authors have identified ten key features using the best-first search method. J48 classifier has shown promising results on the features set (Classification accuracy of 92.88% on four classes).

Lee et al. [20] integrated various filters to detect the abusive text by using unsupervised learning of abusive words. These filters include blacklist [141], n-gram [142], edit-distance, list filtering, and text features. The word2vec skip-gram module is applied in unsupervised learning, and cosine similarity is used as a similarity measure. The best setting produced f-scores on the news article, community comments, and Twitter are 0.869, 0.85, and 0.92, respectively, by trying different combinations of models. In some cases, words only do not carry any positive or negative meaning. But if these words are used in a specific context, they may express harmful meanings. Therefore, Ptaszynski et al. [143] used phrases consisting of morpheme pair in a dependency relation (in Japanese language) to define a harmful polarity score. Therefore, Ptaszynski et al [143] used phrases consisting of morpheme pair in a dependency relation (in the Japanese language) to define a harmful polarity score. However, the precision and recall of this method on abusive and non-abusive phrases are not very significant (70% precision on 50% recall).

Fortuna et al. [144] have conducted an interesting study on the generalization of various language classification models to different datasets. They have investigated Bidirectional Encoder Representations from Transformers (BERT), A Lite BERT (ALBERT), FastText, and support vector machine (SVM) with different settings. BERT comprises of multi-layered bidirectional transformer encoder that can learn general language representations. A sentence or bunch of sentences can be input sequence to the BERT. Nine publicly available datasets (Table II of the reference) from Twitter mainly covering different classes, including racism, sexism, hate, misogynous, aggression toxic, obscene, spam, etc., are used in this study. BERT and ALBERT performed better than other classification models in the intra-dataset classification. In some cases, FastText performed better than others. Although hate speech is not the topic of this paper, these findings can be beneficial in the cyberbullying datasets also. Paul and Saha [85] fine-tuned the BERT model for cyberbullying detection. BERT comes with high computational complexity due to millions of parameters. In this paper, the knowledge distillation method [145], a simpler version of BERT, is used to minimize the computational cost of the process. A fully connected layer is added for classification. The simplified BERT method is tested on three datasets from Twitter, Wikipedia, and FormSpring. F1-score of the proposed method is comparable with the original BERT method with 30 times less computational complexity.

Coi et al. [146] proposed a method to identify the cyberbullies by text mining their comments through the Losada ratio (Positive to negative comments ratio), cyberbullying index (Insulting words rate of the comments), and social network analysis (connection relationship between commenters). A dictionary of insulting words is used to calculate the cyberbullying index. Random forest, logistic regression, and SVM are tested on the features of the comments of 3200 users. Random forest outperformed other classifiers (F-score 0.8, precision 0.81, and recall 0.78). Song et al. [23] implemented the decision tree analysis methods of data mining to predict cyberbullying's risk factors. Causal factors of cyberbullying are used to construct the decision tree. Chi-square Automatic Interaction Detection (CHAID), Exhaustive CHAID, Quick Unbiased Efficient Statistical Tree (QUEST), and Classification and Regression Tree (C&RT) are the decision tree methods used in this study. All methods showed comparable classification accuracy with CR&T showing the best classification accuracy of 74.5%.

Al-Garadi et al. [43] used a more extensive feature set, including network information, activity information, user information, and tweet content, to classify tweets containing cyberbullying or not. Publicly available Twitter data of two months in 2015 within California, USA, is collected through Twitter API. Information gain, chi-square test, and Pearson correlation are used to select the significant features. The top ten features are used to classify cyberbullying or not using LibSVM, random forest, Naïve Bayes, and K-nearest neighbor classifier. The class imbalance was treated by SMOTE (oversampling the minority class). Random forest showed the best results (AUC was 0.943 and f-measure was 0.936).

Yokotani and Takano [147] studied the spread of cyberbullying trends via online social networks. They used a 2-dimensional convolutional neural network, bidirectional long-short term memory (LSTM), and an attention framework to identify the perpetrator-to-victim relationship. Relational patterns, gender/sexual orientation, and Pigg party usage by room owner and visitor are used. A similar classifier is used to identify the victim-to-perpetrator relationship. Data is collected from the users of the Pigg party. The gender/orientation classifier showed 99.69% accuracy. Classifiers for perpetrator-to-victim relationship and victim-to-perpetrator relationship achieved an accuracy of 88.47% and 85.32%, respectively.

Sanchez-Medina et al. [24] studied the causal relationship between dark triad and cyberbullying behaviors. They used “The Dirty Dozen” method [148] to measure the dark triad and questionnaire based on research in [149, 150] to detect cyberbullying and studied the causal relationship between dark triad and cyberbullying behaviors. Various ensemble classifiers, including random forest, bagging, boosting, and logistic regression, are used. Random forest performed best with AUC equals to 0.983. Agarwal et al. [28] identified four classes from different datasets, including Formspring, Twitter, and Wikipedia. Based on features extracted using word embedding, different classifiers are tried on the datasets and found that Bidirectional Long Short-Term Memory (BLSTM) with attention and feature level transfer learning outperformed other classifiers with an f1-score of more than 90%. Chatzakou et al. [33] distinguished cyberbullying from aggressive behaviors on a large dataset collected from Twitter users (1.2 million users, 2.1 million tweets). They have extracted a large set of features including user profile, network-based features, sentiment emotions, and content-based features. Out of these features, important features are selected for the classification. After removing spam tweets, a good classification is obtained between offensive (bullying + aggressive) and normal with AUC of 0.91 using random forest classifier.

Most of the cyberbullying detection is based on training a classifier of a feature set. Cheng et al. [151] introduced an approach of unsupervised cyberbullying detection. Multi-modal features (bag of words representation of text, social network analysis, and time of comments). The proposed learning framework estimates the likelihood of cyberbullying by using the Gaussian mixture model. The performance of the proposed framework (AUC is 0.74) was comparable with other supervised learning methods.

Yuvaraj et al. [46] proposed a framework integrating artificial neural network (ANN) and deep reinforcement learning (DRL) and achieved an accuracy of 98% on the Twitter data (30,384 tweets) with 90% training and 10% testing datasets. Feature sets include content features, user features, network features, and sentiment features. Zhao and Mao [34] developed an extension of a deep learning model called semantic-enhanced marginalized denoising auto-encoder (smSDA). It learns a robust discriminative representation of the text. The proposed method is applied to the Twitter dataset (7321 tweets) and MySpace dataset (800 instances) and obtained an accuracy of 84.9% and 89.7%, respectively. Raisi and Huang [35] proposed a weakly-supervised framework in which an ensemble of two deep learners co-training one and the other. One learner analyzes the content of the message, whereas the other learner considers the user's social structure. Data from Twitter, Instagram and Ask.fm (total 260,800 users and 2,863,801 question-answer pairs) are manually annotated to bullying or not bullying. Experiments showed that the proposed ensemble learner performed better than other classifiers.

The deep neural network model with Gradi-

ent Boosted Decision Trees [42] has shown better accuracy (F1 Score is 0.93) on a similar data set. Zhao et al. [37] proposed a new embedding LSHWE to tackle the problem of obfuscated bullying words and showed that this embedding also improves the computational efficiency of text representation learning on large-scale datasets. Deep learning detector, Bidirectional LSTM with attention, performed better than other LSTM based detectors using cosine similarity distance function in LSHWE (F1-score is 0.8629). Cheng et al. [83] proposed Hierarchical Attention Networks for Cyberbullying Detection (HANCD) containing different layers, word sequence encoder, word-level attention layer, comment sequence encoder, comment level attention layer, contextual layer, and social media attributes embedding layer. [83] proposed Hierarchical Attention Networks for Cyberbullying Detection (HANCD) containing different layers, word sequence encoder, word-level attention layer, comment sequence encoder, comment level attention layer, contextual layer, and social media attributes embedding layer. Finally, a weighted loss function to optimize cyberbullying detection and time interval prediction. The model is compared with other methods on a dataset of Instagram (2218 sessions) and achieved a good F1-score of 0.783. This model may be tested on bigger datasets to prove its efficacy.

Most of the research mentioned above focuses on detecting cyberbullying on datasets collected from one or two social media platforms. Bruwaene et al. [78] collected a text-based dataset from VISR tool of SafeToNet that monitors the social media activity of a child on various social media platforms. They received an F-Score of 0.885 to classify bullying and non-bullying using CNN. Talpur and O'Sullivan [39] classified the severity of cyberbullying as non-cyberbullying, low, medium, and high levels. From the dataset already categorized into types of harassment contents, they ranked the severity of cyberbullying. It is debatable that how the severity of cyberbullying is linked with the types of harassment content. The severity of cyberbullying may depend on various factors, and there is no consistent definition of such classes. After extracting features based on word embedding, sentiment, Lexicon, and semantic orientation, they tried various classifiers to classify cyberbullying severity. Random forest classifier performed the best among these classifiers with a classification accuracy of 91%. Aind et al. [86] used the Q-learning method (a type of reinforcement learning) on the dataset of comments tagged as offensive or non-offensive. Misspelled words are also corrected with the closest words according to their probability. Based on the penalty/reward policy, the Q-learning method learns the sentence as offensive or not offensive (F1-score is 0.86).

Table III summarizes most of the current research conducted in the English-based cyberbullying detections on various social media content data containing more than 5000 tweets, comments, posts/ instances. Deep learning architectures like LSTM and its variants, CNN, and BERT (including its variants) are prevalent in recent years.

Safa et al. [25] built a small Arabic textual corpus by crawling data from Twitter and labeled it with a three-hierarchical annotation scheme. Features are extracted using n-gram, contextual word embedding, and word embeddings. The performance of various machine learning algorithms (SVM, Naïve Bayes, LR, CNN, LSTM, and GRU) on this data is analyzed on multi-classes. Word/ Char-n-grams features with SVM classifier, AUC is 0.84 for hate/offensive and clean classes. For three classes (hate, offensive and clean), the same model produced AUC equals 0.78. These works on the Arabic language can contribute to further research in cyberbullying detection on social media in Arabic-speaking people. Haider et al. [76] collected tweets in the Arabic language from Twitter (35273 tweets) and manually annotated them as bullying and not bullying. F-measures for Naïve Bayes and SVM classifiers are 0.905 and 0.927, respectively. They also enhanced cyberbullying detection performance in the Arabic language using ensemble machine learning [79]. Mouheb et al. [77] detected cyberbullying in the Arabic tweets and YouTube comments (25000 comments and tweets) based on keywords and the Naïve Bayes classifier with an accuracy of 95%. Rachid et al. [36] applied different deep learning methods on a dataset consisting of 32000 comments deleted by a news channel as offensive or obscene. So, their work identifies inappropriate comments that may not be targeted to-

TABLE III
SUMMARY OF RESEARCH IN TEXT BASED CYBERBULLYING DETECTION IN ENGLISH LANGUAGE

Reference	Data Size	Features	Classifier	Performance
[45]	9484 tweets	Features (Personality, net- 10 (work, user	J48	Accuracy = 92.8% (Classes 4)
[146]	650000 posts and comments	Cyberbullying index + Losada ratio + sentiment analysis, social network analysis	Random Forest	F-Score = 0.80 (Classes 2)
[23]	103212 buzzes	Casual factors of cyberbullying	CRT	Accuracy = 74.5%
[43]	10,606 tweets	.Multiple types of Features	Random Forest	F-Score = 0.936 (Classes 2)
[32]	10,000 tweets	User, Text, Network features	Random Forest	Accuracy = 91% (Classes 3)
[28]	12,000 Posts FormSpring	-	CNN	F-Score = 0.93 (Classes 2)
[33]	2.1 Million tweets	User, Text, Network features	Random Forest	F-Score = 0.902 (Classes 3)
[46]	30,384 tweets	Content, user, network, sentiment features	ANN + DRL	Accuracy = 80.7% (Classes 2)
[34]	7,321 Tweets and 800 instances (MySpace)	-	SmSDA	Accuracy = 84%, 89% (Classes 2)
[86]	184,349 comments	Textual features	Reinforcement Learning	Accuracy = 89% (Classes 2)
[37]	9,600 tweets	LSHWE word embedding	BLSTM with attention network	F-Score = 0.863 (Classes 2)
[78]	14,899 Posts	LIWC features	CNN	F-score = 0.885 (Classes 2)
[38]	37,373 tweets	TF-IDF, Word2Vec	LR	F-Score = 0.928 (Classes 2)
[49]	2188 Instagram 7321 Tweets	Context2vec features	LSTM	F-Score = 0.85 (Classes 2)
[85]	16090 tweets (C1) 115854 Wikipedia corpus (C2) 12773 instances Formspring (C3)	-	BERT	F-Score = 0.94 (C1: 4 Classes) F-Score = 0.91 (C2: 2 Classes) F-Score = 0.92 (C3: 2 Classes)

wards anyone. But it can be considered as relevant research towards cyberbullying identification in the Arabic language. On the original dataset, they could not find a good combination to get a better F1-score but on the balanced dataset (number of offensive and non-offensive comments are same), they claimed a good F1-score of 0.84 using RNN/CNN and Fasttext embedding.

Table IV summarizes the research in cyberbullying detection involving Arabic-speaking individuals. Only a few papers have shown considerable success in Arabic language-based cyberbullying detection. A lot of work is needed to create a good corpus of Arabic language related to cyberbullying, modify existing classifiers (possibility of transfer learning in the existing Deep learning architecture for NLP), and perform performance analysis on more extensive datasets. Many valuable works already exist in Chinese [152, 153] and other languages [154, 155].

IV. Multi-modal Cyberbullying Detection

Due to diversity in the types of communication platforms and allowable content types, people are engaging with different kinds of content to maximize their social impact on the community around them. Multi-modal content makes cyberbullying detection and monitoring more challenging for the researchers.

Manuel et al. [40] detected cyberbullying in a publicly available dataset collected from Vine social network, which comprises social media sessions of videos posted, and all the likes and comments associated with this video. The feature set consisted of profile features, media session features, video features, Latent Dirichlet Allocation (LDA) features, bag-of-words similarity, and time aspects of the comments. Random Forest (RF), AdaBoost (AB), Extra Tree (ET), linear Support Vector Classification (SVC), and Logistic Regression (LR) are considered for the binary classification of cyberbullying or no bullying events. Early detection problem is considered as detecting the cyberbullying in the first few comments. A Dual model [156] classifier based on ET produced the best results, and the best F-latency metric was 0.5217. Dual models use different models with different feature sets for each class.

Kumari et al. [157] classified the social media posts containing images into non-aggressive, medium aggressive, and high-aggressive classes using different classifiers. Text features from comments are extracted by an optimized convolutional neural network (CNN), and VGG-18 [127] model is used to extract the features from images. A binary particle swarm (BPSO) algorithm is used to extract the most relevant features. Random forest classifier model with optimized features achieved F1-Score of 0.74. Li et al. [31] collected 3000 images from Instagram and extracted features from the images, captions, and comments. Features are extracted from the comments using Bag of words, Offensiveness level, and Word2vec. Comments-based features produced the best results using SVM with radial basis kernel. Latent Dirichlet allocation (LDA) is used to extract main topics from the image captions' text, Scale-Invariant Feature Transform (SIFT), GIST, color histogram features from the images. But features based on caption and images have not performed well. Kumar et al. [82] pre-

TABLE IV
Summary of Research in Arabic Text Based Cyberbullying Detection

Reference	Data Size	Features	Classifier	Performance
[76]	35273 Tweets	String2Vec, Tweet2Sentiment strength	SVM	F-score = 0.905 (Classes 2)
[79]	35273 Tweets	Word2vec	Meta Classifiers with Bagging	F-score = 0.926 (Classes 2)
[77]	25000 comments and tweets	--	NB classifier	F-Score = 0.927
[36]	32000 comments	Fasttext Embedding	RNN/CNN	F1-Score = 0.84

TABLE V
SUMMARY OF RESEARCH OF MULTI-MODAL CYBERBULLYING DETECTION

Reference	Data Size	Features	Classifier	Performance
[157]	3600 images with comments	Text features/ Image features	Random Forest	F-Score = 0.74
[31]	3000 images with comments	Text features/ Image features	SVM, Deep learning	Accuracy = 95%
[82]	10,000 comments, images	-	CNN	Accuracy = 97%
[41]	733 sessions with 15 posts or more	Textual, Visual, Audio features	LR	AUROC = 0.834 (Classes 2)
[160]	2100 images with comments	Image embedding, Text em-(bedding (TD-IDF	CNN	F-Score = 0.68 (Classes 2)
[49]	2218 sessions (Instagram)	LIWC, Word embedding	HANCD	F-Score = 0.778 (Classes 2)
[50]	969 video sessions with comments	-	Recurrent-CNN ResidualBiLSTM	F-Score = 0.75

sented a combined deep neural network model for textual, visual, and infographics (embedded text in graphics) modalities of social media. CapsNet deep neural network [158] with dynamic routing predicts the cyberbullying from the textual content, and ConvNet [159] predicts the cyberbullying from visual contents. A Late-fusion Perceptron decision layer fuses both decisions. Based on 10,000 comments, the proposed model has achieved AUC equal to 0.98 and an accuracy of 97%.

Multimodal cyberbullying detection, including text (comments, tweets, posts, etc.), images, and videos, is summarized in Table V. More extensive datasets of multimodal content are required to formulate better techniques and performance analysis. Different deep learning architectures are used to tackle the multimodal content's diversity, and few works have shown promising results.

V. AUTOMATED CYBERBULLYING MONITORING AND INTERVENTION SYSTEM

In Japan, the parent-teacher association (PTA) monitors the website activities through net-petrol. Any offensive text is requested to remove from the net-petrol member [143]. But this activity is done manually and hence requires a lot of time and effort. Tommy et al. In Japan, the parent-teacher association (PTA) monitors the website activities through net-petrol. Any offensive text is requested to remove from the net-petrol member [143]. But this activity is done manually and hence requires a lot of time and effort. Tommy et al. [12] described triadic reciprocal relationships between perpetrators, victims, and bystanders. This framework consolidated personal factors, environmental events, and behavioral patterns that influence each other in a triadic reciprocal manner. Perpetrators bully the victims through any social media platform, and victims react to this bullying behavior in many ways. A bystander who is watching this bullying episode through social media platform may confront the perpetrator or remain silent against him. Similarly, he/she may comfort or stand with the victim or remain silent. Moreover, it was found that a higher rate of cyberbullying victims among the peer networks increases the risk of victimization of cyberbullying [147]. Similarly, more perpetrators among

the person's peer network may increase the risk of victimization. Potha et al. [81] recognized bullying temporal patterns in the predator's questions using time series modeling methodology. Jacobs et al. [161] tried to identify automatically different participants in a cyberbullying event from textual cyberbullying traces. [161] tried to identify automatically various participants in a cyberbullying event from textual cyberbullying traces. Although the F1-score of their best model is not very high (56.7%), there is a lot of scope in this area.

Developing a complete automated cyberbullying monitoring and intervention system is the need of the day. More and more people worldwide are joining social media actively, and massive data makes it impossible to monitor social media behaviors manually. Natural language processing algorithms, statistical analysis of social media activity, and deep learning may be the target areas to build such a system. We can train the system on a broader international population, and transfer learning can fine-tune the system for the local population. Interactions among perpetrator, victim, and bystander are very complicated, and a high level of artificial intelligence modeling is required to understand such interactions.

The automated cyberbullying system will focus on the following area:

- Interaction monitoring.
- Mental health monitoring.
- Intervention policies.

A. Interaction Monitoring

Simple detection of cyberbullying is not enough. We should also consider the capacity of the victim to absorb and respond to the bullying. Hence, it is crucial to monitor the interaction among the perpetrator, victim, and bystander to predict the damage done by the bullying.

1) Perpetrator and victim interaction monitoring

Interaction between the perpetrator and the victim will be monitored to detect the presence and absence of cyberbullying. In case of cyberbullying, the system will also assess the severity of the bullying towards the victim. The victim's response to the perpetrator will also be monitored to determine the effect of cyberbullying on the victim. The victim's reaction may be classified as appropriate, not enough, depressing, aggressive, etc. Furthermore, the victim's interaction with his/her social circle, including family, friends, peers, etc., will be monitored to quantify the psychological damage done to the victim's personality.

2) Bystander and victim interaction monitoring

If there is any interaction between bystander (witness of bullying on social media) and victim, either positive or negative should be monitored to assess the severity of cyberbullying on the victim. Positive interaction includes consoling the victim. Negative interaction is joining the perpetrator and bullying the victim, mockery, or spreading the event to others on social media. Blackmailing the victim on this event may also be part of the negative interaction with the victim.

3) Bystander and perpetrator interaction monitoring

Bystanders may be just witnessing the event without any interaction. In some cases, bystanders can interact with the perpetrator in a positive or negative sense. Positive interaction may include stopping the perpetrator and counseling the perpetrator. Bystanders can also become part of the cyberbullying process and encourage the perpetrator. Bystanders can also cyberbully the perpetrator in return for changing his/her behavior towards the victim.

B. Mental Health Monitoring

The victim's mental health also plays a crucial role in an increase in the severity of cyberbullying. Furthermore, monitoring the perpetrator's mental health can help him/her mitigate his/her aggressive behavior towards others. A proper intervention may also improve the overall community environment. Different machine learning algorithms are successfully used to monitor a person's mental health [162-165].

1) Mental state monitoring of the victim.

The mental state of victims should be monitored

to predict depression severity, suicidal tendency, or aggression towards his/her social circle. Victim interaction with family, friends, peers about bullying event or general interaction may also be assessed to predict the severity of the depression if exits. Any aggressive behaviors may result from the cyberbullying to the victim, which he/she shows to his connections.

2) Mental state monitoring of the perpetrator

A person becoming used to perform cyberbullying on others and enjoying the event may cause harm to his/her mental health. Early detection of such behaviors may help the perpetrator to recover from this state.

3) Mental state monitoring of the bystander

Witnessing a cyberbullying act also left unpleasant marks on the psychological state of the bystander. A positive, negative, or no response by such person may affect his/her mental health. So, monitoring bystander mental health is also essential for a healthy society.

C. Intervention policies

Once an event of cyberbullying is detected and its severity is assessed, an intervention is required to ensure the victim and other parties' safety and mental wellness. A decision support system may decide the type of intervention. Different types of interventions are possible. An automatic decision-making system can inform the family member of the victim so that the family member may take corrective actions. Depending on the severity of the cyberbullying, the system can inform law enforcement agencies to intervene in this matter. If the victim, perpetrator, or bystander is already taking some medical treatment, the system may notify the medical staff to intervene.

VI. Conclusions

This study reviewed various aspects of machine learning-based cyberbullying detection. Most of the work is focused on classifying bullying or non-bullying events based on textual, user activity, and network information. Few research works are reported on multi-modal content extracted from social media networks. Deep learning architectures designed for natural language processing are also getting popular to understand the bullying intentions towards the victim. A strong need exists for a comprehensive dataset based on actual events involving victim's perception of the cyberbullying efforts. Unintended biases may exist in the data annotations by a third person not involved in these events. Moreover, the victim's mental state also plays an essential role in the severity of the cyberbullying that victims feel. Therefore, it is desirable to make the cyberbullying detection mechanism more intelligent in understanding the context of the cyberbullying attack and the mental status of the victim and his/her response to such an attack.

VII. Future Directions and Challenges

With the advancement in deep learning algorithms and their high performance in the natural language processing tasks, many new directions are opened for research to monitor social media for harassment, cyberbullying, hate crimes, etc.

Handling of a dynamic corpus: It is observed in this literature review that ways of social interaction and use of language are constantly updating. Many new slang words, intentionally misspelled or words with missing letters, use of new emojis/emoticons are continuously evolving, posing more significant challenges for monitoring systems. Therefore, language corpus should be dynamic, constantly updating with time, and retraining/incremental training of the machine learning algorithms. Transfer learning is an exciting aspect of deep learning algorithms. Pre-trained models on a language can be used for many tasks with further fine-tuning and training. On social media, many users communicate in mixed languages [166]. For example, in Asian countries like India and Pakistan, people mix English with the Urdu language. Hence multilingual identification of cyberbullying is a challenging task. In the last couple of years, few researchers have published papers in this direction [167-169].

Unintended Biasness of the detection system: Machine learning methods may contain an unintended bias towards demographic groups due to biases in the datasets on which these methods

are trained [170]. A cyberbullying detection system is considered biased if it performs better for some demographic groups than others. Reasons for biasedness can be at the feature extraction level (bias in embedding algorithms) or classification level (bias in the human annotators or the machine learning algorithm). Gencoglu [170] suggested that unintended bias in the cyberbullying detection models can be mitigated if the model is trained with fairness constraints. Gencoglu [170] indicated that unintentional bias in the cyberbullying detection models could be mitigated if the model is trained with fairness constraints. Few works are found to solve this problem, and more research is required to produce the cyberbullying detectors using machine learning algorithms that are unbiased and transparent. Several works proposed methods to mitigate the unintended bias in the word embeddings [171-173]. Similarly, measuring and mitigating the classification algorithms' bias is also important for a fairer performance of the classifier [173]. Data is collected and annotated by subjective human annotations. These annotations may have an unintended bias as human annotators who may not feel the victim's pain [174, 175]. Therefore, it is essential to collect real reported data from the victims and properly annotate them. Generalization of machine learning models to various datasets, languages, and countries should also be studied further.

Handling of obfuscated language: Social media users use obfuscated words to bypass the automated screening software for toxic language [176]. Shorter words, conveying the semantic of a sentence, are also used to speed up the typing speed and avoid content moderation [177]. Such words pose a challenge to word embedding methods and understanding the semantic of the sentence for automated systems.

Detection of coordinated bullying towards a specific person: In many cyberbullying cases, more than one user targets a victim for bullying. Such collaborative cyberbullying needs more sophisticated social network analysis and understanding of messages with a common goal towards the victim. Highly coordinated groups create false impressions in the community by flooding objectionable content, making content popular, bullying, and harassing people [178]. Detection of highly coordinated groups is necessary to mitigate the bullying inside a community.

Behavior modifications through virtual learning communities: Helping the perpetrators involved in cyberbullying is also the community's responsibility and the state. Behavior modification of the perpetrators through counseling, using social media, engaging them in constructive ways can be a few of the things done. Recently Nikiforos et al. [179, 180] investigated the possibility of engaging persons in the virtual learning communities who show aggressive behaviors in their physical learning communities. Teacher's early intervention in bullying events is done through automated detection of bullying. They have shown improvement in the behaviors with such a framework. The big question arises, whether artificial intelligence can teach manners or instill moral values in a person?

Social media's ethical and legal issues: Monitoring social media can prevent cyberbullying events, harassment, victimization, etc. But there are specific ethical and legal issues involved in safeguarding the students' privacy and free speech on social media [181]. Absolute privacy of the people on social media may poses threats to the community. Hence, every country decides its ethical and legal guidelines for monitoring social media and respecting their citizens' privacy. Monitoring social media with the perception of privacy can create a healthy environment. R. Clarke [182] coined the term dataveillance in 1988 to systematically use personal data to monitor communication among people. A good description can be found in [183].

Policymaking to mitigate cyberbullying: Big data analytics can give the policymakers the luxury to study various aspects of cyberbullying through automated data analytic tools. Cyberbullying events in different age groups, genders, education level based, workplaces, etc., can be analyzed to create efficient pre-emptive policies that

can reduce cyberbullying. Many societies consist of many racial groups, observing different religions and having different nationalities. In such diverse communities, monitoring cyberbullying becomes essential to foresee any mishaps in the future. Victims of cyberbullying may get depressed and tend toward suicide or may become violent and commit crimes. Therefore, such statistical analysis of the data is a necessity for effective policymaking.

References

[1] T. Vaillancourt, R. Faris, and F. Mishna, "Cyberbullying in children and youth: implications for health and clinical practice," *Can. J. Psychiatry*, vol. 62, no. 6, pp. 368-373, 2017.

[2] C. Burger, D. Strohmeier, N. Spröber, S. Bauman, and K. Rigby, "How teachers respond to school bullying: An examination of self-reported intervention strategy use, moderator effects, and concurrent use of multiple strategies," *Teach. Teach. Educ.*, vol. 51, pp. 191-202, Oct. 2015, doi: 10.1016/j.tate.2015.07.004.

[3] L. Arseneault, L. Bowes, and S. Shakoor, "Bullying victimization in youths and mental health problems:'much ado about nothing'?," *Psychol. Med.*, vol. 40, no. 5, p. 717, 2010.

[4] O. Aluede, F. Adeleke, D. Omoike, and J. Afen-Akpaida, "A review of the extent, nature, characteristics and effects of bullying behaviour in schools," *J. Instr. Psychol.*, vol. 35, no. 2, p. 151, 2008.

[5] I. Rivers, V. P. Poteat, N. Noret, and N. Ashurst, "Observing bullying at school: The mental health implications of witness status," *Sch. Psychol. Q.*, vol. 24, no. 4, p. 211, 2009.

[6] B. Henson, "Bullying beyond the schoolyard: Preventing and responding to cyberbullying," ed: Springer, 2012.

[7] C. Langos, "Cyberbullying: The challenge to define," *Cyberpsychol. Behav. Soc. Netw.*, vol. 15, no. 6, pp. 285-289, 2012.

[8] D. Zhang, E. S. Huebner, and L. Tian, "Longitudinal associations among neuroticism, depression, and cyberbullying in early adolescents," *Comput. Hum. Behav.*, vol. 112, p. 106475, 2020.

[9] M. C. Martínez-Monteagudo, B. Delgado, Á. Díaz-Herrero, and J. M. García-Fernández, "Relationship between suicidal thinking, anxiety, depression and stress in university students who are victims of cyberbullying," *Psychiatry Res.*, vol. 286, p. 112856, 2020.

[10] C. E. Torres, S. J. D'Alessio, and L. Stolzenberg, "The effect of social, verbal, physical, and cyberbullying victimization on academic performance," *Vict. Offender*, vol. 15, no. 1, pp. 1-21, 2020.

[11] A. Oksanen, R. Oksa, N. Savela, M. Kaakinen, and N. Ellonen, "Cyberbullying victimization at work: Social media identity bubble approach," *Comput. Hum. Behav.*, vol. 109, p. 106363, 2020, doi: 10.1016/j.chb.2020.106363.

[12] T. K. Chan, C. M. Cheung, and Z. W. Lee, "Cyberbullying on social networking sites: A literature review and future research directions," *Inf. Manag.*, vol. 58, no. 3, p. 103411, Mar. 2021, doi: 10.1016/j.im.2020.103411.

[13] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv. (CSUR)*, vol. 51, no. 4, pp. 1-30, 2018.

[14] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PLOS One*, vol. 14, no. 8, p. e0221152, 2019.

[15] M. Tomaiuolo, G. Lombardo, M. Mordonini, S. Cagnoni, and A. Poggi, "A survey on troll detection," *Future Internet*, vol. 12, no. 2, p. 31, 2020.

[16] T. Jiang, J. P. Li, A. U. Haq, A. Saboor, and A. Ali, "A Novel Stacking Approach for Accurate Detection of Fake News," *IEEE Access*, vol. 9, pp. 22626-22639, 2021.

[17] A. Kumar, V. Singh, T. Ali, S. Pal, and J. Singh, "Empirical evaluation of shallow and deep classifiers for rumor detection," in *Proc. ICACM 2019*, in Advances in Computing and Intelligent Systems, in Algorithms for Intelligent Systems, 2020, pp. 239-252.

[18] S. R. Sahoo and B. Gupta, "Real-time detection of fake account in twitter using machine-learning approach," in *Proc. CICT 2019*, in Advances in computational intelligence and communication technology, in Advances in Intelligent Systems and Computing, vol. 1086, 2021, pp. 149-159.

[19] E. W. Pamungkas, V. Basile, and V. Patti, "Misogyny detection in twitter: a multilingual and cross-domain study," *Inf. Process. Manag.*, vol. 57, no. 6, p. 102360, 2020.

[20] H.-S. Lee, H.-R. Lee, J.-U. Park, and Y.-S. Han, "An abusive text detection system based on enhanced abusive and non-abusive word lists," *Decis. Support Syst.*, vol. 113, pp. 22-31, 2018.

[21] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop Nat. Lang. Process. Soc. Media*, Spain, Apr. 2017, pp. 1-10, doi: 10.18653/v1/W17-11.

[22] N. Willard, *Cyberbullying and Cyberthreats: responding to the challenge of online social aggression, threats, and distress*. Champaign, IL, USA: Research Press, 2007.

[23] T.-M. Song and J. Song, "Prediction of risk factors of cyberbullying-related words in Korea: Application of data mining using social big data," *Telemat. Inform.*, vol. 58, p. 101524, 2021.

[24] A. J. Sánchez-Medina, I. Galván-Sánchez, and M. Fernández-Monroy, "Applying artificial intelligence to explore sexual cyberbullying behaviour," *Heliyon*, vol. 6, no. 1, p. e03218, 2020.

[25] S. Alsafari, S. Sadaoui, and M. Mouhoub, "Hate and offensive speech detection on arabic social media," *Online Soc. Netw. Media*, vol. 19, p. 100096, 2020.

[26] T. Bosse and S. Stam, "A normative agent system to prevent cyberbullying," in *2011 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, 2011, pp. 425-430, doi: 10.1109/WI-IAT.2011.24.

[27] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. Int. AAAI Conf. Web Soc. Media*, Spain, 2011, vol. 5, no. 1.

[28] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Eur. Conf. Inf. Retr. ECIR 2018*, in Advances in Information Retrieval, in Lecture Notes in Computer Science, vol. 10772, 2018, pp. 141-153.

[29] Noviantho, S. M. Isa and L. Ashianti, "Cyberbullying classification using text mining," in *2017 1st Int. Conf. Inform. Comput. Sci. (ICICoS)*, Indonesia, 2017, pp. 241-246, doi: 10.1109/ICICOS.2017.8276369.

[30] J. Zhang, T. Otomo, L. Li, and S. Nakajima, "Cyberbullying Detection on Twitter using Multiple Textual Features," in *2019 IEEE 10th Int. Conf. Aware. Sci. Technol. (iCAST)*, Japan, 2019, pp. 1-6, doi: 10.1109/ICAwST.2019.8923186.

[31] H. Zhong et al., "Content-Driven Detection of Cyberbullying on the Instagram Social Network," in *Proc. 25th Int. Jt. Conf. Artif. Intell. IJCAI*, July 2016, pp. 3952-3958.

[32] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter," in *Proc. 2017 ACM Web Sci. Conf.*, USA, June 2017, pp. 13-22, doi: 10.1145/3091478.3091487.

[33] D. Chatzakou et al., "Detecting cyberbullying and cyberaggression in social media," *ACM Trans. Web (TWEB)*, vol. 13, no. 3, pp. 1-51, 2019, doi: 10.1145/3343484.

[34] R. Zhao and K. Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 328-339, 1 July-Sept. 2017, doi: 10.1109/TAFFC.2016.2531682.

[35] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection using co-trained ensembles of embedding models," in *2018 IEEE/ACM Int. Conf. Adv Soci. Netw. Anal. Min. (ASONAM)*, Spain, 2018, pp. 479-486, doi: 10.1109/ASONAM.2018.8508240.

[36] B. A. Rachid, H. Azza, and H. H. B. Ghezala, "Classification of Cyberbullying Text in Arabic," in *2020 Int. Jt. Conf. Neural Netw. (IJCNN)*, UK, 2020, pp. 1-7, doi: 10.1109/IJCNN48605.2020.9206643.

[37] Z. Zhao, M. Gao, F. Luo, Y. Zhang, and Q. Xiong, "LSHWE: Improving Similarity-Based Word Embedding with Locality Sensitive Hashing for Cyberbullying Detection," in *2020 Int. Jt. Conf. Neural Netw. (IJCNN)*, UK, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207640.

[38] A. Muneer and S. M. Fati, " A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter," *Future Internet*, vol. 12, no. 11, p. 187, Oct. 29, 2020, doi: 10.3390/fi12110187.

[39] B. A. Talpur and D. O'Sullivan, "Cyberbullying severity detection: A machine learning approach," *PLOS One*, vol. 15, no. 10, p. e0240924, 2020, doi: 10.1371/journal.pone.0240924.

[40] M. F. López-Vizcaíno, F. J. Nóvoa, V. Carneiro, and F. Cacheda, "Early detection of cyberbullying on social media networks," *Future Gener. Comput. Syst.*, vol. 118, pp. 219-229, May 2021, doi: 10.1016/j.future.2021.01.006.

[41] D. Soni and V. K. Singh, "See no evil, hear no evil: Audio-visual-textual cyberbullying detection," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, pp. 1-26, Nov. 2018, doi: 10.1145/3274433.

[42] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion*, Australia, Apr. 2017, pp. 759-760, doi: 10.1145/3041021.3054223.

[43] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Comput. Hum. Behav.*, vol. 63, pp. 433-443, Oct. 2016, doi: 10.1016/j.chb.2016.05.051.

[44] R. I. Rafiq, H. Hosseinmardi, S. A. Mattson, R. Han, Q. Lv, and S. Mishra, "Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network," *Soc. Netw. Anal. Min.*, vol. 6, no. 1, pp. 1-16, 2016.

[45] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using Twitter users' psychological features and machine learning," *Comput. Secur.*, vol. 90, p. 101710, Mar. 2020, doi: 10.1016/j.cose.2019.101710.

[46] N. Yuvaraj et al., "Nature-Inspired-Based Approach for Automated Cyberbullying Classification on Multimedia Social Networking," *Math. Probl. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/6644652.

[47] M. Dadvar, F. d. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proc. Twelfth Dutch-Belgian Inf. Retr. Workshop (DIR 2012)*, Belgium, 2012.

[48] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "PI-Bully: Personalized Cyberbullying Detection with Peer Influence," in *Proc. Twenty-Eighth Int. Jt. Conf. Artif. Intell.*, 2019, pp. 5829-5835, doi: 10.24963/ijcai.2019/808.

[49] N. Rezvani, A. beheshti, and A. Tabebordbar, "Linking textual and contextual features for intelligent cyberbullying detection in social media," in *Proc. 18th Int. Conf. Adv. Mob. Comput. Multimed.*, 2020, pp. 3-10.

[50] S. Paul, S. Saha, and M. Hasanuzzaman, "Identification of cyberbullying: A deep learning based multimodal approach," *Multimed. Tools Appl. (2020)*, pp. 1-20, Sept. 10, 2020, doi: 10.1007/s11042-020-09631-w.

[51] E. Stamatatos, "On the robustness of authorship attribution based on character n-gram features," *J. Law Policy*, vol. 21, no. 2, pp. 421-439, 2013.

[52] E. Raff et al., "An investigation of byte n-gram features for malware classification," *J. Comput. Virol. Hacking Tech.*, vol. 14, no. 1, pp. 1-20, 2018.

[53] W. Khreich, B. Khosravifar, A. Hamou-Lhadj, and C. Talhi, "An anomaly detection system based on variable N-gram features and one-class SVM," *Inf. Softw. Technol.*, vol. 91, pp. 186-197, Nov. 2017, doi: 10.1016/j.infsof.2017.07.009.

[54] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv:1310.4546*, 2013, [Online]Available: https://arxiv.org/abs/1310.4546

[55] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proc. 2013 Conf. N. Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, USA, June 2013, pp. 746-751.

[56] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, no. 1-4, pp. 43-52, 2010, doi: 10.1007/s13042-010-0001-0.

[57] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Qatar, 2014, pp. 1532-1543, doi: 10.3115/v1/D14-1162.

[58] L. Havrlant and V. Kreinovich, "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)," *Int. J. Gen. Syst.*, vol. 46, no. 1, pp. 27-36, May 14, 2017, doi: 10.1080/03081079.2017.1291635.

[59] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135-146, 2017, doi: 10.1162/tacl_a_00051.

[60] V. Nahar, S. Unankard, X. Li, and C. Pang, "Sentiment analysis for effective detection of cyber bullying," in *Asia-Pac. Web Conf.*, in Web Technologies and Applications, in Lecture Notes in Computer Science, vol. 7235, 2012, pp. 767-774.

[61] J. N. Navarro and J. L. Jasinski, "Going cyber: Using routine activities theory to predict cyberbullying experiences," *Sociol. Spectr.*, vol. 32, no. 1, pp. 81-94, 2012, doi: 10.1080/02732173.2012.628560.

[62] L. Corcoran, C. M. Guckin, and G. Prentice, "Cyberbullying or cyber aggression?: A review of existing definitions of cyber-based peer-to-peer aggression," *Soc.*, vol. 5, no. 2, pp. 245-255, 2015, doi: 10.3390/soc5020245.

[63] P. Gradinger, D. Strohmeier, and C. Spiel, "Definition and measurement of cyberbullying," *Cyberpsychology: J. Psychosoc. Res. Cyberspace*, vol. 4, no. 2, 2010.

[64] L. A. Zezulka and K. Seigfried-Spellar, "Differentiating cyberbullies and internet trolls by personality characteristics and self-esteem," *J. Digit. Forensics, Secur. Law*, vol. 11, no. 3, pp. 7-26, Sept. 2016, doi: 10.15394/jdfsl.2016.1415.

[65] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining for gender prediction," in *4th Int. Conf. Adv. Inf. Syst. ADVIS 2006*, in Advances in information Systems, in Lecture Notes in Computer Science, vol. 4243, 2006, pp. 274-283.

[66] M. Sap et al., "Developing age and gender predictive lexica over social media," in *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Qatar, Oct. 2014, pp. 1146-1151, doi: 10.3115/v1/D14-1121.

[67] C. P. Barlett and K. Chamberlin, "Examining cyberbullying across the lifespan," *Comput. Hum. Behav.*, vol. 71, pp. 444-449, June 2017, doi: 10.1016/j.chb.2017.02.009.

[68] H. A. Schwartz et al., "Personality, gender, and age in the language of social media: The open-vocabulary

approach," *PLOS one*, vol. 8, no. 9, p. e73791, 2013, doi: 10.1371/journal.pone.0073791.

[69] O. P. John and S. Srivastava, "The Big-Five trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of Personality: Theory and Research*, L. A. Pervin and O. P. John, Ed, 2nd ed, New York, USA: Guilford, 1999, ch. 4, pp. 102-138.

[70] E. Mitsopoulou and T. Giovazolias, "Personality traits, empathy and bullying behavior: A meta-analytic approach," *Aggress. Violent Behave.*, vol. 21, pp. 61-72, Mar.-Apr. 2015, doi: 10.1016/j.avb.2015.01.007.

[71] D. Azucar, D. Marengo, and M. Settanni, "Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis," *Pers. Individ. Differ.*, vol. 124, pp. 150-159, Apr. 2018, doi: 10.1016/j.paid.2017.12.018.

[72] H. Wei et al., "Beyond the words: Predicting user personality from heterogeneous information," in *Proc. Tenth ACM Int. Conf. Web Search Data Min.*, NY, USA, Feb. 2017, pp. 305-314, doi: 10.1145/3018661.3018717.

[73] N. K. Singh, D. S. Tomar, and A. K. Sangaiah, "Sentiment analysis: a review and comparative analysis over social media," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 1, pp. 97-117, 2020, doi: 10.1007/s12652-018-0862-8.

[74] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," *Artif. Intell. Rev.*, pp. 1-41, Feb. 08, 2021, doi: 10.1007/s10462-021-09958-2.

[75] E. W. Pamungkas, V. Basile, and V. Patti, "A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection," *Inf. Process. Manag.*, vol. 58, no. 4, p. 102544, July 2021, doi: 10.1016/j.ipm.2021.102544.

[76] B. Haidar, M. Chamoun, and A. Serhrouchni, "Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content," in *2017 1st Cyber Secur. Netw. Conf. (CSNet)*, Brazil, Oct. 18-20, 2017, pp. 1-8, doi: 10.1109/CSNET.2017.8242005.

[77] D. Mouheb, R. Albarghash, M. F. Mowakeh, Z. Al Aghbari, and I. Kamel, "Detection of Arabic Cyberbullying on Social Networks using Machine Learning," in *2019 IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA)*, UAE, Nov. 3-7, 2019, pp. 1-5, doi: 10.1109/AICCSA47632.2019.9035276.

[78] D. Van Bruwaene, Q. Huang, and D. Inkpen, "A multi-platform dataset for detecting cyberbullying in social media," *Lang. Resour. Eval.*, vol. 54, no. 4, pp. 851-874, Apr. 06, 2020, doi: 10.1007/s10579-020-09488-3.

[79] B. Haidar, M. Chamoun, and A. Serhrouchni, "Arabic Cyberbullying Detection: Enhancing Performance by Using Ensemble Machine Learning," in *2019 Int. Conf. Internet of Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Soc. Comput. (CPSCom) IEEE Smart Data (SmartData)*, USA, Oct. 21, 2019, pp. 323-327, doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00074.

[80] X. Zhang et al., "Cyberbullying detection with a pronunciation based convolutional neural network," in *2016 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, USA, Dec. 18-20, 2016, pp. 740-745, doi: 10.1109/ICMLA.2016.0132.

[81] N. Potha, M. Maragoudakis, and D. Lyras, "A biology-inspired, data mining framework for extracting patterns in sexual cyberbullying data," *Knowl. Based Syst.*, vol. 96, pp. 134-155, Mar. 15, 2016, doi: 10.1016/j.knosys.2015.12.021.

[82] A. Kumar and N. Sachdeva, "Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network," *Multimed. Syst.*, pp. 1-10, 2021, doi: 10.1007/s00530-020-00747-5.

[83] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the instagram social network," in *Proc. 2019 SIAM Int. Conference Data Min.*, 2019, pp. 235-243, doi: 10.1137/1.9781611975673.27.

[84] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," Multimed. Syst., pp. 1-14, Oct. 13, 2020, doi: 10.1007/s00530-020-00701-5.

[85] S. Paul and S. Saha, "CyberBERT: BERT for cyberbullying identification," *Multimed. Syst.*, pp. 1-8, Nov. 11, 2020, doi: 10.1007/s00530-020-00710-4.

[86] A. T. Aind, A. Ramnaney, and D. Sethia, "Q-Bully: A Reinforcement Learning based Cyberbullying Detection Framework," in *2020 Int. Conf. Emerg. Technol. (INCET)*, India, June 5-7, 2020, pp. 1-6, doi: 10.1109/INCET49848.2020.9154092.

[87] D. A. Freedman, *Statistical models: theory and practice*. New York, NY, USA: Cambridge Univ. Press, 2009.

[88] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. New York, NY, USA: Wiley, 2000.

[89] E. Fix and J. L. Hodges, "Discriminatory analysis: nonparametric discrimination, consistency properties," *Int. Stat. Rev.*, vol. 57, no. 3, pp. 238-247, Dec. 1989, doi: 10.2307/1403797.

[90] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. New York, NY, USA: Springer-Verlag, 2009.

[91] D. Steinberg and P. Colla, "CART: classification and regression trees," in *The top ten algorithms in data mining*, X. Wu and V. Kumar, Ed, Boca Raton, FL, USA: CRC Press, 2009, ch. 10, p. 179.

[92] J. R. Quinlan, *C4.5: PROGRAMS FOR MACHINE LEARNING*, San Rafael, CA, USA: Morgan Kaufmann, 2014.

[93] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *J. R. Stat. Soc. Ser. C-Appl. Stat.*, vol. 29, no. 2, pp. 119-127, 1980, doi: 10.2307/2986296.

[94] D. Biggs, B. De Ville, and E. Suen, "A method of choosing multiway partitions for classification and decision trees," *J. Appl. Stat.*, vol. 18, no. 1, pp. 49-62, 1991, doi: 10.1080/02664769100000005.

[95] W.-Y. Loh and Y.-S. Shih, "Split selection methods for classification trees," *Stat. Sin.*, vol. 7, no. 4, pp. 815-840, Oct. 1997.

[96] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273-297, 1995, doi: 10.1007/BF00994018.

[97] Y. Ma and G. Guo, *Support vector machines applications*, New York, NY, USA: Springer, 2014.

[98] B. Kumar, O. Vyas, and R. Vyas, "A comprehensive review on the variants of support vector machines," *Mod. Phys. Lett. B*, vol. 33, no. 25, p. 1950303, 2019, doi: 10.1142/S0217984919503032.

[99] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front. Comput. Sci.*, vol. 14, no. 2, pp. 241-258, 2020, doi: 10.1007/s11704-019-8208-z.

[100] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123-140, 1996, doi: 10.1007/BF00058655.

[101] R. E. Schapire and Y. Freund, "Boosting: Foundations and algorithms," *Kybernetes*, vol. 42, no. 1, pp. 164-166, 2013, doi: 10.1108/03684921311295547.

[102] R. E. Schapire, "Explaining adaboost," in *Empirical inference*, B. Scholkopf, Z. Luo and V. Vovk, Ed. Berlin, Germany: Springer-Verlag, 2013, ch. 5, pp. 37-52.

[103] Y. Freund, "An adaptive version of the boost by majority algorithm," *Mach. Learn.*, vol. 43, no. 3, pp. 293-318, 2001, doi: 10.1023/A:1010852229904.

[104] A. Torralba, K. P. Murphy, and W. T. Freeman, "Shared features for multiclass object detection," in *Toward Category-Level Object Recognition*, J. Ponce, M. Hebert, C. Schmid and A. Zisserman, Ed. Berlin, Germany: Springer-Verlag, 2006, pp. 345-361.

[105] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *Ann. Stat.*, vol. 28, no. 2, pp. 337-407, Apr. 2000, doi: 10.1214/aos/1016218223.

[106] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Doc. Anal. Recognit.*, Canada, Aug. 14-16, 1995, vol. 1, pp. 278-282, doi: 10.1109/ICDAR.1995.598994.

[107] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[108] A. Prinzie and D. Van den Poel, "Random forests for multiclass classification: Random multinomial logit," *Expert Syst. Appl.*, vol. 34, no. 3, pp. 1721-1732, 2008, doi: 10.1016/j.eswa.2007.01.029.

[109] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved random forest for classification," *IEEE Trans. Image Proc.*, vol. 27, no. 8, pp. 4012-4024, 2018, doi: 10.1109/TIP.2018.2834830.

[110] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24-31, 2016, doi: 10.1016/j.isprsjprs.2016.01.011.

[111] M. Kaur, H. K. Gianey, D. Singh, and M. Sabharwal, "Multi-objective differential evolution based random forest for e-health applications," *Mod. Phys. Lett. B*, vol. 33, no. 05, p. 1950022, 2019, doi: 10.1142/S0217984919500222.

[112] A. B. Shaik and S. Srinivasan, "A brief survey on random forest ensembles in classification model," in *Int. Conf. Innov. Comput. Commun.*, vol. 2, pp. 253-260, 2019, doi: 10.1007/978-981-13-2354-6_27.

[113] E. Scornet, "Random forests and kernel methods," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1485-1500, 2016, doi: 10.1109/TIT.2016.2514489.

[114] A. Davies and Z. Ghahramani, "The random forest kernel and other kernels for big data from random partitions," *arXiv:1402.4293*, 2014, [Online]Available: https://arxiv.org/abs/1402.4293

[115] R. Sarikaya, G. E. Hinton, and A. Deoras, "Application of deep belief networks for natural language understanding," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 4, pp. 778-784, 2014, doi: 10.1109/TASLP.2014.2303296.

[116] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 2, pp. 604-624, Feb. 2020, doi: 10.1109/TNNLS.2020.2979670.

[117] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, USA, June 27-30, 2016, pp. 21-29, doi: 10.1109/CVPR.2016.10.

[118] S. Nam, Y. Kim, and S. J. Kim, "Text-adaptive generative adversarial networks: manipulating images with natural

language," *arXiv:1810.11919*, 2018, [Online]Available: https://arxiv.org/abs/1810.11919

[119] B. Jang, M. Kim, G. Harerimana, S.-u. Kang, and J. W. Kim, "Bi-LSTM model to increase accuracy in text classification: combining Word2vec CNN and attention mechanism," *Appl. Sci.*, vol. 10, no. 17, p. 5841, 2020, doi: 10.3390/app10175841.

[120] M. Bi, Q. Zhang, M. Zuo, Z. Xu, and Q. Jin, "Bi-directional LSTM Model with Symptoms-Frequency Position Attention for Question Answering System in Medical Domain," *Neural Process. Lett.*, vol. 51, no. 2, pp. 1185-1199, 2020, doi: 10.1007/s11063-019-10136-3.

[121] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The handbook of brain theory and neural networks*, M. A. Arbib, Ed. Cambridge, MA, USA: MIT Press, Oct. 1998, pp. 255-258.

[122] J. Sun, X. Cao, H. Liang, W. Huang, Z. Chen, and Z. Li, "New interpretations of normalization methods in deep learning," *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 04, pp. 5875-5882.

[123] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," *arXiv:1602.07868*, 2016, [Online] Available: https://arxiv.org/abs/1602.07868

[124] R. Xiong et al., "On layer normalization in the transformer architecture," in *Proc. 37th Int. Conf. Mach. Learn.*, in Proceedings of machine learning research, vol. 119, July 2020, pp. 10524-10533.

[125] Y. Wu and K. He, "Group normalization," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, in Computer Vision ECCV 2018, in Lecture Notes in Computer Science, vol. 11217, 2018, pp. 3-19.

[126] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84-90, May 2012, doi: 10.1145/3065386.

[127] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014, [Online]Available: https://arxiv.org/abs/1409.1556

[128] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90..

[129] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455-5516, 2020, doi: 10.1007/s10462-020-09825-6.

[130] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[131] O. Melamud, J. Goldberger, and I. Dagan, "context2vec: Learning generic context embedding with bidirectional lstm," in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn.*, Berlin, Germany, Aug. 2016, pp. 51-61, doi: 10.18653/v1/K16-1006.

[132] S. Vosoughi, P. Vijayaraghavan, and D. Roy, "Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder," in *Proc. 39th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, Italy, July 2016, pp. 1041-1044, doi: 10.1145/2911451.2914762.

[133] C. Wang, M. Li, and A. J. Smola, "Language models with transformers," *arXiv:1904.09408*, 2019, [Online] Available: https://arxiv.org/abs/1904.09408.

[134] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, 2018, [Online]Available: https://arxiv.org/abs/1810.04805

[135] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv:1909.11942*, 2019, [Online]Available: https://arxiv.org/abs/1909.11942

[136] Z. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng, and S. Yan, "Convbert: Improving bert with span-based dynamic convolution," *arXiv:2008.02496*, 2020, [Online]Available: https://arxiv.org/abs/2008.02496

[137] S. Geng, P. Gao, Z. Fu, and Y. Zhang, "RomeBERT: Robust Training of Multi-Exit BERT," *arXiv:2101.09755*, 2021, [Online]Available: https://arxiv.org/abs/2101.09755

[138] Z. Liu, G. Li, and J. Cheng, "Hardware Acceleration of Fully Quantized BERT for Efficient Natural Language Processing," *arXiv:2103.02800*, 2021, [Online]Available: https://arxiv.org/abs/2103.02800

[139] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA, USA: MIT press, 2018.

[140] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26-38, 2017, doi: 10.1109/MSP.2017.2743240.

[141] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *2012 Int. Conf. Priv. Secur. Risk Trust And 2012 Int. Conf. Soc. Comput.*, Netherlands, Sept. 2012, pp. 71-80, doi: 10.1109/SocialCom-PASSAT.2012.55.

[142] P. F. Brown, V. J. Della Pietra, P. V. Desouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural

language," *Comput. Linguist.*, vol. 18, no. 4, pp. 467-480, 1992.

[143] M. Ptaszynski et al., "Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization," *Int. J. Child Comput. Interact.*, vol. 8, pp. 15-30, 2016, doi: 10.1016/j.ijcci.2016.07.002

[144] P. Fortuna, J. Soler-Company, and L. Wanner, "How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?," *Inf. Process. Manag.*, vol. 58, no. 3, p. 102524, 2021, doi: 10.1016/j.ipm.2021.102524.

[145] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531*, 2015, [Online]Available: https://arxiv.org/abs/1503.02531.

[146] Y.-J. Choi, B.-J. Jeon, and H.-W. Kim, "Identification of key cyberbullies: A text mining and social network analysis approach," *Telemat. Inform.*, vol. 56, p. 101504, 2021, doi: 10.1016/j.tele.2020.101504

[147] K. Yokotani and M. Takano, "Social contagion of cyberbullying via online perpetrator and victim networks," *Comput. Hum. Behav.*, vol. 119, p. 106719, 2021, doi: 10.1016/j.chb.2021.106719.

[148] P. K. Jonason and G. D. Webster, "The dirty dozen: a concise measure of the dark triad," *Psychol. Assess.*, vol. 22, no. 2, p. 420-432, 2010, doi: 10.1037/a0019265.

[149] B. Çetin, E. Yaman, and A. Peker, "Cyber victim and bullying scale: A study of validity and reliability," *Comput. Educ.*, vol. 57, no. 4, pp. 2261-2271, 2011, doi: 10.1016/j.compedu.2011.06.014.

[150] E. Menesini, A. Nocentini, and P. Calussi, "The measurement of cyberbullying: Dimensional structure and relative item severity and discrimination," *Cyberpsychol. Behav. Soc. Netw.*, vol. 14, no. 5, pp. 267-274, 2011, doi: 10.1089/cyber.2010.0002

[151] L. Cheng, K. Shu, S. Wu, Y. Silva, D. Hall, and H. Liu, "Unsupervised Cyberbullying Detection via Time-Informed Deep Clustering," in *29th ACM Int. Conf. Inf. Knowl. Manag. (CIKM)*, Virtual Event Ireland, Oct. 2020, pp. 185-194, doi: 10.1145/3340531.3411934.

[152] N. Lu, G. Wu, Z. Zhang, Y. Zheng, Y. Ren, and K. K. R. Choo, "Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts," *Concurr. Comput.*, vol. 32, no. 23, p. e5627, 2020, doi: 10.1002/cpe.5627.

[153] P. Zhang, Y. Gao, and S. Chen, "Detect Chinese Cyber Bullying by Analyzing User Behaviors and Language Patterns," in *2019 3rd Int. Symp. Auto. Syst. (ISAS)*, China, 2019, pp. 370-375, doi: 10.1109/ISASS.2019.8757714.

[154] S. Mishra, S. Prasad, and S. Mishra, "Multilingual Joint Fine-tuning of Transformer models for identifying Trolling, Aggression and Cyberbullying at TRAC 2020," in *Proc. Second Workshop Trolling Aggress. Cyberbullying*, France, 2020, pp. 120-125.

[155] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "Deep learning models for multilingual hate speech detection," *arXiv:2004.06465*, 2020, [Online]Available: https://arxiv.org/abs/2004.06465

[156] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro, "Early detection of depression: social network analysis and random forest techniques," *J. Med. Int. Res.*, vol. 21, no. 6, p. e12554, 2019.

[157] K. Kumari, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "Multi-modal aggression identification using Convolutional Neural Network and Binary Particle Swarm Optimization," *Future Gener. Comput. Syst.*, vol. 118, pp. 187-197, May 2021, doi: 10.1016/j.future.2021.01.014.

[158] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *arXiv:1710.09829*, 2017, [Online] Available: https://arxiv.org/abs/1710.09829

[159] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352-2449, 2017, doi: 10.1162/neco_a_00990.

[160] K. Kumari, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach," *Soft Comput.*, vol. 24, no. 15, pp. 11059-11070, 2020, doi: 10.1007/s00500-019-04550-x.

[161] G. Jacobs, C. Van Hee, and V. Hoste, "Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?," *Nat. Lang. Eng.*, pp. 1-26, 2020, doi: 10.1017/S135132492000056X.

[162] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard, and J. Tørresen, "Mental health monitoring with multimodal sensing and machine learning: A survey," *Pervasive Mob. Comput.*, vol. 51, pp. 1-26, 2018, doi: 10.1016/j.pmcj.2018.09.003.

[163] A. B. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," *Psychol. Med.*, vol. 49, no. 9, pp. 1426-1448, 2019, doi: 10.1017/S0033291719000151.

[164] S. D'Alfonso, "AI in mental health," *Curr. Opin. Psychol.*, vol. 36, pp. 112-117, Dec. 2020, doi: 10.1016/j.copsyc.2020.04.005.

[165] M. Arif, A. Basri, and G. Melibari, "Classification of

Anxiety Disorders using Machine Learning Methods: A Literature Review," *Insights Biomed. Res*, vol. 4, no. 1, pp. 95-110, 2020, doi: 10.36959/584/455.

[166] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification with cross-lingual embeddings," *arXiv:2010.05324*, 2020, [Online]Available: https://arxiv.org/abs/2010.05324

[167] A. Kumar and N. Sachdeva, "Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data," *Multimed. Syst.*, pp. 1-15, 2020, doi: 10.1007/s00530-020-00672-7.

[168] A. Malte and P. Ratadiya, "Multilingual cyber abuse detection using advanced transformer architecture," in *TENCON 2019-2019 IEEE Reg. 10 Conf. (TENCON)*, 2019, pp. 784-789, doi: 10.1109/TENCON.2019.8929493.

[169] E. W. Pamungkas and V. Patti, "Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon," *in Proc. 57th Annu. Meet. Assoc. Comput. Linguist. Student Res. Workshop*, Italy, July 2019, pp. 363-370, doi: 10.18653/v1/P19-2051.

[170] O. Gencoglu, "Cyberbullying detection with fairness constraints," *IEEE Internet Comput.*, vol. 25, no. 1, pp. 20-29, Jan.-Feb. 2020, doi: 10.1109/MIC.2020.3032461.

[171] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, and M. Ghassemi, "Hurtful words: quantifying biases in clinical contextual word embeddings," in *Proc. ACM Conf. Health Inference Learn.*, Canada, Apr. 2020, pp. 110-120, doi: 10.1145/3368555.3384448.

[172] S. Dev, T. Li, J. M. Phillips, and V. Srikumar, "On measuring and mitigating biased inferences of word embeddings," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 05, pp. 7659-7666, doi: 10.1609/aaai.v34i05.6267.

[173] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl, "Unintended machine learning biases as social barriers for persons with disabilitiess," *ACM SIGACCESS Access. Comput.*, no. 9, Mar. 2020, doi: 10.1145/3386296.3386305.

[174] M. Wich, H. Al Kuwatly, and G. Groh, "Investigating Annotator Bias with a Graph-Based Approach," in *Proc. Fourth Workshop Online Abuse Harms*, Nov. 2020, pp. 191-199, doi: 10.18653/v1/2020.alw-1.22.

[175] H. Al Kuwatly, M. Wich, and G. Groh, "Identifying and measuring annotator bias based on annotators' demographic characteristics," in *Proc. Fourth Workshop Online Abuse Harms*, Nov. 2020, pp. 184-190, doi: 10.18653/v1/2020.alw-1.21.

[176] J. Li et al., "Textshield: Robust text classification based on multimodal embedding and neural machine translation," in *29th {USENIX} Secur. Symp. ({USENIX} Security 20)*, 2020, pp. 1381-1398.

[177] G. Kovács, P. Alonso, and R. Saini, "Challenges of Hate Speech Detection in Social Media," *SN Comput. Sci.*, vol. 2, no. 2, pp. 1-15, 2021, doi: 10.1007/s42979-021-00457-3.

[178] D. Weber and F. Neumann, "Who's in the Gang? Revealing Coordinating Communities in Social Media," *arXiv:2010.08180*, 2020, [Online]Available: https://arxiv.org/abs/2010.08180

[179] S. Nikiforos, S. Tzanavaris, and K.-L. Kermanidis, "Virtual learning communities (VLCs) rethinking: influence on behavior modification—bullying detection through machine learning and natural language processing," *J. Comput. Educ.*, vol. 7, pp. 531-551, 2020, doi: 10.1007/s40692-0s20-00166-5.

[180] S. Nikiforos, S. Tzanavaris, and K.-L. Kermanidis, "Virtual learning communities (VLCs) rethinking: Collaboration between learning communities," *Educ. Inf. Technol.*, vol. 25, pp. 3659-3675, 2020, doi: 10.1007/s10639-020-10132-4.

[181] I. Fleury and E. Dowdy, "Social Media Monitoring of Students for Harm and Threat Prevention: Ethical Considerations for School Psychologists," *Contemp. School Psychol.*, 2020, doi: 10.1007/s40688-020-00311-y.

[182] R. Clarke, "Information technology and dataveillance," *Commun. ACM*, vol. 31, no. 5, pp. 498-512, 1988, doi: 10.1145/42411.42413.

[183] R. Clarke and G. Greenleaf, "Dataveillance regulation: A research framework," *J. Law Inf. Sci.*, vol. 25, no. 1, 2018.