IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A Comprehensive Review on Cyberbullying detection using Machine Learning Models

[1]Dr. Subhash Bhagavan Kommina, [2]Srinadh Unnava, [3]Puja Deepthi Chowdary Kommina, [4]Gudavalli Giridhar Sai, [5]Kotu Vijay, [6]Rapaka Sandeep
[1]Department of Information Technology,
[1]Sasi Institute of Technology & Engineering, Tadepalligudem, India

Abstract: In the era of web 4.0 social media becomes a trend for every internet user. Everyone sharing their thoughts and emotions through social media which creates some support from their peers at the same time it creates some privacy threats from some of the peers. cyberbullying is important area, which means attacks occurs on social media platforms through comments or tweets. In addition, cyberbullying the person who is bulled will involve into depression, anxiety, and other types of self-harm. A detailed survey is conducted on cyberbullying towards the contributions made by different authors This survey majorly focused on different machine learning techniques and pre-processing techniques collected from popular data sources during the period of 2015 to 2023. This survey reveals that the majority of methods utilized in cyberbullying research are tokenization processing and support vector machines.

Keywords: Cyberbullying, Social Media, Pre Processing, Machine Learning

INTRODUCTION:

Social media sites such as Twitter and Facebook have grown in popularity as a result advancement of internet 3.0 technology, and they play an important part in reshaping human existence. Social media is utilized in daily activities such as education, business, entertainment, government, and online relationship development. By 2023, there will be 3.12 billion monthly active users on social media. This figure will account for almost one-third of the world's population. Twitter is a vital platform and important data source for academics among the many existing social networks. Twitter is a prominent real-time microblogging network where news arrives before it appears in official sources. Twitter has an unfiltered stream as a result of Twitter has greater "freedom of speech" than Facebook. There is a disadvantage by using these social media especially for children and young people. One of major risk raised is “Cyberbullying”, means bullying a person in a social media. The growing number of cyberbullying instances has brought to light the dangers of cyberbullying, especially among children and young people. Children and adolescents take bullying seriously because they do not understand how to manage social concerns, which drives them to vent their feelings on social media in way that might harm others This creates an urgent need to put a halt to cyberbullying .Thus, spotting, cyberbullying on social networking is critical and should be prioritized so that children and society are safeguarded from its consequences. The main thing to do to regulate and prevent cyberbullying is to identify text content using machine learning for foul language identification and categorization. Our goal in this comparative study is to examine the performance and accuracy of machine learning algorithms and pre-processing techniques that detect bulled text and to conclude the best machine learning and pre-processing techniques used for detecting cyberbullying with high accuracy.

LITERATURE SURVEY:

Dadvar et al. [1] examined the gender perspective in the context of cyberbullying detection issue, using MySpace as an example. My Space is a social network that provides an interactive, user-submitted buddy community with individual profiles, blogs, groups, and other features. The writers examined the posts' content, examining it independently of the individuals' personal details. To train a text classifier that is particular to gender, they employed an SVM model. There are around 381.000 posts in the dataset. The gender-based approach's outcomes increased the baseline by 15% in F-measure, 6% in recall, and 39% in precision.

Dinakar et al. [2] achieved 66.7% accuracy by using multiple binary and multiclass classifiers on a manually labeled corpus of YouTube comments. The authors also utilized an SVM learner.

Xu et al. [3] proposed many natural language processing algorithms for detecting bully traces, as well as describing the structure of a bullying event and potential roles. The authors employed Sentiment Analysis to assign roles and Latent Dirichlet Analysis to identify subjects. Cyberbullying detection is treated as a binary (positive/negative) classification issue, with a linear SVM trained on a manually labeled dataset. The results showed an 89% cross validation accuracy, proving that even simple traits and a common classifier may be utilized to detect cyberbullying signs in text.

J. Yadav et al.[4] present a new strategy to detecting Cyberbullying on social media sites is modeled using the BERT model, with a single linear neural network layer acting as a classifier. The model is trained and tested using the Formspring forum and the Wikipedia dataset. The recommended model attained performance accuracy 98% for the Form spring dataset and 96% for the Wikipedia dataset, which is greater than the prior models used. The proposed model worked better on the Wikipedia dataset because of its large size g, which eliminated the need for oversampling, whereas the Form spring dataset needed oversampling.

R. R. Dalvi et al. [5] propose a strategy for recognizing and blocking Internet exploitation on Twitter using supervised categorization machine learning algorithms. Using the Live Twitter API, Tweets are collected, and datasets are generated for this investigation. Support Vector Machine and Naive Bayes are both used to evaluate the collected datasets against the specified model. They used the TFIDF vectorizer to extract the feature. The data show that the cyberbullying model based on the Support Vector Machine has an accuracy of around 71.25%, which is higher than the Naive Bayes model's accuracy of roughly 52.75%.

Trana R.E., et al. [6] created a machine learning strategy to reduce anomalous incidents using text extracted from visual memes. The author created a database using around 19,000 text views published to YouTube. This research assesses the performance of three machine learning models: Uninformed Bayes, Support Vector Machine learning and convolutional neural networks are employed on the YouTube database, and the results are compared to existing Form datasets. In subcategories of the YouTube database, the scientists investigated algorithms for cyberbullying on the internet. In the four categories of race, ethnicity, politics, and generalization, Naive Bayes outperformed SVM and CNN. SVM matched CNN and Naïve Bayes in core body group accuracy, but outperformed both in the same gender group.

G. A. León-Paredes et al.[7] created a model for identifying cyberbullying that employs machine learning (ML) and natural language processing. NB, SVM, and Lare are several machine learning algorithms utilized in the development of the Spanish cyberbullying prevention system (SPC). The dataset utilized in this analysis was obtained from Twitter. Using three separate approaches yielded the best accuracy (93%). The method's average accuracy in detecting cyberbullying is between 80% and 91%. To improve system accuracy, natural language processing techniques such as stemming and lemmatization can be used. If viable, such a model can be used to recognize both native and English languages

P. K. Roy et al.'s work [9], a deep convolutional neural network is used to create a request for the discovery of hate speech on Twitter. A variety of machine learning algorithms, including Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM).

H. P. K. Roy et al.'s work [8], a deep convolutional neural network is utilized to generate a request for the detection of hate speech on Twitter. Machine learning algorithms include LR, RF, NB, and Support Vector Machine (SVM).Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), and K-nearby Neighbors (KNN) were used to detect hate speech related messages on Twitter. The features were then deleted using the tf-idf method. The top machine learning model was SVM, which predicted 53% of hate speech tweets in a 3:1 dataset used to test the train. The data was uneven, which contributed to the low prediction scale. Advanced learning approaches based on convolutional neural networks can provide similar results.. Neural Networks (CNN), Long-Term Memory (LSTM), and Contextual LSTM (CLSTM) are separate distributed databases. Using the proposed DCNN model in combination with 10-fold cross-validation resulted in a very high recall. The score was 0.99 for non-hate speech and 0.88 for hate speech. Test findings show that the k-fold cross-validation process is a preferable alternative for dealing with unequal data. To enhance accuracy in the future, the present database can be expanded

To give a more precise and guided picture of cyberbullying, S. M. Kargutkar et al. [9] created a system that examines material using Convolutional Neural Network (CNN) and Keras techniques. This analysis, which analyzed Twitter and YouTube data, found CNN accuracy to be 87%. In-depth learning-based models have proven successful in identifying instances of digital harassment, overcoming traditional model constraints, and increasing adoption

Rasel, Risul Islam, et al. [10] focus on the removal of social media comments and the investigation of whether these comments have an inappropriate meaning. The replies may be divided into three categories: hate speech, offensive speech, and none of the two. The recommended model correctly classifies more than 93% of the comments on the species. To reduce the amount of input data, Latent Semantic Analysis (LSA) was used as a feature selection approach. To identify relevant notes, TF-IDF was utilized with traditional feature extraction approaches like as tokenization and N-gram. To compute, analyze, forecast, andIn addition, we constructed three different machine learning models: Random Forest, Logistic Regression, and Support Vector Machines (SVMs).calculate, analyze, forecast, and make a mocking comment.

Cynthia Van Hee et al. [11] created a model for identifying cyberbullying in social media posts by analyzing posts from both bullies and victims. The collecting of data from social networking sites such as Ask.fm, two corpora were created. created a preprocessing model including tokenization, PoS- tagging, and lemmatization. To test for linguistic conversion and ensuing correctness, models for English and Dutch were created. The accuracy for the English language was 64%, while the Dutch language was 61%, according to the ML algorithm SVM

A technique to lessen textual cyberbullying was put into place by Mohammed Ali Al-Garadi et al. [12] because this type of hostile behavior has taken over on social networks They employed tokenization lemmatization to build a model after extracting data from Wikipedia, YouTube, Twitter, and Instagram. N- gram was utilized up to five layers to compute TF IDF and count vector for pre-processing. Using SVM, K clustering, Random Forest, and Decision Trees, they conducted a comparative examination of ML algorithms and found that SVM performed the best out of the four machine learning models.

The objective of Sahay, Kshitiz, et al. [13] was to classify and identify bullying in the text by analyzing the traits that differentiate bullies and aggressors from other users. They used Wikipedia, YouTube, and Twitter to get their data. Preprocessing included eliminating URLs and tags from the data collection, as well as running Count Vectors and TF-IDF vectors. For classification, they used LR, RF, SVM, and Gradient Boosting.

Homa Hosseinmardi, et al. [14] proposed applying image processing on Instagram posts to determine test response for text pictures or emotional response. Additionally, they created a mechanism that selects postings based on the terms in the caption that make the short list. A model to automatically identify cyberbullying text on Instagram was proposed by them, based on their model of postings from bullies

Vijay Banerjee et al. [15] developed a cyberbullying detection model utilizing a convolution neural network and compared its accuracy to previous techniques. They exploited the 69874 converted tweets from Twitter into vectors.Accuracy of this model was 93.97%, It was higher than previous versions.

Noviantho, et al. [16] created a cyberbullying classification model utilizing the Naive Bayes method and Support Vector Machine (SVM).They obtained the 1600 Formspring.me discussions with labels derived from questions and responses from Kaggle, which provided the data set they used. Out of the total 12729 data, 11661 are categorized as non-cyberbullying and 1068 as cyberbullying.Terms like "haha," "hehe," "um," and so forth were removed throughout the data cleansing process. To balance the dataset, two classes were created: cyberbullying and non-cyberbullying; four classes were created: non-cyberbullying, cyberbullying with low, middle, and high severity level; and eleven classes were created: non-cyberbullying, cyberbullying with 1–10 severity level. During preprocessing, tokenizations, transfer case, stop word removal, filter token, stemming, and n-gram creation were used.They used SVM and Naive Bayes with linear, poly, and sigmoid kernels for classification. The poly kernel SVM kernel produced the highest average accuracy of 97.11

The goal of H. Watanabe et al.'s study [17] was to identify hate speech on Twitter. Their method is predicated on automatically gathered patterns and unigrams from the dataset. To categorize tweets as clean, provocative, or hateful was their goal. A trio of dataset kinds were employed. Four thousand tweets from Crowdflower were categorized as clean,offensive, and hateful in the1st dataset; fourteen thousand tweets from Crowdflower were categorized as offensive, hateful, and neither in the second dataset; and eight thousand tweets from Github were categorized as sexism, racism, and neither in the third dataset. To create a larger dataset, they merged three smaller ones. They did tokenization, lemmatization, and part-of-speech tagging in addition to removing the URL and tags from the tweets during preprocessing.To find sentiment-based features, semantic features, Unigram features, and pattern features, they employed binary and ternary classification.The accuracy of their suggested model was 87.4% for binary classification, which divided tweets into offensive and non-offensive categories, and 78.4% for ternary classification, which divided tweets into offensive, hateful, and clean categories

D. Chauhan, J. Yadav, and D. Kumar Yadav, Kumar, Chauhan, and colleagues [18] created a cyberbullying classification model based on a pre-trained BERT model. Google researchers have built a learning model called BERT. They do this by using datasets from Wikipedia talk pages, a collaborative knowledge repository, and publicly accessible Formspring, a quality assurance forum. Both datasets had been manually labeled and pre-processed. The Wikipedia dataset has 115864 discussion comments that were carefully marked by ten people, of which 13590 comments are classified as bully. The Formspring dataset has 12773 question-answer pair comments, of which 776 are bully postings. Their model produced an accuracy of 81% for the Wikipedia dataset and 94% for the three-times oversampled Formspring dataset

Aditya Gaydhani, Doma et al. [19] Developed a machine learning model for identifying hate speech and derogatory words on Twitter. The datasets came from GitHub and CrowdFlower. While the GitHub dataset featured columns for tweet id and class—such as sexism, racism, and neither—the Crowd Flower dataset contained tweets with the labels "hateful," "offensive," and "clean." Using the Twitter API, tweets were retrieved by the tweet id. After then, these datasets were merged. Lowercase tweets were created, and stop words, URLs, Twitter mentions, space patterns, and retweet symbols were eliminated. Stemming was used to get rid of word inflectional forms. Next, the dataset was divided into 30% test samples and 70% training samples. After being recovered from the tweets, the N- gram features were weighed based on their TF IDF values. Unigram, Bigram and Trigram features along with L1 and L2 normalization of TF IDF were considered. LR, NB and SVM algorithms are compared. 95% accuracy was obtained using LR with L2 Normalization and n=3

A DL model was employed by AlAjlan et al. [20] to detect cyberbullying. The features are extracted from the input, they employed feature engineering and feature selection approaches. They employed a 39,000-tweet Twitter dataset, from which the duplicates were eliminated during the cleaning process. The model was evaluated on 2700 bully and 6300 non-bully tweets after being trained on 9000 bully and 21,000 non-bully tweets. Their 95% accuracy model outperformed the SVM by a wide margin.

The dataset utilized by Banerjee et al. [21] included 69,874 tweets. They used Glove word embedding to turn the word into vectors. During data preprocessing, stop word accentuation marks were removed and then converted to lowercase. They employed a CNN-based deep learning model on the processed data to identify bullying posts, achieving an accuracy rate of 93.97%.

Using text mining techniques, Cigdem et al. [22] created a model for automatically identifying instances of cyberbullying on social networks. To get the best results, they experiment with several classifier types using feature selection methods. Three distinct social networks were used to obtain the dataset: YouTube, Myspace,and Formspring.me.Two classes i) cyberbullying positive and (ii) cyberbullying negative were created from the dataset. The 13,158 communications in the Formspring.me dataset were contained in an XML file; 892 of those messages dealt with cyberbullying positively, and the other 12,266 dealt with cyberbullying negatively. There are 1753 messages in the Myspace dataset; 357 of them have positive labels, and the remaining 1396 have negative ones.

3464 messages from various individuals were included in the YouTube dataset, and out of which 3047 were negative and 417 were positive. For every dataset, the f-measure value was greater than 0.90 for two classifiers: SGD and MLP

A CNN-based text classification model was presented by Chen et al. [23] for the de facto verbal aggressiveness dataset. Their emotions and feelings were not taken into account when they manually added Facebook comments and Tweets to the datasets. In addition to the hand- labeled remarks, they gathered social network comment information from the "sentiment140 corpus." Following the alteration, the tweets' polarity are classified as aggressive or nonaggressive. During preprocessing, they lowercased and deleted the usernames, which are followed at the rate by stickers on hashed subjects. Using the tf-idf method, features were extracted. The best results were produced by the DL-based CNN model, which had an accuracy of 0.92 and an AUC value of 0.98

The problem of automatically identifying hostility in tweets from the cyber-troll dataset was tackled by Sadiq et al. [24].CNN-LSTM and CNN-BiLSTM models were employed. There are 20,001 cases in the dataset; 7,822 of them are cyber-aggressive, while 12,179 are not. NLTK is first used to preprocess the dataset in order to improve the outcome. Their tf-idf model containing uni- and bi-grams performed better, obtaining an F1-score of 0.90 and an accuracy of 0.92.

Using a dataset from MySpace.com, Nandhini et al. [25] proposed a model that uses the Naïve Bayes machine learning approach and achieved 91% accuracy. They then proposed a second model [4] that used the Naïve Bayes classifier and genetic operations (FuzGen) and achieved 87% accuracy.

Using this method, Romsaiyud et al. [26] obtained 95.79% accuracy on datasets from Slashdot, Kongregate, and MySpace by improving the Naïve Bayes classifier for word extraction and loaded pattern clustering analysis. Their issue is that the cluster processes aren't able to operate concurrently.

In their study, Dinakar et al. [27] collected data from YouTube comment sections with the goal of identifying explicit bullying language related to (1) sexuality, (2) race and culture, and (3) IQ. Following the application of Naïve Bayes and SVM classifiers, SVM produced an accuracy of 66% and Naïve Bayes 63%.

In their study, Di Capua et al. [28] introduced an innovative method for detecting cyberbullying through an unsupervised approach. They employed a variety of classifiers across their dataset, using SVM on FormSpring to achieve 67% recall, GHSOM on YouTube to achieve 60% precision, 69% accuracy, and 94% recall, and Naïve Bayes on Twitter to achieve 67% accuracy. Furthermore, Haidar et al. [10] presented a model to identify cyberbullying; however, they employed the Arabic language and Naïve Bayes, achieving 90.85% precision, and SVM, achieving 94.1% precision. Despite this, they had a large false positive rate.

In order to overcome class imbalance, Zhang et al. [29] employ a unique pronunciation-based convolution neural network (PCNN) to reduce noise and bullying data sparsity. 13,000 messages came via formspring.me and 1313 messages from Twitter. Because the Twitter dataset was unbalanced, its accuracy wasn't calculated. Although they achieved excellent accuracy with a 56% precision score, 78% recall, and 96% accuracy, their dataset was imbalanced, which led to erroneous findings and the 56% precision score.

Zhao et al.'s [30] framework was designed specifically for the detection of cyberbullying. They employed SVM as their primary classifier and achieved 79.4% recall by using word embedding, which creates a list of pre- defined insulting terms and assigns weights to retrieve bullying attributes.

With SVM as their classifier, Chen et al. [31] achieved 77.9% precision and 77.8% recall using a novel feature extraction technique called Lexical Syntactic Feature.

V.Banerjee, et al [32] used CNN Cyberbullying detection using DNN Single data source.

According to Raisi et al. [33], Twitter is one of the social media sites that is visible to the public and where cyberbullying occurs frequently. To the best of our knowledge, given the amount of reviewed publications that have profited from Twitter contents, Twitter is the most accessible resource for scholars studying Natural Language Processing (NLP). The fact that registered users can broadcast brief posts (280 characters each post), the majority of which are textual and provide a direct to point source of data, is one of the reasons this social media is popular among academics checking their suggested algorithm. Furthermore, as users can tweet in other languages on Twitter, it is possible to use Twitter to obtain datasets in languages other than English. Daily usage of Twitter is rising quickly.

With a 78.5% accuracy rate, Reynolds et al. [34] presented a language-based technique for identifying cyberbullying. The writers attained this accuracy by utilizing instance-based training and decision trees. The paper's author has employed personality, sentiment, and emotion as features to enhance the identification of cyberbullying.
Isa et al. [35] presented a method in which they employed Naïve Bayes and SVM classifiers after obtaining their dataset from Kaggle. The average accuracy of the Naïve Bayes classifier was 92.81%, while the SVM with poly kernel produced an accuracy of 97.11%. However, the training and testing dataset sizes were not specified, therefore the findings might not be reliable

Research Questions:

RQ1: What are the most used preprocessing technique in cyberbullying?

RQ2: What are the most used ML algorithm in cyberbullying?

RQ3: What are the most common features that can be indicators of cyberbullying in online conversations

METHODOLOGY:

The procedure of acquiring information for the training and testing of a model is referred to as data collection. This information can be sourced from existing datasets or extracted from comments and messages within specific web-based platforms. To ensure the model is effectively prepared, the obtained data undergoes pre- processing to mitigate noise. This involves the removal of special characters, hashtags, usernames, and irrelevant words such as "as," "what," and "who." Additionally, the identification of cyberbullying within the text is crucial and is achieved through feature extraction methods. Various techniques, such as Bag of Words, TF-IDF Vectorizer, ngrams, Word2Vec, Doc2Vec, and emotion values of messages may be used to study and issustrate the characteristics in the data. Feature selection becomes imperative to enhance accuracy and mitigate overfitting. During this phase, key features are chosen, and methods like Filter, Wrapper, Embedded, and Hybrid approaches are utilized for this purpose .Finally, the classification stage involves employing machine learning algorithms such as Support Vector Machines (SVM), naive Bayes, logistic regression, among others. These algorithms use the selected features to categorize the data into relevant classes, enabling the model to identify and address instances of cyberbullying effectively. The overall process, from data collection to classification, forms a comprehensive pipeline for developing a robust and accurate model for cyberbullying detection

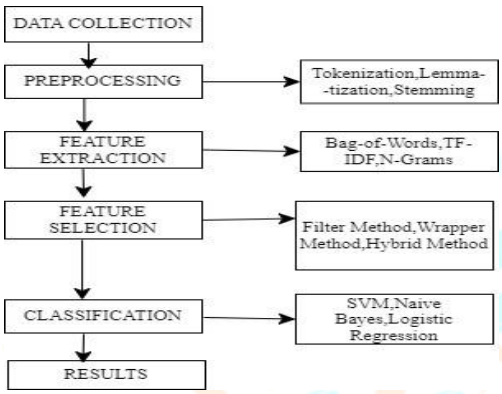

Fig 1: Methodology

Pre-Processing Techniques:

Tokenization: Tokenization is the process of breaking down a text into smaller units, such as words or subwords, called tokens

Bag-Of-Words: It's a way to convert text into a numerical format that can be utilized by machine learning algorithms

Feature Extraction: Feature extraction is a process in machine learning and signal processing where relevant information or features are extracted from raw data

TF-IDF: It is a numerical statistic used in natural language processing and information retrieval to evaluate the importance of a word in a document relative to a collection of documents (corpus)

Sampling: It refers to the process of selecting a subset of data points from a larger dataset.

Stemming: Stemming is a text preprocessing technique in natural language processing (NLP) and information retrieval that involves reducing words to their root or base form.

Lemmatization: Lemmatization is a natural language processing (NLP) technique used in text preprocessing to reduce words to their base or root form, known as the lemma.

Feature Selection: Feature selection is a crucial step in the preprocessing of data for machine learning tasks. It involves choosing a subset of relevant features (input variables) from the original set of features

NGrams: In the context of text preprocessing, particularly in natural language processing (NLP), n- grams are used to represent the frequency and distribution of word sequences within a document.

Table 1: Pre-Processing Techniques

Pre-Processing Techniques	**No. of Authors Worked on it**
Tokenization	11
Bag-of-Words	5
Feature Extraction	8
Sampling	7
TF-IDF	7

Stemming	3
Lemmatization	6
NGrams	5
Feature Selection	3
Count Vector	1
Pos-Tagging	1

Conclusion:

From the above table we observed that from the various preprocessing techniques Tokenization technique is frequently used by maximum number of authors with average accuracy of 83.2%

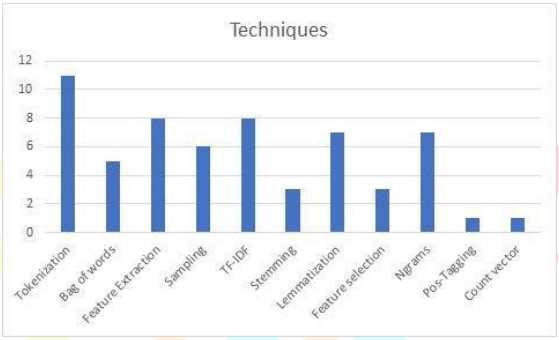

Fig 2:No.of Authors vs Techniques

ML Algorithms:

1.Support Vector Machine: SVM is particularly known for its effectiveness in dealing with high- dimensional feature spaces and for its ability to find a hyperplane that best separates different classes in the feature space. For linear SVM

Formula: $K(Y, Yi) = sum(Y * Yi)$

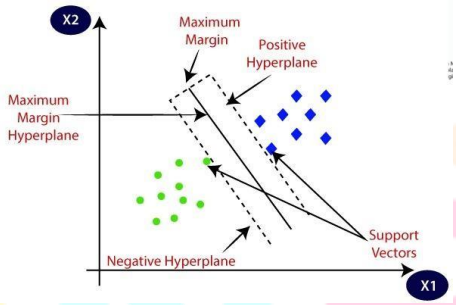

Fig 3: Support Vector Machine

2.Decision Tree: The decision tree classifier can be used in both classification and regression . It can help represent the decision as well as make a decision. The decision tree is a treelike structure where each internal node represents a condition and each leaf node represents a decision. A classification tree returns the class where the target falls. A regression tree yields the predicted value for an addressed input.

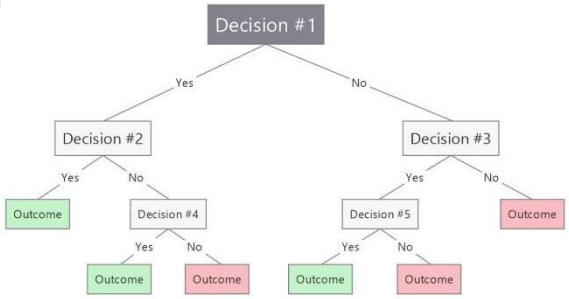

Fig 4:Decision Tree

3.Random Forest:Random Forest is an ensemble learning algorithm that operates by constructing a multitude of decision trees at training time and outputting the class (classification) or mean prediction (regression) of the individual trees for new data points. For

example, if we have two classes namely A and B and the most of the decision tree predict the class label B of any instance, then RF will decides the class label B as follows

$$f(Y) = majority\ vote\ of\ all\ tree\ as\ B$$

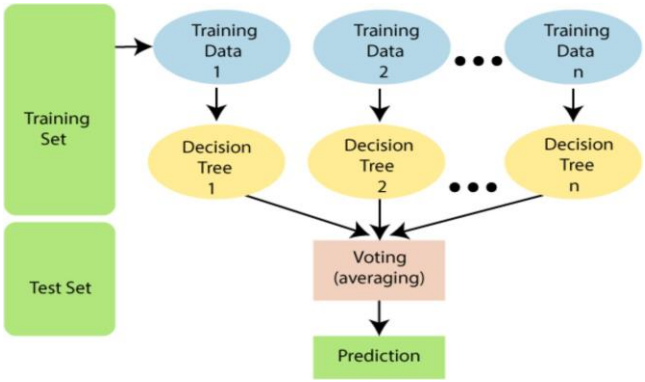

Fig 5:Random Forest

4.K-Means: K-Means is a popular clustering algorithm used in unsupervised machine learning. It's a partitioning method that divides a dataset into k distinct, non-overlapping subsets (clusters). The algorithm works iteratively to assign data points to clusters based on the similarity of their features and then updates the cluster centroids.

Working:

1.Choose the number of clusters k.

2.The first step in k-means is to pick the number of clusters, k.

3.Select k random points from the data as centroids. ...

4.Assign all the points to the closest cluster centroid. ...

5.Recompute the centroids of newly formed clusters. ...

6.Repeat steps 3 and 4

5.BERT: BERT, which stands for Bidirectional Encoder Representations from Transformers, is a pre- trained natural language processing (NLP) model developed by Google. It represents a significant advancement in the field of deep learning for language understanding tasks

6.Convolutional Neural Network (CNN): Convolutional Neural Network, is a type of deep neural network architecture designed for processing and analyzing structured grid-like data.

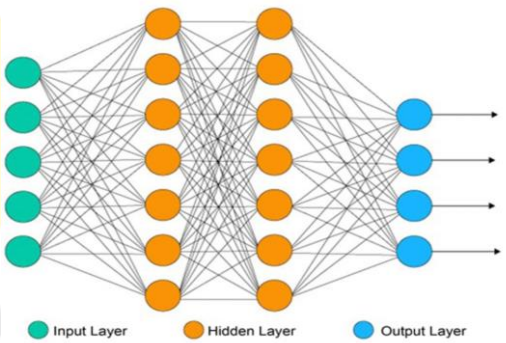

Fig 6: CNN

7.Logistic Regression: It is a statistical method and a fundamental algorithm in machine learning used for binary classification tasks. Despite its name, logistic regression is used for classification, not regression. It's particularly well-suited for problems where the outcome variable is binary (two classes)

Formula: Logistic Regression uses the logistic function (sigmoid function) to model the probability that a given input belongs to a particular class. The logistic function is defined as follows:

$$S(z) = 1/1 + e - z$$

Here, z represents the linear combination of input features and their associated weights, often denoted as θTx, where θ is the vector of weights, and x is the vector of input features.

The hypothesis function hθ(x) in logistic regression is then given by applying the sigmoid function to z:

$$h\theta(x) = S(\theta Tx) = 1/1 + e - \theta Tx$$

This function outputs the predicted probability that the input x belongs to the positive class

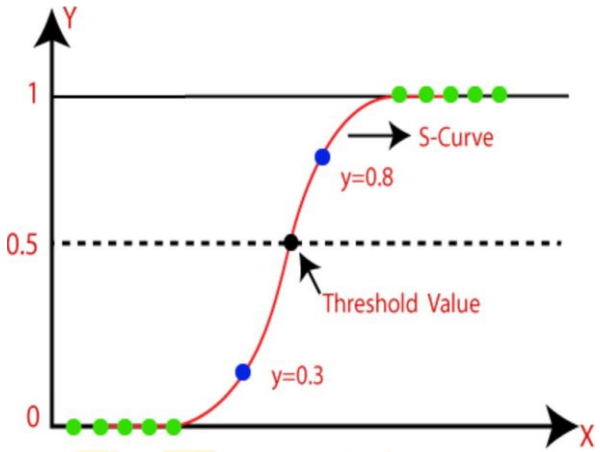

Fig 7: Logistic Regression

8.Gradient Boosting (GB): Gradient Boosting is a powerful machine learning technique used for both regression and classification tasks. It belongs to the ensemble learning family, where multiple weak learners (typically decision trees) are combined to create a strong predictive model.

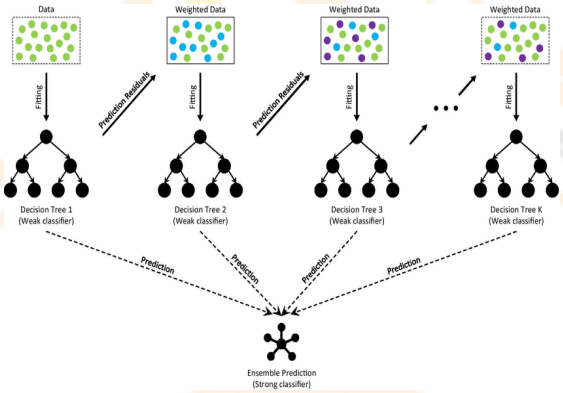

Fig 8: Gradient Boosting

9.Naive Bayes: Naïve Bayes is an efficient machine learning algorithm based on Bayes theorem . The algorithm predicts depending on the probability of an object. The binary and multi-class classification problems can be quickly solved using this technique. Based on Bayes' Theorem it finds the probability of an event occurring given the probability of another event that has already occurred as follows:

$$p(b|A) = p(A|b) \times p(b)$$

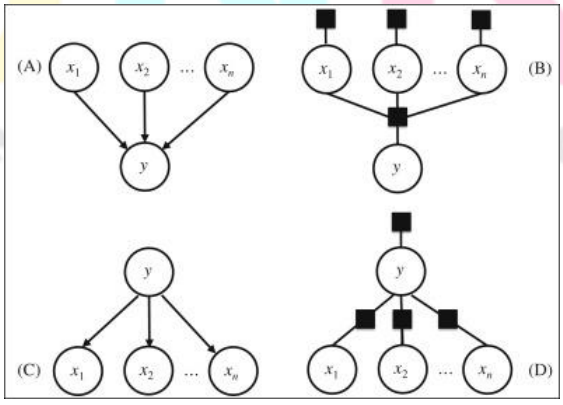

Fig 9: Naïve Bayes

10.KNN:(KNN) is a simple and intuitive algorithm used for both classification and regression tasks in machine learning. It is a type of instance-based learning where the model makes predictions based on the majority class or average target value of its k-nearest neighbors in the feature space

11.K-Clustering: K-means clustering is a type of unsupervised learning algorithm used in machine learning to partition a dataset into K clusters. The goal of K-means clustering is to group data points into K clusters such that each data point belongs to the cluster with the nearest mean (centroid).

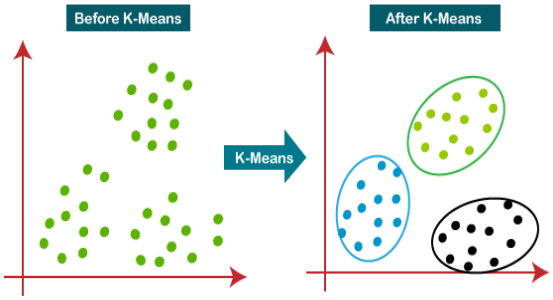

Fig 10: K-Clustering

12.Deep Learning: Deep learning is a subset of machine learning that deals with neural networks containing more than one hidden layer. It is a class of algorithms that uses a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input, allowing the network to learn increasingly abstract features from the data.

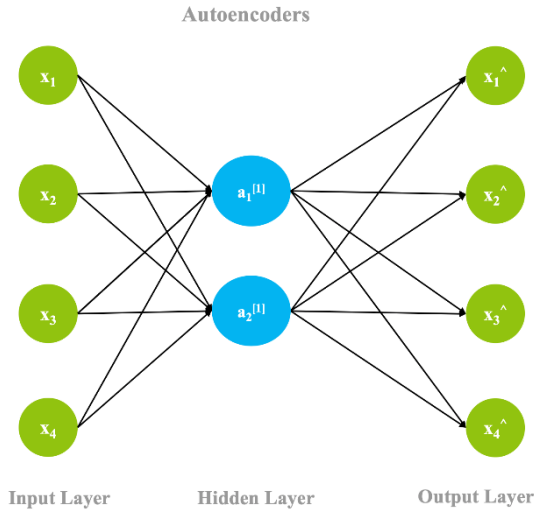

Fig 11: Deep Learning

Table 2: Machine Learning Algorithms

Machine Learning Algorithm	**No. of Authors Worked on it**
Support Vector Machine	22
Decision Tree	6
Random Forest	7
Naïve Bayes	10
K-Means	1
BERT	3
CNN	6
Gradient Boosting	3
Logistic Regression	8
KNN	1
K-Clustering	1

Conclusion:

From the above Figure we observed that from the various ML algorithms support vector machine (SVM) algorithm is frequently used by maximum number authors with average accuracy of 80.5%

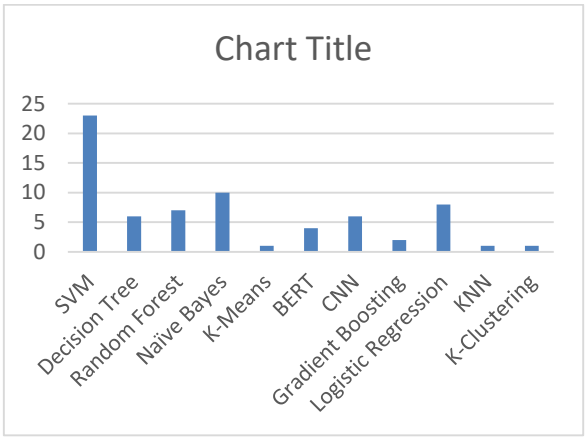

Fig 12:No.of.Authors Vs ML Algorithms

Conclusion:

This survey majority focus on to detect the cyberbullying in various social media platforms by the different authors .It was observed based on the above research data the mostly used preprocessing technique is tokenization and most Used machine learning algorithm is support vector machine. Usage of tokenization preprocessing technique given accuracy of 85% & SVM algorithm gives accuracy of 86.05%.Using the tokenization preprocessing technique and SVM algorithm gives accuracy of 86.08%.

REFERENCES:

1.M. Dadvar. 2012.”Cyberbullying detection: astep toward a safer internet yard”. In Proceedings of the 21st International Conference on World Wide Web(WWW ’12 Companion). ACM, New York, NY, USA, 121-126

2. Dinakar “Modeling the detection of Textual Cyberbullying.” The Social Mobile Web 11.02(2011):11-17

3. J. Xu, “Learning from Bulling Traces in Social Media,"Proc.2012 Conf.North Am. Chapter Assoc. Comput. Linguist. Hun. Lang. Technol, pp. 656-666,2012

4. J. Yadav, D. Kumar and D. Chauhan,Cyberbullying Detection using Pre-TrainedBERT Model, ICESC, pp.1096-1100,doi:10.1109/ICESC48915.2020.9155700. (2020)

5. . R.R. Dalvi, S. Baliram Chavan and A. Halbe, Detecting A Twitter Cyberbullying Using Machine Learning, ICICCS, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893. (2020)

6. Trana R.E., Gomez C.E., Adler R.F. (2021)Fighting Cyberbullying: An Analysis ofAlgorithms Used to Detect Harassing Text Found on YouTube. In: Ahram T. (eds)Advances in Artificial Intelligence, Softwareand Systems Engineering. AHFE 2020.Advances in Intelligent Systems andComputing, vol 1213. Springer, Cham.https://doi.org/10.1007/978-3-030-51328-3_2.(2020)

7. . A. León-Paredes et al., Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language, CHILECON pp. 1-7, doi: 10.1109/CHILECON47746.2019.8987684. (2019)

8. P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, A Framework for Hate Speech Detection Using Deep Convolutional Neural Network, in IEEE Access, vol. 8, pp. 204951-204962,

9. S. M. Kargutkar and V. Chitre, A Study of Cyberbullying Detection Using Machine Learning Techniques, ICCMC, pp. 734-739,

doi:10.1109/ICCMC48092.2020.ICCMC-00013 7. (2020)

10. Rasel, Risul Islam & Sultana, Nasrin & Akhter, Sharna & Meesad, Phayung, Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach. 37-41. 10.1145/3278293.3278303. (2018)

11. Van Hee C, et al. Automatic detection of cyberbullying in social media text. PLoS ONE. 2018;13(10):e0203794.

12. Mohammed, A. (2019). Social media cyberbullying detection using machine learning. International Journal of Advanced Computer Science and Applications, 10(5), 703-707

13. Sahay, 2018. Detecting cyberbullying and aggression in social commentary using nlp and machine learning. International Journal of Engineering Technology Science and Research 5

14. Hosseinmardi, H.2015. Detection of cyberbullying incidents on the instagram social network. arXiv preprint arXiv:1503.03909 .

15. Banerjee, V,2019. Detection of cyberbullying using deep neural network, in: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), IEEE. pp. 604–607.

16. Noviantho, 2017. Cyberbullying classification using text mining, in: 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), pp. 241–246. doi:10.1109/ICICOS. 2017.8276369.

17. Watanabe, 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE access 6, 13825–13835.

18. Yadav, J., Kumar, D., Chauhan, D., 2020. Cyberbullying detection using pre-trained bert model, in: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE. pp. 1096– 1100.

19. Gaydhani, A., Doma,2018. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. arXiv preprint arXiv:1809.08651 .

20. Al-Ajlan "Deep learning algorithm for cyberbullying detection," International Journal of Advanced Computer Science and Applications, vol. 9, no. 9, 2018

21. Banerjee, V, 2019. Detection of cyberbullying using deep neural network, in: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), IEEE. pp. 604–607

22. Cigdem A, Çürük E, E¸ssiz ES (2019) Automatic detection of cyberbullying in formspring.me, myspace and Youtube social networks. Turk J Eng 3(4):168–178

23. Chen J (2020) Verbal aggression detection on twitter comments: convolutional neural network for short-text sentiment analysis. Neural Comput Appl 32(15):10809–10818

24.Sadiq S (2021) Aggression detection through deep neural model on Twitter. Future Gener Comput Syst 114:120–129

25. B Nandhini Cyberbullying detection and classification using information retrieval algorithm. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), page 20. ACM, 2015.

26.Romsaiyud,Automated cyberbullying detection using clustering appearance patterns. In Knowledge and Smart Technology (KST), 2017 9th International Conference on, pages 242– 247. IEEE, 2017.

27. Karthik Dinakar, Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3):18, 2012.

28. Michele Di Capua,. Unsupervised cyber bullying detection in social networks. In Pattern Recognition (ICPR), 2016 23rd International Conference on, pages 432–437. IEEE, 2016m

29. Xiang Zhang. Cyberbullying detection with a pronunciation based convolutional neural network. In 2016 15th EEE International Conference on Machine Learning and Applications (ICMLA), pages 740–745. IEEE, 2016.

30. Zhao. Automatic detection of cyberbullying on social networks based on bullying features. In Proceedings of the 17th international conference on distributed computing and networking, page 43. ACM, 2016

31. Chen, Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), pages 71–80. IEEE, 2012

32. V. Banerjee, "Detection of Cyberbullying Using Deep Neural Network," 2019 5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019, pp. 604–607, 2019, doi: 10.1109/ICACCS.2019.8728378.

33. . Raisi "Weakly supervised cyberbullying detection using co-trained ensembles of embedding models," in 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2018, pp. 479–486.

34. Reynolds K(2011) Using machine learning to detect cyberbullying. In 10th International Conference on Machine learning and applications and workshops, vol. 2, pp. 241–244, IEEE