CSAD-CYBER SECURITY HYBRID LEARNING MODELS FOR DETECTING AGGRESSION IN SOCIAL MEDIA

Raja Ram S[1]
[1]Research Scholar, Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India-627012.
E-mail: er.rajaram@gmail.com

Dr. B. Balakumar [2]
[2]Assistant Professor, Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India-627012
E-mail: balakumarcite@msuniv.ac.in

Dr. Parasuraman Kumar[3]
[3]Assistant Professor, Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India-627012
E-mail: kumarcite@gmail.com

Dr. M. Fathu Nisha[4]
[4]Associate Professor, Department of Electronics and Communication Engineering, Rathinam Technical Campus, Coimbatore, Tamilnadu, India-641021
E-mail: nishahameed2016@gmail.com

Dr. Chokka Anuradha[5]
[5]Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India-522 302
E-mail: dranuradha@kluniversity.in

Abstract:

Cyber security data aggression detection (CSAD) is the most important in social media. Unquestionably, people rely too heavily on social media to communicate effectively. However, there is no suitable restriction on who can participate in communication. Therefore, anonymous senders of irrelevant and occasionally hostile messages undermine the core goal of successful communication. Its influence on society grows along with its popularity, shifting from largely positive to negative. Online aggression, which referred to as the deliberate information use technology to harass, threat, defamation, or otherwise damage a further party, has a negative effect. There is a growing need for automatic filters to find and delete these undesired messages due to the volume of tweets, text messages, and other forms of rewets. However, the majority of current approaches only take into

account NLP-based feature extractors, such as TF-IDF and Word2Vec, without taking into account emotional aspects, making them less efficient for detecting cyber violence. In this study, we retrieved eight new emotional variables to recognize hostile remarks using the newly created a three-layer deep neural network. Dataset for Cyber-Troll and ASR was used to test the suggested HDNN design. Embedding words and 8 different emotional variables mixed together and fed into to the HDNN significantly increase recognition while maintaining a straightforward and computationally less taxing approach. Our suggested model outperforms the rivals by a wide amount, scoring an F1 score of 97% when compared to the most recent models.

Highlights

- To measure the importance and relevance of unstructured language, which represents the user's subjective importance, we created a novel method in this study.
- We believe that by using our approach, the challenge of carrying out a substantial quantity of unstructured text for situational awareness in cyber safety may be resolved.
- Previously designated pertinent articles using linguistic similarity and correlating the detected entities using the to create a knowledge graph for cyber security assess the text's subjective significance.
- we developed a Cybersecurity Information Graph containing 215,202 semantic tuples and coached a unique Model of Named Entity Recognition with more than 17 million words
- To provide characteristics that would represent them relevance as well as subjective importance, their significance and personal relevance.

Keywords: Deep learning, processing language naturally, and detection of aggression, ASR

1. OVERVIEW

Internet usage has demonstrated to be a superb device for facilitating public involvement. On via social media sites like Twitter and Facebook, many important discussions take place online. Additionally, the social media primary means that of communication for most individuals, thereby turning the entire world into a global village. But this quick transition to digitalization raises several issues, one of which is cyberbullying. Cyberbullying is the use of social media platforms to harass, demean, threaten, or otherwise bother others. Due to anonymous profiles, it is difficult to reduce cyberbullying. Online bullying is riskier than physical bullying since it causes psychological and emotional illnesses in people. 8% of social media users, many of whom may be bullied, are unaware of cyberbullying, according to [1].

Different sorts of cyber bullying, including sexism, racism, and cyberaggression, can be categorised. [2] defines cyber aggression as hostile or violent behaviour towards people on digital media. Hatred is thought to be the fundamental driving force for cyber aggression [3][1]. Hatred is based on racial or ethnic origin or religion, colour, other, and gender reasons. Regardless of age, online bullying can result in problems that are identical for everyone. A recent study [4] found that about 3.3 million Facebook postings and 4.5 million tweets are made per minute. These numbers, however, are rising daily. You can notice these are massive numbers of Facebook postings or tweets frequently contain inappropriate content by analysing and gathering feedback on specific topics, goods or famous people.

[5] asserts that cyber bullying poses a severe risk to consumers of social media since its victims are more prone to developing issues including lack of confidence, nervousness, and fear, rage, both even suicidal ideation. One out of three teenagers experience online threats, and 25% of internet users report being bullied, according to statistics gathered by [6]. Additionally, 36% of these social media users will have experienced bullying by the year 2020, according to the 145 million active users who tweet on Twitter every day about a variety of issues [7]. Although it is challenging to avoid online abuse on social media, there are some clever ideas that could help. Models for deep learning and machine learning are obvious when the needs for automatic recognition of cyberbullying contents are taken into consideration.

The detection of cyber violence is an essential challenge for processing of natural language (NLP). Tokenizing the data and preprocessing it to remove any unnecessary text are also crucial steps. The data is then processed through feature extraction and selection to get it ready for deep learning or machine learning models. There has been a lot of study utilising traditional computer learning models. We employed conventional machine learning algorithms in the study in [8] to perform text classification. As classification models, they employed the naïve Bayes (NB), the support vector machine (SVM), and logistic regression (LR). The computer learning models were also subjected to a comparative study by the authors of [9]. Naive Bayes and SVM were employed in a different study [10] to find instances of cyberbullying in Arabic literature. The drawback of [11] is that they only applied simple machine learning models and textual characteristics. However, using conventional (TF-IDF) term frequency inverse document frequency feature extraction methods), such as Word2Vec, it is challenging to make a substantial gains in detecting cyber assault. Even without vocal tones and facial expressions, recognising emotions is still a difficult process. Because of the low Twitter text size and the prevalence of online slang and humour, it might be challenging to infer emotions from textual data. Strong models used in deep learning, like the long short-term memory (LSTM), recurrent neural network (RNN), and convolutional neural network (CNN), multilayer perceptron (MLP), and dense neural network (HDNN), are used to solve this challenge. Some variations of these networks do better in text categorization because Using deep learning, you can capacity learn deeply complex material characteristics [11].

Deep learning models were the main tool we employed in our study to identify cyber violence. In Section 3, which illustrates the whole flow of the suggested technique, we will provide our framework for the proposed model. The framework for detecting violence is built using a combination of Word2Vec features and unique emotional features.

These are the contributions:

1. To test their contributions to the identification of hostile comments, the suggested model yields eight fresh emotions variables based on the text tweets.
2. Compared to its rivals, the HDNN model uses fewer (perhaps less) layers and is optimised with fine-tuned parameters to detect hostility.

3. Using multiple assessment metrics, In the suggested HDNN model, contrasted using the current models that are modern.

4. For improved performance, a hybridised model that involves manually selecting features, feeding these features, and feature extraction characteristics the use of a neural network tweet behaviour evaluation is given.

2. WORK PROCESS

study in the area of NLP is now possible because to the'social web''s quick development. Researchers have made using a a multitude several techniques, including deep learning, Natural language processing (NLP), machine learning, and determine the polarity of the opinions voiced on different social media platforms, taking into account that social media is an active field of study. platforms for social media like Twitter, weblogs, Facebook, as well others have evolved into multilingual spaces where users of various languages, ethnicities, and cultures leave daily comments on a variety of subjects, including photographs, celebrities, and other issues. As a result, the variety of individuals on social media may result in threads containing bullying, hostility, and hate speech. The intricacy of other languages makes it difficult to manage this kind of online behavior.

Researchers have used terms including bullying online [12] [13] [14], inflammatory profanity [15], racism [16], hate speech [17] [18] [19], and language [20] [21] to describe violence. [22] compares from online attackers to passive users, and examines whether a passive person seeks psychological help more frequently versus a cyber aggressor. A Twitter dataset was utilised by the authors of [23] to identify hostility Using user-based, network, and text-based aspects. Additionally, they hypothesised that attackers are often participate in internet forums with the desire to spread negative language, whereas bullied individuals generally Publish fewer posts and engage in fewer discussions. The Lexicon-based technique was employed in a different study [24] to recognise hate speech. They performed a subjective analysis using sentiment lexicon and received an F1 score of 70%. The authors of [25] demonstrated how selecting useful text elements, like verbs, nouns,

and adjectives, a variety of Besides others, emoticons, might enhance the effectiveness of abusive language identification.

The majority of computational linguistics research frequently concentrates on languages with a wealth of resources, like English. However, due to a lacking resource, including ready-made datasets, academics have given resource-poor languages less attention. To identify objectionable social media posts' language in a variety language group, different regional academics have employed machine learning algorithms. For instance, the authors of [26]] discussed their advances in the field of violence detection in the regional Indian languages of Marathi and Hindi. They combined bag-of-words (BOW) with SGD, or stochastic gradient descent, and linear regression (categorization using LR) methods. They were not employing any models for deep learning and only used a short dataset. Similar to this, ref. [8] conducted their studies on Twitter, in Arabic while taking into account the potency of bullying remarks. Using cutting-edge methods like SVM and KNN, [5] Article concentrated based on an examination of user credibility categorise Text messages from Indonesia as bullying or not. Furthermore, ref.[27] used TF-IDF, a single feature, and an easy approach, MLP, to conduct their trials using Twitter data to identify aggressiveness. The elements suggested in this research can further enhance the method described in [27] . a model for machine learning foul language identification text in German was developed in a study by [28]. Another study [29] classified German tweets using several machine learning algorithms to look for objectionable material. In the study [30], a classifier ensemble was utilised to find objectionable text on Portuguese-language websites. Additionally, [31] used an n-gram to identify objectionable words in Arabic text from internet comments. vector of support machines, decision tree and simplistic Bayes containing n-gram characteristics of words in groups were employed in the study by [32] a method to find offensive content on Indonesian social media. Similar tests were been out to identify offensive material in Indonesian by other researchers [33] [34] . The majority of researchers conducted thorough tests on datasets in English to detect cyberbullying. Below, a number of studies are discussed.

The authors of [35]conducted studies to find "bully" in the social network, traces. They employed traces of "bully" on the Twitter dataset to detect using semantic and syntactic aspects for textual information using emojis. They also evaluated their model

using data from other social media platforms, like Form spring and YouTube. In order to recognise sarcasm in tweets, Reference [36] developed a three-part model practical User tagging, hostile and uplifting emoticons are features. To further understand the challenge of sarcasm recognition, they also compared the performance of a system for detecting sarcasm in human speech. In [37], the author conducted experiments to identify irony and sarcasm in a sample of English-language tweets. The TF-IDF vector and emoji approaches converting text to vectors were features. Also, the author covered the connection between irony, Arrogance and online bullying. The writers of [38] gathered English -language tweets about cyberbullying through Twitter, in order to forecast the tweets, they developed an automatic model for detecting cyberbullying identification based on readability, emotion score, and content. Additionally, the writers determined the The text has an emotive tone by counting the instances of unfavorable an emotional tweet and using a vocabulary of curse words. The English dataset was also used in Ref. [17] to discover aggression detection using the Bert model.

New applications NLP applications of deep learning have also gained attention. as robust learning has gained popularity across several application fields. It's been established that some robust learning strategies are more effective than others in machine learning. Experiments were conducted by scientists in [21] to distinguish between offensive language and swearing. They used learning in groups, which had an accuracy rate of 87%. This work can be enhanced by fine-tuning meta classifier hyper parameters. Additionally, reference [39]employed CNN to locate textual data with hate speech. They employed the feature extraction methods Word2Vec and character four-gram. By applying LSTM and BiLSTM to determine the sequential nature of the data, this work can be made better. Character level classification was carried out by ref. [7] using logit-boost and LSTM to learn high level classification. Similar to this, ref. [40] used LSTM, BiLSTM, and GRU to conduct trials on the Bangla text to detect cyberbullying. [41] conducted an empirical investigation to assess how well deep learning fared in identifying cyberbullying on social media. For experiments, RNN, GRU, BiLSTM, LSTM, and are employed.

3. METHODOLOGY

The models for deep learning and machine learning that have accustomed to identify aggression are all thoroughly studied in this section. The collection utilised used in the tests will be described sections 3 before selection and extraction of features are covered correspondingly in Sections 3.4 and 3.5, to start off the technique. After that, we go into detail about the outcomes on page 4. The outcome is finally explained sections 5 in great detail.

3.1. Metrics for Evaluation

Using the average of the F1-score, recall, accuracy, and recall, we assess our models. These metrics' Calculated values included The terms "true positive" (TP), "false positive" (FP), "true negative" (TN), and "false negative" (FN) are used. True positives (TP) what the appropriately labelled cyber-aggressive tweets, whereas false positives (FN) are mislabeled tweets that aren't hostile online. Tweets with accurate information identified as a passive, non-aggressive TN, but those who are mistakenly categorised as identified as such are referred to as FP.

The percentage of accurately identified hostile and non-aggressive tweets is known as accuracy.

Accuracy is defined as

$$((TP + FN + TN + FP)/(TP + TN)) \quad (1)$$

The percentage of tweets correctly classified as cyber -aggressive among all tweets is known as precision.

Precision is equal to TP divided by

$$(TP + FP) \quad (2)$$

That's the remembrance that percentage of hostile tweets in the collection as a whole.

Recall is equal to

$$(TP + FN)/TP \quad (3)$$

Your classifier's A F1 score is gauge of how effectively it balances recall and precision.

the F1-score equal to

$$(2\ (P + R)\ /\ (P + R) \quad (4)$$

3.2. Dataset

The publicly accessible Cyber-Troll dataset (accessed on 9 February 2022 at https://www.kaggle.com/ datasets/dataturks/dataset-for-detection-of-cybertrolls) was utilised in this study. This dataset was produced by Data-Turk with the purpose of detecting hostility. The dataset includes tweets in the English language that have been divided into two categories by the Data-Turk society: both aggressive (CA) and passive (NCA) behaviour online. The messages in aggressive online tweets are meant to offend or harm a person online. The other hand hand, non-cyber-aggressive tweets are those that have no malicious intent and do not hurt other people. As indicated in Table 1, 20,001 tweets total in the sample, 12,179 of which are NCA and 7822 of which are CA. As seen in Figure 1, the NCA accounts for 61% of the data, while CA tweets make up around 39% of the total.

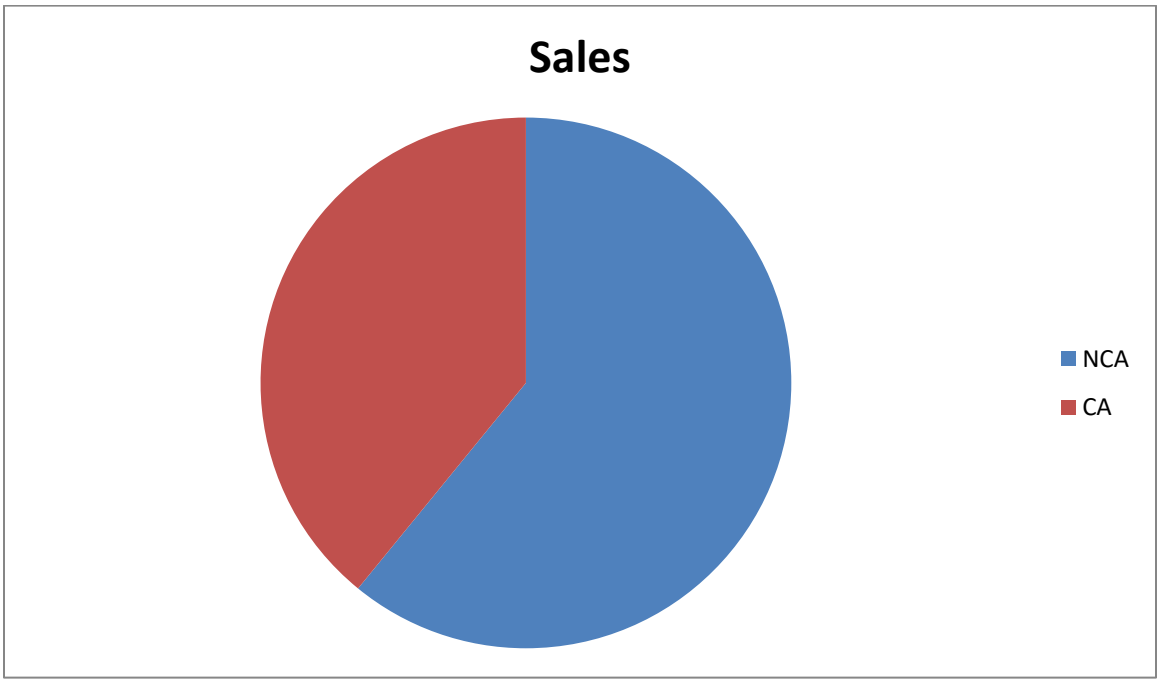

Figure 1 shows the Cyber-Troll dataset's class distribution in a pie chart.

Table1. Dataset on Cyber-Trolls

Table Name	overall tweets	Aggressive	Not combative
Internet Troll	20,001	7822	12,179

3.3. Preprocessing

Preprocessing enhances outcomes Natural language processing is a growing area activity carried out on social media data, according to prior studies. By deleting unnecessary information, valuable features can be extracted from text. This dimensionality reduction speeds up computation and results in a model that is shorter. Our dataset was subjected to these procedures utilising the NLTK (Natural Language Toolkit). Additionally, models were implemented using the programming language Python, the deep learning framework Keras (accessed on 9 February 2022 at https://keras.io), and a Journal for Python.

3.4. Features Extraction

A raw dataset that has not been processed condensed towards a more useful form of future processing using the feature extraction process. Most of the time, through the removal of pointless data, It makes the data smaller in size information, ensuring which the smaller dataset accurately describes the original without dataset losing any important details. Machine learning models' calculation times are shortened through feature extraction. When working with text data, feature extraction must be utilised to get it into vector form. Typically, the text data's feature vector is obtained using TF-IDF and Word2Vec and ASR. also retrieved fresh emotional features for Word2Vec-based aggressiveness detection. The features that were taken out that are effort below are described.

3.4.1. Discrete separate emotions

Word choice was employed by a team of researchers at the Canadian National Research Council (NRC) [42] [43] [44] to pinpoint specific both happy and sad feelings in

tweets. Eight emotional qualities were gleaned from the NRC: apprehension, happiness, surprise, trust, worry, grief, wrath, and disgust. The investigation showed that fear, grief, sadness, rage, and disdain viewed as distinct unpleasant emotions, but Trust, excitement, delight, and expectation are thought to have favourable, distinct feelings. To the our best knowledge, previous research does not yet applied these innovative aspects to cyberbullying issues. The The following mathematical equation was used to extract all emotional components.

$$F_{x_i} = \frac{(W_{mt} * W_n)}{100} \quad (5)$$

where F xi the psychological evaluation of the particular emotional characteristic X, and i is a member of the set of emotional features, which also includes emotions like expectation, surprise, faith, anxiety, sadness, rage, and disgust. Wmt is the overall word count for a sentence, while Wn is the word count. in the dictionary that match a certain attribute, such as "anger," "sad," etc.

3.4.2. Word2Vec and ASR

Word2Vec is an algorithm [45], in particular, offered new concepts which had a significant influence on processing of natural language. among the most often used methods for text processing is Word2Vec. Due to Word2Vec's word embedding foundation, it produces vectors while working with words. The embedding of Word2Vec method creates the potential for using identify connections among the words with similar meanings when used in conjunction with other words. The TF-IDF embedding method was employed by certain researchers [45] [46] – [47]. Although straightforward and effective, TF-IDF does not adequately represent a number of fundamental language ideas, including synonymy and polysemy [48]. Additionally, it fails to adequately convey the meaning of the phrases [49]. Utilizing Word2Vec is encouraged because it has been prepared for a sizable database and lowers computing difficulty. The concept behind it is that words that are similar to one another tend to have similar properties. On the Skip-gram model, it is based. which was trained using using negative sampling to anticipate words around a particular word centre word. Thus, the text can be visualised in vector space as points. Before being fed to deep learning models, Word2Vec attributes are mixed with emotional features.

3.5. Features Choice

Among the key issues with there are numerous machine learning and deep learning algorithms. over fitting [50]. locating the ideal combined with several attributes of create classifiers effective for resolving a particular issue is the feature selection process. The goal of feature choice is to eliminate additional characteristics that could unnecessary, due to the fact that they are sufficiently modelled other characteristics, or that the client does not require. To select the ideal collection of features from all of the combined features, we employed selecting built on shared knowledge. The mutual information method was thoroughly reviewed by the author in [51]. They also talked about some of Mi's variations. Correlation coefficient mutual information (CCMI)), a brand- new feature choice technique built based on mutual knowledge, was introduced by the author in [52]. The shared knowledge and To determine the relationship between two features, one uses the coefficient of correlation. In relation to the intrusion detection system, [53] combined the correlation coefficient with the mutual information approach. Mutual information, which measures how much information there is acquired around Y by understanding X's value, is an analysis of the relationship using statistics between Features X and the target Y are two variables (which often depict unpredictable variables) [54] . Calculating mutual information involves

$$I(X;Y) = \int_X \int_Y P(x,y) \log \frac{P(x,y)}{P(x)P(y)} dxdy \qquad (6)$$

Here, P(x, y) represents the the probability density functions for the variables X and Y, while P(X), P(y), respectively, stand for Marginal density processes. The resemblance the difference the difference between the joint distribution P(x, y) and the marginal distributions product P(x, y) is quantified by the mutual understanding. P(x,y) is the result of will be P(x, y), and their integration will be P(x, y). zero Should X and Y be distinct from one another). When choosing features, we want to ensure that the desired parameter y is relevant to the top predictive feature subset (XS). By maximising the shared knowledge of the characteristic and we can reach the goal Y achieve the top group of XS characteristics.

$$S = argmax\ I(XS; y), \quad s.t.|S| = k \tag{7}$$

The 25 best embedding features out of all of them were chosen, and they were coupled has eight emotional characteristics being fed to deep studying and machine discovering algorithms.

3.6. Learning Machines Models

K-nearest, using logistic regression, and neighbouring support vector machines (KNN, DT, NB, NB, and GB are examples of decision trees, naïve bayes, and gradient boosting.) were the seven fundamental machine learning methods we employed. In order to train these automated learning techniques, Word2Vec and emotional features were fed to them.

3.7. Models of deep learning

CNN, HDNN, BiLSTM, LSTM, and BiLSTM model deep learning were to create this investigation. We first transformed text input into a vector of numbers it is feedable any robust educational model. With the Word2Vec technique, employed the ability to embed text in order to train the suggested HDNN. Additionally, we manually collected emotional characteristics from the tweets. Both feature types were then mixed and given to the model later. The exact hyper-parameters for each deep neural network model are listed from Table 2.

Table 2 lists the deep learning models' parameters.

Inverse Parameters	HDNN	CNN	LSTM	BiLSTM
LSTM devices	-	-	2	2

occult neurons	-	-	512,256	512,256
thick layers	3(256,128,1)	1 (3)	2 (256,2)	2 (256,2)
optimum pooling	-	4	-	
Operate on the secret Layer	ReLU	ReLU	ReLU	ReLU
Perform work on the final layer	Softmax	Softmax	Softmax	Softmax
Epochs	100	200	200	200
Sample size	128	128	128	128
Optimizer	Adam	Adam	Adam	Adam

SUGGESTED MODEL

As a module for classification, we employed HDNN, which is made up of dense layers that are fully connected. Through numerous experimentation based on trial and error, our HDNN model was improved. In addition, the model's effectiveness was assessed using kNN-fold validation. Once adjusted, HDNN only utilises three substantial layers; The sigmoid activation function is used in the output layer. while with the first two the activation function is layered for linear rectified units (ReLU).

With 128 samples in the batch, we chose dropout 0.2. Experiments were carried out using LSTM and BiLSTM, and cross-binary entropy was used to calculate losses. Each sentence in the dataset is converted into one of two different types of features using the HDNN method. The Word2Vec-embedding model produced one feature set, and the

emotional analysis of the tweet produced the other. The HDNN model was fed both types of characteristics in order to evaluate if The tweet was obnoxious or not. The Formula 1 used in this essay decides whether a sentence is aggressive or not after receiving one sentence at a time for input. various symbols used by the algorithm include: tx: a single tweet, FE: emotional features, Text that has been pre-processed is abbreviated as Pt, or Fm token words, or Tokens. Combination characteristic, or Cf. Additionally, Figure 2 displays the HDNN model's framework.

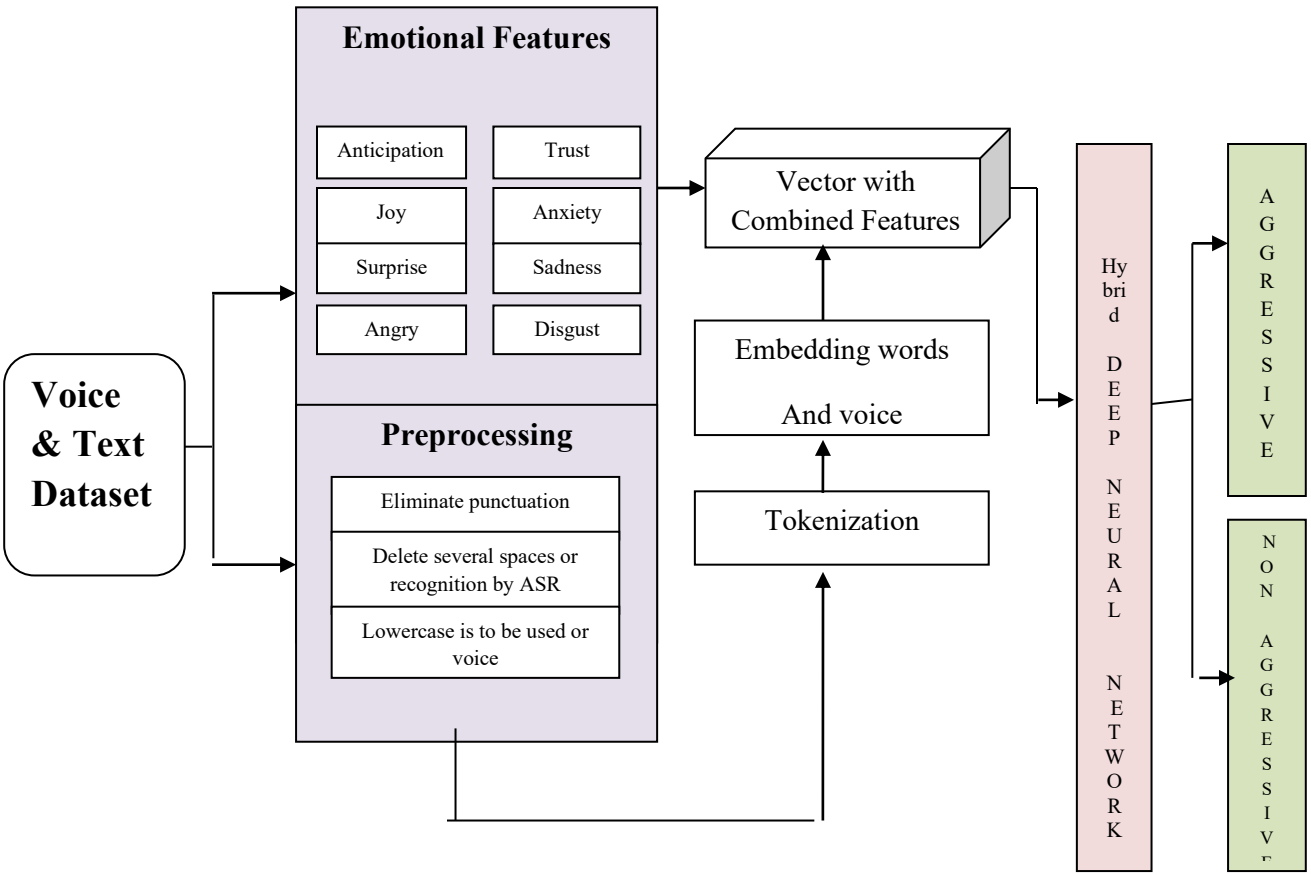

Figure 2 shows the suggested model's framework for detecting cyber violence using HHDNN.

Algorithm 1: Feature engineering phase proposed algorithm
Textual dataset of data Result: Initialization of aggressive or passive text and voice; Do the Emotional Features(fg) for each text and voice. 8 manually derived emotional

characteristics within every tweet tx and voice does
(P t) processing a tweet Tweet preprocessing (tx)
WTokens, or word tokens The Embedding Matrix for P 's (F m) Tokenization
every tweet, embed two WTokens as F m
Feature vector F g + F m combined
Dense layer C f (D1)
(D1)Hybrid Dense layer C f
Layer of density (sigmoid) C f
Whether one is aggressive or not in text or voice

4. RESULTS

The outcomes of applying a number models for the deep learning and machine learning identification of aggressiveness in twitter datasets are shown in this section. Only Word2vec characteristics were used to feed the models, and they were coupled manually with acquired emotive qualities. In the experiments, run, and the prototypes underwent a cross-validation of 10 times process. Following that is contrasting the suggested deep learning model to current best practises. Finally, by using hypothesis testing, the importance of the emotional qualities was proven.

4.1. Word2vec and Automatic speech recognition experimentation

The effectiveness of classification for various traditional when Word2vec features are introduced into machine learning models is shown in Table 3. In terms of performance across all classification metrics, the NB classifier performed the poorest. GB classification methodology, on the other hand, beat all other classifiers and recorded the highest results across the board. The DT classifier came in second place. The GB classifier's highest F1 score was 77%, demonstrating that the traditional machine learning methods could not be relied upon to accurately detect violence.

Table 3 lists the evaluation metrics for the machine learning models that were utilised and fed Word2vec data. Average percentages accustomed to represent the measurements.

Models	Precision	Recall	F1-Score	Accuracy
SVM	67.21 0.23	67.74 0.85	64.24 0.74	62.17 0.36
LR	62.36 0.18	49.24 0.73	44.96 0.19	58.14 0.90
NB	62.75 0.22	49.74 0.69	44.09 0.74	48.27 0.18
KNN	67.36 0.77	67.88 0.37	67.08 0.74	66.23 0.88
GB	77.82 0.58	76.66 0.69	77.10 0.27	76.71 0.05
DT	75.08 0.41	74.47 0.81	74.33 0.38	73.13 0.61
LDA	69.94 0.04	69.94 0.82	68.31 0.36	68.55 0.54

Table 4 displays the capacity models of deep learning for classification developed making use of Word2vec features. In comparison to all other deep learning models, suggested HDNN model obtained the greatest assessment metrics, followed by the BiLSTM. The new HDNN outperforms the GB machine learning F1-score model with 87%. This finding emphasises how Machine learning is outperformed by deep models for learning in the job of detecting hostility. The F1-score values, however, demonstrate Word2vec-driven features' insufficiency for accurate classification of utilising models for both deep learning and machine learning to analyze hostility. As a result, Word2vec and ASR features and manually extracted features were integrated and added within the replicas.

Table 5 lists the evaluation metrics for the models for deep learning that were employed and given the Word2vec and ASR features. Percentages accustomed to represent the measurements.

Models	Precision	Recall	F1-Score	Accuracy
LSTM	80.25 0.24	82.15 0.47	80.65 0.85	82.15 0.15
BiLSTM	83.17 0.74	85.84 0.21	84.76 0.86	85.17 0.48
CNN	81.14 0.47	80.38 0.16	79.96 0.84	82.14 0.37
HDNN	86.28 0.08	87.74 0.19	87.11 0.83	88.34 0.17

Word2vec 4.2 and ASR emotional feature fusion experiments

These are tests, eight sensitive characteristics manually crafted retrieved based on practise data that had been annotated, and they characteristics of Word2vec were merged with classify aggression.

The machine learning classifiers' performance metrics with integrated features are displayed the Table 5. The classifier for GB noted the greatest Metric numbers in contrast compared using more machine learning models, much like in the trials described in Section 4.1. The GB's F1-score with the combined features was 86%, nevertheless. This finding emphasises how important it is to combine emotional information with Word2vec and ASR features in order to enhance classification performance.

Table 6 shows the categorization results supplied with the data from the deep learning classifiers, combined collection of emotive and word2vec characteristics. With the greatest F1-score value of 97%, the suggested HDNNS improved all other deep learning models in terms of performance. Additionally, it was discovered that when the merged feature set was delivered to the classifiers as opposed to the isolated Word2vec features, performance metrics for each deep learning model increased. For the traditional machine learning classifiers, the same finding was observed. This finding indicates that the emotional traits improved the computers' capacity to differentiate between hostile and conciliatory tweets.

Table 5 lists the evaluation metrics for artificial intelligence algorithms that were utilised and fed the combined (Word2vec + emotive) features. Average percentages accustomed to represent the measurements

Models	Precision	Recall	F1-Score	Accuracy
SVM	80.15 0.85	80.69 0.21	74.37 0.96	78.48 0.65
LR	78.17 0.73	59.67 0.09	61.36 0.19	68.48 0.54
NB	82.18 0.38	59.64 0.63	54.19 0.37	56.90 0.71
KNN	79.15 0.81	79.37 0.94	79.16 0.09	76.38 0.73
GB	83.28 0.54	84.08 0.66	84.18 0.75	86.06 0.77
DT	77.17 0.19	78.85 0.84	78.25 0.07	75.37 0.96
LDA	78.27 0.21	78.18 0.29	77.39 0.74	75.69 0.24

.

Figure 3 displays a bar chart showing the F1-scores of all designs created by combining the collection of features so that the effectiveness of machine learning and deep learning models can be compared. Since the data classes were unbalanced and The F1 rating considered recall and accuracy, a models of machine learning and deep learning are compared were taken into consideration. It is evident which the suggested Model HDNN outperformed all more models for machine learning and deep learning when it was trained using the combined features.

Models	Precision	Recall	F1-Score	Accuracy
LSTM	88.41 0.67	86.96 0.75	87.47 0.67	88.14 0.27
BiLSTM	91.21 0.74	90.86 0.41	90.96 0.96	91.43 0.17
CNN	88.75 0.56	87.34 0.63	86.41 0.38	84.74 0.68
HDNN	98. 17 0.49	97. 73 0.93	98. 18 0.29	97. 07 0.29

Table 6 lists the measures for evaluating the deep learning algorithms that were utilised and fed the combined (Word2vec + emotive) features. Average percentages accustomed to represent the measurements.

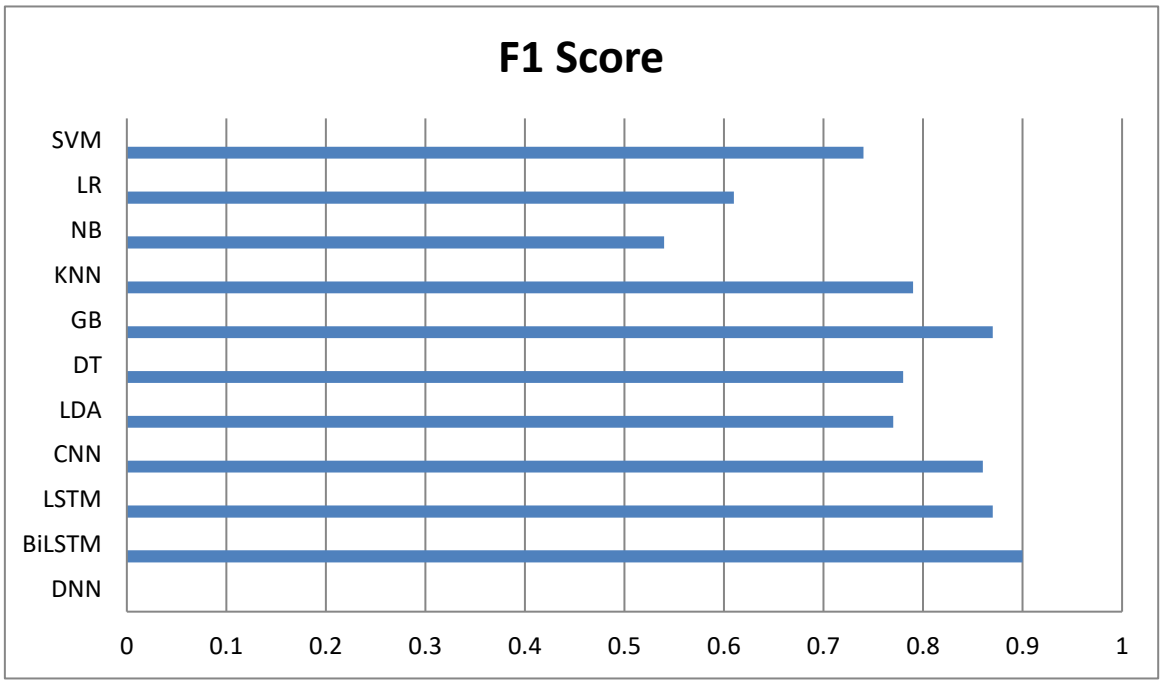

Figure 3 shows a F1 score in a bar graph to support deep learning and machine learning models that were trained with the combined an emotive feature set and word2vec.

This study's key discovery is that adding emotional variables to word embedding features enhances the ability to distinguish between classes that are aggressive and passive in the Twitter dataset for cybertrolls. According to the classification results measurements on the aggregate characteristics, emotional features are crucial for correctly classifying hostile tweets. Additionally, the innovative investigation on the dataset in use demonstrates that adding Word2vec to emotionally derived manually variables enhanced the classifiers' classification accuracy. This result implies that characteristics like wrath, anxiety, disgust, and anticipation are crucial in distinguishing hostile content. A hypothesis was tested done to assess the significance of the psychological traits to be able to support this conclusion.

4.3. Testing Hypotheses and the Importance of the Emotional Features

Table 7 displays the outcomes of the statistical that was analysed done on the emotional characteristics to confirm their importance in the tweets that were gathered. N

(the number of records in the data), the average, the average deviation, highest and lowest, the quarter-one median (Q1)), and the 75th percentile (Q3) are all presented in Table 7's entries as a rundown of the data records' characteristic statistics. The distribution of such properties can be explained by drawing boxplots and histograms, as shown in Figures 4 and 5. The height of each bin in the histogram, which a bar graph of a numerical variable that has been binned, represents the proportion of instances falling into that bin. It is used to get a sense of the distribution of an attribute. the vast bulk of the factors exhibit Gaussian distributions, as seen in Figure 4.

Plot lines are like whiskers show The attribute's allowable range, the median is represented by the line, and centre half of the data points are represented by the box in the boxplot. Any dots that are not within the whiskers are probably outliers. Figure 6 shows the connection between all features and results of the prediction. The anxiety, disgust, and melancholy features show a strong link with the anger feature, however the pleasure feature indicates a strong association with the elements of surprise, anticipation, and trust. Figure 6 shows that there is a low association between the outcome and the emotional feature. How strongly two continuous variables are related to one another is determined by the correlation coefficient. In contrast, the input features are continuous In this research whereas The category response variable (aggressive and non-aggressive). Logistic regression can therefore be used to extract the a connection exists between the reaction and the input features [55]. As an illustration the classification method to forecast a classification answer, consider binomial regression logistic. In this case, the likelihood that a sensitive trait is associated with a particular result (aggressive or non-aggressive) is estimated using logistic regression. an analytical analysis It is based on a binomial Using a logistic regression model, fitted to forecast the result in accordance with emotional feature values in the twitter data records is shown in Table 8. For each coefficient the regression model provided, Table 8 entries give a prediction, Z-score, common deviation, significance, and the p-value. Asterisks indicate the significance level, and more asterisks indicate a coefficient's high significance. The majority of coefficients exhibit a high level of importance, as seen in Table 8, for example. The logistic regression model's first five probabilities are 0.46, 0.37, 0.53, 0.39, and 0.37, which are all near to 50%. A hypothesis testing was done to evaluate the efficacy of the logistic regression analysis, taking into

account the H(0) has B n = 0, which follows B 1 = 0, B 2 = 0, etc is the null hypothesis denotes a null association between the emotional factor and the response variable variables as input capabilities. In the event that any 6 = 0 the null hypothesis, learning model can be regarded as effective. The ANOVA test was used to examine the validity of the hypotheses, and it produced The residual and null are two performance measurements deviation. The void deviation demonstrates the predictability of the response variable using one that only has an intercept phrase. However, the residual deviation demonstrates how the emotional qualities of the input predictors might aid in an accurate the response variable's forecast. A learning model that uses emotional features performs better when the remaining deviation is smaller unlike the null deviation. The numbers 20,888 and 21,407, respectively, for the residual and null deviance, were received. The link contrasting the emotional characteristics and the grouping of the gathered tweets as whether someone is assertive or not is therefore ensured, and the null hypothesis may be rejected.

Table 7: A statistical summary of emotional characteristics

Emotional Aspect	N	Mean	SD	Min	Q1	Q3	Max
Anger	20,001	3.014	5.964	0	0	4.348	75
Anticipation	20,001	1.472	3.708	0	0	0	50
Anxiety	20,001	2.162	4.874	0	0	3.333	100
Disgust	20,001	3.109	6.235	0	0	4.348	100
Joy	20,001	1.532	4.097	0	0	0	100
Sadness	20,001	2.292	5.038	0	0	3.571	100
Surprise	20,001	0.748	2.753	0	0	0	100
Trust	20,001	1.485	3.864	0	0	0	100

A logistic binomial regression employing empathetic qualities is statistically summarized in Table 8.

	Estimate	Typical Error	Value z	Pr(>|z|)	Significance
Intercept	0.5531	0.0205	26.9310	$<2x10^{-16}$	***
Anger	0.0003	0.0057	0.0520	0.9588	
Anticipation	0.0023	0.0059	0.3970	0.6914	
Anxiety	0.0088	0.0066	1.3270	0.1846	
Disgust	0.0474	0.0054	8.7020	$<2x10^{-16}$	***
Joy	0.0380	0.0067	5.7090	$1.13x10^{-8}$	***
Sadness	0.0228	0.0064	3.5660	0.0004	***
Surprise	0.0061	0.0079	0.7690	0.4420	
Trust	0.0140	0.0063	2.2390	0.0252	***

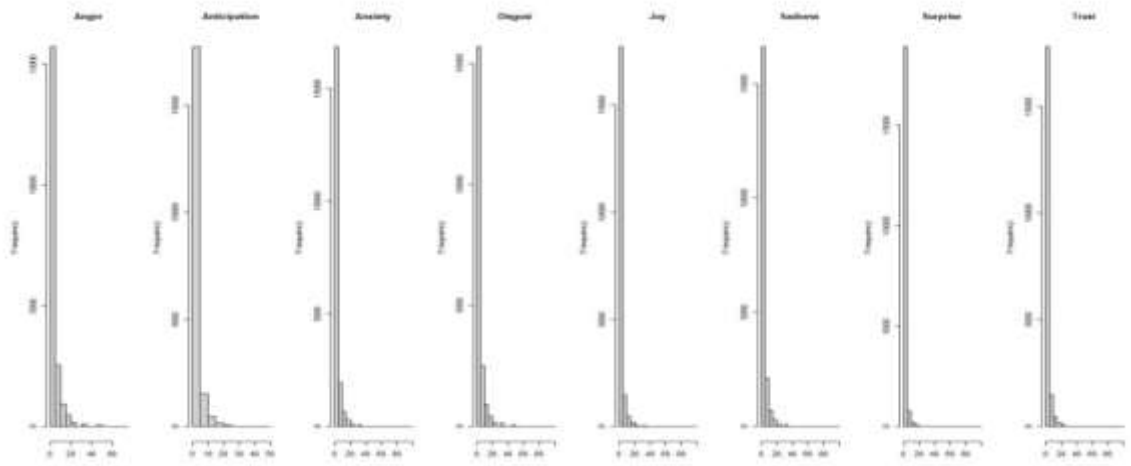

Figure 4. No of times user-specified ranges of the emotional features

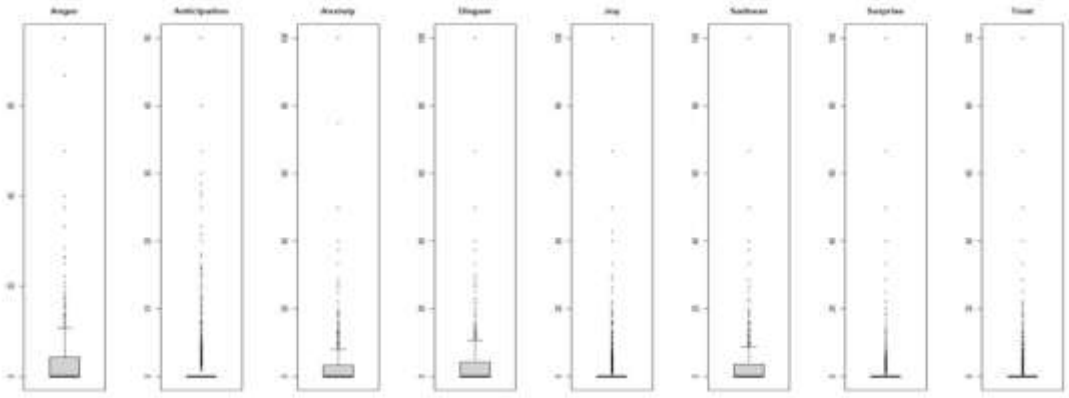

Figure 5. The emotional feature boxplots

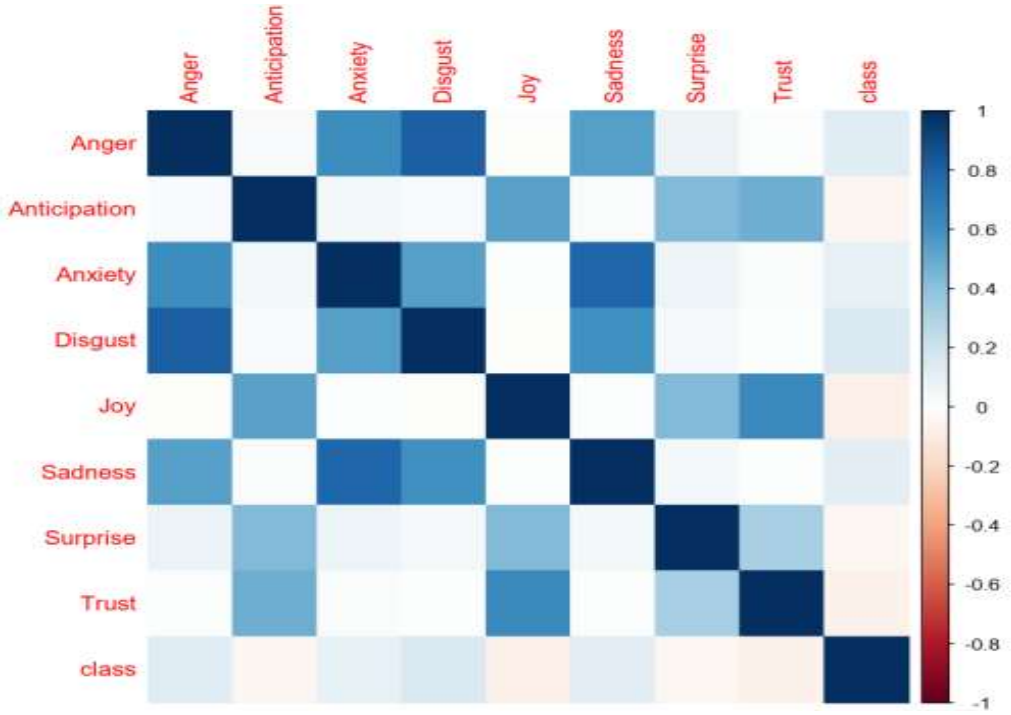

Figure 6.The relationship between emotional characteristics and the outcome predicted

4.4. A comparison of the suggested model with the existing best practices

Two investigations were contrasted with the suggested HDNN model that was on the combined feature set, trained. While the work in [56] made use of a CNN classifier and their own dataset, [30] using the the identical Cyber-Troll dataset as an classifier with MLP. The FD-IDF features were used in both experiments. Table 9's comparison of the two models demonstrates how the suggested HDNN with the combined characteristics beats the most recent models.

Table 9 compares the proposed HDNN model's classification performance to that of the current models.

Author	Dataset	Features	Models	Results
Sadiq Saim (30)	Cyber-Troll	TF-IDF	MLP	F1 = 0.90
57 Chen Junti	Personal data	TF-IDF 2D	CNN	Acc = 0.92
proposed approach	Cyber-Troll	Combined	HDNN	F1 = 0.97

4.5. Result Analysis

Cyber security data aggression detection (CSAD) In the discipline the field of natural language processing, recognising violence from textual material is particularly difficult. The model of the neural network advanced in recent years grown in popularity for problems involving natural language processing. Its capacity for learning has greatly boosted its significance in NLP. these channels require several layers in order to obtain more advanced characteristics and understand dependencies over time; this makes the calculation time-consuming and laborious. The solution utilised in this research was straightforward and effective. Our research shows that the HDNN with less layers can learn practical data representations that may be applied to the detection of violence. The suggested model uses word embedding features and eight emotional factors to create a classifier that can recognise hostile language in text. Utilising the dataset of Cyber-Trolls received an F1 ranking of 97%, we assessed the suggested HDNN model. a number of deep studying and machine educating models were outperformed by the HDNN model. In comparison to earlier approaches, the suggested framework produced superior outcomes with shorter practise sessions and fewer potential layers. the present idea might expand this specific ability to take into account a multilingual environment that is common in many nations that are not English-speaking ones. We intend to test the detection of hostility with larger datasets in the future. We also want to expand the databases to include social networking sites like Facebook and languages like Arabic, Urdu, Hindi, etc. This will be useful in the home, but as more people use it, there could be instances where individuals are unfairly stalking allegations or experiences because of disagreements among friends. By spotting these instances, social network studies of violence detection may be improved.

5. Comparison of the suggested model with the existing best practises

Two investigations were contrasted with the suggested HDNN model that was on the combined feature set, trained. While the work in [57] made use of a CNN classifier and their own dataset, [30] using the the identical Cyber-Troll dataset as an classifier with MLP.

The FD-IDF features were used in both experiments. Table 9's comparison of the two models demonstrates how the suggested HDNN with the combined characteristics beats the most recent models.

Table 9 compares the proposed HDNN model's classification performance to that of the current models.

Author	Dataset	Features	Models	Results
Sadiq Saim (30)	Cyber-Troll	TF-IDF	MLP	F1 = 0.90
57 Chen Junti	Personal data	TF-IDF 2D	CNN	Acc = 0.92
proposed approach	Cyber-Troll	Combined	HDNN	F1 = 0.97

6. CONCLUSIONS:

Cyber security data aggression detection (CSAD) In the discipline the field of natural language processing, recognising violence from textual material is particularly difficult. The model of the neural network advanced in recent years grown in popularity for problems involving natural language processing. Its capacity for learning has greatly boosted its significance in NLP. these channels require several layers in order to obtain more advanced characteristics and understand dependencies over time; this makes the calculation time-consuming and laborious. The solution utilised in this research was straightforward and effective. Our research shows that the HDNN with less layers can learn practical data representations that may be applied to the detection of violence. The suggested model uses word embedding features and eight emotional factors to create a classifier that can recognise hostile language in text. Utilising the dataset of Cyber-Trolls received an F1 ranking of 97%, we assessed the suggested HDNN model. a number of deep studying and machine educating models were outperformed by the HDNN model. In comparison to earlier approaches, the suggested framework produced superior outcomes with shorter practise sessions and fewer potential layers. the present idea might expand this specific ability to take into account a multilingual environment that is common in many nations that are not English-speaking ones. We intend to test the detection of hostility with larger datasets in the future. We also want to expand the databases to include social networking

sites like Facebook and languages like Arabic, Urdu, Hindi, etc. This will be useful in the home, but as more people use it, there could be instances where individuals are unfairly stalking allegations or experiences because of disagreements among friends. By spotting these instances, social network studies of violence detection may be improved.

Funding Statement: There is no Research Funding Program for this article. **Conflicts of Interest**: The authors declare that they have no conflicts of interest to report regarding the present study.

Availability of data and material: Executing Encrypted Data in the Database-Service Provider Model

Code availability: Using the python code we are develop the demo

Declaration:

We certify that all named authors have read, reviewed, and approved the paper and that no other individuals who meet the requirements for authorship but are not listed have contributed to the work. We also reaffirm that we all approved of the order in which the authors are listed in the manuscript. We certify that we have taken all necessary precautions to preserve the intellectual property connected to this work and that there are no intellectual property-related barriers to publishing, including the date of release. By completing this, we attest to having complied with our institutions' intellectual property rules.

REFERENCES

[1] H. Nurrahmi and D. Nurjanah, “Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility,” in *2018 International Conference on Information and Communications Technology (ICOIACT)*, Mar. 2018, pp. 543–548. doi: 10.1109/ICOIACT.2018.8350758.

[2] H. Machackova, “Bystander reactions to cyberbullying and cyberaggression: individual, contextual, and social factors,” *Curr. Opin. Psychol.*, vol. 36, pp. 130–134, Dec. 2020, doi: 10.1016/j.copsyc.2020.06.003.

[3] O. Oriola and E. Kotzé, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," *IEEE Access*, vol. 8, pp. 21496–21509, 2020, doi: 10.1109/ACCESS.2020.2968173.

[4] "Visualizing Eight Years Of Twitter's Evolution: 2012-2019 – The GDELT Project." https://blog.gdeltproject.org/visualizing-eight-years-of-twitters-evolution-2012-2019/ (accessed Aug. 22, 2023).

[5] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Prediction of cyberbullying incidents in a media-based social network," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2016, pp. 186–192. doi: 10.1109/ASONAM.2016.7752233.

[6] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artif. Intell. Rev.*, vol. 52, no. 3, pp. 1495–1545, Oct. 2019, doi: 10.1007/s10462-017-9599-6.

[7] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alex. Eng. J.*, vol. 60, no. 3, pp. 3401–3409, Jun. 2021, doi: 10.1016/j.aej.2021.02.009.

[8] B. Haidar, C. Maroun, and A. Serhrouchni, "A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 2, pp. 275–284, Dec. 2017, doi: 10.25046/aj020634.

[9] M. Khairy, T. M. Mahmoud, and T. Abd-El-Hafeez, "Automatic Detection of Cyberbullying and Abusive Language in Arabic Content on Social Networks: A Survey," *Procedia Comput. Sci.*, vol. 189, pp. 156–166, Jan. 2021, doi: 10.1016/j.procs.2021.05.080.

[10] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural Language Processing Advancements By Deep Learning: A Survey." arXiv, Feb. 27, 2021. doi: 10.48550/arXiv.2003.01200.

[11] I. Lauriola, A. Lavelli, and F. Aiolli, "An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools," *Neurocomputing*, vol. 470, pp. 443–456, Jan. 2022, doi: 10.1016/j.neucom.2021.05.103.

[12] C. Van Hee *et al.*, "Detection and fine-grained classification of cyberbullying events," Sep. 2015.

[13] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, Sep. 2012, pp. 71–80. doi: 10.1109/SocialCom-PASSAT.2012.55.

[14] 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.

[15] S. Malmasi and M. Zampieri, "Challenges in Discriminating Profanity from Hate Speech." arXiv, Mar. 14, 2018. doi: 10.48550/arXiv.1803.05495.

[16] A. Alotaibi and M. H. Abul Hasanat, "Racism Detection in Twitter Using Deep Learning and Text Mining Techniques for the Arabic Language," in *2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, Nov. 2020, pp. 161–164. doi: 10.1109/SMART-TECH49988.2020.00047.

[17] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Comparing pre-trained language models for Spanish hate speech detection," *Expert Syst. Appl.*, vol. 166, p. 114120, Mar. 2021, doi: 10.1016/j.eswa.2020.114120.

[18] G. B. Herwanto, A. Maulida Ningtyas, K. E. Nugraha, and I. Nyoman Prayana Trisna, "Hate Speech and Abusive Language Classification using fastText," in *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Dec. 2019, pp. 69–72. doi: 10.1109/ISRITI48646.2019.9034560.

[19] 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, May 2021, doi: 10.1016/j.ipm.2021.102524.

[20] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, and M. T. Sadiq, "Automatic Detection of Offensive Language for Urdu and Roman Urdu," *IEEE Access*, vol. 8, pp. 91213–91226, 2020, doi: 10.1109/ACCESS.2020.2994950.

[21] R. Kumar, B. Lahiri, and A. Kr. Ojha, "Aggressive and Offensive Language Identification in Hindi, Bangla, and English: A Comparative Study," *SN Comput. Sci.*, vol. 2, no. 1, p. 26, Jan. 2021, doi: 10.1007/s42979-020-00414-6.

[22] M. Garaigordobil, J. P. Mollo-Torrico, J. M. Machimbarrena, and D. Páez, "Cyberaggression in Adolescents of Bolivia: Connection with Psychopathological

Symptoms, Adaptive and Predictor Variables," *Int. J. Environ. Res. Public. Health*, vol. 17, no. 3, p. 1022, Feb. 2020, doi: 10.3390/ijerph17031022.

[23] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean Birds: Detecting Aggression and Bullying on Twitter," in *Proceedings of the 2017 ACM on Web Science Conference*, in WebSci '17. New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 13–22. doi: 10.1145/3091478.3091487.

[24] D. Njagi, Z. Zuping, D. Hanyurwimfura, and J. Long, "A Lexicon-based Approach for Hate Speech Detection," *Int. J. Multimed. Ubiquitous Eng.*, vol. 10, pp. 215–230, Apr. 2015, doi: 10.14257/ijmue.2015.10.4.21.

[25] D.-S. Zois, A. Kapodistria, M. Yao, and C. Chelmis, "Optimal Online Cyberbullying Detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 2017–2021. doi: 10.1109/ICASSP.2018.8462092.

[26] R. Pawar and R. R. Raje, "Multilingual Cyberbullying Detection System," in *2019 IEEE International Conference on Electro Information Technology (EIT)*, May 2019, pp. 040–044. doi: 10.1109/EIT.2019.8833846.

[27] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B.-W. On, "Aggression detection through deep neural model on Twitter," *Future Gener. Comput. Syst.*, vol. 114, pp. 120–129, Jan. 2021, doi: 10.1016/j.future.2020.07.050.

[28] G. I. Sigurbergsson and L. Derczynski, "Offensive Language and Hate Speech Detection for Danish." arXiv, Mar. 23, 2023. doi: 10.48550/arXiv.1908.04531.

[29] J. Schneider, R. Roller, P. Bourgonje, S. Hegele, and G. Rehm, "Towards the Automatic Classification of Offensive Language and Related Phenomena in German Tweets," 2018. Accessed: Sep. 13, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Towards-the-Automatic-Classification-of-Offensive-Schneider-Roller/8602ea0fcfd0e04944b93d5b01f4155415f7a3da

[30] R. Pelle, C. Alcântara, and V. P. Moreira, "A Classifier Ensemble for Offensive Text Detection," in *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, in WebMedia '18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 237–243. doi: 10.1145/3243082.3243111.

[31] B. Haidar, C. Maroun, and A. Serhrouchni, "A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 2, pp. 275–284, Dec. 2017, doi: 10.25046/aj020634.

[32] M. O. Ibrohim and I. Budi, "A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media," *Procedia Comput. Sci.*, vol. 135, pp. 222–229, Jan. 2018, doi: 10.1016/j.procs.2018.08.169.

[33] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," *Proc. Third Workshop Abus. Lang. Online*, pp. 46–57, 2019, doi: 10.18653/v1/W19-3506.

[34] "Abusive Language Detection on Indonesian Online News Comments | IEEE Conference Publication | IEEE Xplore." https://ieeexplore.ieee.org/document/9034620 (accessed Sep. 14, 2023).

[35] M. Di Capua, E. Di Nardo, and A. Petrosino, "Unsupervised cyber bullying detection in social networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec. 2016, pp. 432–437. doi: 10.1109/ICPR.2016.7899672.

[36] R. González-Ibáñez, S. Muresan, and N. Wacholder, "Identifying Sarcasm in Twitter: A Closer Look," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 581–586. Accessed: Sep. 14, 2023. [Online]. Available: https://aclanthology.org/P11-2102

[37] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, "Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection," *Inf. Process. Manag.*, vol. 58, no. 4, p. 102600, Jul. 2021, doi: 10.1016/j.ipm.2021.102600.

[38] P. J. Lee, Y. H. Hu, K. Chen, J. M. Tarn, and L. E. Chen, "Cyberbullying detection on social network services," in *Proceedings of the 22nd Pacific Asia Conference on Information Systems - Opportunities and Challenges for the Digitized Society: Are We Ready?, PACIS 2018*, Association for Information Systems, 2018. Accessed: Sep. 14, 2023. [Online]. Available: https://scholars.ncu.edu.tw/en/publications/cyberbullying-detection-on-social-network-services

[39] M. Al-Ajlan and M. Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning," Apr. 2018, pp. 1–5. doi: 10.1109/NCG.2018.8593146.

[40] "(PDF) Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study." https://www.researchgate.net/publication/329813815_Cyberbullying_Detection_in_Social_Networks_Using_Deep_Learning_Based_Models_A_Reproducibility_Study (accessed Sep. 14, 2023).

[41] M. S. I. Malik and A. Hussain, "Helpfulness of product reviews as a function of discrete positive and negative emotions," *Comput. Hum. Behav.*, vol. 73, pp. 290–302, Aug. 2017, doi: 10.1016/j.chb.2017.03.053.

[42] "The Biology of Emotions | Introduction to Psychology." https://courses.lumenlearning.com/waymaker-psychology/chapter/the-biology-of-emotions/ (accessed Sep. 14, 2023).

[43] "NRC Emotion Lexicon." https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (accessed Sep. 14, 2023).

[44] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality." arXiv, Oct. 16, 2013. doi: 10.48550/arXiv.1310.4546.

[45] N. Kulkarni and E. Al, "A comparative study of Word Embedding Techniques to extract features from Text," *Turk. J. Comput. Math. Educ. TURCOMAT*, vol. 12, no. 12, Art. no. 12, May 2021.

[46] 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, Aug. 2020, doi: 10.1007/s00500-019-04550-x.

[47] M. Khairy, T. M. Mahmoud, A. Omar, and T. Abd El-Hafeez, "Comparative performance of ensemble machine learning for Arabic cyberbullying and offensive language detection," *Lang. Resour. Eval.*, Aug. 2023, doi: 10.1007/s10579-023-09683-y.

[48] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Gener. Comput. Syst.*, vol. 117, pp. 47–58, Apr. 2021, doi: 10.1016/j.future.2020.11.022.

[49] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.

[50] "(PDF) A systematic review on overfitting control in shallow and deep neural networks | Mehdi Ghatee - Academia.edu." https://www.academia.edu/86352421/A_systematic_review_on_overfitting_control_in_shallow_and_deep_neural_networks (accessed Sep. 14, 2023).

[51] J. Vergara and P. Estevez, "A Review of Feature Selection Methods Based on Mutual Information," *Neural Comput. Appl.*, vol. 24, Jan. 2014, doi: 10.1007/s00521-013-1368-0.

[52] H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient," *Appl. Intell.*, vol. 52, no. 5, pp. 5457–5474, Mar. 2022, doi: 10.1007/s10489-021-02524-x.

[53] F. Amiri, M. Rezaei Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," *J. Netw. Comput. Appl.*, vol. 34, no. 4, pp. 1184–1199, Jul. 2011, doi: 10.1016/j.jnca.2011.01.002.

[54] J. Che, Y. Yang, L. Li, X. Bai, S. Zhang, and C. Deng, "Maximum relevance minimum common redundancy feature selection for nonlinear data," *Inf. Sci.*, vol. 409–410, pp. 68–86, Oct. 2017, doi: 10.1016/j.ins.2017.05.013.

[55] R. Aggarwal and P. Ranganathan, "Common pitfalls in statistical analysis: Linear regression analysis," *Perspect. Clin. Res.*, vol. 8, no. 2, pp. 100–102, 2017, doi: 10.4103/2229-3485.203040.

[56] "(PDF) Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis." https://www.researchgate.net/publication/323882643_Verbal_aggression_detection_on_Twitter_comments_convolutional_neural_network_for_short-text_sentiment_analysis (accessed Sep. 14, 2023).

[57] "Deployment of Machine Learning and Deep Learning Algorithms in Detecting Cyberbullying in Bangla and Romanized Bangla text: A Comparative Study | IEEE Conference Publication | IEEE Xplore." https://ieeexplore.ieee.org/document/9392608 (accessed Sep. 14, 2023).

Author's Biography

S. Rajaram received B.E Degree in Electronics and Communication Engineering from Anna University, Chennai India in 2006, He finished M.Tech degree in Computer and Information Technology from Manonmaniam Sundaranar University in 2011. Currently doing research in cyber security system.

Dr. B. Balakumar received his B.Tech degree in Electronics and Communication Engineering from National Institute of Technology (NIT), Hamirpur, Himachal Pradesh, India in the year 2003.He obtained his M.Tech degree in the field of Computer and Information Technology from Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India in the year 2006. He received his Doctor of Philosophy (Ph.D) in Computer and Information Technology in the area of Medical Image Processing from Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India in the year 2019. Presently he is working as a Assistant Professor (Grade III) in Centre for Information Technology & Engineering, Manonmaniam Sundaranar University, Tirunelveli, India. He has published many research papers in International Journals and Conferences. He has reviewed articles in International Conference Proceedings and Journals. He is a Regional Editor and Reviewer of International Journals. He is a Member of many Medical Imaging Journals including BMC journals, SAGE Journals and Journal of the National Cancer Institute. He is also a Life Member of Professional Bodies like ISTE, CSI etc. His area of interest includes Digital Image Processing, Cloud Computing, Computer Networks, Cyber Security, Data Science & Analytics, Etc. He is guiding the Ph.D. Research Scholars in the areas of Digital Image Processing, Cloud computing, Computer Networks, Cyber Security, Big Data Science & Analytics, Robotics etc.

Dr. Parasuraman Kumar received the Master of Technology (M.Tech.) degree in Computer and Information Technology in 2008 and the Doctor of Philosophy (Ph.D.) in Information Technology-Computer Science and Engineering in 2012 from Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, Tamil Nadu, India and Master of Business Administration (M.B.A.) degree in Systems from Alagappa University, Karaikudi, Tamil Nadu,

Dr. M. Fathu Nisha working as Associate Professor in the Department of Electronics and Communication Engineering, Rathinam Technical Campus, Coimbatore, having 18 years of Teaching experience. Completed Ph.D in Anna University, Chennai with specialization of Digital Image Processing and Completed M.E in Communication systems and B.E in Electronics and Communication Engineering. Research interest is Textile Image Processing, Medical image Processing and Machine learning.

Dr. Chokka Anuradha, Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur District, Andhra Pradesh