IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

REVIEW OF AUTOMATIC DETECTION OF HATE SPEECH FROM TEXT USING MACHINE LEARNING AND DEEP LEARNING CLASSIFICATION TECHNIQUES

[1] Prof.Babita V.Kasar,[2]Prof. Dipti D. Mehare,[3]Prof. KavitaK. Nagariya, [4]Prof. Deepali R. Gadbail

[1]Assistant professor,[2]Assistant professor, [3]Assistant professor, [4]Assistant professor,
1Department of Computer Science & Engineering
[1]P. R. PotePatil College of Engineering and Management, Amravati,Maharashtra

Abstract:Cyber environments have been prone to unethical activity in the past few years. Social media platforms are becoming more and more prone to the identification of hate speech (HS) in modern culture. For online platforms to eliminate hate speech content in real-time they need powerful hate speech identification algorithms. A range of techniques exists to detect HS, including ML, Dictionary, DL, and so forth. The purpose of this study is to demonstrate an extensive understanding of DL and ML classification techniques, with their advantages and applications.

Index Terms **- Hate Speech (HS), HS Techniques: Machine learning (ML) and Deep Learning (DL), Comparison of ML and DL, Advantages, and Application of Classification Techniques.**

I. INTRODUCTION

A social media account allows you to express yourself openly and communicate with them. Sentiment analysis of content from social media platforms is a significant area of natural language processing and is necessary for many applications.Human life has become increasingly reliant on the internet. Thousands of persons are troubled by others, often through hate speech in politics, racism, and other forms of discrimination. A Pew Research Centre Organization study conducted among 4,248 US adults in 2017 found that 41% of Internet handlers have experienced some type of online harassment, and 66% reported witnessing this behavior in others. A majority of Americans, i.e., 18% of handlers, have been subjected to serious harassment like threats of violence, sexual harassment, or stalking [1,2].Occurrences like these are typically caused by the sharing of unsuitable photos, insulting comments, and posts. Researchers have focused their attention on identifying these kinds of contents and removing them for years. The level of automation of this task has grown in popularity in recent years, along with a heightened significance in the detection of online HS.

II. BACKGROUND

Researchers have been concentrating on tackling the challenges associated with identifying and minimizing harmful content, which has led to a significant focus on hate speech detection in social media. Many studies have explored various machine-learning and deep-learning algorithms to improve the efficiency of HS detection.

Hate Speech Analysis Process

Step 1: Documents are input from the social networking site (SNS)

It is possible to collect information from Internet Social Networks using the Application Programming Interface (API).

Step 2: Process of pre-processing text

As part of pre-processing, eliminating undesirable punctuation, stop words are removed, stemming is performed, tagging parts of speech, and the score is calculated.

Step 3: Indexing of documents

Determine which opinion words that most accurately characterize the setting in which we are working.

Step 4: The Selection of Features

Data is automatically selected for the attributes most relevant to work by an automatic process.

Step 5: The algorithm for classification

HS detection techniques are classified based on different approaches, including machine learning, deep learning, and others. It is common to classify a grouping of particular features for a certain goal, and then test it on another to see if it can identify or extract the correct features and provide the right classification.

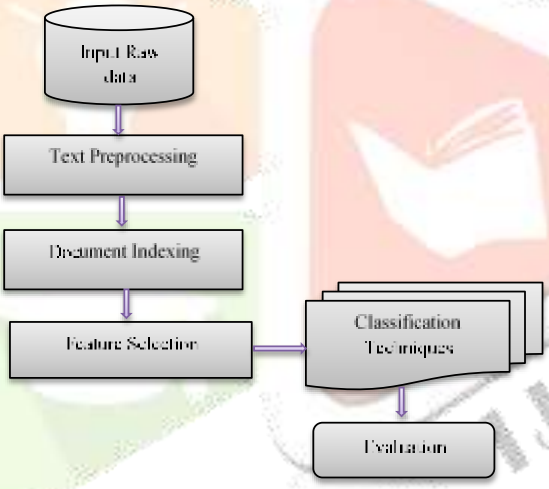


Figure 1: Hate Speech Analysis Process

This research presents various Machine Learning (ML) and Deep Learning (DL) based techniques for HS content identification from text. This paperis arranged as follows,

Section 1. Introduction. **Section 2.** Discussion of Background. **Section 3.** Discussion of Classification Techniques. **Section 4.** The comparison between classification techniques. **Section 5.** The advantages of these techniques. **Section 6.** Applications. **Section 7.** Concluding this paper.

III. HS CLASSIFICATION TECHNIQUES

Finding and removing hate speech posts or comments manually is time-consuming and computationally expensive. Due to these challenges and the increasing amount of harmful content on social media, there is a strong argument for automating the detection of hate speech. It has become common for Natural Language Processing (NLP) to develop systems that can automatically identify abusive and offensive language.

HS detection can be carried out using many different methods, some of which are:

- ML approach
- DL approach

1. ML APPROACH

An ML algorithm uses training data to build a mathematical model that can make predictions or decisions without the need for explicit programming. ML involves learning a classifier model by gathering training data and then evaluating the model's performance using a test data set[3].

Types of machine learning can be categorized as:

- Unsupervised machine learning
- Supervised machine learning

1.1 Unsupervised ML

Rodriguez et al. (2019) proposed a method for identifying hate speech content on Facebook via sentiment analysis. The Facebook post and comments were removed using Graph API. VADER and JAMMIN were employed to eliminate the irrelevant texts. TFIDF was used to transform pre-processed documents into vectors [4]. The generated matrix is used as an input matrix for the k-means clustering algorithm. Based on sentiment and emotion analysis, responses and articles with the highest negative sentiments and emotions were identified.Sylvia Jaki et al. (2019) demonstrated an approach to detect hate speech content using unsupervised learning on Twitter [5]. With Twitter API, they gathered over 50,00 data sets using the NLP method to classify words into clusters. By clustering spherical k-means and skip-gramming., they grouped the top 250 most biased terms into three clusters.Consequently, they received an F1 score of 84.21%.

Michele Di Capua et al. (2019) outlined a method for detecting cyberbullying using unsupervised methods. A total of 54,000 data sets were gathered from YouTube and all were manually annotated. With the SOM-Toolbox-2 platform, the GHSOM network algorithm has been implemented. They used the K-fold approach, with K = 10, to train and test GHSOM. Hence, they have reached 64% accuracy [6].

1.2 Supervised ML

P. Sari 2019, proposed a technique to detect HS on Twitter by utilizing logistic regression. The data were gathered from an online environment (Twitter) and employedFiltering,Tokenizing, and Case Folding techniques during thepreprocessing phase. The Logistic Regression (LR) was utilized later in feature engineering, and it was found to have 84% accuracy [7].

OluwafemiOriolaet al. developed a method for detecting offensive comments on Twitter in 2020. Data was gathered via Twitter and marked into 2-parts: Free-speech "FS" and Hate-speech "HT" and then performed pre-processing phase. TF-IDF is the technique employed to convert text into feature vectors during the feature engineering phase. Using an enhanced support vector machine with n-grams has proven to be an effective method with 89.4% accuracy [8].

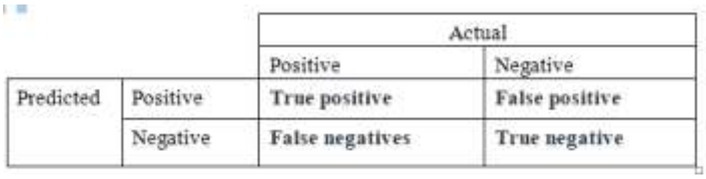
<u>Confusion Matrix</u>:

In machine learning, use a confusion matrix to measure the performance of a classification model.It is a means of displaying the number of accurate and inaccurate instances based on the model’s predictions.

The matrix displays the number of instances produced by the model on the test data.

- **True positives (TP):** You predicted a positive value, and it is correct.
- **True negatives (TN):** You predicted a negative value, and it is correct.
- **False positives (FP):** You predicted a positive value, and it is negative.
- **False negatives (FN):** You predicted a negative value, and it is a positive

Table1. <u>Confusion Matrix</u>



		Actual	
		Positive	Negative
Predicted	Positive	**True positive**	**False positive**
	Negative	**False negatives**	**True negative**

To find how accurate classification model is, use the **following metrics:**

1) **Accuracy:** The accuracy is used to find the portion of correctly classified values.

Accuracy= (TP+TN)/(TP+TN+FP+FN)

2) **Error rate:** it defines how often the model gives the wrong predictions.

Error Rate = (FP+FN) /(TP+TN+FP+FN)

3) **Precision:** It is the true positives divided by the total number of predicted positive values.

Precision = (TP)/(TP+FP)

4) **Recall:** It is used to calculate the model's ability to predict positive values.

Recall: = (TP)/(TP+FN)

5) **F1-Score:** It is the harmonic mean of Recall and Precision.

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

<u>For example/-</u>

Table2. Confusion Matrix Example

		Actual	
		Positive	Negative
Predicted	Positive	3	2
	Negative	7	8

From the above diagram, see that:

- True Positives (TP) = 3
- True Negatives (TN) = 8
- False Positives (FP) = 2
- False Negatives (FN) =7

1. **Accuracy=** (TP+TN)/(TP+TN+FP+FN) =(3+8)/(3+8+2+7)=11/20=**0.55**
2. **Error Rate:**(FP+FN) /(TP+TN+FP+FN)=(2+7)/(3+8+2+7)= 9/20=**0.45**
3. **Precision =** (TP)/(TP+FP)=(3)/(3+2)=3/5=**0.6**
4. **Recall: =** (TP)/(TP+FN)=3/(3+7)=3/10=**0.3**
5. **F1-Score:** (2*0.6*0.3)/(0.6+0.3)=0.36/0.9=**0.4**

2. DL APPROACH

Detecting hate speech has shifted attention to DL approaches over the past few years. Deep Learning is a new field that falls under the much broader field of Machine Learning, where convolutional neural networks (CNN) and recurrent neural networks (RNN) are often employed.

2.1 Convolutional Neural Network (CNN)

The bi-GRU-CNN-LSTM Model was proposed by T.V Huynh et al. (2019) for the detection of HS. They gathered data from Twitter and classified it as offensive, hateful, or clean (OFFENSIVE, HATE, and CLEAN). They then used three neural network models, including BiGRU-LSTM-CNN, TextCNN, and BiGRU-CNN to identify HS from the cleaned data. As a result, they scored 70.57% of F1 [9].

In a study published in 2020, Raghad et al. showed that detected automatic HS on Saudi Twitter by an algorithm. In this study, authors examined 3- neural network designs: CNN and GRU-based classifiers, along with a hybrid classifier that uses both. The authors evaluated BERT, a new language representation approach, on the task of detecting hate speech. For this study, a community dataset of Nine thousand three hundred and sixteen (9316) tweets was categorized as Hateful, Abusive, and Normal. A comparison and evaluation of 4-models has also been done: CNN, GRU, CNN + GRU, and BERT, and CNN outperformed than further models, in terms of F1 and AUROC score (0.79 and 0.89)[10].

2.2 Recurrent Neural Network (RNN)

NLP analyses words and sentences, where individuals in a sentence are influenced by the one preceding and following word. Such dependency problemsare handled by RNN. The network stream is regulated by the tanh activation function. The tanh squishes the input value between −1 and 1. Vectors pass through many such math operations while going through a neural network, and may lose their actual meaning as a result. The vanishing gradient issues then arise. Because of this, gradient-updated layers tend to stop learning. Thus, RNNs have short-term memory as it forgets data for long sequences[10].

IV. COMPARISION BETWEEN ML APPROACH VS DL APPROACH

Factors	Machine Learning Approach	Deep Learning Approach
Data Requirements	Training can be done with less data.	A lot of data is needed to fully understand it.
Hardware dependencies	Requires low-end machines.	High-end machines are needed for deep learning. GPUs are therefore required as well.
Execution time	Training with machine learning takes less time, from seconds to hours.	Training takes more time than machine learning. It takes so long because there are so many parameters in deep learning algorithms.
Hyperparameter Tuning	There are numerous options.	Tuning proficiencies are limited.

V. ADVANTAGES AND DISADVANTAGES OF ML AND DL APPROACH

Classification Techniques	Advantages	Disadvantages
Machine Learning Approach	**1) Scope of Improvement** Machine learning is a rapidly evolving field that offers numerous opportunities for improvement and has the potential to become the leading technology of the future. Significant research and innovation are taking place in this area, contributing to enhancements in both software and hardware.	**1) Data Acquisition** The entire premise of machine learning revolves around identifying valuable data. If a reliable data source is not provided, the results will be inaccurate. The quality of the data is also crucial. Therefore, machine learning is heavily dependent on data and its quality.
	2) Wide Range of Applicability This technology has a wide range of applications. Machine learning plays a role in almost every field, including hospitality, ed-tech, medicine, science, banking, and business, thus creating more opportunities.	**2)Time and Resources** Machines need time for their algorithms to adjust to the environment and learn. Trial runs verify accuracy and reliability, but require significant resources and expertise. They are costly in terms of time and expenses.
Deep Learning Approach	**1) Handling huge and complicated datasets** Deep learning algorithms can handle enormous and complex datasets, making it a powerful tool for gaining insights compared to traditional methods.	**1)Requires a Large Amount of Data** Deep learning relies on massive datasets for training, which requires high-quality data and significant time and resources for acquisition.
	2)Handling Varied Types of Data Deep Learning algorithms have the capacity to handle structures as well as unstructured data like Images, texts, and audio.	**2) Extensive computing Needs** Deep learning has a major disadvantage: it requires more computing resources to train specific models with large datasets, such as powerful central processors, graphics processing units, and large storage and memory.

VI. APPLICATIONS OF DEEP LEARNING APPROACH & MACHINE LEARNING APPROACH

[1] Medical imaging knowledge are increasingly used for detecting tumors and other malignancies in the body.
[2] Time-series forecasting based on Machine Learning is being used in Marketing.
[3] In the development of industrial robots, Deep Learning is playing an important role.
[4] Self-driving cars use machine learning algorithms to navigate to their destinations.
[5] Several industries use NLP to analyze customer reviews and determine their sentiment.

VII. CONCLUDING REMARKS

As HS has become a critical issue in online media, this paper introduces an overview of how HS has been instinctively detected in text over the years. The survey includes noteworthy activities in two sections: ML and DL techniques for HS identification. Here, discuss different conceptions behind DL and ML techniques and understand their operational differences, analogies, and applications. Also, the confusion matrix provides a summary of how well the classification model is performing, with labels indicating true positives, true negatives, false positives, and false negatives, helping to assess the accuracy and errors of the model.Furthermore, it outlines the potential of deep learning algorithms in exploring and analyzing unknown structures to derive useful representations through feature learning and continued evolution with additional data input.

REFERENCES

[1] Duggan M. Online harassment 2017. Washington: Pew Research Center; 2017. Search in Google Scholar.

[2] Emon EA, Rahman S, Banarjee J, Das AK, Mittra T, "A Deep Learning Approach to Detect Abusive Bengali Text". 2019 7th International Conference on Smart Computing & Communications (ICSCC), pp. 1–5. 10.1109/ICSCC.2019.8843606. Search in Google Scholar.

[3] Mohiyaddeen and Dr. Shifaulla Siddiqi (2021), "Automatic Hate Speech Detection: A Literature Review", International Journal of Engineering and Management Research, Volume-11, Issue-2 (April 2021), pp.116-121, 2021.

[4] A. Rodriguez, C. Argueta, & Y. L. Chen, "Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis", 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 169–174, 2019.

[5] S. Jaki& T. De Smedt. (2018). Right-wing German hate speech on twitter : Analysis and automatic detection, pp. 1–31.

[6] M. Di Capua, E. Di Nardo, & A. Petrosino, "Unsupervised Cyber Bullying Detection in Social Networks", 23rd International Conference on Pattern Recognition (ICPR) Cancún Center, Cancún, México, pp. 427–432, December 4-8, 2016.

[7] Purnama Sari Br Ginting; BudhiIrawan; CasiSetianingsih, "Hate speech detection on twitter using multinomial logistic regression classification method", 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), pp. 105-111,2019.

[8] O. Oriola& E. Kotze. (2020), "Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets", IEEE Access, Volume 8, pp. 21496 – 21509, 2020.

[9] T. Van Huynh, D. Nguyen, K. Van Nguyen, N. L. Nguyen, & A. G. Nguyen., "Hate speech detection on vietnamese social media text using the Bi-GRU-LSTM-CNN Model", arXiv:1911.03644v3 [cs.CL], 22 Dec 2019.

[10] RaghadAlshalan * and Hend Al-Khalifa*, "A Deep Learning Approach for Automatic Hate Speech Detection in The Saudi Twittersphere", Journal/Article-Applied Sciences, Volume 10, Issue 23., Page No1-16.,1 December 2020.

[11]AkankshaBisht, Annapurna Singh, H. S. Bhadauria, JitendraVirmani and Kriti, "Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model", © Springer Nature Singapore Pte Ltd. 2020, Recent Trends in Image and Signal Processing in Computer Vision, Advances in Intelligent Systems and Computing, Volume 1124, pp. 243-26