International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 897-903 www.ijaem.net ISSN: 2395-5252

Hate Speech Detection

Prof. Sudhansh Shekhar Pandey, Ishita Chhabra, Rishav Garg, Saksham Sahu

Assistant Professor, KIET Group Of Institutions,Delhi NCR.
KIET Group Of Institutions, Delhi NCR.
KIET Group Of Institutions,Delhi NCR.
KIET Group Of Institutions, Delhi NCR.

Date of Submission: 10-04-2023 Date of Acceptance: 20-04-2023

ABSTRACT - Hate speech is becoming a major issue as social media platforms grow in popularity. Hate speech commonly focuses on gender, race, or religion to spread hatred and violence. The goal of this project was to develop a machine learning model that could automatically detect hate speech in online text. To achieve this, the paper used a Deep Convolutional Neural Network (CNN) architecture. The model was trained on a dataset of tweets containing hate speech, offensive language, or neither, labeled , and showed high performance of in terms of accuracy, recall, and F1 score. I was able to achieve it.One of the main advantages of using a -deep CNN for this task is its ability to capture complex patterns in the data. By using multiple layers of convolutional filters, the model was able to learn features related to hate speech detection.

Keywords - Deep Convolutional Neural Network, Precision, Recall, F1-Score, Accuracy, Bi-LSTM, MLP.

INTRODUCTION

Hate speech is a serious problem online, leading to the spread of harmful and discriminatory language that can cause significant harm to individuals and communities. Efforts have been made to address this issue through moderation and community-led efforts, but there is a need for automated tools that can help detect and mitigate hate speech at scale. Identifying hate speech on social media platforms such as Twitter is important because hate speech can negatively impact individuals and communities. It leads to feelings of fear, anxiety and isolation, creating an environment in which it is difficult for people in marginalized communities to participate fully.

Twitter receives nearly 500 million tweets daily, making it impossible to create a human-based hate detection system.

AI systems are in place to flag the text of tweets, but one of the biggest challenges is to reduce false positives (flagging non-hate things as hateful) so that these systems can be more expressive. To be able to detect hate speech without infringing on freedom. This project proposes an ML-based model trained to detect hate speech. This paper is based on hate speech detection using flat techniques such as random forests, support vector machines (SVM), and K Nearest Neighbors.

This study is based on CNN. One reason researchers can use CNNs to detect hate speech is that CNNs are particularly good at learning hierarchical representations of data. This helps you understand the context and meaning of words and phrases in your sentences.

CNNs can also be trained on large datasets. This helps them learn to recognize patterns and traits associated with hate speech. One of the main advantages of using deep CNNs for this task is their ability to capture local and global patterns in the data. Convolutional filters can learn features related to hate speech detection by examining the relationships between words and phrases in the text. Additionally, using multiple

-layer convolutional filters allows the model to learn more abstract features that capture higher-level concepts.

This paper includes CNN models with various filter sizes and Deep Convolutional Neural Network models such as Long Short-Term Memory (LSTM) and Bi-LSTM for higher accuracy.

The Twitter dataset is used to classify tweets into three classes: hate speech, offensive language, or neither.

SCOPE OF WORK

This white paper contains research on machine learning models using shallow and deep learning techniques. This white paper covers

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 897-903 www.ijaem.net ISSN: 2395-5252

implementations of random forests, K nearest neighbors, support vector machines, logistic regression, decision trees, and naive Bayes algorithms. We trained a CNN model with different filter sizes. Word embedding methods such as LSTM and Bi-LSTM models are used. Comparisons of precision, recall and F1 scores are performed using the method of ref. [3].

LITERATURE SURVEY

Hate speech detection has become a major topic for researchers in recent years. He has two main methods used by researchers. One is a traditional machine learning approach and the other is a deep learning approach. A lot of research has been done on detecting hate speech using machine learning techniques. Here are some examples of research using machine learning for this purpose:

• This [1] study used a combination of machine learning techniques such as support vector machines (SVM) and logistic regression to classify text as hate speech or not. hate speech. The study found that the machine learning model achieved an accuracy of about 73% on his dataset of Twitter messages.

• This [2] study used a combination of machine learning techniques such as random forest and logistic regression to classify text as either hate speech or non-hate speech. The study found that the machine learning model achieved an accuracy of about 84% on his dataset of Twitter messages.

A sample study using a deep convolutional neural network (DCNN) follows.

• In [4], researchers used deep CNNs to classify tweets as either hate speech or non-hate speech. The model was trained on a dataset of over 15,000 tweets and achieved an accuracy of 95.

Five%. The model used a combination of convolutional and fully connected layers and was trained with the Adam optimizer with a learning rate of 0.001.

• In a published article in [5], researchers developed a detailed CNN model for detecting hate speech in online comments. The model was trained on a dataset of over 100,000 comments and achieved his F1 score of 0.

93 on test devices. The model uses a combination of convolutional and recurrent layers and was trained with the Adam optimizer with a learning rate of 0.001.

• [6] In a study published in the journal IEEE Access in 2020, researchers used deep CNNs to classify social media posts as either hate speech or non-hate speech. The model was trained on a dataset of over 60,000 posts and achieved an accuracy of 94.6%。

MATERIALS AND METHODS

This paper includes a baseline machine learning model and a DCNN model and compares the results of the baseline paper dataset with the selected dataset. In this paper, we used different convolutional layers and filter sizes to improve the accuracy of hate tweets. The model was trained on Google Colab. The dataset is pulled from www.kaggle.

com and consists of 24783 tweets from Twitter. This record was tagged with three classes: Hate_Speech, Offensive_Language, and neither. There were 1430 hate tweets (5.77% of the total) and 19,190 abusive tweets (77.43% of the total), both of which were 4163 (as of December 20).

80% of the total amount).

Datasets Used -

https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset

A CNN consists of several layers, each with a specific Execute the processing function. data entered.

A. Embedding Planes

Embedding planes are planes used to represent input data in a low-dimensional space. The purpose of the embedding layer is to store the raw input data. B. Convert the text or image pixels into a numerical representation that can be processed by the CNN. Input Dimension, Output Dimension, Weight and Input Length are used as parameters. An embedding layer is typically the first layer in a CNN, taking raw input data as input and producing a numerical representation of that data as output. The numeric representation produced by the embedding layer is often called the embedding vector.

B. Pooling Layer

The purpose of the pooling layer is to reduce the size of the feature maps produced by the convolutional layers while preserving the important information contained in those feature maps. There are different types of pooling layers. B. Max pooling and average pooling. Max pooling works by selecting the largest value from a group of adjacent values in the map of features

, whereas average pooling works by taking the average of the values within the group. I used global max pooling.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 897-903 www.ijaem.net ISSN: 2395-5252

C. Dense Layers

The purpose of dense layers is to make predictions based on features extracted from the convolutional layers of a CNN. A dense layer consists of a set of neurons connected to all neurons in the previous layer. Each neuron receives input from all neurons in the previous layer and produces an output based on that input. The output of a dense layer is the predictions made by the CNN based on features extracted from the input data.

D. Impairment Layers

Impairment layers work by randomly zeroing out some of the neurons in the network during training. This has the effect of "dropping" these neurons out of the network, preventing the network from being overly dependent on a particular neuron or group of neurons. A dropout layer is usually used after one or more of his dense layers in a CNN and applied to the output of those layers. Dropout rate is a hyperparameter that determines the probability of a neuron going to zero.

A high dropout rate means more neurons are failing, which can make the network more robust, but can also reduce its ability to learn.

E. Classification Layer

The classification layer is the layer used to make predictions about the classes or categories of the input data samples. A classification layer takes as input features extracted from the convolutional layers of a CNN and produces predictions as output. I used softmax layer paper.

Softmax layers are used for multiclass classification where the input data belongs to one of several classes. Produces a probability distribution over the classes, with each class assigned a probability between 0 and 1. The class with the highest probability is chosen as the final prediction. The classification layer is primarily the last layer in the convolution process.

Evaluation Techniques

Some common techniques used in this article:

Classification Metrics: These metrics are used to evaluate the performance of a CNN's classification layer.

Common classification metrics include accuracy, which measures the percentage of correct predictions of the CNN, and accuracy, which measures the percentage of positive predictions that are actually correct.

Confusion Matrix: The Confusion Matrix is a table showing the number of true positive, true negative, false positive, and false negative predictions of the CNN. It can be used to calculate various evaluation metrics such as: B. Precision, recall, and F1 score.

ROC Curve: A Receiver Operating Characteristic (ROC) curve is a graphical representation of a CNN's true positive and false positive rates. It can be used to visualize the CNN's sensitivity-specificity trade-off and choose an appropriate threshold for classification prediction.

Metrics are as follows:

- Precision (P): Precision is defined as the number of true positive predictions made by the classifier divided by the total number of positive predictions made by the classifier.

Precision = $TP_0 / (TP_0 + FP_O)$

- Recall (R): It is the fraction of tweets that have been identified from the total number of hate speech tweets present.

Recall = $TP_0 / (TP_0 + FN_e)$

- F1-Score (F1): It is the harmonic mean of Precision(P) and Recall(R)

F1-Score = 2 * (P * R) / (P + R)

RESULTS AND CONCLUSION

A. MACHINE LEARNING MODELS

This paper has implementation of machine learning models such as Random Forest, K Nearest Neighbours, Support Vector Machine, Logistic Regression, Decision Tree, Naive Bayes and compared accuracy with our base paper.

Precision : As shown in table 1:

TABLE 1 : Precision Parameter Results of Baseline Models

	Results
Random Forest	0.97
K Nearest Neighbors	0.98
Support Vector Machine	0.99
Logistic Regression	0.98
Decision Tree	0.95
Naive Bayes	0.46

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 897-903 www.ijaem.net ISSN: 2395-5252

Recall: As shown in table 2.

TABLE 2: Recall Parameter Results of Baseline Models

	Results
Random Forest	0.95
K Nearest Neighbors	0.95
Support Vector Machine	0.94
Logistic Regression	0.95
Decision Tree	0.95
Naive Bayes	0.91

F1-Score: As shown in table 3

TABLE 3: F1-Score Parameter Results of Baseline Models

	Results
Random Forest	0.95
K Nearest Neighbors	0.96
Support Vector Machine	0.96
Logistic Regression	0.96
Decision Tree	0.96
Naive Bayes	0.61

B. DEEP LEARNING MODELS

Deep learning models were implemented by adding multiple convolutional layers and different filter sizes. The dataset used for this study is taken from Kaggle.com. The following models were implemented:

1. 1-CNN MODEL

One layer CNN model was implemented with filter size equals to 5. Layers included Embedding layer (main layer), and sublayers - global max-pooling layer, dense and dropout layer. The following graph was plotted between accuracy and loss as shown in figure 1.

FIGURE 1 - 1-CNN filter size 5 model results

2. 2-CNN MODEL

Two layer CNN model was implemented with filter size equal to 4, 3 and 2. Layers included Embedding layer (main layer), convolutional layer 1 and 2 and sublayers - global max-pooling layer, dense and dropout layer. The following graphs were plotted between accuracy and loss as shown in figure 2, 3, 4.

FIGURE 2 - 2-CNN filter size = 4 model results

FIGURE 3 - 2-CNN filter size = 3 model results

FIGURE 4 - 2-CNN filter size = 2 model results

3. LONG TERM SHORT MEMORY(LSTM) MODEL

The LSTM model was implemented with layers including Embedding layer, LSTM layer and sublayers - global max-pooling layer, two dense and dropout layers each. The following graphs was plotted between accuracy and loss as shown in figure 5

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 897-903 www.ijaem.net ISSN: 2395-5252

FIGURE 5 - LSTM Model results

4. Bi-LONG TERM SHORT MEMORY(LSTM) MODEL

The Bi-LSTM model was implemented with layers including Embedding layer, Bi-LSTM layer and sublayers - global max-pooling layer, two dense and three dropout layers as shown in figure 6.

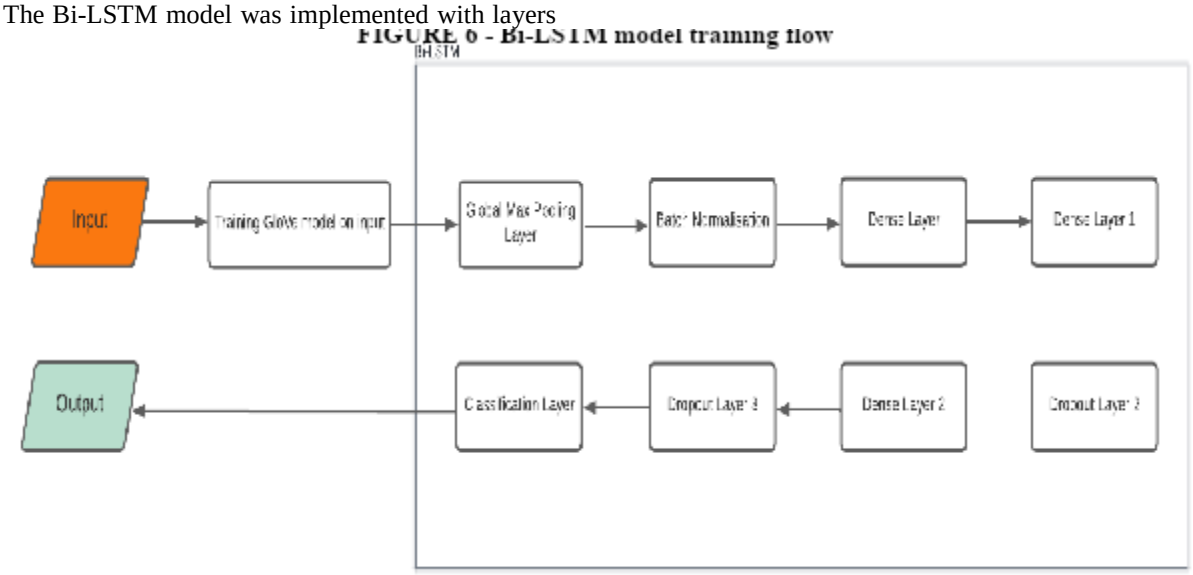

FIGURE 6 - Bi-LSTM model training flow

The following graphs were plotted between accuracy and loss as shown in figure 7.

FIGURE 7 - Bi LSTM model results

5. MULTILAYER PERCEPTRON LAYER(MLP) MODEL

The MLP model was implemented with layers including Embedding layer, Flattening layer and sublayers - dense, classification and dropout layers each as represented in figure 8.

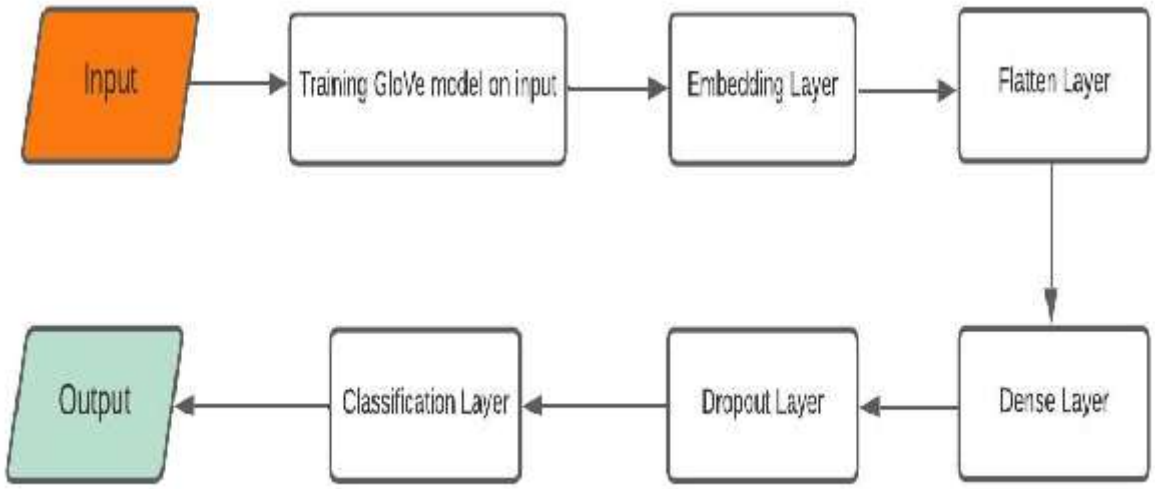

FIGURE 9 : MLP model training flow

The following graph was plotted between accuracy and loss as shown in figure 9.

FIGURE 9 - MLP model results

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 897-903 www.ijaem.net ISSN: 2395-5252

DEEP LEARNING MODELS RESULTS

Precision: Comparison of base paper results and proposed paper results including new models trained. As shown in table 4.

TABLE 4 - Precision Parameter Results of Deep Learning Models

	Base Paper Results	**Proposed Paper Results**
1 CNN (5g)	0.67	0.64
2 CNN(4g, 3g, 2g)	0.62	0.69
LSTM	0.64	0.66
New Models Results		
Bi-LSTM	0.72	
MLP	0.63	

Recall: Comparison of base paper results and proposed paper results including new models trained. As shown in table 5.

TABLE 5 - Recall Parameter Results of Deep Learning Models

	Base Paper Results	**Proposed Paper Results**
1 CNN (5g)	0.53	0.67
2CNN(4g, 3g, 2g)	0.55	0.70
LSTM	0.53	0.75
New Models Results		
Bi-LSTM	0.55	
MLP	0.66	

F1-Score Comparison of base paper results and proposed paper results including new models trained.: As shown in table 6

TABLE 6 - F1-Score Parameter Results of Deep Learning Models

	Base Paper Results	**Proposed Paper Results**
1 CNN (5g)	0.59	0.84
2CNN(4g, 3g, 2g)	0.57	0.87
LSTM	0.53	0.79

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 897-903 www.ijaem.net ISSN: 2395-5252

New Models Results	
Bi-LSTM	0.84
MLP	0.84

CONCLUSION

This research addresses the problem of detecting hate speech on Twitter using deep convolutional neural networks. Initially, machine learning-based classifiers such as Logistic Regression, Random Forest, Naive Bayes, Support Vector Machines, Decision Trees, K-Nearest Neighbors were used to identify his HS-related tweets on Twitter with traits .

Deep learning-based CNNs (1 and 2 CNNs with different filter sizes), MLP, LSTM, and their combined Bi-LSTM models also yield similar results on fixed partitioned datasets.

Current research only addresses HS issues with text data. However, images are also often used for this. Therefore, in the future, researchers can insert images with text or analyze video datasets to collect more HS-related posts from Twitter.

This study only used tweets written in English, but it can be further expanded by mixing other languages such as Hindi, Tamil, and French. Real-time data analytics can also be performed on live data from the Twitter API. Regarding future scope, there are several potential directions for hate speech detection projects using deep learning techniques. For example, projects could focus on improving the accuracy and reliability of models, or developing new techniques to deal with more complex or nuanced cases of hate speech. Additionally, the project could explore ways to integrate the model into real-world applications such as social media platforms and online forums to automatically identify and flag hateful or harmful content. increase.

To create a general framework with a deep learning model, you need enough examples in your training data set. In the future, the current dataset can be extended to improve accuracy. Deep learning techniques such as convolutional neural networks (CNN) can be powerful tools for detecting hate speech from text and audio. However, it is important to carefully consider the limitations and potential biases of the model and the data it is training on. This is because these can have a significant impact on the accuracy and reliability of the model.

Overall, deep learning techniques can be a valuable tool for detecting hate speech and promoting online safety, but we should approach the issue carefully and avoid potential limitations and biases of the models. is important to consider.

REFERENCES

[1] Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das., " A Framework for Hate Speech Detection Using Deep Convolutional Neural Network" IEEE Access (Volume: 8), pp. 10 November 2020.

[2] P. Charitidis, S. Doropoulos, S. Vologiannidis, I. Papastergiou, and S. Karakeva,''Towards countering hate speech against journalists on social media,'' Online Social Netw. Media, vol. 17, pp. 2020.

[3] L. Gao and R. Huang, ''Detecting online hate speech using context aware models,'' 2017, arXiv:1710.07395. [Online]. Available: http://arxiv.org/abs/1710.07395

[4] O. Oriola and E. Kotze, ''Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets,'' IEEE Access, vol. 8, pp. 21496–21509, 2020.

[5] C. Stokel-Walker, ''Alt-right's 'Twitter' is hate-speech hub,'' New Scientist, vol. 237, no. 3167, p. 15, Mar. 2018.

[6] P. Charitidis, S. Doropoulos, S. Vologiannidis, I. Papastergiou, and S. Karakeva, ''Towards countering hate speech against journalists on social media,'' Online Social Netw. Media, vol. 17, pp. 1–10, May 2020.

[7] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, ''Deep learning for hate speech detection in tweets,'' in Proc. 26th Int. Conf. World Wide Web Companion - WWW Companion, 2017, pp. 759–760.

[8] Z. Zhang, D. Robinson, and J. Tepper, ''Detecting hate speech on Twitter using a convolution-GRU based deep neural network,'' in Proc. Eur. Semantic Web Conf. Heraklion, Greece: Springer, 2018, pp. 745–760.