International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

A Machine Learning Based Approach for Hate Speech Detection

Yogesh Jadhaw[1], Prof. Ashish Tiwari[2]

Abstract: In recent years, the proliferation of social media and digital communication platforms has led to an exponential increase in user-generated content. While these platforms offer numerous benefits, they have also become breeding grounds for hate speech. Identifying and mitigating hate speech is crucial to maintaining a safe and inclusive online environment. Traditional methods of content moderation, which rely heavily on human moderators, are increasingly inadequate due to the sheer volume of content and the nuanced nature of hate speech. The proposed work presents an artificial intelligence based technique for detection of hate speech. The proposed approach uses the Deep BayesNet for identifying potential hate speech. It has been shown that the proposed approach attains higher classification accuracy compared to existing work in the domain.

Keywords:- Social Media, Hate Speech BayesNet, Classification Accuracy.

I.Introduction

Hate speech, defined as any communication that belittles or discriminates against individuals based on their race, religion, ethnicity, gender, sexual orientation, or other characteristics, presents a significant challenge for moderation [1]. The scale of the problem is immense; millions of posts, comments, and messages are generated daily across various platforms. Moreover, hate speech is often subtle and context-dependent, making it difficult for traditional keyword-based filters to detect effectively [2]. This complexity necessitates the use of advanced ML algorithms capable of understanding context, sentiment, and implicit meanings. Machine learning offers several advantages over traditional methods in identifying hate speech [3]. Firstly, ML models can be trained on vast datasets, enabling them to recognize patterns and nuances that human moderators might miss. These models can learn from context, differentiating between hate speech and benign content that might contain similar keywords. Secondly, ML systems operate at a scale and speed unattainable by human moderators, allowing for real-time monitoring and intervention. This rapid response is crucial in mitigating the spread of harmful content before it reaches a large audience. [4]

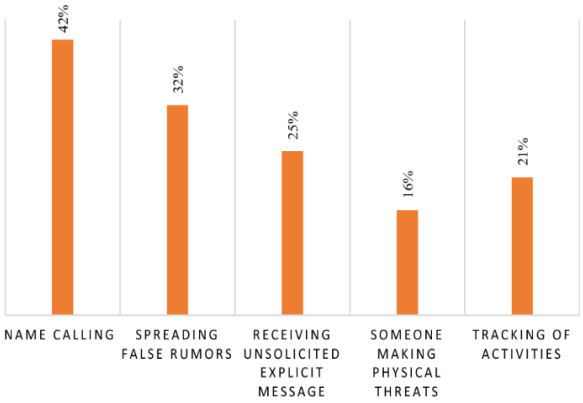


Fig.1 Categorization of Hate Speech Cases (Source: https://link.springer.com/article/10.1007/s13278-022-00951-3)

Moreover, hate speech is often subtle and context-dependent, making it difficult for traditional keyword-based filters to detect effectively. This complexity necessitates the use of advanced ML algorithms capable of understanding context, sentiment, and implicit meanings [5].

II. Existing Challenges

Despite its potential, implementing ML for hate speech detection is fraught with challenges. One major issue is the bias in training data, which can lead to models that unfairly target certain groups while overlooking others. Ensuring the diversity and representativeness of training datasets is essential to

mitigate this risk. Additionally, the dynamic nature of language, including slang, memes, and coded language used to evade detection, requires continuous updating and retraining of ML models. Privacy concerns also arise, as the collection and analysis of user data must comply with regulations and respect individual rights [6].

While ML can significantly enhance the identification of hate speech, it is not a standalone solution. Ethical considerations, such as the risk of over-censorship and the potential for false positives, necessitate a hybrid approach where human moderators work alongside ML systems. Human oversight is crucial to review contentious cases and provide context that automated systems might lack. This collaborative approach helps balance the need for effective moderation with the protection of free speech and user rights [7].

III. System Design using Regression Learning based Bayesian Regularized ANN

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. Other advantages include [8]:

1. **Adaptive learning:** An ability to learn how to do tasks based on the data given for training or initial experience.
2. **Self-Organization:** An ANN can create its own organization or representation of the information it receives during learning time.
3. **Real Time Operation:** ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability [9].

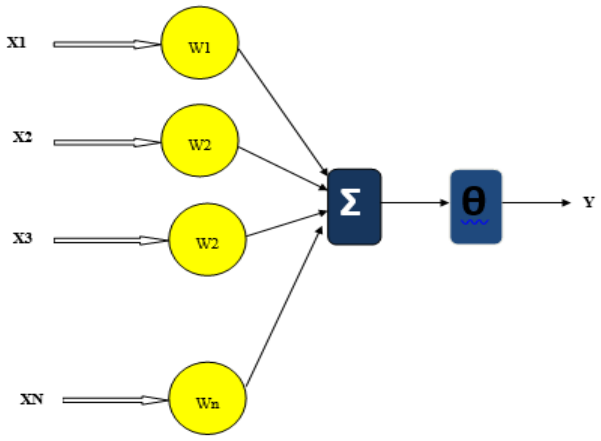


Fig.2 Mathematical Model of Neural Network

The output of the neural network is given by:

$$\sum_{i=1}^{n} XiWi + \Theta \quad (1)$$

Where,
Xi represents the signals arriving through various paths,
Wi represents the weight corresponding to the various paths and
Θ is the bias. It can be seen that various signals traverssing different paths have been assigned names X and each path has been assigned a weight W. The signal traverssing a particular path gets multiplied by a corresponding weight W and finally the overall summation of the signals multiplied by the corresponding path weights reaches the neuron which reacts to it according to the bias Θ. Finally its the bias that decides the activation function that is responsiblefor the decision taken upon by the neuralnetwork. The activation function φ is used to decide upona the final output. The learning capability of the ANN structure is based on the temporal learning capability governed by the relation [10]:

$$w(i) = f(i, e) \quad (2)$$

Here,
w (i) represents the instantaneous weights
i is the iteration
e is the prediction error

The weight changes dynamically and is given by:

$$W_k \overset{e,i}{\rightarrow} W_{k+1} \quad (3)$$

Here,
W_k is the weight of the current iteration.

W_{k+1} is the weight of the subsequent iteration.

(i) Regression Learning Model

Regression learning has found several applications in supervized learning algorithms where the regression analysis among dependednt and independent variables is eeded [11]. Different regression models differ based on the the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a relationship between x (input) and y(output). Mathematically [12],

$$y = \theta_1 + \theta_2 x \quad (4)$$

Here,
x represenst the state vector of inut variables
y rperesenst the state vector of output variable or variables.
Θ1 and Θ2 are the co-efficents which try to fit the regression learning models output vector to the input vector.

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y). The cost function J is mathematically defined as:

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_1 - y_i)^2 \quad (5)$$

Here,
n is the number of samples
y is the target
pred is the actual output.

(ii) Gradient Descent in Regression Learning

To update θ1 and θ2 values in order to reduce Cost function (minimizing MSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values and then iteratively updating the values, reaching minimum cost. The main aim is to minimize the cost function J [13].

(iii) Bayesian Regularization

The Bayesian Regularization (BR) algorithm is a modified version of the LM weight updating rule with an additional advantage of usingthe Baye's theorem of conditional probability for a final classification [14].
The weight updating rule for the Bayesian Regularizationis given by:

$$w_{k+1} = w_k - (J_k J_k^T + \mu I)^{\wedge} - 1(J_k^T e_k) \quad (6)$$

Here,
w_{k+1} is weight of next iteration,
w_k is weight of present iteration
J_k is the Jacobian Matrix
J_k^T is Transpose of Jacobian Matrix
e_k is error of Present Iteration
μ is step size
I is an identity matrix.

The decision making approach of the Bayesian Classifier can be understood graphically using the graph theory approach. The approach for computing the probability among different disjoint sets can be understood using the set theory approach shown in the subsequent steps. The figures clearly depict the decision to be taken in cases of different overlapping data value categories [15].

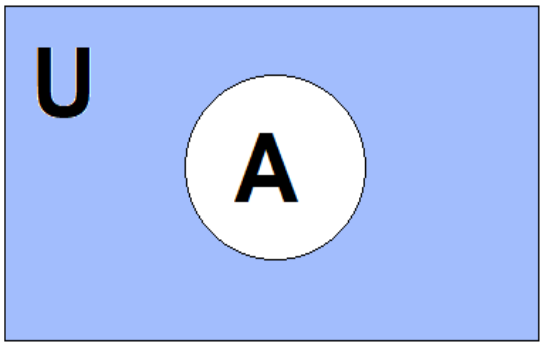


Fig.3 Universal Set Containing a Subset 'A'

Let us assume that the Bayesian Regularization algorithm needs to categorize the set A among multiple subsets in the superset U, for the time being in which only A exists exclusively.

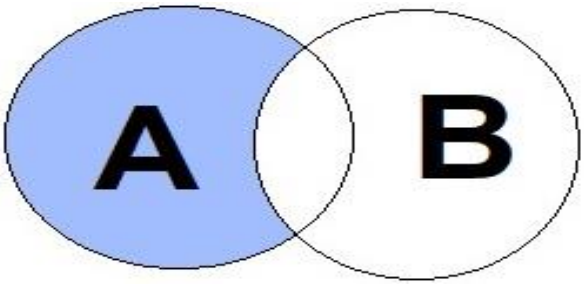


Fig.4 Probability of Exclusive Occurrence of 'A'

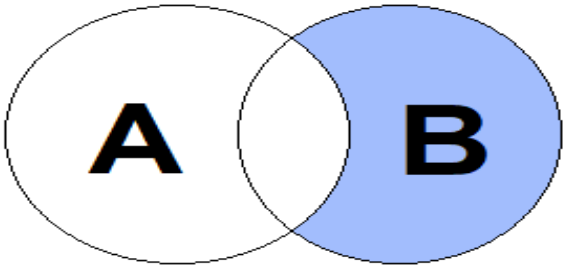


Fig.5 Probability of Exclusive Occurrence of 'B'

Figures 4 and 5 depict the probability of exclusive occurrence of events A and B respectively.

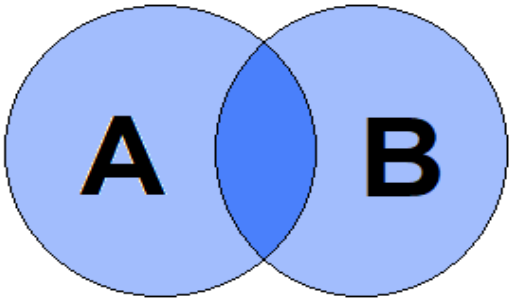


Fig.6 Probability of Union of A and B

Moreover for the predictive classification of ant data set, the Baye's Rule is followed, which is given by:

$$P\frac{A}{B} = \frac{P(A).P\frac{B}{A}}{P(B)} \quad (7)$$

Here,

$P\frac{A}{B}$ is the probability of occurrence of A given B is true.

$P\frac{B}{A}$ is the probability of occurrence of B givenA is true.

$P(B)$ is the probability of occurrence of B

$P(A)$ is the probability of occurrence of A

In the present case the, 70% of the data has been taken for training and 30% of the data has been taken for testing.

The conditional probability of the sentiment can be also seen as an overlapping event with the classification occurring with the class with maximum conditional probability. The mathematical formulation for the above mentioned probabilistic approach can be understood as follows:
Let there be 'N' classes of data sets available in the sample space 'U'.
Let the conditional probability of each of such sets be given by:

$$P(\frac{A}{U}),\ P(\frac{B}{U}),\ \ldots\ldots\ P(\frac{N}{U}). \quad (8)$$

The BR algorithm tries to find out the maximum among the probabilities:

$$P(\max) = \begin{matrix} P(\frac{A}{U}) \\ P(\frac{B}{U}) \\ \vdots \\ \vdots \\ P(\frac{N}{U}) \end{matrix} \quad (9)$$

The maximum value of the probability decides the classification of a dataset into a particular category. Assuming that X attains the maximum in such a sample space:

$$P_{max} = X \quad (10)$$

Where,

$$P\left(\frac{X}{U}\right) = P\ \frac{X}{\prod_{i=1}^{i=n} U_i} \quad (11)$$

Here,
$\prod_{i=1}^{i=n} U_i$ represents the conditional probability cumulative for all possible data set classes in the sample space U
X is the maximum probability corresponding to a particular data set and n is the total number of classes of categorization.

IV. Evaluation Parameters

Since errors can be both negative and positive in polarity, therefore its immaterial to consider errors with signs which may lead to cancellation and hence inaccurate evaluation of errors. Therefore we consider mean square error and mean absolute percentage errors for evaluation. The system accuracy can be evaluated in terms of the mean square error which is mathematically defined as:

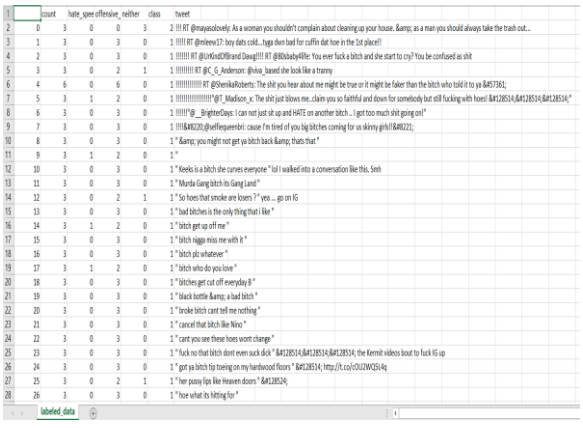
$$mse = \frac{1}{n}\sum_{I=1}^{N}(X - X')^2 \qquad (12)$$

Here,
X is the predicted value and
X' is the actual value and n is the number of samples.

V. Results:

The data has been collected from Kaggle.

Data Normalization: Canonization (normalization) of the text is the process of bringing to a single format, convenient for further processing. When working with large amount of information, it is necessary to exclude from the document all non-informative parts of speech (prepositions, particles, conjunctions, etc.).



	count	hate_speech	offensive_language	neither	class	tweet
2	0	3	0	0	3	2 !!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. & as a man you should always take the trash out...
3	1	3	0	3	0	1 !!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
4	2	3	0	3	0	1 !!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80babydiffe: You ever fuck a bitch and she start to cry? You be confused as shit
5	3	3	0	2	1	1 !!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
6	4	6	0	6	0	1 !!!!!!!!!!!!! RT @SherikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya 
7	5	3	1	2	0	1 !!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows me..claim you so faithful and down for somebody but still fucking with hoes! 😂😂😂"
8	6	3	0	3	0	1 !!!!!!"@__BrighterDays: I can not just sit up and HATE on another bitch .. I got too much shit going on!"
9	7	3	0	3	0	1 !!!!“@selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!!”
10	8	3	0	3	0	1 " & you might not get ya bitch back & thats that "
11	9	3	1	2	0	1 "
12	10	3	0	3	0	1 " Keeks is a bitch she curves everyone " lol I walked into a conversation like this. Smh
13	11	3	0	3	0	1 " Murda Gang bitch its Gang Land "
14	12	3	0	2	1	1 " So hoes that smoke are losers ? " yea ... go on IG
15	13	3	0	3	0	1 " bad bitches is the only thing that I like "
16	14	3	1	2	0	1 " bitch get up off me "
17	15	3	0	3	0	1 " bitch nigga miss me with it "
18	16	3	0	3	0	1 " bitch plz whatever "
19	17	3	1	2	0	1 " bitch who do you love "
20	18	3	0	3	0	1 " bitches get cut off everyday B "
21	19	3	0	3	0	1 " black bottle & a bad bitch "
22	20	3	0	3	0	1 " broke bitch cant tell me nothing "
23	21	3	0	3	0	1 " cancel that bitch like Nino "
24	22	3	0	3	0	1 " cant you see these hoes wont change "
25	23	3	0	3	0	1 " fuck no that bitch dont even suck dick " 😂😂😂😂 the Kermit videos bout to fuck IG up
26	24	3	0	3	0	1 " got ya bitch tip toeing on my hardwood floors " 😂 http://t.co/cOU2WQ3L4q
27	25	3	0	2	1	1 " her pussy lips like Heaven doors " 😌
28	26	3	0	3	0	1 " hoe what its hitting for "

labeled_data

Fig.7 Raw Data

Figure 7 renders a sample screenshot of the raw data.

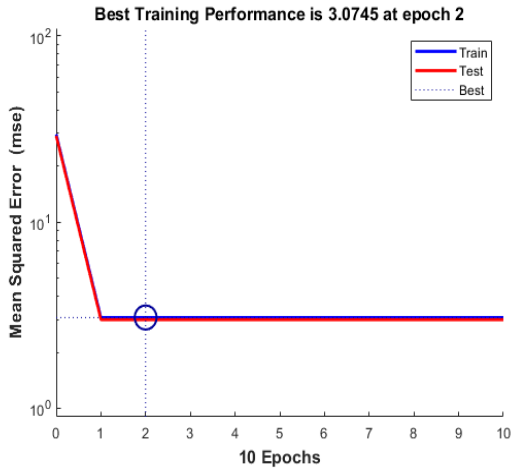


Fig.8 Variation of MSE

The variation of the mean squared error as a function of the number of epochs is shown in the above figure. It can be seen that the MSE stabilizes at a value of 3.0745.

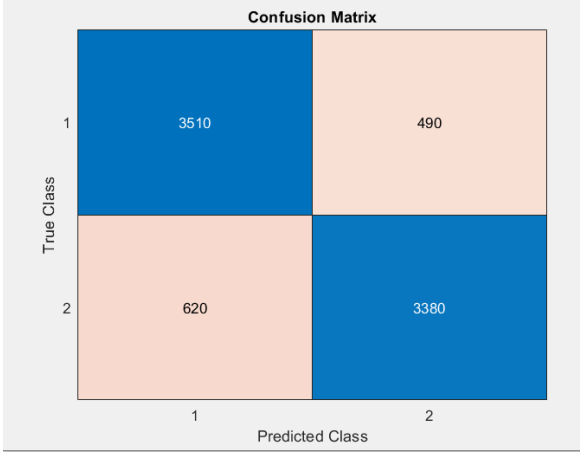


Fig.9 Confusion Matrix

Figure 9 depicts the confusion matrix for the proposed work. It can be observed that the proposed work attains a classification accuracy of around 86% which is significantly higher than that of previous work Bigoulaeva, et al. [16] which is around 78%.

VI. Conclusion:

In conclusion, the integration of machine learning into hate speech detection frameworks is essential to address the scale and complexity of the problem. ML offers powerful tools to identify and mitigate harmful content in ways that traditional methods cannot. However, to maximize its effectiveness and minimize potential pitfalls, it must be implemented thoughtfully, with attention to ethical concerns and the inclusion of human oversight. As technology and language continue to evolve, ongoing research and development in ML for hate speech detection will be crucial in maintaining safe and inclusive online spaces. Its has been shown that the proposed approach attains higher classification accuracy compared to existing work in the domain.

References

[1] M. F. Wright, B. D. Harper, and S. Wachs, ``The associations between cyberbullying and callous-unemotional traits among adolescents:The moderating effect of online disinhibition," *J. Personality Individual Differences*, vol. 140, pp. 41_45, Apr. 2019.

[2] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz and L. Plaza, "Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data," in *IEEE Access*, vol. 8, pp. 219563-219576, 2020, doi: 10.1109/ACCESS.2020.3042604.

[3] S. Khan *et al.*, "HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional Gated Recurrent Unit With Capsule Network," in *IEEE Access*, vol. 10, pp. 7881-7894, 2022, doi: 10.1109/ACCESS.2022.3143799.

[4] R. Singh *et al.*, "Deep Learning for Multi-Class Antisocial Behavior Identification From Twitter," in *IEEE Access*, vol. 8, pp. 194027-194044, 2020, doi: 10.1109/ACCESS.2020.3030621.

[5] Singh, T., Kumari, M. Burst: real-time events burst detection in social text stream. *J Supercomput* **77**, 11228–11256 (2021). https://doi.org/10.1007/s11227-021-03717-4

[6] Singh, T., Kumari, M. & Gupta, D.S. Real-time event detection and classification in social text steam using embedding. *Cluster Comput* **25**, 3799–3817 (2022).

[7] D. K. Jain, R. Jain, Y. Upadhyay, A. Kathuria, and X. Lan, ``Deep re_nement: Capsule network with attention mechanism-based system for text classification," *Neural Comput. Appl.*, vol. 32, no. 7, pp. 1839_1856,Apr. 2020.

[8] P. K. Jain, R. Pamula, and S. Ansari, ``A supervised machine learning approach for the credibility assessment of user-generated content," *Wireless Pers. Commun.*, vol. 118, no. 4, pp. 2469_2485, Jun. 2021.

[7] 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. Cham, Switzerland: Springer, 2018, pp. 745_760.

[8] A. R. Gover, S. B. Harper, and L. Langton, ``Anti-Asian hate crime during the COVID-19 pandemic: Exploring the reproduction of inequality," *Amer. J. Criminal Justice*, vol. 45, no. 7, pp. 647_667, 2020.

[9] **https://www.ohchr.org/en/statements/2023/01/freedom-speech-not-freedom-spread-racial-hatred-social-media-un-experts**.

[10] J. Langham and K. Gosha, ‘‘The classification of aggressive dialogue in social media platforms,’’ in Proc. ACM SIGMIS Conf. Comput. People Res., Jun. 2018, pp. 60–63.

[11] P. Fortuna and S. Nunes, ‘‘A survey on automatic detection of hate speech in text,’’ ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, 2018.

[12] W. Dorris, R. Hu, N. Vishwamitra, F. Luo, and M. Costello, ‘‘Towards automatic detection and explanation of hate speech and offensive language,’’ in Proc. 6th Int. Workshop Secur. Privacy Anal., Mar. 2020, pp. 23–29.

[13] A. Alrehili, ‘‘Automatic hate speech detection on social media: A brief survey’’ in Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA), Nov. 2019, pp. 1–6.

[14] S. Modi, ‘‘AHTDT—Automatic hate text detection techniques in social media’’ in Proc. Int. Conf. Circuits Syst. Digit. Enterprise Technol. (ICCSDET), Dec. 2018, pp. 1–3.

[15] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, ‘‘Machine learning techniques for hate speech classification of Twitter data: State of the-art, future challenges and research directions’’ Comput. Sci. Rev., vol. 38, Nov. 2020, Art. no. 100311.

[16] I Bigoulaeva, V Hangya, I Gurevych, A Fraser, “Label modification and bootstrapping for zero-shot cross-lingual hate speech detection”, Language Resources and Evaluation, Springer 2023, vol.57, pp. 515–1546.