International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 02 | February - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

A Collaborative Learning Approach for Identifying Hateful Speech on Social Media Applications

Akshay Baviskar[1] Dr. Manish Vyas[2]

Abstract: Social media, hailed for connecting people across the globe, has become an integral part of our daily lives. However, the same platforms designed to foster communication and community have become breeding grounds for hateful content. Hate speech, discrimination, and online harassment are pervasive issues that threaten the fabric of digital society. . Machine learning models are indispensable from identification of radical content due to the sheer size and complexity of the data. In this approach, a collaborative learning based approach has been proposed for identifying hateful content. It has been shown that the proposed approach attains higher classification accuracy compared to existing work in the domain.

Keywords:- Hateful content, social networks, text analysis, Collaborative Learning, Classification Accuracy.

I.Introduction

The rise of hateful content on social media is undeniable. Platforms that were meant to be spaces for dialogue and expression are now witnessing the proliferation of hate speech, fueled by anonymity, tribalism, and the echo-chamber effect [1]. Hateful content can take various forms, including discriminatory language, cyberbullying, threats, and even the dissemination of extremist ideologies. Its impact is far-reaching, affecting individuals, communities, and societies on a global scale [2]. The consequences of hateful content on social media are multifaceted. First and foremost, it can lead to psychological harm for those targeted, causing emotional distress, anxiety, and even depression [3]. Moreover, the spread of hate speech can contribute to real-world violence, as seen in instances where online radicalization has led to offline acts of terror. Social cohesion is also undermined, as hate speech fosters division and animosity between different groups, eroding the sense of community that social media platforms aim to foster [4].

Addressing hateful content on social media is a complex challenge. The sheer volume of user-generated content, coupled with the evolving nature of language and communication, makes it difficult to develop foolproof automated systems for detection. The fine line between freedom of speech and preventing hate speech further complicates the task, as platforms grapple with the need to moderate content while respecting users' rights to express their opinions [5].

As the data to be analyzed is extremely large and complex, machine learning models are needed to filter radical and non-radical content based on certain level of sophistication [6].

II. Mitigation Strategies

The common mitigation strategies are presented in this section [7]-[8]:

Advanced Content Moderation Algorithms: Social media platforms must invest in and continuously improve content moderation algorithms. Machine learning models can be trained to identify and flag potentially hateful content, helping human moderators in their efforts to maintain a safer online environment.

User Education and Awareness: Promoting digital literacy and educating users about the consequences of hate speech can be a powerful preventive measure. Encouraging responsible online behavior and fostering empathy can contribute to a more positive online culture [9].

Community Reporting and Moderation: Empowering users to report instances of hateful content encourages a sense of shared responsibility. Social media platforms can establish transparent reporting mechanisms and involve the community in content moderation processes [10].

Stricter Enforcement of Policies: Platforms need to enforce their community guidelines rigorously. Clear and unequivocal consequences for violating policies can act as a deterrent, sending a strong message against hate speech.
Collaboration with Civil Society [11]-[12]:

Collaboration between social media platforms, governments, and civil society organizations can lead to the development of comprehensive strategies to combat hateful content. Joint efforts can include research, policy development, and community outreach [13].

III. Proposed Algorithm

The proposed algorithm to be used for classification of hateful content is collaborative learning [14].
Training an optimized model is of major importance before it can be used to predict the outcome of the data inputs [16]. Neural Networks can be used for a variety of different purposes such as pattern recognition in large and complex data pattern sets wherein the computation of parameters would be extremely daunting for conventional statistical techniques. The weights or the equivalents of experiences are evaluated and updated based on the data patterns which are fed to the neural networks for training. The framework of collaborative learning consists of three major parts: the generation of a population of classifier heads in the training graph, the formulation of the learning objective, and optimization.

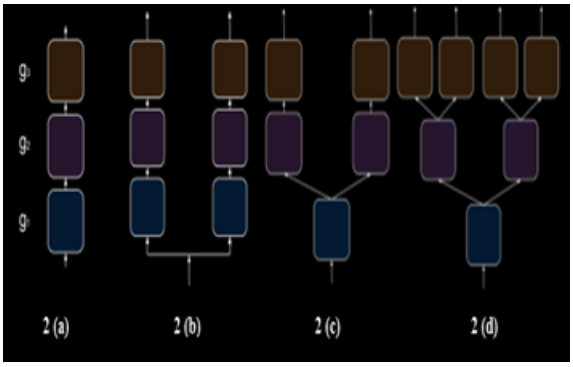

Fig.1 (a) Target Network (b) Multiple Instances (c) Simple ILR Sharing (d) Hierarchical ILRs haring

The proposed collaborative model is discussed next:
At inference time, only the original network is kept and all added parts are discarded. Unlike auxiliary training, each classifier head here has an identical network to the original one in terms of graph structure. This approach leads to advantages over auxiliary training in terms of engineering effort minimization. First, it does not require to design additional networks for the auxiliary classifiers [17]. Second, the structure symmetry for all heads does not require additional different weights associated with loss functions to well balance injected backpropagation error flows, because an equal weight for each head's objective is optimal for training. Mathematically:
The target network to be trained is given by:

$$\boldsymbol{z = g(x, \theta)} \quad \textbf{(1)}$$

Here,
g is determined by the graph architecture
θ represents the network parameters.
The term g can also be represented as the cascade of the following sub-nets, given mathematically by:

$$\boldsymbol{g(x, \theta) = g3(g2(g1(x1, \theta1), \theta2), \theta3)} \quad \textbf{(2)}$$

Here,
$g(x, \theta)$ denotes the generator polynomial.

The cascade of the network is often termed as **Ensemble Neural Network (ENN).**
Here,

$$\boldsymbol{\theta = [\theta1, \theta2, \theta3]} \quad \textbf{(3)}$$

In general, it is observed that that the training memory size is roughly proportional to the number of layers/operations. With the multi-instance pattern, the number of parameters in the whole training graph is proportional to the number of heads. Obviously, ILR sharing can proportionally reduce the memory consumption and speed up training, compared to multiple instances without sharing.

Back propagation's popularity stems from the fact that:

1) Its stable

2) Its fast.

The mathematical treatment of the weight updating rule can be given by:

Let $\Delta\boldsymbol{\omega}$ be the amount by which the weight is updated in every iteration. Then $\Delta\boldsymbol{\omega}$ is mathematically computed as:

$$\Delta\boldsymbol{\omega} = \boldsymbol{J^T J} + \boldsymbol{\mu I^{-1} J^T e} \qquad \textbf{(4)}$$

where

$\boldsymbol{\omega}$ the weight vector,

I is the identity matrix,

$\boldsymbol{\mu}$ is the combination coefficient,

(Q X R) X N the Jacobian matrix

Multilayer networks typically use sigmoid transfer functions in the hidden layers. These functions are often called "squashing" functions, because they compress an infinite input range into a finite output range [18]. Sigmoid functions are characterized by the fact that their slopes approach zero, as the input gets large. This causes a problem when we use steepest descent (gradient decent/ back propagation) to train a multilayer network with sigmoid functions, because the gradient can have a very small magnitude and, therefore, cause small changes in the weights and biases, even though the weights and biases are far from their optimal values [19].

To compute the features of all the tweets $(\boldsymbol{n_m})$, an additional summation is to be used and the final expression becomes:

$$\min(\boldsymbol{x_1 \ldots . x_{n_m}}) \sum_{i=1}^{n_m} \left[\sum_{j:r(i,j)=1}^{n} \left[\boldsymbol{\theta_j^T x^i - y^{i,j}}\right]^2 \Big/ \boldsymbol{n} \right] + \frac{\boldsymbol{\lambda}}{2} \sum_{i=1}^{n_m} \sum_{i=1}^{n} [\boldsymbol{x_i}]^2 \qquad \textbf{(5)}$$

The above cost function is to be minimized. So, given values of θ, we can estimate x. Equation 5 depicts the collaborative learning model for the neural architecture.

IV. Results and Discussions:

The system has been implemented on MATLAB using the Deep Learning Toolbox.

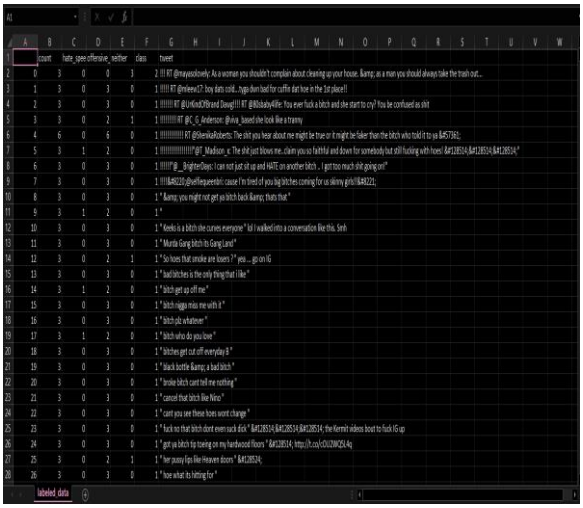

Fig.2 Annotated Data

The figure depicts the annotated dataset for classification.

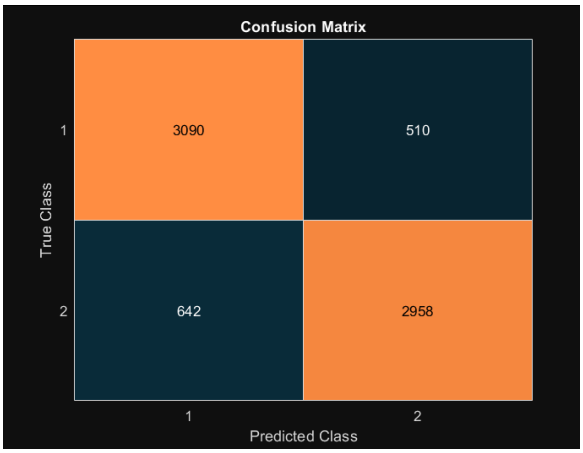

Fig.3 Confusion Matrix

The confusion matrix renders information about the TP, TN, FP and FN rates [20].

The accuracy of classification is computed as:

$$Ac = \frac{TP + TN}{TP + TN + FP + FN}$$

Substituting values,

$$Ac = \frac{3090 + 2958}{3090 + 2958 + 510 + 642} = 84\%$$

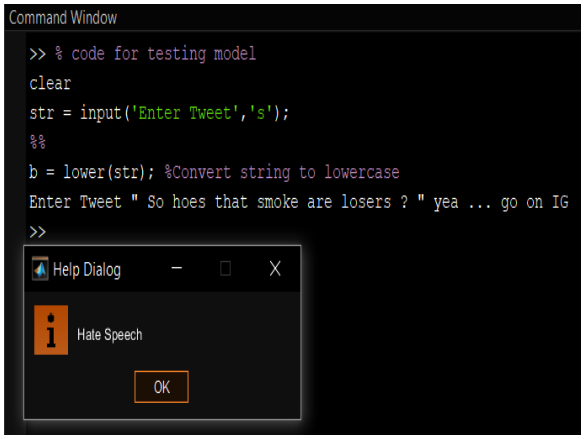

Fig.4 Prototype Testing

Figure above depicts the prototype testing.

The summary of parameters is presented in table 1, for ready reference.

S.No.	Parameter	Value
1	Dataset	Kaggle
2	No of samples	24784
3	Model	Collaborative Learning
4	Algorithm	Resilient Back Prop
5	Classification Accuracy	84%
6	Accuracy of Previous Work [18]	78.11% (best case)

It can be observed that the proposed approach attains higher classification accuracy compared to existing models.

Conclusion:

Previous discussions suggest that Hateful content on social media is a pressing issue that demands collective action. As we navigate the complexities of the digital age, it is essential to strike a balance between freedom of expression and the need for responsible online behavior. By implementing advanced moderation technologies, fostering user education, and promoting collaborative efforts, we can strive to create a digital landscape that is inclusive, respectful, and conducive to meaningful connections. The responsibility lies not only with social media platforms but with each individual user who contributes to the rich tapestry of online discourse. The proposed approach presents a collaborative learning model for hate speech classification. It has been shown that the proposed approach outperforms existing models in terms of classification accuracy.

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] A. Arango, J. Pérez, and B. Poblete, ‘‘Hate speech detection is not as easy as you may think: A closer look at model validation (extended version),’’ Inf. Syst., vol. 105, Mar. 2022, Art. no. 101584.

[17] A. Schmidt and M. Wiegand, ‘‘A survey on hate speech detection using natural language processing’’ in Proc. 5th Int. Workshop Natural Lang. Process. Social Media, 2017, pp. 1–10

[18] 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, Art.no.1198.

[19] F Husain, O Uzuner, “Investigating the effect of preprocessing arabic text on offensive language and hate speech detection”, ACM Transactions on Asian and Low Resource Language Information Processing vol.21, no.4, pp.1-20

[20] M Lansley, F Mouton, S Kapetanakis, “SEADer++: social engineering attack detection in online environments using machine learning”, Journal of Information and Telecommunication, Taylor and Francis, 2020, vol.4, no.3, pp.346-362