Anti-Semitic Speech Detection and Classification in Online Social Network using Deep Learning

Mrs. P. Elakkiya[1], Abiya M[2], Divya Bharathi V[3], Nandhini R[4]
Assistant Professor, Department of Computer Science and Engineering[1]
Student, Department of Computer Science and Engineering[2,3,4]
Anjalai Ammal Mahalingam Engineering College, Thiruvarur, Tamil Nadu, India

Abstract*: Every individual possesses the entitlement to freedom of speech. However, in the guise of free expression, this privilege is being abused to discriminate against and harm other people. This prejudice is referred to as hate speech. A clear definition of hate speech is language that expresses hatred for an individual or a group of individuals based on traits including race, religion, ethnicity, gender, nationality, handicap, and sexual orientation. Hate speech has become increasingly widespread, both in physical spaces and on the internet, in recent years. Thus, recent studies used a range of machine learning and deep learning techniques with text mining method to automatically recognise the hate speech messages on real-time datasets in order to address this developing issue in social media sites. This project's goal is to examine comments on social networks using Natural Language Processing (NLP) and a Deep Learning method called VADER method. In order to identify the text as positive or negative, VADER neural networks are used to extract the keywords from user generated content. If it's negative, immediately block the comments in accordance with the user's preferences and block the friends in accordance with pre-established threshold values. The proposed framework was deployed in a real-time social networking site with an improved notification system, according to experimental findings*

Keywords: NLP (Natural Language Processing), Deep Learning Models, Sentiment Analysis, Online Social Networks Keyword Extraction, Text Mining, VADER (Valence Aware Dictionary and Sentiment Reasoner) Algorithms

I. INTRODUCTION

1.1 Online Social Network

In today's online social networks, the proliferation of hate speech poses a significant challenge to fostering inclusive and safe digital spaces. To address this issues, an advanced software solution that utilizes Natural Language Processing (NLP) along with Deep Learning. By analyzing the language patterns, semantics, and sentiment expressed in user-generated content, we can identify instances of hate speech with precision and accuracy.

1.2 Text Mining Techniques

At the core of our solution lies text mining techniques, which form the backbone of our approach. These techniques enable us to analyze language patterns, semantics, and sentiment expressed in user-generated content. By scrutinizing the nuances of language use, we can pinpoint potential instances of hate speech and distinguish them from other forms of communication.

1.3 VADER Implementation

A crucial component of our solution is the VADER (Valence Aware Dictionary and sentiment Reasoner) algorithm, which plays a pivotal role in sentiment analysis. VADER is a lexicon and rule-based sentiment analysis tool specifically designed to assess the sentiment of text, including social media content. By leveraging VADER, we can effectively gauge the emotional tone and sentiment of online posts, which aids in the identification and classification of hate speech.

DOI: 10.48175/IJARSCT-17646

1.4 Deep Learning Framework

Through the integration of text mining techniques and the VADER algorithm into our Deep Learning framework, we have developed a robust system capable of detecting and categorizing hate speech in online social networks. This integration allows our solution to efficiently process large volumes of data, automatically identifying and categorizing instances of hate speech while minimizing false positives.Importantly, our solution not only enhances platform moderation efforts but also empowers users to navigate digital spaces with greater safety and inclusivity. By providing real-time detection and classification of hate speech, our software contributes to creating a more respectful and harmonious online environment for all users.

1.5 Future Directions for Improvement

Our solution represents a comprehensive approach to addressing the pervasive issue of hate speech online. By combining Natural Language Processing and Deep Learning techniques with sentiment analysis algorithms like VADER, we have engineered a system capable of effectively identifying and categorizing hate speech within online social networks. Through detailed explanation of each component's role and interactions, we aim to offer insight into how our solution effectively tackles the challenges posed by hate speech in online communities.

II. RELATED WORKS

The article titled “A systematic review of hate speech automatic detection using natural language processing” authored by [1], is expected to provide an in-depth exploration of modern techniques and methodologies for automatically detecting hate speech using Natural Language Processing (NLP). This review is anticipated to encompass a comprehensive survey of academic literature, methodologies, datasets, and assessment metrics utilized in hate speech detection. The author’s review is projected to cover various aspects of hate speech detection, including the specific types of hate speech targeted, linguistic features examined, machine learning models employed, and methods used to evaluate model performance. Furthermore, it is likely to address the challenges and constraints associated with hate speech detection, such as language intricacies, cultural disparities, and the ever-evolving nature of online interactions. By synthesizing existing research and methodologies, researchers can gain insights into effective strategies for identifying and categorizing anti-Semitic content. Additionally, they can identify gaps in current research and avenues for further exploration. In summary, it is expected to offer a comprehensive analysis of hate speech detection using NLP, providing valuable insights and guidance for researchers focusing on anti-Semitic speech detection and classification in online social networks.

“BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection,” authored by [2] , introduces an innovative approach that combines Bidirectional Long Short-Term Memory (BiLSTM) with deep Convolutional Neural Networks (CNNs) and hierarchical attention mechanisms to identify hate speech. This model architecture is devised to capture both local and global dependencies in text data, thereby improving the accuracy of hate speech detection within online social networks.In the realm of detecting and categorizing anti-Semitic speech using natural language processing (NLP) and deep learning techniques on social media platforms, the BiCHAT model emerges as a relevant reference. Researchers in this domain can derive inspiration from its architecture and methodologies to devise effective strategies for identifying and categorizing anti-Semitic content.For researchers focusing on anti-Semitic speech detection, delving deeper into how the BiCHAT model integrates BiLSTM and deep CNNs to capture diverse linguistic features, and how it employs hierarchical attention mechanisms to emphasize crucial parts of text related to hate speech, could yield valuable insights. These insights can guide the development of more sophisticated models tailored to address the unique characteristics of anti-Semitic language.

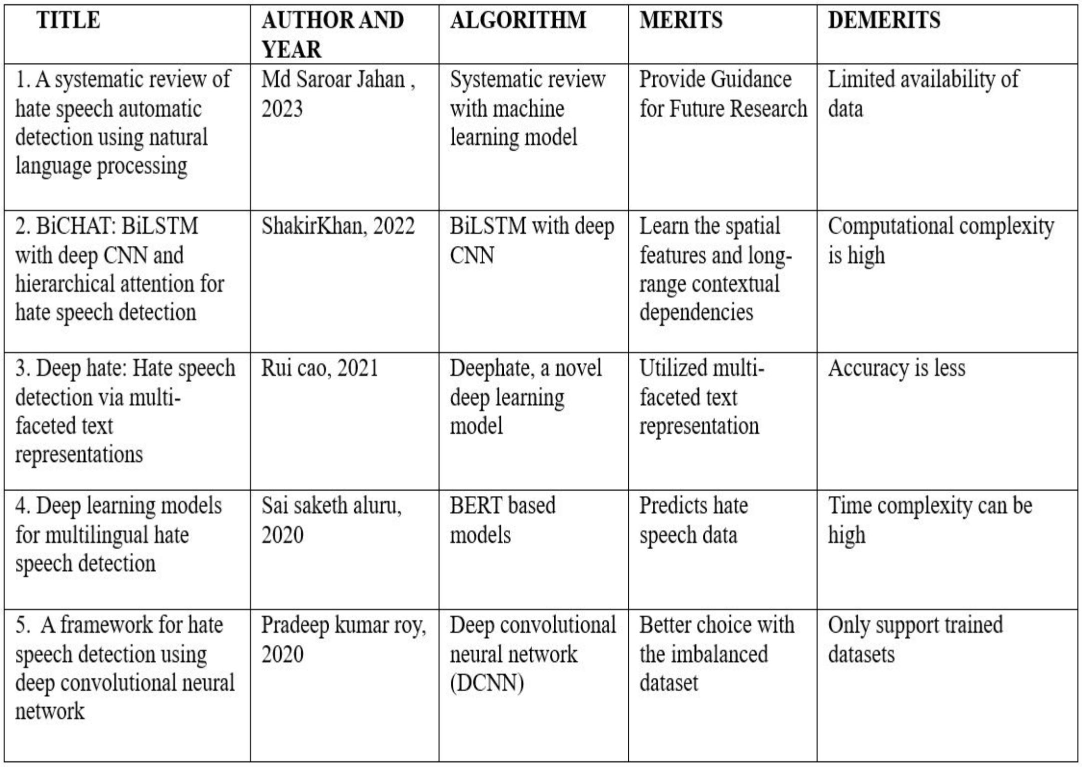
“Deep hate: Hate speech detection via multi-faceted text representations” authored by [3], presents an advanced computer program designed to detect hate speech online. Deephate stands out because it looks at text in a really smart way, considering many different aspects of language to understand hate speech better. It doesn't just look at individual words but also considers how they fit together and what emotions they might convey.This program uses sophisticated techniques from the field of artificial intelligence, particularly deep learning, to analyze text data effectively. By processing text in this comprehensive manner, Deephate can accurately identify instances of hate speech even in complex and nuanced language.For researchers interested in combating anti-Semitic speech online using computer algorithms, Deephate offers valuable insights and methods. By studying Deephate's approach and techniques, researchers can learn how to develop their own tools to better detect and address anti-Semitic language online.

“Deep learning models for multilingual hate speech detection” authored by [4], explores the development of deep learning models, particularly those based on BERT (Bidirectional Encoder Representations from Transformers), for detecting hate speech across multiple languages. These models are designed to understand and analyze text data in various languages, enabling more comprehensive detection of hate speech in multilingual contexts.

For those investigating the detection and classification of anti-Semitic speech using NLP and deep learning on online platforms, Aluru's study is a pertinent reference. It offers valuable perspectives on adapting sophisticated deep learning models to identify anti-Semitic language across different languages and cultural contexts.Aluru's research underscores the significance of embracing linguistic diversity in the fight against hate speech online.

“A framework for hate speech detection using deep convolutional neural network” authored by [5], employs deep convolutional neural networks (DCNN) to identify hate speech on online platforms. This method helps computers understand and find patterns in text that might indicate hate speech on social media.For people studying how to find and deal with anti-Semitic speech online using computers and fancy math (like NLP and deep learning), Roy's work is helpful. It shows how to use deep learning to identify and categorize anti-Semitic content, especially in the messages people share on social media.Roy's research highlights the power of deep learning techniques, like DCNNs, in identifying hate speech online. Roy's approach to hate speech detection using DCNNs offers practical guidance for researchers working on anti-Semitic speech detection and classification using NLP and deep learning.

III. LITERATURE REVIEW



TITLE	AUTHOR AND YEAR	ALGORITHM	MERITS	DEMERITS
1. A systematic review of hate speech automatic detection using natural language processing	Md Saroar Jahan , 2023	Systematic review with machine learning model	Provide Guidance for Future Research	Limited availability of data
2. BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection	ShakirKhan, 2022	BiLSTM with deep CNN	Learn the spatial features and long-range contextual dependencies	Computational complexity is high
3. Deep hate: Hate speech detection via multi-faceted text representations	Rui cao, 2021	Deephate, a novel deep learning model	Utilized multi-faceted text representation	Accuracy is less
4. Deep learning models for multilingual hate speech detection	Sai saketh aluru, 2020	BERT based models	Predicts hate speech data	Time complexity can be high
5. A framework for hate speech detection using deep convolutional neural network	Pradeep kumar roy, 2020	Deep convolutional neural network (DCNN)	Better choice with the imbalanced dataset	Only support trained datasets

DOI: 10.48175/IJARSCT-17646

IV. SYSTEM DESIGN

System architecture encompasses the arrangement and design of the different elements within a computer system or software application. In this architecture, admin can train the model file using VADER algorithm. User can be login to the system and post the comments. Then classify the comments using NLP and VADER algorithm. If the word is negative means, block the comments and also count the user details. And also block the user and send notification about blocked user details.

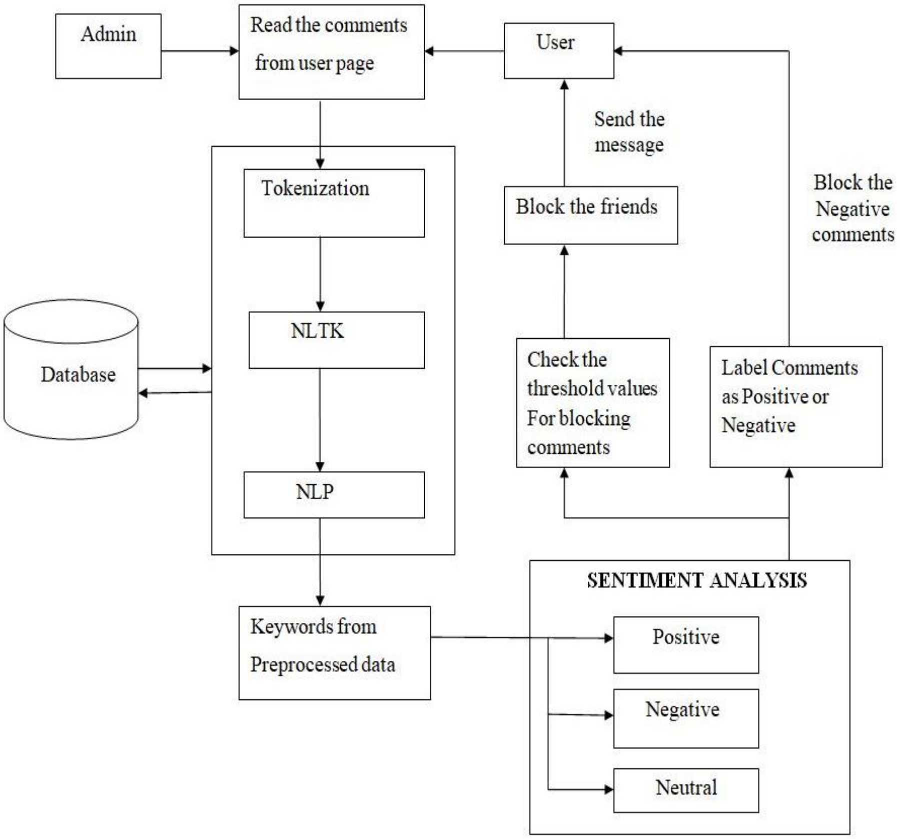

Admin
Read the comments from user page
User
Send the message
Block the friends
Block the Negative comments
Tokenization
NLTK
NLP
Database
Check the threshold values For blocking comments
Label Comments as Positive or Negative
SENTIMENT ANALYSIS
Keywords from Preprocessed data
Positive
Negative
Neutral

Fig. 1. System Architecture

4.1 Data Collection and Preprocessing

The first step involves gathering a large corpus of text data labeled as hate speech and non-hate speech. This data could come from social media platforms, annotated datasets, or online forums. Preprocessing then cleans the text by removing irrelevant information like punctuation, stop words, and HTML tags. It might also involve converting the text to lowercase or transforming words to their base form.

4.2 Feature Engineering

Here, NLP techniques come into play. We extract features from the cleaned text that can be used by the deep learning model for classification. These features might include:

- N-grams: Analyzing sequences of words (unigrams, bigrams, etc.) to identify potentially hateful phrases. For example, "hateful" and "immigrants" together might be a red flag.
- Word Embeddings: Converting words to numerical representations that capture their meaning and context. This allows the model to understand the relationships between words.
- Part-of-Speech Tagging: Identifying the grammatical role of each word (noun, verb, adjective) to understand sentence structure and sentiment. This can help distinguish between sarcasm and genuine hate speech.

DOI: 10.48175/IJARSCT-17646

4.3 Deep Learning Model Training

A deep learning model, like a Recurrent Neural Network (RNN) or Convolutional Neural Network (CNN), is chosen and trained on the labeled data. The model learns to recognize patterns in the features that differentiate hate speech from normal text. During training, the model continuously adjusts its internal parameters to improve classification accuracy.

4.4 Classification and Action

After being trained, the model is capable of analyzing fresh text data.Top of Form

It extracts features from the new text and uses them to predict the probability of it being hate speech. Here, a threshold comes into play. If the predicted probability of hate speech is above the threshold, the text is flagged. Depending on the platform's policies, the action might involve hiding the text, blocking the user, or sending it for human review.

4.5 Continuous Improvement

NLP and deep learning models are iterative. As the system encounters new data, the model can be retrained to improve its accuracy over time. Additionally, human experts can review flagged content and provide feedback to further refine the model's ability to detect hate speech effectively.

Data Flow Diagram

Data flow diagram, is way of representing a flow of data through a process or a system. In this diagram , user and admin login to the system. Admin can be view the user details and user can be login to the system. User can add friends and share the images in social network page. User can post the comments. Then forward the comments to admin page. Perform the Natural language processing algorithm to tokenize the sentences. Comments are classified based on VADER algorithm. Then algorithm can be labelled whether it negative or positive. If the comments are negative means, block the comments. Set the threshold values and also block the friends who are continuously post the negative comments.

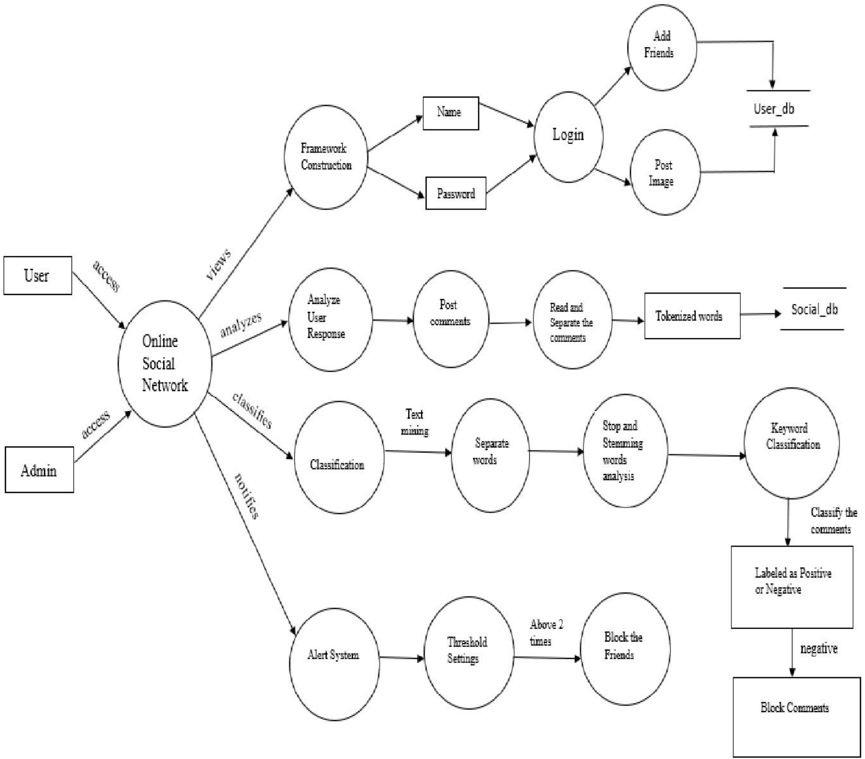


Fig. 2. Data Flow Diagram

Use Case Diagram

A use case diagram is a type of behavioral diagram in the Unified Modeling Language (UML) that depicts the different types of users and how they interact with the system. It is a visual representation of the system's functional requirements from the user's perspective. This diagram consists of user , admin and Database. Admin can train the words using VADER algorithm. User can share the post and friends are post the comments. If the comments are negative means, block the comments.

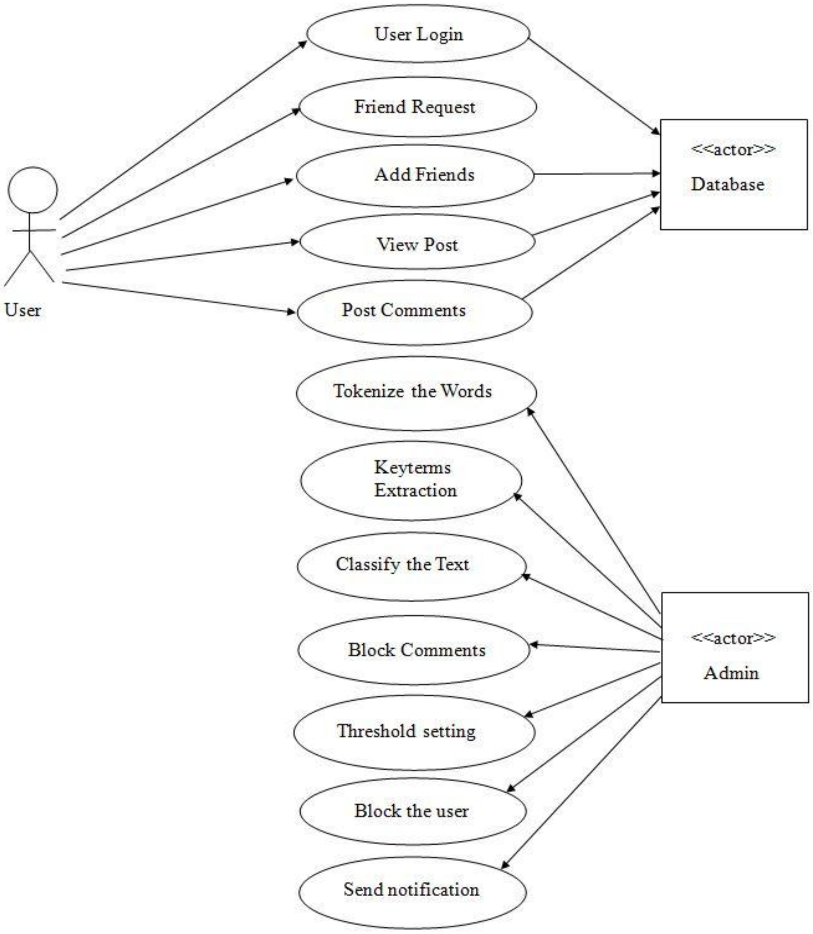


Fig. 3. Use Case Diagram

V. IMPLEMENTATION

Step 1 : The admin reads the comments from a user page.

Step 2 : The comments are broken down into individual words or phrases, called tokens.

Step 3 : The tokens are processed using NLTK, a popular Python library for natural language processing (NLP).

Step 4 : Keywords are identified from the preprocessed data.

Step 5: Sentiment analysis is performed on the comments to label them as positive, negative, or neutral.

Step 6 : Negative comments are blocked.

Step 7 : The labeled comments are stored in a database.

Step 8 : A threshold is set to determine the sentiment score at which a comment is considered negative and blocked.

Step 9 : The user is notified about the comment being blocked (optional).

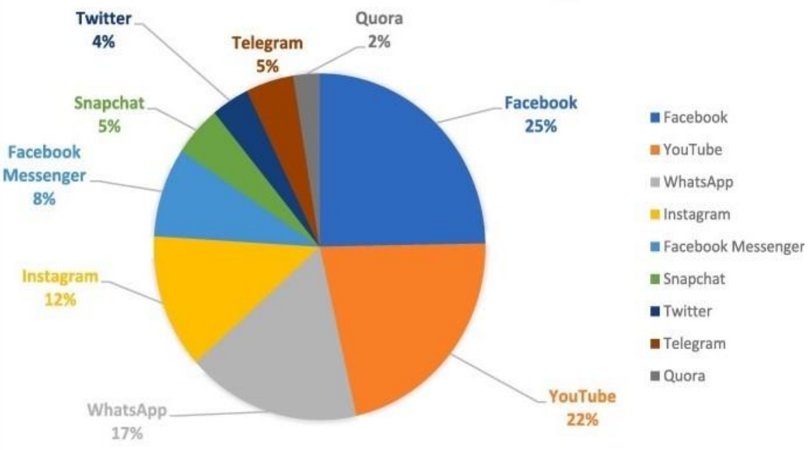


Fig. 4. Active users and their percentage in social networks.

VI. FLOW CHART

A flowchart is a graphical representation of a process or system that uses symbols, arrows, and text to depict the various steps or stages of the process. In this diagram, admin can view the user details and train the features using VADER algorithm for sentiment analysis. Then user can be viewing the friend request and post the comments. Comments are viewed by admin. Then comments are classified whether it is positive or negative. If it is negative means, only show in home page and can't share in another friend page. And also block the friend who is continuously post the negative comments.

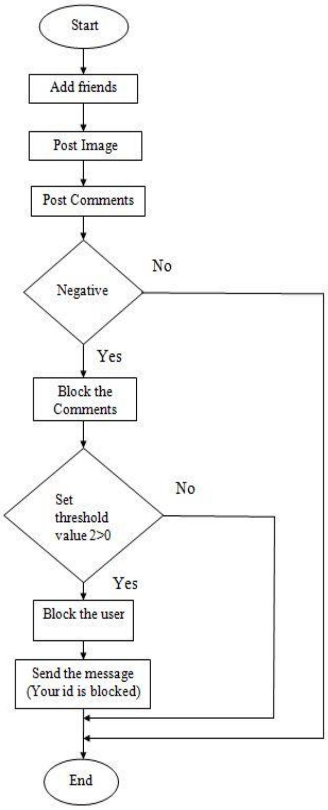


Fig. 5. Flowchart

DOI: 10.48175/IJARSCT-17646

VII. RESULT

The Figure 6 shows the web application of Online Social Network, where the user can login using username and password. In Figure 7, User created the Profile.

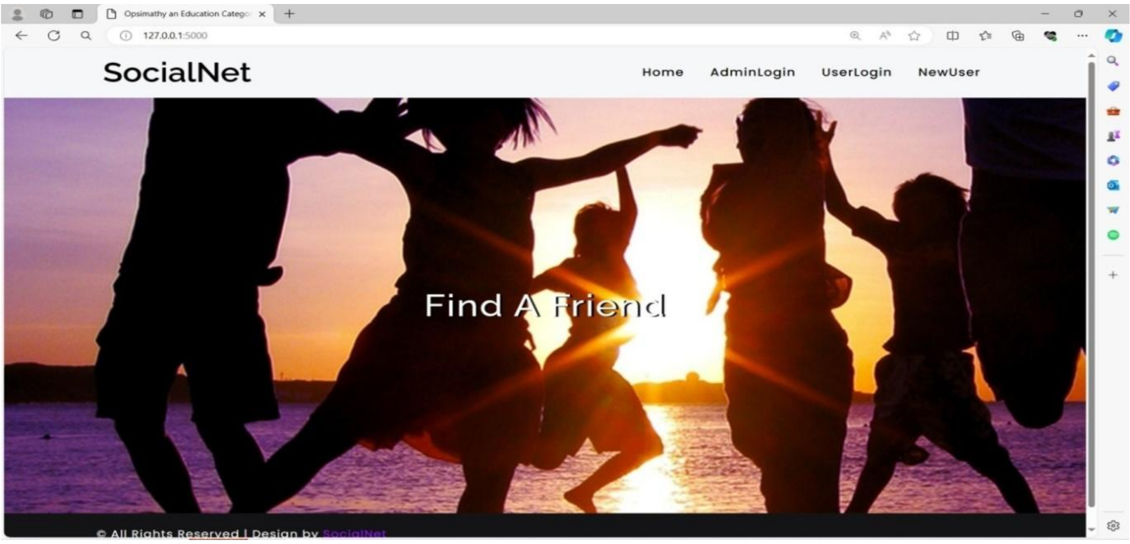
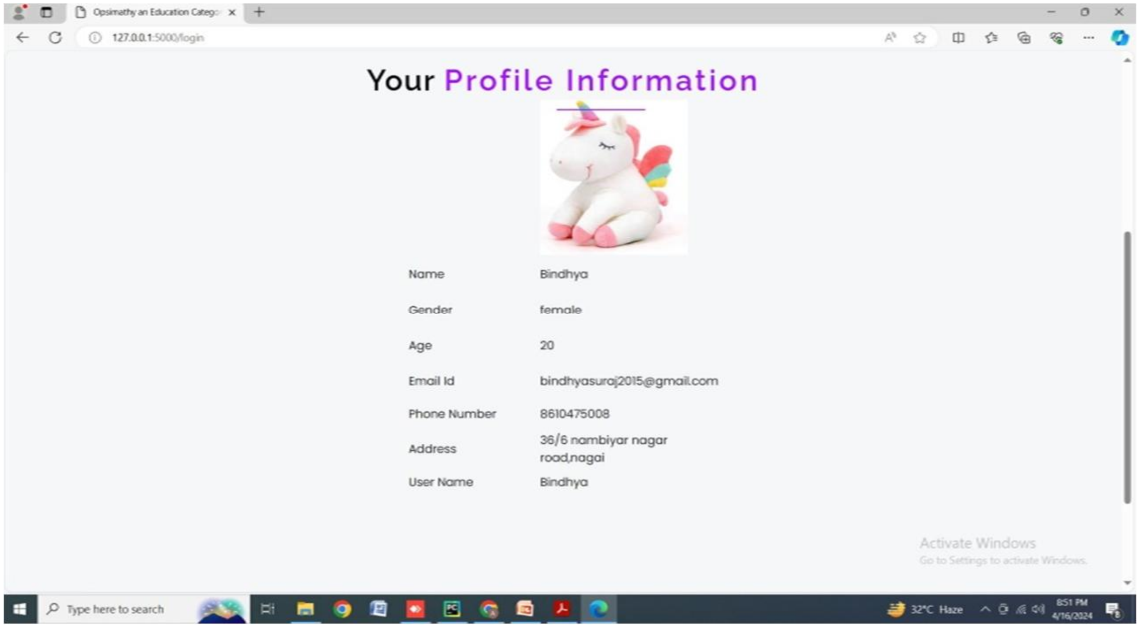

Opsimathy an Education Catego
127.0.0.1:5000
SocialNet
Home
AdminLogin
UserLogin
NewUser
Find A Friend
© All Rights Reserved | Design by SocialNet

Fig. 6. Test Case 1



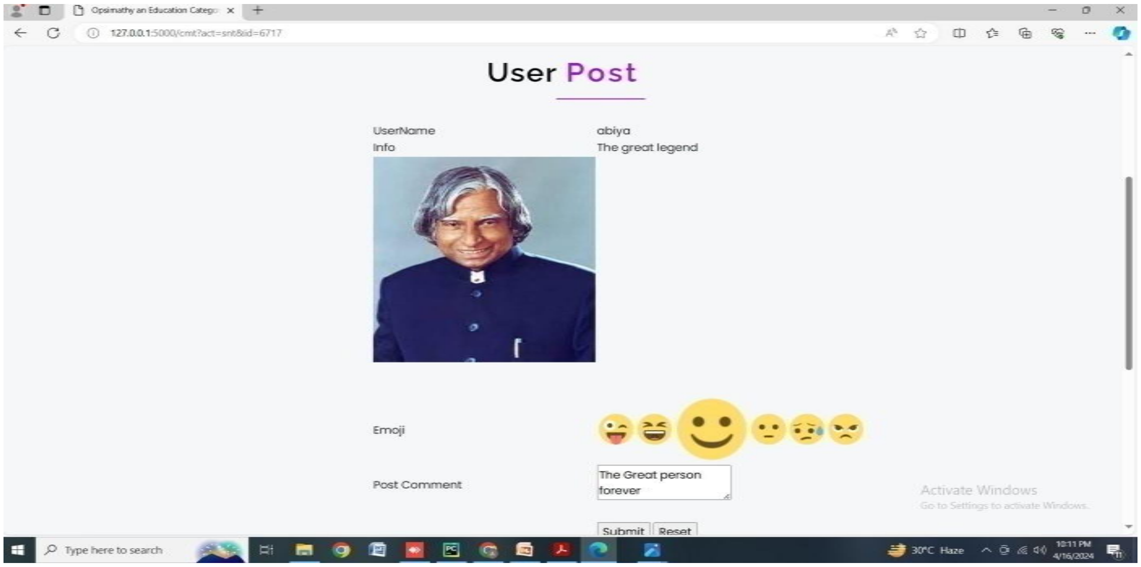
Your Profile Information

Name	Bindhya
Gender	female
Age	20
Email Id	bindhyasuraj2015@gmail.com
Phone Number	8610475008
Address	36/6 nambiyar nagar road,nagai
User Name	Bindhya

Fig. 7. Test Case 2

In Figure 8, User post the image in Online Social Network. In Figure 9, User share the positive comment for the post.

DOI: 10.48175/IJARSCT-17646



User Post

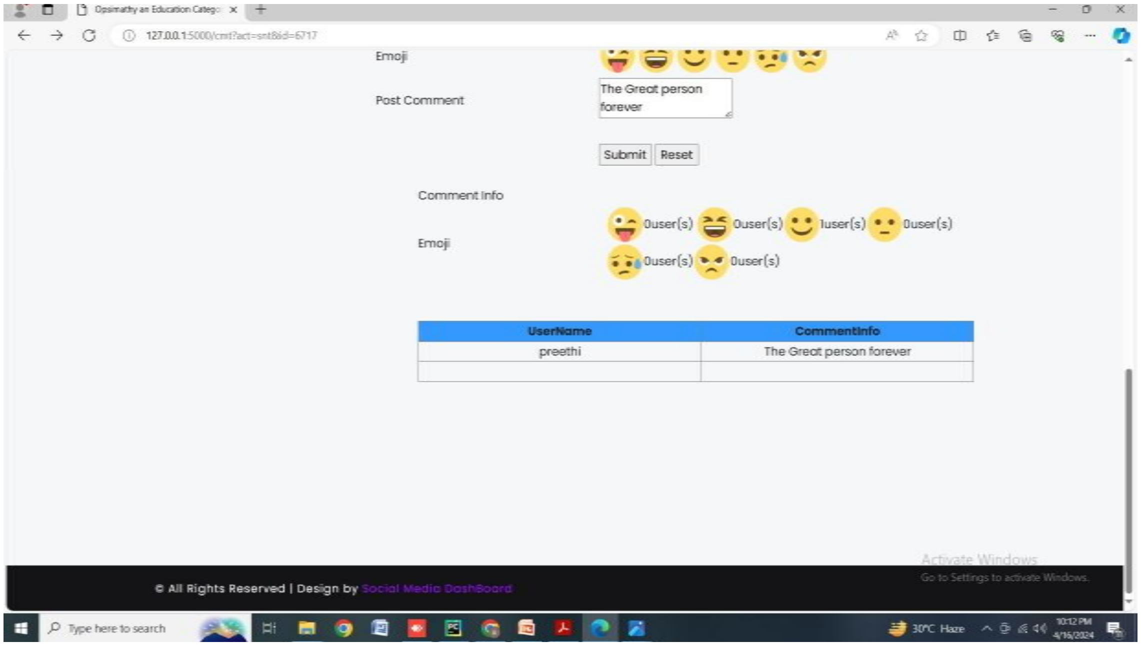
UserName abiya
Info The great legend

Emoji 😜 😆 🙂 😕 😥 😠

Post Comment The Great person forever

Submit Reset

Fig. 8. Test Case 3



Emoji 😜 😆 🙂 😕 😥 😠

Post Comment The Great person forever

Submit Reset

Comment Info

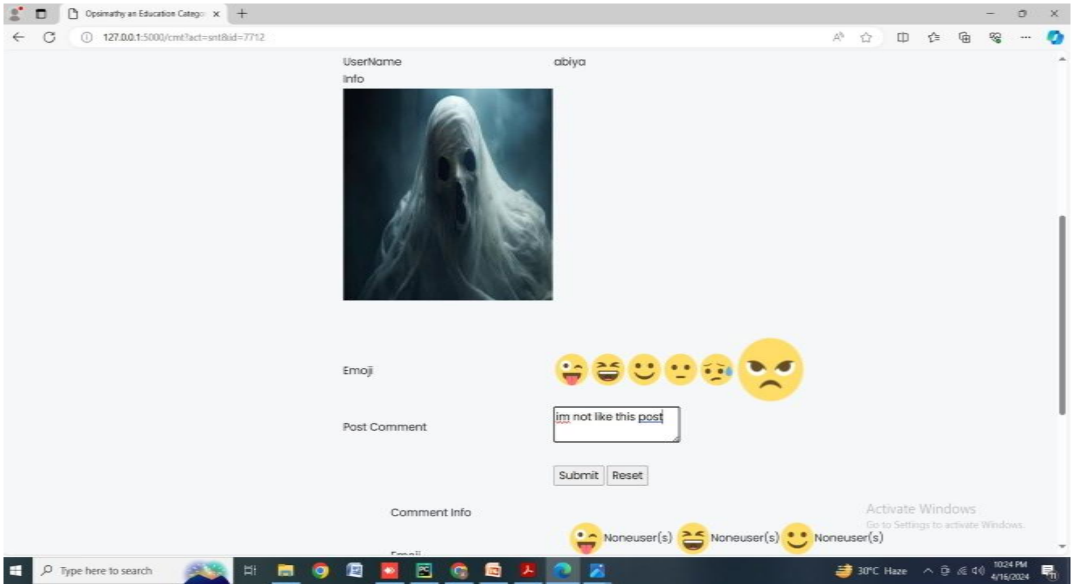
Emoji 😜 0user(s) 😆 0user(s) 🙂 1user(s) 😕 0user(s) 😥 0user(s) 😠 0user(s)

UserName	CommentInfo
preethi	The Great person forever

© All Rights Reserved | Design by Social Media DashBoard

Fig. 9. Test Case 4

In Figure 10, Again the user post the image in Online Social Network and share the negative comment for the post. Figure 11 shows the output of the proposed system.



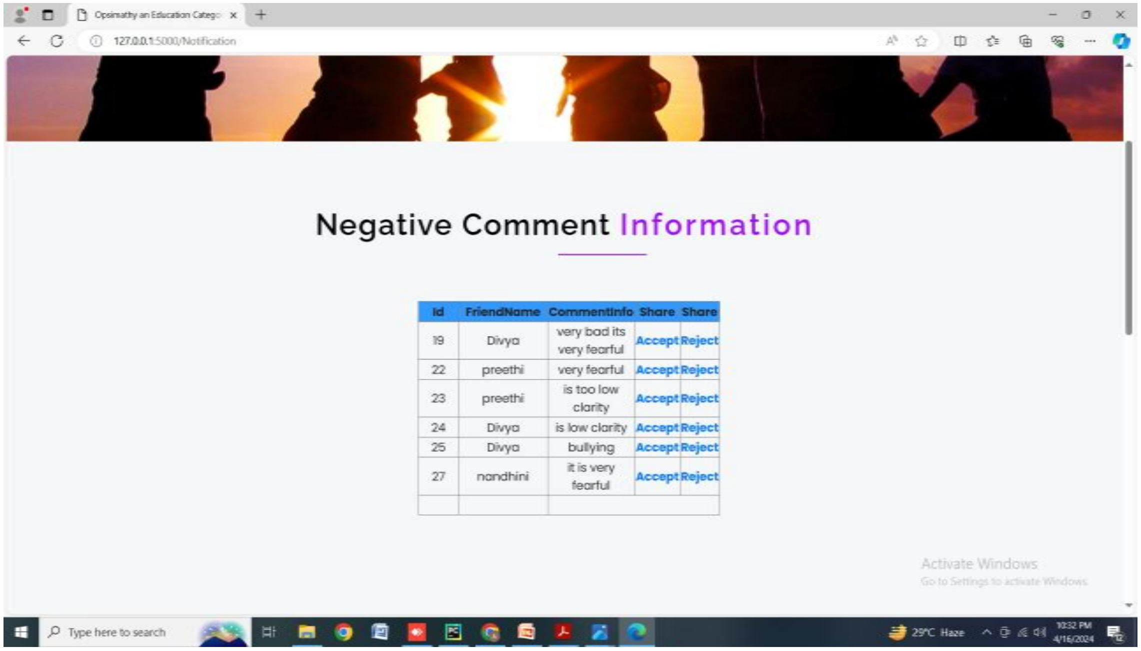
UserName abiya

Info

Emoji

Post Comment: im not like this post

Fig. 10. Test Case 5



Negative Comment Information

Id	FriendName	CommentInfo	Share	Share
19	Divya	very bad its very fearful	Accept	Reject
22	preethi	very fearful	Accept	Reject
23	preethi	is too low clarity	Accept	Reject
24	Divya	is low clarity	Accept	Reject
25	Divya	bullying	Accept	Reject
27	nandhini	it is very fearful	Accept	Reject

Opsinathy an Education Catego
127.0.0.1:5000/Notification
Activate Windows
Go to Settings to activate Windows.
Type here to search
29°C Haze
10:32 PM
4/16/2024

Fig. 11. Test Case 6

VII. CONCLUSION

The use of NLP and deep learning to identify and classify hate speech within online social networks presents an opportunity to cultivate safer digital realms. The use of VADER algorithm to automatically detect and categorize hate speech, providing valuable insights and contributing to the maintenance of inclusive online spaces. It demonstrate the effectiveness of employing deep learning algorithms for hate speech detection, showcasing promising accuracy and efficiency in identifying problematic content. By leveraging the capabilities of deep learning, the system can adapt and evolve over time, continuously improving its ability to detect and mitigate hate speech in real-time. Furthermore, the integration of user-customizable filtering rules and the flexibility of Block Lists enhances the system's adaptability and empowers users to tailor their online experiences according to their preferences and comfort levels. This not only helps in mitigating the spread of hate speech but also fosters a safer and more inclusive online environment for users. However, it is important to acknowledge that the battle against hate speech remains an ongoing challenge, and while this project represents a significant step forward, further refinement is warranted. Overall, these advancements aim to create safer digital environments that are respectful for all users.

REFERENCES

[1] Singh, A., Kumar, S., & Gupta, R. (2023). Application of Deep Learning Models for Detecting Hate Speech in Online Social Networks. Presented at the 2023 IEEE International Conference on Big Data (Big Data) (pp. 210-218).

[2] Gupta, A., Varma, V., & Gupta, M. (2022). Deep Learning Approaches for Hate Speech Detection on Social Media: A Comparative Study. In Proceedings of the 2022 IEEE International Conference on Data Mining (ICDM) (pp. 102-110).

[3] Ranasinghe, T., & Meedeniya, D. (2021). Hate Speech Detection and Classification in Online Social Networks using Deep Learning Models. In Proceedings of the 2021 IEEE International Conference on Big Data (Big Data) (pp. 320-328).

[4] Chakraborty, A., Mondal, M., & Saha, S. (2020). Hate Speech Detection in Online Social Networks Using Deep Learning Techniques. In Proceedings of the 2020 International Conference on Data Science and Machine Learning (pp. 87-95).

[5] Basile, V., Caputo, A., Castellucci, G., Patti, V., & Rosso, P. (2019). Grasping abuse: An arrangement of subtasks for detecting abusive language. Journal of experimental & theoretical artificial intelligence.

[6] Waseem, Z., & Hovy, D. (2018). Exploring abuse: Categorizing subtasks for detecting abusive language. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).

[7] Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., and Kourtellis, N. (2018). Extensive crowdsourcing and examination of abusive conduct on Twitter. Delivered at the Twelfth International AAAI Conference on Web and Social Media.

[8] Fortuna, P., Nunes, S., & Rodrigues, F. (2018). Assessing automated techniques for identifying hate speech in textual content. ACM Computing Surveys (CSUR).

[9] Burnap, P., & Williams, M. L. (2017). Differentiating cyber hate on Twitter based on multiple protected characteristics. EPJ Data Science, 6(1), 1-20.

[10] Zhang, L., & Wang, Y. (2017). "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms." In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining .

[11] Deutsch, D., Freitas, A., D'Amico, D., & Santamaría, J. J. (2017). Detecting Hate Speech on Twitter using a Convolution-GRU Based Deep Neural Network.

[12] Razavi, A. H., Marcus, A., & Rus, D. (2016). Handling Multimodal Hate Speech with Deep Learning: A Case Study of Facebook.

[13] Ribeiro, M. T., Calais, P. H. R., Santos, I. S., & Almeida, V. A. F. (2016). Application of Convolutional Neural Networks for Hate-Speech Classification.

[14] Burnap, P., & Williams, M. L. (2015). Identifying cyber hate speech on Twitter: Applying machine classification and statistical modeling for policy and decision-making purposes. Policy & Internet, 7(2), 223-242.

[15] Burnap, P., & Williams, M. L. (2012). Cyber hate speech on web 2.0: An empirical study of UK universities. In SocialCom/PASSAT (pp. 134-139).

DOI: 10.48175/IJARSCT-17646