JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING MODEL FOR HATE SPEECH DETECTION IN ONLINE CONTENT.

Vinay.N ,Ponmali.S ,Prasanna.S

Department of Computer Science and Engineering

Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College Avadi-600062

I. ABSTRACT

People now utilize the internet as a major medium to voice their opinions. Blogging and social networking websites allow users to post whatever they desire. There are times, though, when specially designed content for a group of individuals to create hatred or discrimination is brought to the forefront, leading to issues in society. This kind of content is known as hate speech. Hate speech can diminish social peace and harmony to a large extent. On some occasions, it has led to extremism and social conflict. Explained are the various types of hate speech—e.g., racism, sexism, religious hate speech, etc.—and the ways used to fight them. We also define the problem and the solution for hate speech detection on the public internet. Consequently, internet hate speech needs to be tracked. In this case, we review related work in the hate speech detection field.

In addition to identifying hate speech on the internet, our suggested methodology classifies it as "hate speech," "offensive speech," "fair speech," etc. As data collected is processed, hate speech reports could be created, making internet usage more accessible to users of all types.

KEYWORDS: Hate Speech, Social media platforms, Discrimination, Extremism, Racism, Sexism, Religion.

II. INTRODUCTION

In recent years, there have been numerous studies attempting to break the harmful impact of hate speech on social media. It is not unusual to disagree on what constitutes hate speech and what is a hostile message. Hate speech refers to speech that denigrates or disparages groups based on their gender identity, religion, physical characteristics, or other factors. It can be clever or use high-brow language, among others. This assertion underscores how subjective it is to define hatred. Whether a message is insulting or discriminatory or not relies on how it is received by the recipient.

Hate Speech Detection (HSD) automatically identifies hate speech via machine learning and natural language processing. Our intention is to remove harmful content from the internet and social media so that everyone can utilize them in a safer and more friendly platform. With text datasets labeled with training models to identify hate speech features, HSD is often a binary hate-versus-non-hate text classification problem. The widespread use of transformer-based machine learning models was revealed by a recent text-based hate speech detection system evaluation. They are powerful models that have been shown to perform exceptionally well on a variety of natural language processing tasks.

This research offers a summary of detecting hate speech through machine learning techniques and highlights possible challenges for forthcoming studies in this quickly developing field.

III. LITERATURE REVIEW

A. A framework for hate speech detection using deep convolutional neural network.

In their 2020 paper, Roy and colleagues introduced a way to spot hate speech using deep Convolutional Neural Networks (CNNs). They started by turning words into numbers, a process called word embedding, so the computer could understand the text. Then, they used CNNs to automatically pick out important patterns from the text. Think of it like a detective looking for clues within the words themselves. This helped the computer learn what kind of language is often used in hate speech, without needing someone to manually point out every single clue. Finally, the computer used this knowledge to decide if a piece of text was hateful or not, showing that CNNs are pretty good at figuring out what makes hate speech stand out.

B. Advances in machine learning algorithms for hate speech detection in social media

In their 2021 study, Mullah and Zainon reviewed the progression of machine learning techniques for identifying hate speech on social media. Their research encompassed both traditional methods, such as Support Vector Machines and Naive Bayes, and more modern deep learning approaches, including CNNs, RNNs, and Transformers. The review provided a critical examination of the strengths and weaknesses of these various algorithms, with a focus on data preprocessing, feature engineering strategies, and the evolution of methodologies.

C. Hate speech detection using convolutional and Bi-directional bgated recurrent unit with Capsule network

Khan et al. (2022) proposed HCovBi-caps, a hybrid deep learning model that combines Convolutional Neural Networks (CNNs), Bi-directional Gated Recurrent Units (Bi-GRUs), and Capsule Networks for hate speech detection. Their methodology involved using CNNs to extract local features, Bi-GRUs to capture sequential dependencies, and Capsule Networks to model hierarchical relationships and spatial information. This integrated approach aimed to capture both local and global contextual information, as well as the hierarchical structure of hate speech, thereby improving detection accuracy.

D. A systematic review of Hate Speech automatic detection using Natural Language Processing.

In their 2023 paper, Jahan and Oussalah offer a thorough systematic review of the field, with a specific focus on the automatic detection of hate speech using Natural Language Processing (NLP) techniques. This review carefully examines the evolution and current state of NLP-driven methodologies, covering a wide range of approaches from traditional feature-based methods to modern deep learning models. By synthesizing a large body of research, the authors identify key trends, challenges, and future directions in hate speech detection. Their work highlights the increasing reliance on sophisticated models like Transformers while also emphasizing the ongoing need for robust datasets, effective multi-lingual solutions, and the critical integration of ethical considerations in system development. This systematic overview serves as a valuable resource, providing a structured understanding of the advancements and ongoing complexities in this critical research area.

IV. PROPOSED METHODS

Current techniques for detecting hate speech employ various technologies and strategies to identify and mitigate harmful content shared on online platforms. One common approach is Natural Language Processing (NLP), which allows systems to analyze text data and recognize patterns indicative of hate speech. NLP techniques utilize tools such as sentiment analysis, topic modeling, and word embeddings to comprehend the meaning and context of language used in online discussions.

A. Machine learning models:

Hate speech detection is typically carried out using supervised learning methods, including support vector machines (SVM), logistic regression, and deep learning models such as recurrent neural networks and convolutional neural networks. These models are trained on examples of both hate speech and non-hate speech to improve their accuracy and performance over time..

B. Social network analysis (SNA):

Social Network Analysis (SNA) investigates the structure of social media platforms to identify individuals or groups that spread hate speech. By analyzing patterns of interaction, relationships, and the dissemination of harmful information within online groups, SNA can help identify and monitor the sources of hate speech.

C. Human-in-the-loop Approaches:

Human moderators are crucial in hate speech detection systems. They analyze flagged content and make decisions based on context and nuance. By integrating human judgment and expertise into automated detection systems, we can improve accuracy and better address the complexities of context-dependent hate speech.

D. Natural language processing (NLP):

Face recognition algorithms, such as Eigenfaces, and natural language processing (NLP) approaches utilize text data to identify linguistic patterns associated with hate speech. This includes techniques like sentiment analysis, topic modeling, and word embedding.

NLP models can recognize harsh language, offensive phrases, and discriminatory speech by analyzing the context and semantics of the text.

V. Logistic regression :

Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes. It is extensively used in predictive modelling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.

FOR EXAMPLE: 0 – represents a negative class; 1 – represents a positive class. Logistic regression is commonly used in binary classification problems where the outcome variable reveals either of the two categories (0 and 1).

VI. Multinomial Logistic Regression :

Multinomial Logistic Regression is an extension of binary Logistic Regression used for multi-class classification problems where the dependent variable has three or more nominal categories.

The target variable ('class') in the dataset has three distinct classes:

- 0: Neutral or irrelevant
- 1: Hate speech or offensive content
- 2: Non-hate speech or normal content

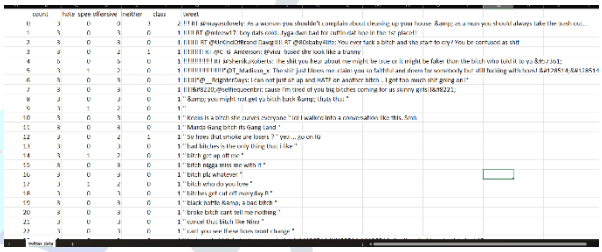


Fig: Labeled dataset

1. **Log-Odds** (for each category relative to "neither"):
   - Category 0 (Hate Speech):
   - $\log(P(Y=0) / P(Y=2)) = \beta_{00} + \beta_{10}X_1 + \beta_{20}X_2 + ...$
   - This formula models the log-odds of a tweet being classified as "hate speech" compared to "neither," based on the features extracted from the tweet.
   - β_{00} is the intercept for the "hate speech" category.
   - β_{10}, β_{20}, etc., are the coefficients
   - that show how each feature (e.g., X_1, X_2) influences the log-odds.

2. **Category 1** (Offensive Language):

   • $\log(P(Y=1) / P(Y=2)) = \beta_{01} + \beta_{11}X_1 + \beta_{21}X_2 + ...$

   • This formula models the log-odds of a tweet being classified as "offensive language" compared to "neither," based on the features.

   • β_{01} is the intercept for the "offensive language" category

   • β_{11}, β_{21}, etc., are the coefficients for the features.

3. **Probabilities** (using the softmax function):

   3.1 Probability of Hate Speech (Y=0):

   • $P(Y=0) = \exp(\beta_{00} + \beta_{10}X_1 + \beta_{20}X_2 + ...) / (1 + \exp(\beta_{00} + \beta_{10}X_1 + \beta_{20}X_2 + ...) + \exp(\beta_{01} + \beta_{11}X_1 + \beta_{21}X_2 + ...))$

   3.2 Probability of Offensive Language (Y=1):

   • $P(Y=1) = \exp(\beta_{01} + \beta_{11}X_1 + \beta_{21}X_2 + ...) / (1 + \exp(\beta_{00} + \beta_{10}X_1 + \beta_{20}X_2 + ...) + \exp(\beta_{01} + \beta_{11}X_1 + \beta_{21}X_2 + ...))$

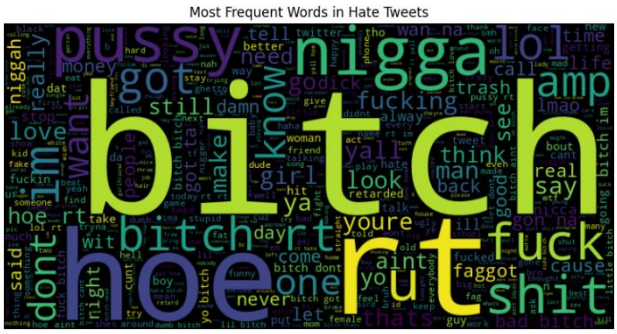
   3.3 Probability of Neither (Y=2):

   • $P(Y=2) = 1 / (1 + \exp(\beta_{00} + \beta_{10}X_1 + \beta_{20}X_2 + ...) + \exp(\beta_{01} + \beta_{11}X_1 + \beta_{21}X_2 + ...))$

VII. Observations from the Model:

1. word cloud

- A word cloud (also known as a tag cloud or weighted list) is a visual representation of text data where the size of each word indicates its frequency or importance.



2. Confusion Matrix:

- A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

```
# Train Logistic Regression model
logreg = LogisticRegression(max_iter=1000)
logreg.fit(X_train, y_train)

# Evaluate model
y_pred = logreg.predict(X_test)
print("Test Accuracy: {:.2f}%".format(accuracy_score(y_test, y_pred) * 100))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("Classification Report:\n", classification_report(y_test, y_pred))
```

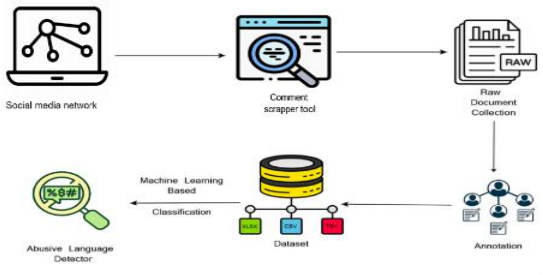
```
Test Accuracy: 88.90%
Confusion Matrix:
 [[  47  215   28]
 [  40 3684  108]
 [   2  157  676]]
Classification Report:
              precision    recall  f1-score   support

           0       0.53      0.16      0.25       290
           1       0.91      0.96      0.93      3832
           2       0.83      0.81      0.82       835

    accuracy                           0.89      4957
   macro avg       0.76      0.64      0.67      4957
weighted avg       0.87      0.89      0.87      4957
```

Model accuracy :88.9%

VIII. BLOCK DIAGRAM



1) **Define Objectives and Scope**:
Clearly define the objectives of hate speech detection within the context of your organization or community. Determine the scope of hate speech detection, including the

types of platforms or channels where detection will be implemented.

2) **Comment Scrapper tool**:
A "scraper tool" is a software application designed to extract information from websites, typically in an automated fashion. These tools can be useful for various purposes, such as gathering data for research, monitoring competitors' websites, or aggregating content for analysis.

3) **Raw Document Collection**:
Gather a diverse dataset of text, audio, or visual content containing examples of hate speech and non-hate speech.

4) **Annotation**:
Annotate the dataset with labels indicating whether each example contains hate speech or not.

5) **Dataset**:
Train the selected models using the annotated dataset, adjusting hyperparameters and optimizing performance through iterative training.

6) **Evaluation and Validation**:
Evaluate the performance of trained models using appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score.
Validate the models using cross validation techniques to ensure robustness and generalization to unseen data.

IX. OUTPUT:

By using TF-IDF Vectorization and multinomial Logistic Regression. We trained a model that can predict and detect Hate Speech.

```
[18]: # Define prediction function
def predict_hate_speech(model, vectorizer):
    while True:
        user_input = input("Enter a sentence (or type 'exit' to quit): ")
        if user_input.lower() == "exit":
            break
        processed_input = preprocess_text(user_input)
        input_vector = vectorizer.transform([processed_input])
        prediction = model.predict(input_vector)[0]
        print("Hate Speech" if prediction == 1 else "Not Hate Speech")

# Test the function
predict_hate_speech(logreg, vectorizer)

Enter a sentence (or type 'exit' to quit):  she is soo beautiful
Not Hate Speech
Enter a sentence (or type 'exit' to quit):  you are not allowed to talk to that bitch
Hate Speech
Enter a sentence (or type 'exit' to quit):  he said that his wife is a whore
Hate Speech
Enter a sentence (or type 'exit' to quit):  he ia a smart niggah in white pants
Hate Speech
Enter a sentence (or type 'exit' to quit):  exit
[ ]:
```

X. CONCLUSION

As hate speech continues to be a social issue, the necessity for automated hate speech detection systems arises. the journey of hate speech detection has shown significant progress and ongoing challenges. We've witnessed the shift from manual to automated content monitoring thanks to NLP and machine learning. These methods offer real-time detection, scalability, and unbiased moderation. However, the evolving nature of hate speech requires continuous development to adapt to changing tactics and languages. We showed existing techniques for this issue, as well as an innovative system with respectable accuracy. We also presented a new technique that can outperform existing systems on this challenge while improving interpretability. More study is needed to address the remaining obstacles, both technically and practically.

XI. FUTURE WORK

The future of hate speech detection holds several

promising directions and evolving techniques. Some notable future scopes and techniques include:

1. **Advanced NLP Models**: Future hate speech detection systems will likely leverage advanced natural language processing (NLP) models, such as transformer-based architectures (e.g., GPT-4, BERT). These models can capture intricate contextual nuances in language, improving the accuracy of hate speech identification.

2. **Multimodal Approaches**: Hate speech often extends beyond text to include images, audio, and video content. Future techniques will likely focus on developing multimodal systems that can effectively analyze and detect hate speech across various media types.

3. **Contextual Understanding**: Enhancements in contextual understanding will be crucial for distinguishing between genuine expressions, humor, and sarcasm from actual hate speech. Techniques that better grasp the context of conversations will contribute to more accurate detection.

XII. ACKNOWLEDGEMENT

We would like to extend out heartfelt thanks to our project supervisor, colleagues, friends for their unwavering support and encouragement. Their valuable guidance and input were instrumental in the successful completion of thisproject, and we are grateful for their contributions.

XIII. REFERENCE

[1] Roy, Pradeep Kumar, et al. "A framework for hate speech detection using deep convolutional neural network." IEEE Access 8 (2020): 204951-204962.

[2] Mullah, Nanlir Sallau, and Wan Mohd Nazmee Wan Zainon. "Advances in machine learning algorithms for hate speech detection in social media: a review." IEEE Access 9 (2021): 88364-88376.

[3] Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep learning based fusion approach for hate speech detection. IEEE Access, 8, 128923-128929.

[4] Khan, Shakir, et al. "HCovBi-caps: hate speech detection using convolutional and Bi-directional gated recurrent unit with Capsule network." IEEE Access 10 (2022): 7881-7894.

[5] Boishakhi, Fariha Tahosin, Ponkoj Chandra Shill, and Md Golam Rabiul Alam. "Multi-modal hate speech detection using machine learning." 2021 IEEE International Conference on Big Data (Big Data). IEEE, 2021.

[6] Abro, Sindhu, et al. "Automatic hate speech detection using machine learning: A comparative study." International Journal of Advanced Computer Science and Applications 11.8 (2020).

[7] Maity, Krishanu, et al. "A deep learning framework for the detection of Malay hate speech." IEEE Access (2023).

[8] Mozafari, Marzieh, Reza Farahbakhsh, and Noel Crespi. "A BERT-based transfer learning approach for hate speech detection in online social media." Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8. Springer International

Publishing, 2020.
[9] Corazza, Michele, et al. "A multilingual evaluation for online hate speech detection." ACM Transactions on Internet Technology (TOIT) 20.2 (2020): 1-22.
[10] Kovács, György, Pedro Alonso, and Rajkumar Saini. "Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources." SN Computer Science 2 (2021): 1-15.
[11] Kapil, Prashant, and Asif Ekbal. "A deep neural network based multi-task learning approach to hate speech detection." Knowledge-Based Systems 210 (2020): 106458.
[12] Modha, Sandip, et al. "Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance." Expert Systems with Applications 161 (2020): 113725.
[13] Mollas, Ioannis, et al. "Ethos: an online hate speech detection dataset." arXiv preprint arXiv:2006.08328 (2020).
[14] Gomez, Raul, et al. "Exploring hate speech detection in multimodal publications." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020.
[15] Jahan, Md Saroar, and Mourad Oussalah. "A systematic review of Hate Speech automatic detection using Natural Language Processing." Neurocomputing (2023): 126232.