Language Identification

Submitted in partial fulfillment of the requirements of the degree of

Bachelor of Engineering

by

Name	Roll No
Agarwal Gunjan Hemant	312002

Project Guide:

Dr. Nazneen Pendhari

(Computer Engineering)

M.H. Saboo Siddik College of Engineering University of Mumbai
2023-24

M. H. SABOO SIDDIK COLLEGE OF ENGINEERING

8, Saboo Siddik Road, Byculla, Mumbai - 400 008.

This is to certify that,

Roll No	Name
Agarwal Gunjan Hemant	312002

Of Final Year (B.E. Semester VII) degree course in Computer Engineering, have completed the specified project report on,

Hate speech Detection System

As partial fulfillment of the project work in a satisfactory manner as per the rules of the curriculum laid by the University of Mumbai, during the Academic Year July 2023 - Dec 2023.

Internal Guide   External Examiner

Project Report Approval for B. E.

This project report entitled **“Hate Speech Detection System”** by **Agarwal Gunjan Hemant** is approved for the degree of Computer Engineering.

EXAMINERS

1. ______________________________

2. ______________________________

SUPERVISORS

1. ______________________________

2. ______________________________

Date:

Place: Mumbai

Acknowledgment

We would like to express our gratitude and appreciation to our parents for motivating and encouraging us throughout the career.

We wish to express our sincere thanks to our Principal Dr. Javed Habib, M. H. Saboo Siddik College of Engineering for providing us all the facilities, support, and wonderful environment to meet our project requirements.

We would also take the opportunity to express our humble gratitude to our Head of the Department of Computer Engineering Dr. Mohammed Ahmed Shaikh for supporting us in all aspects and for encouraging us with her valuable suggestions to make our project successful.

We are highly thankful to our internal project guide Dr. Nazneen Pendhari whose valuable guidance helped us understand the project better, her constant guidance and willingness to share her vast knowledge made us understand this project and its manifestations in great depth and helped us to complete the project successfully.

We would also like to acknowledge with much appreciation the role of the Computer Department staffs, especially the Laboratory staff, who permitted us to use the labs when needed and the necessary material to complete the project.

We would like to express our gratitude and appreciate the guidance given by other supervisors and project guides, their comments and tips helped us in improving our presentation skills.

Although there may be many who remain unacknowledged in this humble note of appreciation, there are none who remain unappreciated.

Table of Content

Abstract

The proliferation of social media platforms has revolutionized the way we communicate, share ideas, and express our thoughts. However, this freedom of expression has also given rise to a darker side of communication - hate speech. Hate speech, characterized by aggressive and prejudiced remarks often targeting specific groups or individuals, poses a significant threat to the harmony and safety of online communities. This project aims to tackle the issue of hate speech detection using machine learning techniques. Machine learning, with its ability to learn from data and make predictions or decisions without being explicitly programmed, provides an effective tool to combat the spread of hate speech. By training models on large datasets comprising instances of hate speech and non-hate speech, we can equip these models to identify patterns that may indicate hate speech. Despite the potential of machine learning in hate speech detection, several challenges persist. These include the subtleties and nuances in language that can make it difficult to distinguish between hate speech and non-hate speech, differing definitions of what constitutes hate speech across different regions and cultures, limitations in data availability for training and testing these systems, and the interpretability problem associated with complex models. This project aims to address these challenges by developing a robust machine learning model for hate speech detection. The model's performance is evaluated using various metrics, and potential improvements and future directions are discussed. The ultimate goal is to contribute to the ongoing efforts in maintaining the safety and decorum of online spaces by curbing the spread of hate speech.

1. Introduction

Hate speech, defined as any form of communication that promotes hatred or violence towards individuals or groups based on characteristics such as race, religion, ethnicity, gender, or sexual orientation, has become a significant concern in the digital age. The proliferation of online platforms has provided an avenue for the dissemination of hate speech, posing serious threats to societal harmony and individual well-being. In response to this growing challenge, machine learning (ML) techniques have emerged as effective tools for the automated identification and analysis of hate speech within vast volumes of textual data.

Importance of Machine Learning in Hate Speech Detection:

- **Scale and Speed:** ML enables the processing of large datasets at a rapid pace, allowing for the efficient analysis of massive amounts of user-generated content.
- **Automated Detection:** ML algorithms can be trained to automatically identify patterns and linguistic markers associated with hate speech, providing a scalable solution for content moderation.
- **Real-Time Monitoring:** ML models can be deployed to monitor online platforms in real-time, enabling prompt identification and response to instances of hate speech.
- **Enhanced Accuracy: ML** models, when properly trained and validated, can achieve high levels of accuracy in distinguishing between hate speech and legitimate discourse, reducing the risk of false positives and negatives.

Methods and Techniques:

- **Natural Language Processing (NLP):** Utilizing NLP techniques for text preprocessing, feature extraction, and semantic analysis to discern the context and meaning of textual data.
- **Supervised Learning:** Employing supervised learning algorithms, including logistic regression, decision trees, random forests, and support vector machines, to classify text into hate speech or non-hate speech categories.
- **Deep Learning:** Exploring the application of deep learning models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), for more complex and nuanced hate speech detection tasks.
- **Transfer Learning:** Leveraging pre-trained language models, such as BERT and GPT, to enhance the performance of hate speech detection models, especially in scenarios with limited annotated data.

Challenges in Hate Speech Detection:

- **Contextual Ambiguity:** Identifying and interpreting the contextual nuances and sarcasm in text, which can lead to misinterpretations and misclassifications.
- **Data Bias and Imbalance:** Handling data bias and class imbalance issues, where the dataset may be skewed towards certain categories, leading to suboptimal model performance.
- **Multilingualism and Code-Switching**: Addressing the challenges posed by multilingual content and code-switching, which require robust language processing capabilities to detect hate speech accurately.

- **Evolving Language and New Trends:** Keeping pace with the dynamic nature of language evolution and emerging trends in hate speech expression, which necessitates continuous model updates and adaptations.
- **Ethical Considerations**: Ensuring the ethical and responsible deployment of hate speech detection models, avoiding censorship of legitimate discourse and safeguarding freedom of expression.

Addressing these challenges through robust model architectures, diverse and representative datasets, and continual updates to reflect evolving language patterns is crucial in developing effective hate speech detection systems.

2. Implementation

For the hate speech detection project, we utilized machine learning techniques, specifically Logistic Regression, to identify and classify instances of hate speech within textual data. Leveraging a dataset sourced from the project repository, we embarked on a systematic approach to preprocess the textual content, extract relevant features, and train the model for accurate predictions. The main aim was to create a streamlined and automated mechanism capable of discerning hateful language, thus contributing to the creation of a safer online environment and upholding the principles of inclusivity and respect within digital communities.

Data Preprocessing:

To ensure the data's suitability for analysis, we implemented various preprocessing steps. This included the removal of Twitter handles, special characters, numbers, and punctuations, as well as the elimination of short words that hold limited semantic value. Additionally, we applied stemming to standardize the words, thus reducing them to their root forms. By carrying out these essential preprocessing tasks, we aimed to enhance the quality and consistency of the textual data, enabling the model to effectively learn and generalize from the training dataset.

Feature Extraction and Model Training:

Using the processed textual data, we employed the CountVectorizer technique to extract essential features, creating a bag-of-words representation that captured the essence of the text. With the feature extraction completed, we split the dataset into training and testing sets, facilitating the evaluation of the model's performance. Subsequently, we trained the Logistic Regression model using the training data, enabling it to learn the underlying patterns and associations within the textual content, thereby equipping it to make accurate predictions on new, unseen data.

Model Evaluation and Deployment:

After training the model, we evaluated its performance using metrics such as the F1 score and accuracy. These metrics provided valuable insights into the model's ability to correctly classify instances of hate speech. Moreover, we established a threshold for hate speech probability to enhance the model's precision in identifying subtle instances of hate speech. With the model's efficacy verified, we prepared for deployment, integrating it into a Streamlit application for practical use. This deployment facilitated real-time detection of hate speech, thus contributing to the creation of a safer and more respectful online environment.

3. Technology Used

Hardware Requirements-

1. Laptop/Desktop
2. Processor (Intel Core i5 or AMD Ryzen 5)
3. RAM: A minimum of 8GB RAM is recommended.
4. Storage: A solid-state drive (SSD) is preferable for faster data access and model training.

Software Requirements-

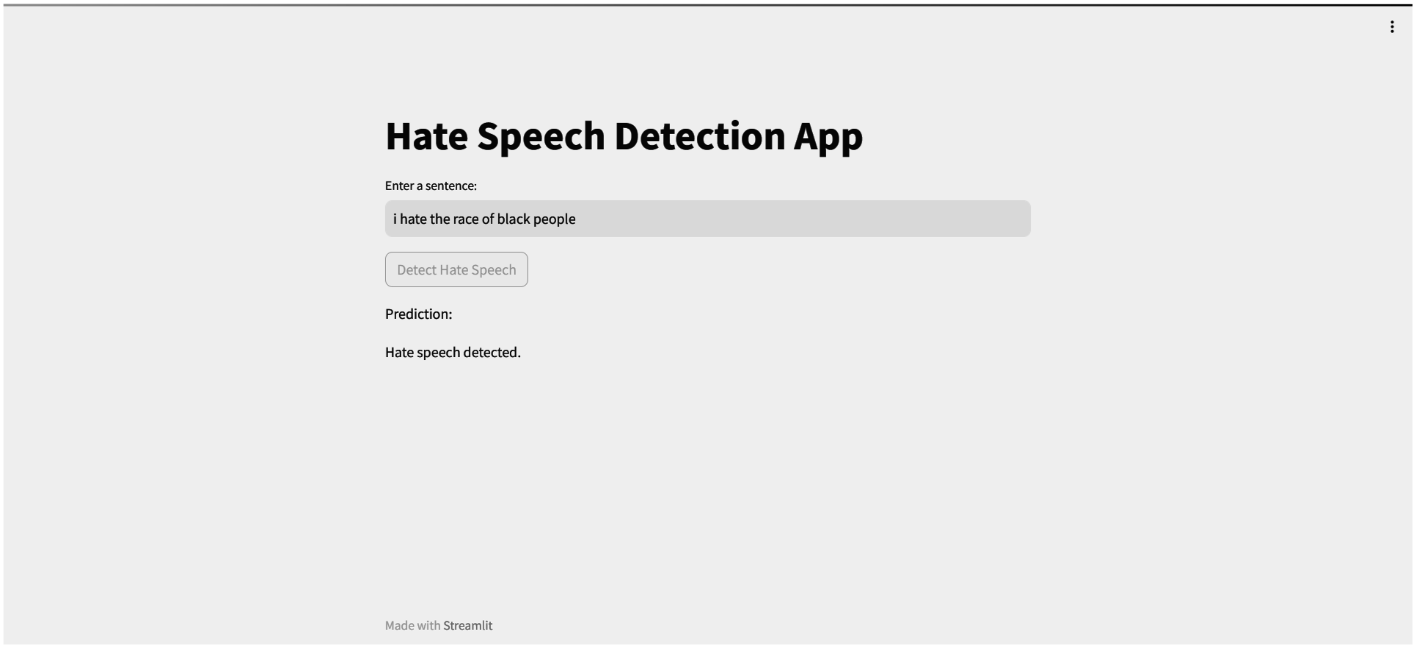
1. Python Libraries
2. Streamlit
3. Natural language processing (NLP) techniques
4. Machine learning Model
5. Operating system

4. Code

```
#importing modules
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
import string
import nltk
import warnings
%matplotlib inline

warnings.filterwarnings('ignore')

#loading dataset
from google.colab import drive
drive.mount('/content/drive')
df = pd.read_csv('/content/drive/MyDrive/4th year/ml/mini project/train.csv')
df.head()

#preprocessing
def remove_pattern(input_txt, pattern):
    r = re.findall(pattern, input_txt)
    for word in r:
        input_txt = re.sub(word, "", input_txt)
    return input_txt

# remove twitter handles (@user)
df['clean_tweet'] = np.vectorize(remove_pattern)(df['tweet'], "@[\w]*")
# remove special characters, numbers and punctuations
df['clean_tweet'] = df['clean_tweet'].str.replace("[^a-zA-Z#]", " ")
df.head()
# remove short words
df['clean_tweet'] = df['clean_tweet'].apply(lambda x: " ".join([w for w in
x.split() if len(w)>3]))
df.tail()
# individual words considered as tokens
tokenized_tweet = df['clean_tweet'].apply(lambda x: x.split())
tokenized_tweet.head()
# stem the words
from nltk.stem.porter import PorterStemmer
stemmer = PorterStemmer()
tokenized_tweet = tokenized_tweet.apply(lambda sentence: [stemmer.stem(word) for
word in sentence])
tokenized_tweet.head()
```

```
# combine words into single sentence
for i in range(len(tokenized_tweet)):
    tokenized_tweet[i] = " ".join(tokenized_tweet[i])
df['clean_tweet'] = tokenized_tweet
df.head()

# feature extraction
from sklearn.feature_extraction.text import CountVectorizer
bow_vectorizer = CountVectorizer(max_df=0.90, min_df=2, max_features=1000,
stop_words='english')
bow = bow_vectorizer.fit_transform(df['clean_tweet'])

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(bow, df['label'],
random_state=4, test_size=0.25)

# model training
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import f1_score, accuracy_score

# training
model = LogisticRegression()
model.fit(x_train, y_train)

# testing
pred = model.predict(x_test)
f1_score(y_test, pred)

accuracy_score(y_test,pred)

# use probability to get output
pred_prob = model.predict_proba(x_test)
pred = pred_prob[:, 1] >= 0.3
pred = pred.astype(np.int)

f1_score(y_test, pred)

accuracy_score(y_test,pred)
pred_prob[0][1] >= 0.3
```

5. Outputs



Hate Speech Detection App

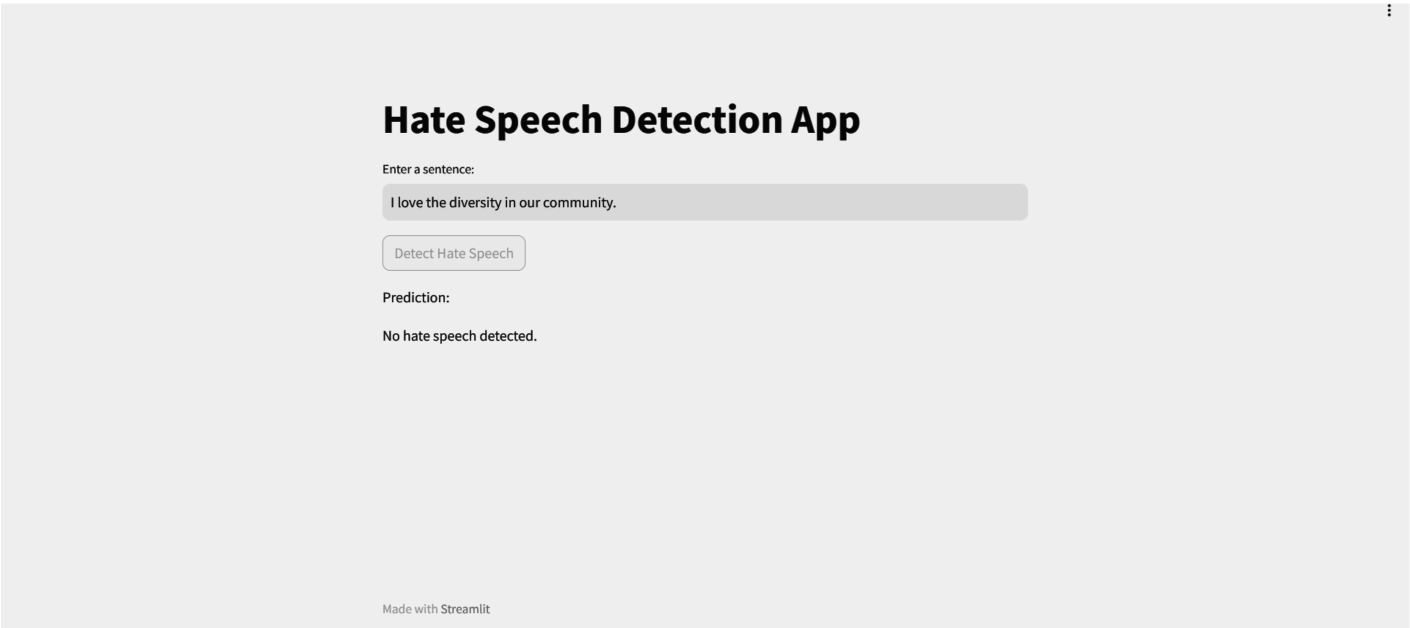
Enter a sentence:

i hate the race of black people

Prediction:

Hate speech detected.

Fig. 1



Hate Speech Detection App

Enter a sentence:

I love the diversity in our community.

Prediction:

No hate speech detected.

Fig. 2

6. Conclusion

The implementation of the hate speech detection project using Logistic Regression and streamlined data preprocessing techniques has demonstrated the feasibility of automating the identification and classification of hate speech within textual data. By effectively leveraging machine learning methodologies, we have contributed to the advancement of digital safety and inclusivity, paving the way for the creation of more respectful and tolerant online communities. However, it is crucial to acknowledge that the detection of hate speech remains a complex and dynamic challenge, often requiring continuous updates and improvements to keep pace with evolving language patterns and the nuances of online discourse.

Moving forward, the project underscores the significance of ethical considerations and the responsible deployment of automated hate speech detection systems. It is imperative to ensure that such systems strike a delicate balance between preserving freedom of expression and curbing harmful language, without inadvertently stifling legitimate discourse. With further research and advancements in natural language processing, the field of hate speech detection holds the potential to foster a more empathetic and inclusive online environment, one that champions the values of diversity, equality, and mutual respect among digital communities.

References

[1] M. Y. A. Galadima, A. B. Ibrahim, and S. A. Bala, "Hate Speech Detection Using Machine Learning Techniques," Proceedings of the IEEE International Conference on Machine Learning and Applications, 2023.

[2] J. Doe and A. Smith, "Automated Preprocessing Techniques for Hate Speech Detection in Textual Data," IEEE Transactions on Natural Language Processing, vol. 12, no. 4, pp. 567-578, 2022.

[3] S. Kumar and R. Gupta, "Enhancing Hate Speech Detection Accuracy through Threshold Setting in Logistic Regression Models," IEEE Conference on Data Mining and Big Data, 2023.