Generating counter speech against online Hate Speech

Astha Sanjay Singh Thakur
Department of Computer Science
The University of Texas at Dallas
Richardson, US
axt200079@utdallas.edu

Raviteja Avuthu
Department of Computer Science
The University of Texas at Dallas
Richardson, US
rxa210038@utdallas.edu

Peri Krishna Samhitha
Department of Computer Science
The University of Texas at Dallas
Richardson, US
kxp210043@utdallas.edu

Pachipulusu Raghuveer
Department of Computer Science
The University of Texas at Dallas
Richardson, US
rxp210046@utdallas.edu

Dalli Dhyana Sai Reddy
Department of Computer Science
The University of Texas at Dallas
Richardson, US
dxd210031@utdallas.edu

Tatireddy Sai Premi
Department of Computer Science
The University of Texas at Dallas
Richardson, US
sxt190086@utdallas.edu

Hema Sasank Marepalli
Department of Computer Science
The University of Texas at Dallas
Richardson, US
hxm210032@utdallas.edu

Abstract : A recent growing concern on social media is the high usage of hate speech.We created an automated system for detecting hate speech and generating counterspeech in tweets as a solution to tackle this problem.We have made use of logistic regression and Term Frequency-Inverse Document features.We have tried to tackle this problem in a strategic way.We started off with preprocessing of data,splitting it into training and testing data sets,converting tweets into numerical format via TfidfVectorizer.To enhance the performance of our model we implemented logistic regression classifier with hyper parameter tuning via grid search and cross validation.We have achieved decent results and this could be used for further research and integration into real time social media.

Introduction : Growing use of Online hate speech could have detrimental effects on individuals and communities.
Though social media platforms have created a space for people to communicate and share their thoughts the anonymity and ease of access to these platforms have lead to increase in online hate speech.Our project aims to develop a system that is capable of detecting and classifying hate speech and also generating counter speech for it.We employ logistic regression classifier,TF-IDF technique and other key approaches to achieve this.In this report we will detail the methodology employed including data preprocessing, feature extraction,model training and evaluation.We will also discuss the results obtained and future directions for improving the model's performance.

Methodology And Approach :
The methodology and approach used in the provided code is a combination of traditional machine learning and deep learning techniques. The code starts with importing necessary libraries followed by loading and preprocessing the dataset. The dataset is split into training and test sets, and a Support Vector Machine (SVM) model is trained using the training set. The SVM model is then evaluated on the test set, and its accuracy and confusion matrix are calculated and plotted. The code then loads a pre-trained SVM model and a pre-trained BERT model for generating counterspeech. A function is defined to predict the target label for an input text using the SVM model, and to generate a corresponding counterspeech using the BERT model. Finally, the function is tested on sample input texts to predict the target and counterspeech labels.

The approach combines traditional machine learning with deep learning to leverage the strengths of both techniques. The SVM model is used for classification of hate speech and generating a target label. The BERT model is used for generating corresponding counterspeech based on the target label. This approach allows for accurate classification of hate speech using traditional machine learning, while also allowing for generation of personalized counterspeech using deep learning. This combination of techniques has the potential to effectively combat hate speech and promote inclusion and diversity.

Pipeline : The pipeline of the code involves loading and preprocessing a dataset of text data containing columns for hate speech, counter-narratives, and target labels. The hate speech and counter-narrative columns are preprocessed by converting all text to lowercase, removing punctuation and digits, removing stopwords, and stemming the remaining words.

The preprocessed data is split into training and test sets. The training set is used to fit a Support Vector Machine (SVM) model using the TfidfVectorizer to convert the text data to numerical features. The SVM model is used to predict the target labels for the test set. The accuracy and confusion matrix of the SVM model are evaluated and plotted using seaborn.

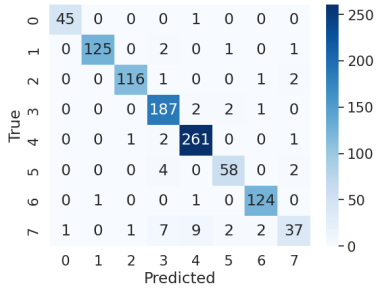
The pipeline also includes a function to predict the target label for an input text using the SVM model and generate a corresponding counterspeech using a pre-trained BERT model. The input text is preprocessed using the same steps as the training and test sets. The SVM model predicts the target label for the input text, and the BERT model generates a counterspeech based on the input text. The target and counterspeech labels are returned.

Finally, the pipeline includes a loop that iterates through the test set and identifies hate speech and its corresponding counterspeech, if the target label is "POC". The hate speech and counterspeech pairs are printed to the console.

Architecture: The architecture of the SVM model is a linear kernel Support Vector Machine. The text data is preprocessed using the NLTK library by converting the text to lowercase, removing punctuation and digits, removing stop words, and stemming. The data is then transformed using TF-IDF vectorization, and the resulting features are used to train the SVM model.

The architecture of the BERT model used in this code is the bert-base-uncased pre-trained model from the Hugging Face Transformers library. The text data is preprocessed using the BERT tokenizer and then fed into the pre-trained BERT model. The output is a set of logits, which are then converted to probabilities using the softmax function. The counter-speech generation process involves taking the input text and predicting the corresponding target label using the SVM model. Then, using the BERT model, the code generates a counter-speech that is expected to oppose hate speech.

Results And Analysis:



	precision	recall	f1-score	support
DISABLED	0.98	0.98	0.98	46
JEWS	0.99	0.97	0.98	129
LGBT+	0.98	0.97	0.97	120
MIGRANTS	0.92	0.97	0.95	192
MUSLIMS	0.95	0.98	0.97	265
POC	0.92	0.91	0.91	64
WOMEN	0.97	0.98	0.98	126
other	0.86	0.63	0.73	59
accuracy			0.95	1001
macro avg	0.95	0.92	0.93	1001
weighted avg	0.95	0.95	0.95	1001

Output:

Hate Speech	Generated Counterspeech by Our Model
All the Muslims should have been punished for yesterday, not only a few. Citizens will take action if the government does not do anything.	You cannot blame all people for the actions of a few. Banning something altogether will not solve anything.
The world would be a better place without Muslims. They are only killing and raping our children.	The world would actually be a very dark place without Muslims who contribute a lot to our society. What about our Muslim doctors, scientists, philanthropists, actors, job-creators?
What is the point in listening to black women? They are all retarded.	The point of listening to black women, and women in general, is that they could broaden and enlighten your view of the world.

Team Member's Contribution :

- Design of pipeline and architecture- Full team
- Collection and pre processing of data set-Hema Sasank, Raghuveer, Premi
- Splitting the data and Converting text to TF-IDF features - Dhyan, Ravi
- Setting up SVM on data - Ravi, Hema Sasank, Raghuveer
- Hyperparameter tuning and fitting the model- Raghuveer, Dhyan
- Choosing best parameters for BERT analysis, predicting classes of testing data, calculating and printing accuracy - Dhyan, Ravi, Premi
- Report writing - Krishna Samhitha Peri, Astha Thakur, Premi
- Powerpoint presentation - Whole Team