Hate Speech Detection Chatbot

Conversational AI system designed to identify and flag instances of hate speech in real-time communications

Go to the page

Team Members (Grp 8)

- Anouska Srivastava
- Anish Borkar
- Ankur Kaushal
- Dyuti Dasmahapatra
- D Veera Harsha Vardhan Reddy
- Godavarthi Sai Nikhil
- Himanshu Sharma
- Nichenametla Karthik Raja
- Vandamansu Sai Sumanth

Problem Statement

- Hate speech is a growing concern on the internet and social media platforms, as it can lead to harm, marginalization, and discrimination towards individuals and communities based on their ethnicity, race, religion, gender, sexual orientation, or other characteristics.
- The challenge is to develop a solution that can accurately identify hate speech in real-time and mitigate its harmful effects.
- The goal of hate speech detection is to provide a safe and inclusive online environment by reducing the spread of hate speech and promoting respectful communication.
- To develop a system that can accurately identify and respond to instances of hate speech in real-time communications, while also respecting cultural and contextual nuances and protecting freedom of expression.

Literature Review

	Paper Title	Citation	Methodology	Dataset	Inferences
1	• **Detection of hate speech text in Hindi-English code-mixed data**	• **K Sreelakshmi** et al. *(Procedia Computer Science 171,737-744, 2020)*	• Support Vector Machine (SVM)-Radial Basis Function (RBF) classifier using Random Forest and fastText	• A dataset of 10000 code-mixed text samples collected from different sources ,Twitter API	• **F1-Score :** 0.8580
2	• **Date expansion using back translation and paraphrasing for hate speech detection**	• **Djamila R.** et al. *(Online Social Networks and Media 24, 100153 , 2021)*	• Back-Translation Method, Paraphrasing technique for data augmentation with LSTM and CNN, fastText	• AskFm corpus, Formspring, Warner & Waseem ,Olid & Wikipedia toxic comments dataset	• **F1-Score :** 0.8580
3	• **Arabic offensive and hate speech detection using a cross-corpora multi-task learning model**	• **Wassen A.** et al. *(Informatics 8 (4), 69, 2021)*	• Multi-task learning (MTL) model on a pre-trained Arabic Language model	• OSACT , L –HSAB , T-HSAB, Abusive dataset provided by Haddad et al.	• **F1-Score :** 0.8580
4	• **SWE2: SubWord Enriched & Significant Word Emphasized Framework for Hate Speech Detection**	• **Guanyi M.** et al. *(Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 1145-1154, 2020)*	• Phonetic & character level embedding with an **LSTM + attention**-based word-level feature extraction method	• Waseem, Davidson, HateLingo datasets ; alongwith Legitimate messages from twitter by Twitter API	• **F1-Score :** 0.8580

Literature Review

	Paper Title	Citation	Methodology	Dataset	Inferences
1	• **Detection of hate speech text in Hindi-English code-mixed data**	• **K Sreelakshmi** et al. *(Procedia Computer Science 171,737-744, 2020)*	• Support Vector Machine (SVM)-Radial Basis Function (RBF) classifier using Random Forest and fastText	• A dataset of 10000 code-mixed text samples collected from different sources ,Twitter API	• **F1-Score :** 0.8580
2	• **Date expansion using back translation and paraphrasing for hate speech detection**	• **Djamila R.** et al. *(Online Social Networks and Media 24, 100153 , 2021)*	• Back-Translation Method, Paraphrasing technique for data augmentation with LSTM and CNN, fastText	• AskFm corpus, Formspring, Warner & Waseem ,Olid & Wikipedia toxic comments dataset	**F1 score** *(with CNN model)* : 96.9 , 98.9, 99.4, 94.4 , 99.3 resp.
3	• **Arabic offensive and hate speech detection using a cross-corpora multi-task learning model**	• **Wassen A.** et al. *(Informatics 8 (4), 69, 2021)*	• Multi-task learning (MTL) model on a pre-trained Arabic Language model	• OSACT , L –HSAB , T-HSAB, Abusive dataset provided by Haddad et al.	**F1 score** : 92.34, 88.73, 87.18, 80.50
4	• **SWE2: SubWord Enriched & Significant Word Emphasized Framework for Hate Speech Detection**	• **Guanyi M.** et al. *(Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 1145-1154, 2020)*	• Phonetic & character level embedding with an **LSTM + attention**-based word-level feature extraction method	• Waseem, Davidson, HateLingo datasets ; alongwith Legitimate messages from twitter by Twitter API	**F1-Score:** 1. (No Adversarial Attack) : 0.953 2. (Extreme Adversarial attack) : 0.934

Dataset Identification

Dataset identification is a critical aspect as the **quality and representativeness of the data** used to train machine learning models can have a significant impact on their performance.

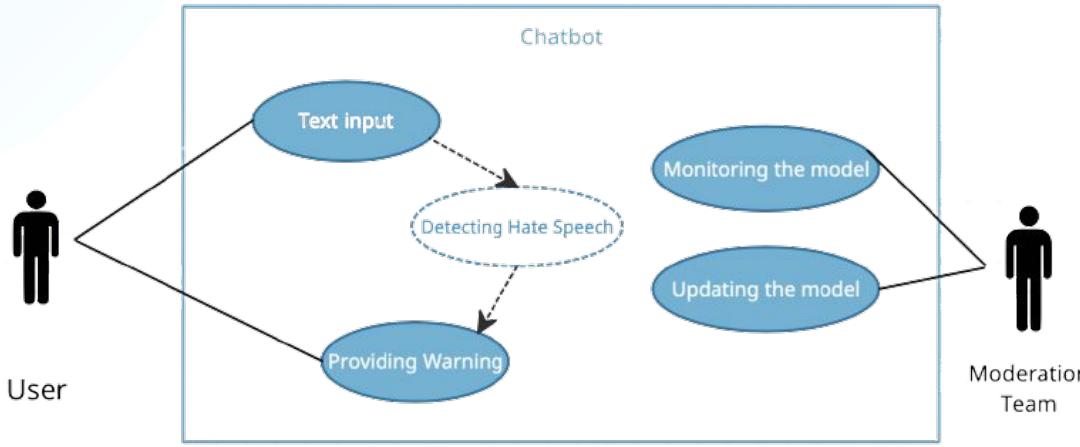
Dataset taken from github with **25,296 rows of data** .

CSV file contains **6 columns**:

1. **count** = number of CrowdFlower users who coded each tweet
2. **hate_speech** = number of CF users who judged the tweet to be hate speech
3. **offensive_language** = number of CF users who judged the tweet to be offensive.
4. **neither** = number of CF users who judged the tweet to be neither offensive nor non-offensive
5. **class** = class label for majority of CF users. 0 - hate speech, 1 - offensive language , 2 - neither
6. **tweet** = the tweet or comment that was posted on social media

	count	hate_speech	offensive_lar	neither	class	tweet
0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't complain about cl
1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe
2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever f
3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
4	6	0	6	0	1	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me migh
5	3	1	2	0	1	!!!!!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows me..claim you
6	3	0	3	0	1	!!!!!!"@__BrighterDays: I can not just sit up and HATE on another l
7	3	0	3	0	1	!!!!“@selfiequeenbri: cause I'm tired of you big bitches com
8	3	0	3	0	1	" & you might not get ya bitch back & thats that "
9	3	1	2	0	1	"
10	3	0	3	0	1	" Keeks is a bitch she curves everyone " lol I walked into a conversat
11	3	0	3	0	1	" Murda Gang bitch its Gang Land "
12	3	0	2	1	1	" So hoes that smoke are losers ? " yea ... go on IG
13	3	0	3	0	1	" bad bitches is the only thing that i like "
14	3	1	2	0	1	" bitch get up off me "
15	3	0	3	0	1	" bitch nigga miss me with it "
16	3	0	3	0	1	" bitch plz whatever "
17	3	1	2	0	1	" bitch who do you love "
18	3	0	3	0	1	" bitches get cut off everyday B "
19	3	0	3	0	1	" black bottle & a bad bitch "
20	3	0	3	0	1	" broke bitch cant tell me nothing "
21	3	0	3	0	1	" cancel that bitch like Nino "

High Level Solution Design



Use Case Diagram

01

Data Collection

Selecting and using an existing dataset to train the machine learning model

02

Data Preprocessing

Data preprocessing involves cleaning, transforming, and encoding text data for modeling.

03

Model Training/ Model Deployment:

Train a machine learning model, such as logistic regression, CNN or a RNN, to identify patterns & make predictions.

04

Input Processing

Text data tokenization and padding to equal length for input processing.

05

Output Generation

Based on the prediction, the chatbot will generate an appropriate response.

06

Continuous Improvement

The system should be continuously monitored.This may involve updating the data used for training

THANK YOU