Hate Burst

Lovepreet Singh
IIIT-Delhi
Delhi, India
lovepreet21044@iiitd.ac.in

Syed Mohd Gulam Baquer
IIIT-Delhi
Delhi, India
syed21100@iiitd.ac.in

Sristi Sharma
IIIT-Delhi
Delhi, India
sristi21098@iiitd.ac.in

Kanak Panwar
IIIT-Delhi
Delhi, India
kanak21037@iiitd.ac.in

Parul Jain
IIIT-Delhi
Delhi, India
parul21064@iiitd.ac.in

1 PROBLEM FORMULATION

"Visual content is getting worse but it can be made better using a portal by filtering out hateful content from visual and textual data"

Filtering out comments posted on social media platforms by millions of people every second, which contain hateful speech in the form of text, audio, or video.

2 MOTIVATION

In this growing age of social media, hate speeches and comments are one of the main problems growing substantially. And we are witnessing the adverse effect of this problem, like people are setting their narrative based on what they heard or saw. So figuring out which thing is suitable for their platform and which is not is one of the significant problems social media giants face. These social media giants produce millions of data every second, and monitoring that data is not an easy task. Although they have a moderation team, they are inefficient and take a lot of time to filter out inappropriate content. So here comes our website "`Fake Burst`", which will flag out whether the content is appropriate or not. Not only textual comments, but our website will also tell whether the *video*(in comments) you are watching or the *audio*(in comments) you are listening to is suitable for the community or not. Within a single click, you can verify the content.

3 DATASET

Below are the bar charts for English and Hindi Dataset respectively.

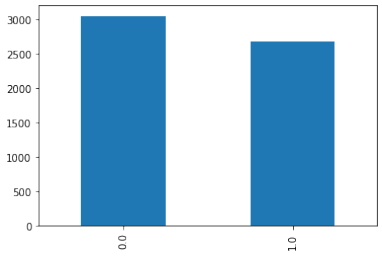

Figure 1: English Dataset Bar Chart

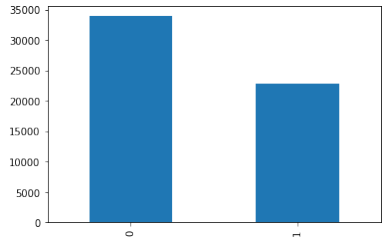

Figure 2: Hindi Dataset Bar Chart

- 0 on X-axis indicates non-hateful comment.
- 1 indicates hateful comment.

4 LITERATURE REVIEW

4.1 Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TF-IDF based Approach

This paper aims to classify tweets into 3 categories: hateful, offensive, and clean. Here they have used n-grams and TF-IDF as preprocessing and have used multiple machine learning models to train the data. This proposed to automatically classify tweets into three classes which consists of hateful, offensive and clean. N-grams was generated from the dataset as features and their corresponding TF-IDF values to the multiple machine learning algorithms and further a comparative analysis of these was made. Based on the comparative analysis of Logistic Regression, Naive Bayes and Support vector machine(SVM), Logistic regression performed better as compared other as it gave a ngram range of 1 to 3 for the L2 normalization of TF-IDF and on the test data, the model gave an accuracy of 95.6%.

Link: https://arxiv.org/pdf/1809.08651.pdf

4.2 Offensive Video Detection

This paper emphasizes filtering the hate speech from videos. The identification of offensive material can be performed automatically using machine learning. They took a video in Portuguese language and converted it into a textual form using some ensemble model technique. They achieved the final result based on the classifiers like the random forest, TF/IDF, etc.

Link: https://aclanthology.org/2020.lrec-1.531.pdf

---

lll

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00

4.3 Detecting Hate tweets: Twitter Sentiment Analysis

This paper focuses on the toxic content we receive from Twitter as people of different cultures and backgrounds use it. This paper classifies Twitter tweets into two classes Hate Speech and Non-Hate Speech. They have used the Twitter database. They performed experiments by leveraging a bag of words and the term frequency-inverse Documents Frequency(TFIDF). And to achieve this, they applied the Logistic Regression , Naive Bayes, decision Tree, Random Forest. The results show that Logistic Regression performs Comparatively better with the TF IDF approach.

Link: https://towardsdatascience.com/detecting-hate-tweets-twitter-sentiment-analysis-780d8a82d4f6

4.4 Hate Speech detection in Audio

This Paper aims towards the detection of hate speech in audio like clubhouse or voice chat rooms, in this detection system, hate speech will be detected through audio on voice based social media platforms. It is seen that voice has more impact on society rather than written content, people more likely to raise their voices in the context of race, religion, gender, nationality, sexual orientation, and spread hatred on the basis of these contexts. This often seems like hate speech comes in a sophisticated way, when hatred in text is less emphasized people often use their voice to spread hatred. In this paper, audio messages are collected as data from different platforms and processed them by extracting the text features using three methods, Bag of Words, Word2Vec,and Perspective Scores and then used different base classifiers to compare the extracted features.

Link: https://arxiv.org/pdf/2106.13238.pdf

4.5 Detection of Hate Speech in Videos Using Machine Learning

This Paper aims towards the detection of hate speech in the video. It is converting video speech to audio and then audio to text. It uses YouTube videos as a dataset. To get the audio from video, it has used FFmpeg API. FFmpeg API is a multimedia framework that allows users to encode, decode and convert media between different formats. In this research, they have used the TextBlob python library to perform Sentiment Analysis on the dataset. It has implemented Naïve Bayes Classifier, Random Forest Classifier, Linear Support Vector Machines (SVM) model and Recurrent Neural Network (RNN) model. Two different kinds of experiments were conducted. The first one is classifying the videos into normal or hateful videos. The second one is classifying the videos into normal, racist or sexist videos. Among all the models, Random Forest gives higher accuracy of 0.9464 and 0.8571 for experiments 1 2, respectively.

Link: https://ieeexplore.ieee.org/document/9458005/citations?tabFilter=papers#citations

4.6 Challenges in Hate speech detection

There is a big challenge in detecting what speech is considered to be hate speech. There are several terms that do not come under hate speech but are used as hate to humiliate others.

The similarity in letters in the comments, the tendency in some posts to replace letters with similar-looking numbers, e.g. “E”s with 3s, or “l”s with 1s, and so on). Another challenge is imbalanced data, there is an imbalance between hateful and hateless comments which becomes challenging in identifying the hateful comments.

Hatebase: This paper contains the labeled data of 24783 tweets, among which 1430 comes under hate speech, 19190 comes under offensive language, and 4163 comes under neither hate nor offensive. This paper used the CNN-LSTM model to train the data.

Link: https://link.springer.com/article/10.1007/s42979-021-00457-3

4.7 Hate and Offensive Speech Detection in Hindi and Marathi

Dataset: HASOC 2021 Hindi and Marathi hate speech datasets Dataset has binary labels. It has used various deep learning models, BERT, and basic models based on LSTM are used with fast text word embeddings. The transformer-based models performed the best and even the basic models along with FastText embeddings give a good performance. **Link:** https://arxiv.org/pdf/2110.12200.pdf

4.8 Towards generalisable hate speech detection: a review on obstacles and solutions :

This paper focuses on genralisable of hate speech , Generalization can help to detect different kind of hate speech with same model and same acccuracy , Different type of abusive and hateful comment are closly related , and some time it is very tough to get the detection of hateful speech which is not as offensive as others becuase of use of words. This research paper Gives overview of detection of hateful comment regardless of its speaker , way of saying, or target. Hate speech may be different from absuive r offfensive speech so It is necessary to get the generalization of methd which detect hate speech. This research paper tells how can be train moodel fo genralizatiion , Example – CNN-GRU Model. **Link:** https://arxiv.org/pdf/2110.12200.pdf

5 BASELINE RESULTS

We have 2 base line models one for English and other for Hindi language. We have used Keras RNN LSTM to train our both the datasets. Our English Hindi dataset consists of 50k 5k tweets respectively. On the English dataset model we are getting an accuracy of 92 percent and 91 percent on the Hindi dataset. Currently we are working on extending our model and dataset towards the visual data.

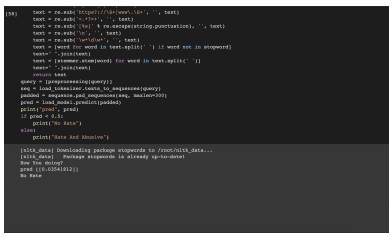

Figure 3: for English Dataset

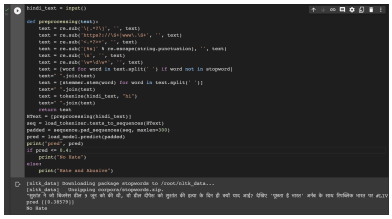

Figure 4: for Hindi Dataset

6 PROPOSED METHODS

6.1 Overview

We are aiming to create a platform where a user can give input in any form it can be text/ audio/ video based. And our website will process this user input into a text format with labels and the label value will be flagged out to show whether the given input was hateful or not.

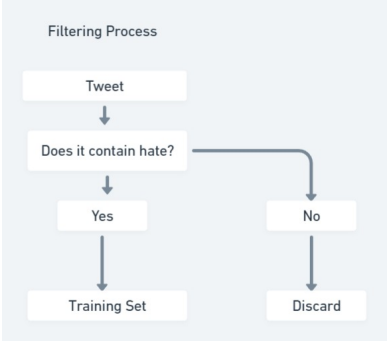

Figure 5: Plan Of Work

6.2 Detailed Description

- **Preprocessing on the data includes:**
  - removed stop words
  - removed hashtags
  - removed special characters from tweets
  - lemmatization Stemming
- **Dataset Split**
  - Split the dataset into the ratio of 80:20 for train and test as well.
- **Model training:**
  - **Try 1:** We trained our data using Xgboost but did not get satisfying results.
  - **Try 2:** After that we tried to find the tokenizer of the dataset and train our model using Keras LSTM, using sigmoid as an activation function and loss a binary cross-entropy and we used 5 epochs to train our model. And we got an accuracy of 94 percent for the English dataset. We followed the same steps for the Hindi dataset as well. The only difference is, in the Hindi dataset for preprocessing we are using the iNLTK open-source library and used 4 epochs to train the model. Accuracy we got is approx 92 percent.