Project Synopsis

Cyber Trolling Detection System

Submitted By

Dikesh Kurve (A-31)

Saurabh Barse (A-54)

Yash Solanke (A-64)

Uder the guidance of

Prof D Bhagat

Department of AI

Department of Artificial Intelligence

G.H. Raisoni College of Engineering

(Autonomous Institute Affiliated to Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur)

Near CRPF gate, Digdoh Hills, Nagpur

Session 2022-23

INDEX

ABSTRACT

Hate Speech and harassment are widespread in online communication, due to users' freedom and anonymity and the lack of regulation governed by social media. Due to this cyber trolling and bullying is major issue in a society. To overcome this problem we can use the ability of machine learning to hate speech detection to capture common properties from topic generic datasets and transfer this knowledge to recognize specific manifestations of hate speech using NLP, ML, Video Processing and Analysis. Our main goal is to Apply the model on both text and video-based datasets. We use different machine learning and deep learning technique including multi modal approaches. We use dataset that is divided into topic-specific like misogyny, sexism, racism, xenophobia, homophobia. Training a model on a combination of several (training sets from several) topic-specific datasets is more effective than training a model on a topic-generic dataset. Dataset can be gathered from various sources like from YouTube API, Twitter API, web-scrapping or from various government sources. Our aim is to perform exploratory data analysis on collected data and derive conclusion from it, using machine learning tools and techniques. Our main objective is to detect the violent and hatred content from social media. Our work provides a promising solution to the problem of building real-time cyber trolling systems that have constrained software or hardware capabilities

1. INTRODUCTION

Nowadays, people increasingly use social networking sites, not only as their main source of information, but also as media to post content, sharing their feelings and opinions. Social media is convenient, as sites allow users to reachpeople worldwide, which could potentially facilitate a positive andconstructive conversation between users. However, this phenomenon has a downside, as there are more and more episodes of hate speech and harassmentin online communication which resembles trolling. The Oxford Dictionary describes trolling as making "a deliberately offensive or provocative online posting with the aim of upsetting someone or eliciting an angry response fromthem". ". In a current society Cyber Trolling is increasing day by day, whichindirectly or directly impacting the society in a negative way. Troll against person, organization or religion creates the negative impact and leads to social dissonance. To tackle the cyber trolling immediate measure should be taken without violating to the freedom of speech. This is due especially to the freedom and anonymity given to users and to the lack of effective regulationsprovided by the social network platforms. There has been a growing interest in using artificial intelligence and Natural Language Processing (NLP) to address social and ethical issues. Using Machine learning tools and techniquewe can detect the cyber trolling and avoid the negative outcomes in future.

(fig. 1.1) Forms of trolling

2. OBJECTIVES

Phase: 1

In this phase our aim is to build the cyber trolling detection system for only text based input. In this phase NLP (Natural Language Processing) will play key role to achieve objective. Text processing, Sentence summarizing, analysing the semantics of input will play key role to train the model. After performing EDA (Exploratory Data Analysis) we are going to use various machine learning models like Naïve Bayes, Random Forest to get the end result. In a glance our main goal will be to build a machine learning pipeline for only text based input for cyber trolling detection system.

Phase: 2

In this phase our aim is to detect violent and hatred content in a form of video base inputs. In this phase video processing like audio to text translations, signal processing to make filtered content input, computer vision to analyse and detectthe actions and emotion in video content can be used. After gathering the data. Data formatting and data processing will be the major task to create the accuratedataset. After this EDA (Exploratory Data Analysis) is performed to obtain the insight and to proceed with the further decision making approach. In the final stage we will be using various machine learning model to train and test the datasetand predict the outcome of input dataset.

Our end output will be in the form of UI (User Interface) which will take input as*text* as well as *video* and gives various prediction regarding giving input data likewhether content is resembling the trolling or not in various scales and factors.

3. PROPOSED METHODOLOGY/ SYSTEM ARCHITECTURE

The main objective is to create a machine learning piplining workflow which keeps the project in a right track and increase the efficiency of end result. Our main aim is to create multi model system to detect the troll on a social media in a form of both text as well as video. To this objective we are going to follow certain steps as follow

- **Data collection & processing**

In a initial phase we are going to collect the data for text and video from various source like from YouTube API, Twitter API or from some data providing services. After collecting data our goal will be to apply methods to formulate well structured dataset. Various tools and technique are used such as numpy, pandas to make a collected data well structured. After structuring the data various Exploratory Data Analysis methods applied to create intermediate results. A well structured and error free data is forwarded in pipeline.

- **NLP on DATA**

Collected data is already present in a form of text on which we can perform various NLP functions such as Sentiment Analysis, Named Entity Recognition, Summarization, Topic Modeling, Text Classification, Keyword Extraction. Lemmatization and stemming. Which create a intermediate results which can be useful to train the model. NLP is not only applied on text data but also applied on

converted data which are generated from video dataset. Video to text conversion is done using various signal processing and mathematical libraries like numpy, scipy and etc.

- **Computer vision on data**

Our main dataset is divided into two types first is text-based data and second is video based data. This video-based data is used to identify the type of content basedon actions, expression, language and etc. which can further increase the accuracy ofmodel. To derive the certain conclusion from data we can use libraries like OpenCVwhich helps to process video data and creates the certain conclusion based on data.one of the application of OpenCV can be to detect the actions view in a video whether the action perform is violent or peaceful another use case can be to detect the expression from given video data whether it is good or violent.

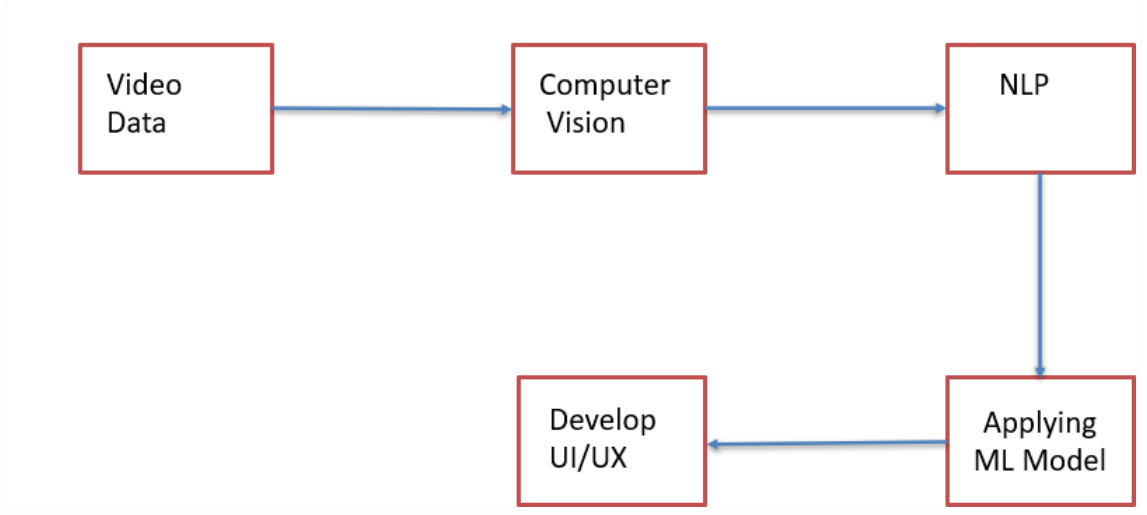
- **Applying ML Model**

In this stage we are applying some feature selection and feature extraction methodsto calculate dominating features and remove the extra feature which can act as noiseand can impact the final results of a model. After this various model of machine learning like naïve bayes, random forest, ensemble models are applied andaccording to end accuracy final model is selected. we are also using some deep learning model to learn certain trends in a dataset. deep learning model such as CNN, LSTM can be used for certain predictions. In this method we need a best algorithm which are less computational and create

the effective results. after some trial and error we can lead to well optimized machine learning model which gives useffective accuracy.

- **Developing User Interface**

This is the final phase in which we are creating the front end for user interaction withmodel. This frontend act as a input which gets input from outside and give this to backend model which gives response in terms of predictions. In a backend service we are using flask micro framework or similar to flask. On front end we are using HTML/CSS and JavaScript to make front end more interactive and user friendly.

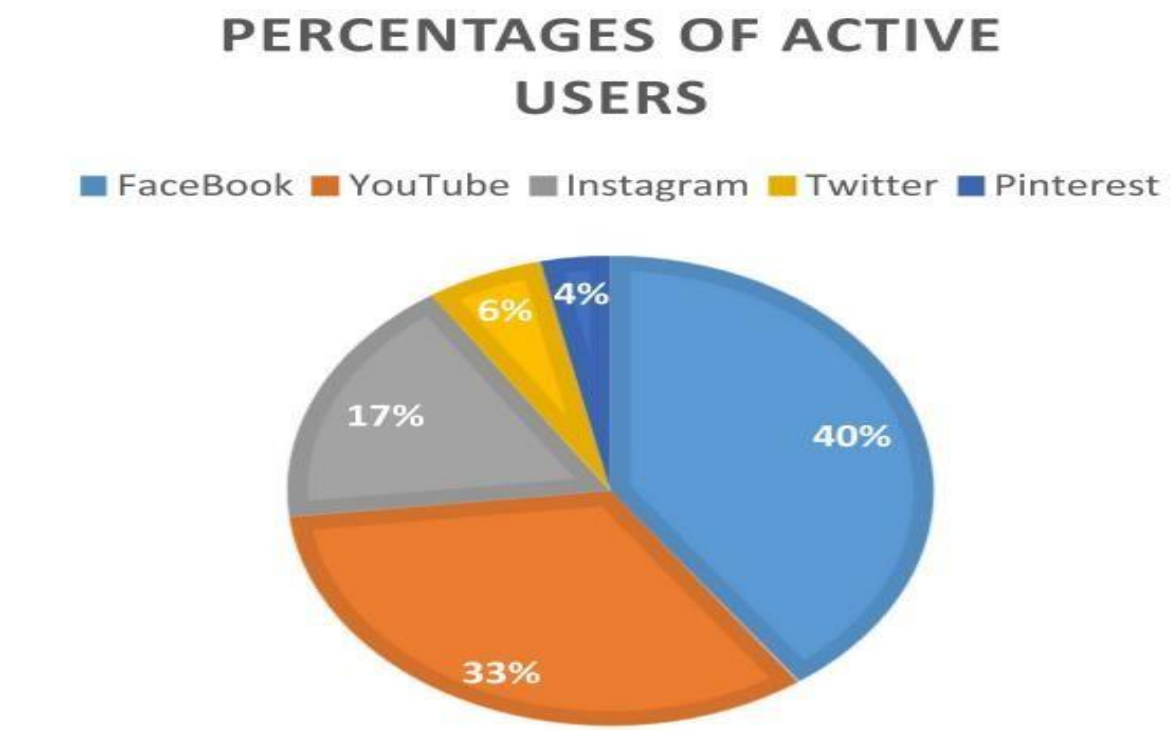


(fig 2.1) Workflow for the project

4. LITERATIRE SURVEY OF EXISTING SOLUTIONS/ PRODUCTS

Abusive messages in social media is a complex phenomenon with a broad range of overlapping modes and goals. Cyberbullying and hate speech are typical examples of abusive languages that researchers have put more interest in the past few decades due to their negative impacts in our societies. Several research have been conducted to automatically detect these undesirable messages among other messages in social media. The automatic detection of hate speech using machine learning approaches is relatively new, and there are very limited review papers on techniques for automatic hate speech detection [. The recent and related survey papers available on review of hate speech detection methods during this research work were few. The following were the available traditional literature review related to automatic detection of hate speech using MLA.

ML algorithms have contributed immensely in hate speech detection and SM content analysis generally. Offensive comments such as HS and cyberbullying are the most researched areas in NLP in the past few decades. ML algorithms have been of great help in this direction in terms of SM data analysis for the identification and classification of offensive comments. The advances in ML algorithms researches have made significant impacts in many fields of end eavour which led to some important tools and models for analysing a large amount of data in real-world problems like SMNs content analysis.



(fig 3.1) Active users on social media.

5. HARDWARE / SOFTWARE SPECIFICATION

- Python 3 version (Updated preferred)
- Jupyter ,Colab Notebook.
- Pandas, numpy, NLTK Toolkit, scikit-learn library.
- Various visualization library like matplotlib, seaborn, etc

- OpenCV for computer vision.
- Scipy.signal for signal processing.
- Keras for deep learning applications.
- Graphical Processing Unit (GPU).
- Flask micro-framework for backend.
- Other backend service might get preferred instead of flask.
- HTML/CSS/Javascript for frontend.
- Git/GitHub for version control and source contribution.
- Heroku cloud for deployment.

6. CONCLUSION

The problem statement itself is very unique and multi-dimensional which can be solved in various way using various tools and technique which can further optimizedaccording to future scope. Our main motivation behind selecting this problem is its wide application which can benefit the society in positive way. Our motive is to notonly solve the given problem but to solve it in more optimized way even the adaptedproblem statement is in research area in a computing domain we are willing to add new ways or approach to solve the problem in a efficient way.

7. REFERENCES

NANLIR SALLAU MULLAH, (Member, IEEE), AND WAN MOHD NAZMEE

WAN ZAINON: “Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review”

LINK: https://ieeexplore.ieee.org/document/9455353

Automatic Hate Speech Detection using Machine Learning: A Comparative Study(IJACSA) International Journal of Advanced Computer Science and Applications LINK:https://thesai.org/Downloads/Volume11No8/Paper_61-Automatic_Hate_ Speech_Detection.pdf

Detection of Hate Speech in Videos Using Machine Learning
2020 International Conference on Computational Science and Computational Intelligence (CSCI)
LINK:https://american-cse.org/sites/csci2020proc/pdfs/CSCI2020-6SccvdzjqC7bKupZxFmCoA/762400a585/762400a585.pdf

Hatred and trolling detection transliteration framework using hierarchical LSTM in code-mixed social media text:-
LINK:https://link.springer.com/article/10.1007/s40747-021-00487-7

Multimodal Sentiment Analysis on Video Streams using Lightweight Deep Neural Networks
LINK: https://www.scitepress.org/Papers/2021/103044/103044.pdf

Emotionally Informed Hate Speech Detection: A Multi-target Perspective LINK:https://link.springer.com/article/10.1007/s12559-021-09862-5

ETHOS: a multi-label hate speech detection dataset

LINK:https://link.springer.com/article/10.1007/s40747-021-00608-2