INSTANDER :A HATE SPEECH DETECTION SYSTEM FOR ONLINE SOCIAL NETWORKING

A

Synopsis Report

BACHELOR OF TECHNOLOGY

in

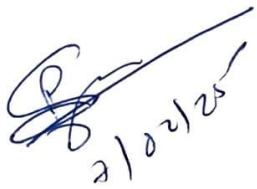
COMPUTER SCIENCE & ENGINEERING

by

S.No	Student's Name	Roll Number	SAP ID
1	Vaishnavi Dubey	R2142220195	500101976
2	Rishi Raj Jain	R2142220150	500105417
3	Parth Soni	R2142220125	500102209
4	Chitransh Soni	R2142220061	500102177

under the guidance of

Dr. Mohammad Ahsan



School of Computer Science

UPES

Via Prem Nagar, Dehradun, Uttarakhand

February – 2025

INDEX

Abstract

This project focuses on developing "INSTANDER," an advanced system designed to detect and mitigate hate speech on online social networking platforms. The system utilizes natural language processing (NLP) and machine learning techniques to analyze user-generated content, identifying and flagging potentially harmful speech in real time.

A key feature of the system is its ability to classify text into various categories of offensive language, ensuring accurate detection while minimizing false positives. INSTANDER integrates a robust database management system (DBMS) to store flagged content, user reports, and moderation actions, enabling efficient content moderation. By automating the detection process and reducing reliance on manual moderation, INSTANDER enhances online safety, fostering a more inclusive and respectful digital environment. The application aims to provide a reliable and scalable solution for combating hate speech, ensuring healthier interactions across social media platforms.

Introduction

Online social networking platforms have become a crucial space for communication and expression. However, the rise of hate speech poses significant challenges, leading to toxic environments, online harassment, and misinformation. Manual content moderation is time-consuming and often ineffective, highlighting the need for an intelligent, automated solution.

INSTANDER is designed to address these challenges by providing a comprehensive system for detecting and mitigating hate speech in real time. Leveraging advanced natural language processing (NLP) and machine learning techniques, the platform accurately classifies offensive content while minimizing false positives.

The system automates the monitoring process, reducing the burden on human moderators and enhancing platform safety. By integrating real-time analysis and a dynamic moderation interface, INSTANDER improves content regulation and fosters healthier digital interactions. Backed by a robust database management system (DBMS), the platform efficiently stores flagged content, user reports, and moderation actions, ensuring effective hate speech management across online communities.schedules, providing an efficient and well-organized sports festival experience for all involved.

Literature Review

Perspective API

Description:

Developed by Google, Perspective API is an AI-powered tool that analyzes and scores text-based content for toxicity. It helps platforms identify harmful speech and provides moderation assistance.

Limitations:

While effective in detecting offensive language, it struggles with contextual understanding, often misclassifying sarcastic or coded hate speech. Additionally, it lacks a real-time monitoring system that adapts to emerging hate speech patterns.

Hatebase

Description:

Hatebase is a multilingual hate speech database that helps organizations track and identify hate speech trends worldwide. It provides structured data for improving machine learning models in hate speech detection.

Limitations:

It primarily focuses on keyword-based detection, which can lead to high false positive rates.

Facebook's Community Standards Enforcement

Description:

Facebook employs AI-driven moderation tools to detect and remove hate speech before it spreads. The platform also relies on user reports and human moderators to refine its detection process.

Limitations:

The system often faces challenges in accurately detecting nuanced hate speech, especially in different languages and cultural contexts. It also depends on manual intervention, which delays response time and allows harmful content to remain accessible.

Problem Statement

INSTANDER is a hate speech detection system leveraging NLP and machine learning for real-time identification and moderation of harmful content on social networks. It ensures context-aware detection, automated intervention, and efficient database management, fostering a safer and more inclusive digital environment.

Objective

The goal of INSTANDER is to develop a robust system for detecting and mitigating hate speech on online social networking platforms. The platform addresses key challenges such as ineffective content moderation, lack of contextual understanding, high false positives, and delayed intervention. INSTANDER utilizes advanced NLP, machine learning algorithms, and a scalable database management system to provide an integrated approach for real-time monitoring, automated moderation, and comprehensive record-keeping. The ultimate objective is to ensure a safe, inclusive, and well-regulated digital environment for all users, minimizing harmful interactions while preserving free expression.

Methodology

Requirement Gathering and Analysis:

- Engage with stakeholders such as social media platform administrators, content moderators, and cybersecurity experts to understand challenges in hate speech detection.
- Define key functionalities, including real-time text analysis, contextual hate speech detection, automated moderation, and database management for flagged content.

System Design and Architecture:

- Design a **modular and scalable** system architecture to handle real-time processing efficiently.
- Create class diagrams, entity-relationship diagrams (ERDs), and data flow diagrams (DFDs) for each module.

Module Development:

- Text Processing and NLP Analysis: Implement machine learning-based classification models to analyze and categorise text as hate speech or non-offensive content.
- Contextual Understanding: Use deep learning and sentiment analysis to improve accuracy in detecting implicit and coded hate speech.
- Automated Moderation System: Develop a system to flag, warn, or remove detected content based on severity and user reports.

Database Design and Integration:

- Design a normalized relational database schema to store flagged content, user reports, and moderation actions.
- Implement CRUD operations and indexing for fast retrieval and efficient moderation workflows.

User Interface Development:

- Start with a console-based interface for initial testing and model evaluation.
- Plan for a web-based dashboard using frameworks like React or Flask to allow administrators to review flagged content and take appropriate actions.

Testing and Debugging:

- Unit testing for individual modules to validate NLP accuracy and database operations.
- Integration testing to ensure seamless interaction between detection models, database, and user
- Performance testing to assess real-time processing capabilities under high data loads.

Algorithms

Datasets

- HateSpeechDataset, Stormfront, Twitter Hate Speech, Facebook Hateful Memes.

Data Structures

- Text tokenization and embedding: Word2Vec, BERT embeddings as vector representations.
- Trie and Hash Tables: Fast lookup of offensive words and phrases.
- Graph structures: Social network analysis for identifying hate speech spreaders.
- Priority Queue (Heap): Ranking flagged messages based on severity.
- Bloom Filters: Efficient detection of previously flagged content.
- Sparse Matrices: TF-IDF and NLP feature extraction for classification.

SWOT Analysis

Strengths:
- High accuracy in hate speech detection using advanced NLP and deep learning models such as BERT and LSTMs.
- Real-time moderation enabled by efficient data structures like Trie, Hash Table, and Graphs for fast content filtering.
- Context-aware detection that can analyze implicit and coded hate speech, reducing false negatives.
- Scalable system that supports large-scale platforms with distributed processing and cloud integration.

Weaknesses:
- High computational costs due to the processing power and storage required for deep learning models.
- Potential for false positives and bias, as AI models may misclassify content or reflect biases from training data.
- Dependency on large datasets for training, which are necessary to maintain high accuracy.
- Limited multilingual support, making it difficult to detect hate speech across multiple languages and dialects.

Opportunities:
- Integration with social media platforms like Facebook, Twitter, and YouTube for enhanced content moderation.
- Adaptive learning and AI improvements through continuous reinforcement learning to enhance detection accuracy.
- Expansion into video and audio moderation, incorporating speech and visual content analysis.
- Support from regulatory frameworks as governments impose stricter regulations on online hate speech.

Threats:
- Evasion techniques where users modify spelling, use memes, or coded language to bypass detection.
- Legal and ethical challenges related to balancing content moderation with free speech concerns.
- Data privacy concerns requiring compliance with regulations such as GDPR and CCPA.
- Competition from established AI moderation tools developed by major tech companies like Google and Meta

AREA OF APPLICATION

. Social Media Platforms – Helps platforms like Facebook, Twitter, Instagram, and YouTube detect nd moderate hate speech in posts, comments, and messages.

. Online Forums and Communities – Ensures safe discussions on platforms like Reddit, Discord, and uora by identifying and filtering toxic content.

News Websites and Blogs – Monitors user-generated comments on news articles and blogs to event the spread of hateful and offensive language.

Educational Platforms – Prevents cyberbullying and harassment in online learning environments ch as Coursera, Udemy, and university discussion boards.

Gaming Communities – Moderates in-game chats, voice communications, and forums in ultiplayer online games to reduce toxicity and abusive behavior.

nclusion

STANDER is a comprehensive and intelligent hate speech detection system designed to nce online safety and content moderation. By leveraging advanced natural language essing (NLP), deep learning models, and efficient data structures, it ensures real-time ntification and filtering of offensive content across various digital platforms. The system grates scalable machine learning algorithms with robust database management, enabling less text processing, contextual analysis, and automated moderation.

applications spanning social media, online forums, gaming communities, educational rms, and law enforcement, INSTANDER aims to create a safer and more inclusive l environment. Despite challenges such as evasion techniques, computational costs, and l concerns, the system's adaptive learning capabilities and continuous model vements help enhance its accuracy and effectiveness.

ddressing hate speech detection with an innovative and data-driven approach, ANDER stands as a crucial tool in combating online toxicity, ensuring responsible digital unication, and promoting healthier user interactions across the internet.

nces

Nazmine, M. K. Tareen, H. K. Tareen, S. Noreen, and M. Tariq, "Hate Speech and Social Media: A iatic Review," Turkish Online Journal of Qualitative Inquiry (TOJQI), vol. 12, no. 8, pp. 5285–ul. 2021.

Kovács, P. Alonso, and R. Saini, "Challenges of Hate Speech Detection in Social Media," SN ter Science, vol. 2, p. 95, Feb. 2021, doi: 10.1007/s42979-021-00457-3.

Al-Hassan and H. Al-Dossari, "Detection of Hate Speech in Social Networks: A Survey on ngual Corpus,"

K. Vishu Tyagi and S. Das, "Deep Learning for Hate Speech Detection in Social Media," in Proc. EEE 4th Int. Conf. Comput., Power Commun. Technol. (GUCON), Kuala Lumpur, Malaysia, Sep.

SOC 2024: Hate Speech and Offensive Content Identification in English and Bangla.