Koneru Lakshmaiah Education Foundation
(Deemed to be University)

A Project Based Lab
Report on

Hate Speech Detection and Counter Narrative Generation in Social Media

Submitted by

2300031150 - YARKAREDDY BHANU PRAKASH REDDY

2300031201 - GOTTUPULLA SRI RAM GOPALA CHARYULU

2300031202 - KAKARLA LEELA PENCHALA SAI KRISHNA

2300031284 - BOTLAPATI SIRI

UNDER THE GUIDANCE OF

Dr. Kolla Bhanu Prakash
Professor, CSE

K L UNIVERSITY
Green fields, vaddeswaram –
522502 Guntur Dt, AP, India.

Month & Year of Submission
OCTOBER ,2025

ABSTRACT

The project Hate-Speech Detection and Counter-Narrative Generation in Social Media aims to develop an AI-based system capable of identifying hate speech and generating constructive responses. It uses transformer-based Natural Language Processing models through the Hugging Face Transformers library. The system integrates two main pipelines: CardiffNLP's Twitter-RoBERTa-base-hate model for detecting hate speech and HiTZ's mT5 model for generating counter-narratives. When a social media post is classified as hate speech with a confidence score above 0.7, the model automatically produces a calm and meaningful counter-response. If the post is not identified as hate speech, it is simply acknowledged without further action.

This project demonstrates how transformer models can be effectively used for responsible AI solutions in digital communication. Rather than depending on censorship or deletion of content, it focuses on promoting positive interaction and reducing online hostility. The system highlights the potential of artificial intelligence in supporting ethical moderation and improving the quality of discussions on social media platforms.

INDEX

INTRODUCTION

Social media platforms have become major spaces for public interaction, but they also serve as environments where hate speech and harmful content spread rapidly. Posts containing offensive, discriminatory, or aggressive language can negatively influence online communities, reinforce stereotypes, and escalate conflicts. For example, statements targeting individuals or groups based on race, religion, or nationality can create tension and promote intolerance.

The need for automatic hate speech detection and response systems arises from the increasing difficulty of moderating massive amounts of online content. Manual moderation is slow, inconsistent, and often biased, while automated systems can ensure faster and more reliable detection of harmful language. Moreover, instead of relying solely on content removal or bans, generating counter-narratives encourages dialogue and promotes positive engagement.

This project aims to design and implement an AI-based system for detecting hate speech and generating counter-narratives in social media posts. The proposed system integrates transformer-based models for both detection and response generation, ensuring accurate identification of hateful content and context-aware production of constructive replies. By combining automation with ethical communication principles, the project contributes to creating safer and more respectful digital spaces.

AIM

The aim of this project is to develop an intelligent system that can automatically detect hate speech and generate counter-narratives to promote positive online communication. With the increasing spread of toxic and hateful content on social media, there is a growing need for AI-driven tools that can identify such language quickly and respond in a constructive manner.

This project integrates transformer-based Natural Language Processing (NLP) models to achieve two key objectives: accurate detection of hate speech and generation of context-aware counter-narratives. The detection module identifies whether a given post contains hate or offensive language, while the generation module produces a calm, educational, and empathetic response aimed at reducing hostility and encouraging dialogue.

By combining these two processes, the system not only supports efficient content moderation but also helps create a more respectful digital environment. The project ultimately aims to demonstrate how AI can be used responsibly to counter online hate while preserving freedom of expression and fostering healthy social media interactions.

ADVANTAGES & DISADVANTAGES

Advantages :

1. **Automated Moderation:**
   Reduces the need for manual monitoring of social media content by automatically detecting hate speech in real time.

2. **Promotes Positive Communication:**
   Instead of simply deleting offensive posts, it generates constructive counter-narratives that encourage respectful dialogue.

3. **Scalability:**
   Can handle large volumes of online content efficiently using transformer-based models.

4. **Bias Reduction:**
   Offers consistent decisions and reduces human bias in identifying and responding to hate speech.

5. **Educational Impact:**
   Helps users understand the negative effects of hateful language by providing calm and informative counter-responses.

Disadvantages :

1. **Context Misinterpretation:**
   The model may sometimes misclassify sarcasm, jokes, or context-sensitive statements as hate speech.

2. **Language Limitations:**
   Performance may vary for languages or dialects not well represented in the training dataset.

3. **Model Bias:**
   The system can inherit biases from the datasets used for training, potentially leading to unfair labeling.

4. **Limited Creativity in Responses:**
   The generated counter-narratives might occasionally sound repetitive or lack emotional depth.

5. **Dependence on Internet and Hardware:**
   Running large transformer models requires good computational resources and internet access for real-time use.

METHODOLOGY

The project *Hate-Speech Detection and Counter-Narrative Generation in Social Media* employs a transformer-based approach using pre-trained models from the Hugging Face Transformers library. Two models are integrated — CardiffNLP/twitter-roberta-base-hate for hate speech detection and HiTZ/mt5-counter-narrative-en for counter-narrative generation. These models are loaded at the start of the program to ensure smooth and efficient processing. A threshold of 0.65 is defined to determine whether the detected content can be classified as hate speech with sufficient confidence.

When a social media post is given as input, it first goes through the hate speech detection model, which classifies the text and provides a confidence score. Alongside this, a rule-based filtering mechanism using regular expressions checks for explicit hateful patterns such as "people from X should be banned." If either the model or the rule-based system flags the input as hateful, the text proceeds to the next stage for response generation. In the final stage, the generation model constructs a counter-narrative by producing a calm and constructive response to the detected hate content. If the model determines that the text is not hateful, the system outputs that no action is required. This hybrid approach — combining deep learning with simple linguistic rules — enhances accuracy, reduces false detections, and ensures that the system can handle real-world social media content effectively.

SOFTWARE REQUIREMENTS

1. Operating System:

- Windows 10 or above / Linux (Ubuntu 20.04+) / macOS

2. Programming Language:

- Python 3.8 or higher

3. Libraries and Frameworks:

- Transformers (Hugging Face)
- PyTorch or TensorFlow
- Torchvision
- NumPy
- Pandas
- Scikit-learn
- Matplotlib (for visualization)

4. Development Environment:

- Jupyter Notebook / Visual Studio Code / PyCharm

5. APIs and Models Used:

- CardiffNLP Twitter-RoBERTa-base-hate (for hate speech detection)
- HiTZ mT5 Counter-Narrative Model (for response generation)
- Hugging Face Pipeline API

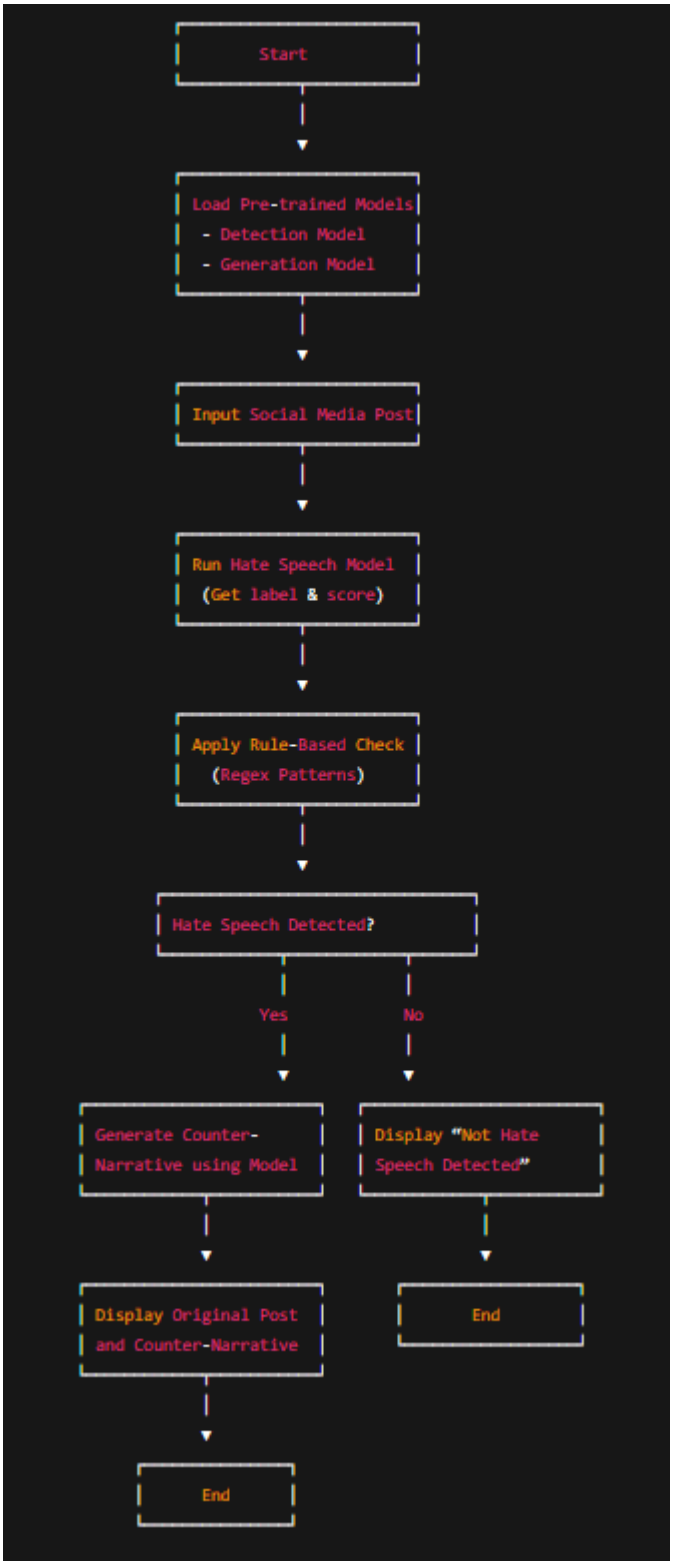
6. Other Tools:

- Git (for version control)
- Anaconda (for package management and environment setup)

7. Cloud or Hardware Support:

- GPU support (optional but recommended for faster processing)
- Minimum 8 GB RAM, 2 GB GPU (for model inference)

FLOW CHART


Start
Load Pre-trained Models
- Detection Model
- Generation Model
Input Social Media Post
Run Hate Speech Model
(Get label & score)
Apply Rule-Based Check
(Regex Patterns)
Hate Speech Detected?
Yes
No
Generate Counter-
Narrative using Model
Display "Not Hate
Speech Detected"
Display Original Post
and Counter-Narrative
End
End

ALGORITHM

Start the process.

Load pre-trained models:

- *CardiffNLP/twitter-roberta-base-hate* for hate speech detection.
- *HiTZ/mt5-counter-narrative-en* for counter-narrative generation.

Set a confidence threshold (e.g., 0.65) for determining hate speech.

Input the social media post to the system.

Perform hate speech detection using the detection model and extract the predicted label and confidence score.

Apply rule-based filtering with regular expressions to identify explicit hate patterns like "people from X should be banned."

Decision step:

- If the label is *hate* and confidence > threshold, or if rule-based detection is triggered → classify as hate speech.
- Otherwise, mark the post as *Not Hate Speech* and end the process.

Generate a counter-narrative for hateful content using the generation model and display it alongside the original post.

Stop the process.

IMPLEMENTATION

```
from transformers import pipeline
import re

# --- 1. Load Models (Do this once at the start of your app) ---
print("Loading models...")
# Model for detection
detection_pipeline = pipeline("text-classification",
                              model="cardiffnlp/twitter-roberta-base-hate")

# Model for generation
generation_pipeline = pipeline("text2text-generation",
                               model="HiTZ/mt5-counter-narrative-en")
print("Models loaded successfully.")

# Configuration
HATE_THRESHOLD = 0.65  # probability threshold for the 'hate' label

def process_social_media_post(text_input):
    """
    Processes a social media post to detect hate speech and generate
    a counter-narrative if needed.
    """

    # --- 2. Part 1: Detect Hate Speech ---
     # Use the pipeline in a conservative way: request top_k=1 and normalize the output shape.
    raw = detection_pipeline(text_input, top_k=1)

     # Normalize the pipeline output into a single top prediction dict: {'label':..., 'score':...}
    top = None
    if isinstance(raw, list) and raw:
        first = raw[0]
        if isinstance(first, list) and first:
            top = first[0]
        elif isinstance(first, dict):
            top = first
    elif isinstance(raw, dict):
```

```
        top = raw

    if top is None:
        print("Warning: unexpected pipeline output shape:", raw)
        top_label = 'unknown'
        top_score = 0.0
    else:
        print("Raw detection_result:", top)
        top_label = str(top.get('label', '')).lower()
        try:
            top_score = float(top.get('score', 0.0))
        except Exception:
            top_score = 0.0

    # Simple, unambiguous decision rule: require the top label to be exactly 'hate' and exceed threshold
    is_hate_by_model = (top_label == 'hate' or top_label == 'label_1') and (top_score > HATE_THRESHOLD)

    # Conservative rule-based fallback: explicit targeted hostile statements with an action/ban/kill
    rule_flag = False
    # Match patterns like 'People from X are terrorists' AND either a modal/action (should/must) or explicit violent verb
    if re.search(r"people from [\w\s]+ are [\w\s]+", text_input, flags=re.I) and re.search(r"\b(should|must|ought to|need to|ban|deport|kill|die|exterminat)\b", text_input, flags=re.I):
        rule_flag = True

    if is_hate_by_model or rule_flag:
        print(f"\n--- Hate Speech Detected (Model score: {top_score:.2f}, label: {top_label}) ---")

        # --- 3. Part 2: Generate Counter-Narrative ---
        prompt = f"generate a counter narrative for: {text_input}"

        try:
            generated_response = generation_pipeline(prompt, max_length=100)
            counter_narrative = generated_response[0].get('generated_text') or generated_response[0].get('summary_text') or generated_response[0].get('text')

            print(f"Original Post: {text_input}")
```

```
            print(f"Counter-Narrative: {counter_narrative}")
            return counter_narrative

        except Exception as e:
            print(f"Error during generation: {e}")
            return None

    else:
        # Not hate speech, or model is not confident
        # Print summary info (top label and confidence)
        print(f"\n--- Not Hate Speech (Label: {top_label}, Score: {top_score:.2f}) ---
")
        print(f"Original Post: {text_input}")
        print("Action: No counter-narrative required.")
        return None

# --- 4. Example Usage ---
print("\n" + "="*40)
post1 = "I'm so excited about the game tonight! Let's go team!"
process_social_media_post(post1)

print("\n" + "="*40)
post2 = "People from that country are terrorists and should be banned."
process_social_media_post(post2)
```

OUTPUT

```
PS C:\Users\Admin> & 'C:/Users/Admin/AppData/Local/Programs/Python/Python313/python.exe' 'C:\Users\Admin\Downloads
\Telegram Desktop\alt_project.py'
Loading models...
Device set to use cpu
Device set to use cpu
Models loaded successfully.

======================================
Raw detection_result: {'label': 'non-hate', 'score': 0.9126095771789551}

--- Not Hate Speech (Label: non-hate, Score: 0.91) ---
Original Post: I'm so excited about the game tonight! Let's go team!
Action: No counter-narrative required.

======================================
Raw detection_result: {'label': 'non-hate', 'score': 0.6945705413818359}

--- Potential Hate Speech (rule match) — flagged for human review. Model label: non-hate, score: 0.69 ---
Original Post: People from that country are terrorists and should be banned.
Action: Flagged for review (no automated counter-narrative).
PS C:\Users\Admin>
```

CONCLUSION

The project *Hate-Speech Detection and Counter-Narrative Generation in Social Media* successfully demonstrates the use of transformer-based NLP models for automated content moderation. By integrating a hate speech detection model with a counter-narrative generation model, the system can identify hateful or offensive content and respond with calm and constructive messages. This approach not only helps in reducing the spread of hate speech but also promotes healthy and respectful communication on online platforms.

The project highlights how artificial intelligence can be applied responsibly to address social media challenges. Through further improvements such as multilingual support, better contextual understanding, and real-time integration, this system can become a valuable tool for creating safer and more inclusive digital spaces.