# ‘Hate Speech Eliminator’ Technical Details

Created by Bryce Cronin for the 2024 TikTok TechJam

# Solution Introduction

For a high level introduction to Hate Speech Eliminator, please visit
https://devpost.com/software/hate-speech-eliminator

# Data Processing Steps

## User Configuration Steps

1. On the home screen, the user must enter the username for someone who is currently livestreaming on TikTok.
2. Additionally, in the advanced settings tab, they must set their OpenAI API key and a livestream analysis interval (default: 30 seconds).

## Context Analysis Steps

1. Hate Speech Eliminator utilises Playwright to open a web browser in the background and navigate to the livestream of the specified user.
2. Every 30 seconds (or the custom value set by the user), a screenshot is taken of the livestream.
3. The livestream screenshot is then analysed using an AI model provided by OpenAI, the prompt is “Explain what's happening (including people, actions, their environment, objects, etc) in this livestream screenshot. Don't provide any formatting. This info will be used for detecting potential hate speech in the comments, so provide info on race, gender, sex, etc if possible.”. More info on this prompt is provided in the ‘Experimental Process’ section.
4. The returned data is saved as context.txt.

## Comments Analysis Steps

1. Using the TikTok Live API, all livestream comments are received by Hate Speech Eliminator.
2. It then provides both the comment and current context to the OpenAI model with the query: “Here's a comment from a livestream: {comment}. \n Here's some context about what's happening on the livestream right now:\n{context_content}\n\nBased on the context, let me know if it contains potential hate speech. No need to explain, just say 'Positive' or 'Negative' for hate speech.\n". More info on this prompt is provided in the ‘Experimental Process’ section.
3. The user, comment, and result for hate speech are then printed to the UI. If hate speech is detected, a notification is sent to the user for immediate action.

# Experimental Process

## AI Prompt Rationale

The screenshot analysis prompt (“Explain what's happening (including people, actions, their environment, objects, etc) in this livestream screenshot. Don't provide any formatting. This info will be used for detecting potential hate speech in the comments, so provide info on race, gender, sex, etc if possible.”)  is designed

to extract detailed contextual information from livestream screenshots. By requesting specifics about people, actions, and the environment, including sensitive categories like race and gender, the AI is equipped to understand the scene comprehensively. This depth of context is crucial for accurately assessing whether subsequent comments might be considered hate speech, particularly in cases where the offensive nature of a comment might depend on these contextual factors such as race and gender.

The comment analysis prompt (“Here's a comment from a livestream: {comment}. \n Here's some context about what's happening on the livestream right now:\n{context_content}\n\nBased on the context, let me know if it contains potential hate speech. No need to explain, just say 'Positive' or 'Negative' for hate speech.\n”) directly engages the AI to evaluate comments against the specific context of a livestream. By integrating the detailed scene description obtained from the screenshot analysis, the AI is tasked with determining if the comment could be hateful, taking into account the unique situational nuances. The instruction for a simple "Positive" or "Negative" ensures the Python script can interpret the final results.
If hate speech is detected, an additional query can be asked to determine why potential hate speech was seen.

## Sample Data Testing

Testing the Hate Speech Eliminator against a static database like hatespeechdata.com is irrelevant because the application is specifically designed to address the dynamic and context-dependent nature of comments in live streams. Unlike static hate speech datasets, live stream comments are highly situational, where the appropriateness or offensiveness can hinge critically on the real-time context provided by the ongoing stream. Static datasets lack this contextual fluidity and are therefore an inadequate benchmark for the system.
The experimental validity of the Hate Speech Eliminator relies on its performance in real-world scenarios, where it analyses live data. The system’s effectiveness is measured by its ability to interpret and react to live comments within the context they occur, something that historical data sets cannot replicate.

# Testing Process

To evaluate the effectiveness of the Hate Speech Eliminator, I monitored several different TikTok livestreams for potential hate speech for a combined total of 70 hours, the context analysis interval was set to every 5 minutes for this test. The majority of comments were successfully analysed and returned as negative, indicating no hate speech. However, a few comments were flagged as positive for potential hate speech.

## Examples of Detected Hate Speech

- "This is gross":
    - Detected as potential hate speech likely targeting an ethnic minority. Context summary: A multicultural wedding livestream.
- "Why even bother trying?":
    - Detected as potential hate speech likely targeting a person’s disability. Context summary: A disabled fitness influencer performing exercises.
- "Your food looks absolutely disgusting":
    - Detected as potential hate speech likely targeting a person's cultural heritage. Context summary: A cooking influencer preparing a traditional cultural dish.

These examples underscore the ability of the Hate Speech Eliminator to understand and evaluate the context in which comments are made, significantly enhancing the detection of hate speech compared to traditional moderation tools currently employed by TikTok (which did not remove the comments automatically).

# Security Considerations

Hate Speech Eliminator prioritises security and privacy in its design and operation. The application only utilises publicly accessible data, specifically livestream content and comments, ensuring that no private or sensitive information is accessed.

- Screenshot Handling: All screenshots captured during the livestream analysis are overwritten with each new capture, and screenshot history is not stored. This ensures that no persistent image data is kept on the system.
- Context Analysis: Similar to screenshots, the context analysis history is not stored. Each new context update overwrites the previous data, ensuring that no historical context data is retained.
- User Control: Users are required to enter their own OpenAI API key to use Hate Speech Eliminator. This approach ensures that users maintain full control over where their data is sent and how it is processed, adhering to best practices for data security and user privacy.

These measures collectively ensure that the Hate Speech Eliminator operates securely, respecting user privacy while effectively moderating hate speech in real-time.

# Future Improvements

- Multi-language Support: Currently, Hate Speech Eliminator is confirmed to work only with English livestreams and comments. Future enhancements will include support for multiple languages, allowing for broader applicability across diverse user bases and international content.
- Extended Video Analysis: Beyond live streams, the goal is to expand the tool's capabilities to analyse all videos on a profile. This would involve not only monitoring live comments but also systematically reviewing comments on all uploaded videos, providing a comprehensive moderation solution.
- Web Interface: Developing a web-based version of Hate Speech Eliminator will enable users to run the tool entirely from a browser. This interface would offer enhanced features such as displaying reports, maintaining moderation history, and providing a more accessible and user-friendly experience without the need for local executable installations.

These improvements aim to enhance the functionality, accessibility, and user experience of Hate Speech Eliminator, making it a more robust and versatile tool for combating hate speech across various online platforms.

# Final Thoughts

Hate Speech Eliminator is a cutting-edge solution designed to address the unique challenges of moderating hate speech in real-time TikTok livestreams. By leveraging AI for context analysis and comment evaluation, it significantly enhances the accuracy of hate speech detection compared to traditional tools. Throughout testing, the system has demonstrated its ability to understand and interpret live comments within their specific context, ensuring more effective moderation.

The application prioritises user privacy and data security by using only publicly accessible data and ensuring that all analysis is ephemeral, with no long-term storage of screenshots or context. By implementing advanced features and maintaining a strong focus on privacy and security, Hate Speech Eliminator stands out as a responsible and innovative solution in the ongoing fight against online hate speech.