CSCE 5290: Natural Language Processing

Project Proposal

Group-11

Chandana Vallabhaneni - 11680559

Lakshmi Nanditha Reddy sura - 11684611

Bala Naga Narasimha Reddy Machha -11671010

Sai Akhil Pinni -11703725

GitHub repository Link: https://github.com/orgs/NLP-Group-11/teams/collaborators/repositories

Title: Web-Based Multilingual Speech Text Hate Speech Detection: Identifying Harmful Content in Transcribed Speech Across Languages

1. Motivation

The proliferation of speech-based content on digital platforms, including transcripts from podcasts, voice messages, and videos, has created a significant challenge in identifying harmful speech across different languages. The main issue this project seeks to address is the need for a unified system capable of processing transcribed speech in multiple languages and identifying toxic content, which is essential for maintaining safe and inclusive online environments. By leveraging an existing multilingual hate speech dataset, we aim to create a robust solution that can effectively detect hate speech in transcribed content across various languages.

2. Significance

This project holds considerable significance in the field of content moderation and online safety, particularly for platforms dealing with transcribed speech content. By developing a tool that can analyze transcribed spoken language input in various languages and detect harmful content, we can:

- Enhance the efficiency and accuracy of hate speech detection for transcribed speech across linguistic boundaries
- Contribute to creating safer online spaces by promptly identifying toxic speech in transcripts
- Support platforms that handle transcribed content in managing multilingual moderation

- Advance hate speech detection technologies specifically for the nuances of transcribed spoken language
- Promote inclusivity by ensuring moderation capabilities across diverse languages and cultures
- Address the unique challenges posed by transcribed speech, such as informal language and context-dependent meanings

3. Objectives

The primary objectives of this project are:

1. Develop a web-based hate speech detection system that processes transcribed speech input in multiple languages
2. Implement robust text processing techniques to handle the unique characteristics of transcribed speech
3. Utilize the existing multilingual hate speech dataset to train an effective detection model
4. Build a user-friendly interface for inputting or uploading transcribed speech for analysis
5. Achieve near real-time processing and analysis of transcribed speech content
6. Attain a minimum accuracy of 90% in hate speech detection across supported languages.

In this It will use two main components: Speech Recognition and Text Classification, each with two models.

For speech recognition, we will use Model M1 (ex: Data Speech) and Model M2 (ex: Wav2Vec 2.0). These models will convert spoken language into text. The output from either M1 or M2 will then be used as input for the text classification stage.

For text classification, the transcribed text will be processed by Model M3 (ex: XLM-R) and Model M4 (ex: mBERT - Multilingual BERT). These models will analyze the text and detect hate speech, classifying it as offensive or neutral.

Success will be measured by:

- The system's ability to process transcribed speech in all supported languages
- Accuracy rates of hate speech detection in transcribed speech content
- Processing speed and near real-time capabilities
- Robustness in handling various speech patterns and colloquialisms
- User feedback on the web interface and overall system functionality

4. Features

Key technical features and deliverables of the project include:

1. Text Processing for Transcribed Speech:
    - Handling of speech disfluencies and informal language ○ Context-aware processing to capture speech-specific nuances
2. Multilingual Text Analysis:
    - Support for multiple languages based on the existing dataset
    - Handling of code-switching and mixed-language text
3. Hate Speech Detection Model:
    - Training using the existing multilingual hate speech dataset
    - Fine-tuning for transcribed speech characteristics
    - Classification of content into appropriate categories (e.g., offensive, neutral)
4. Web Application:
    - User-friendly interface for text input or transcript upload
    - Real-time processing and result display
5. Integration Pipeline:
    - Seamless connection between text processing and hate speech detection modules
    - Optimization for low-latency processing of large text inputs
6. Detailed Reporting:
    - Generation of comprehensive analysis reports for transcribed content
    - Visualization of detection results with text highlights

5. Dataset

The project will utilize the existing multilingual hate speech dataset:

- Content: Labeled examples of hate speech and neutral content across multiple languages Data Preparation:

- Preprocessing to normalize text and handle speech-specific characteristics
- Potential augmentation techniques to enhance model performance on transcribed speech
- Splitting into training, validation, and test sets to ensure robust model evaluation

6. Visualization

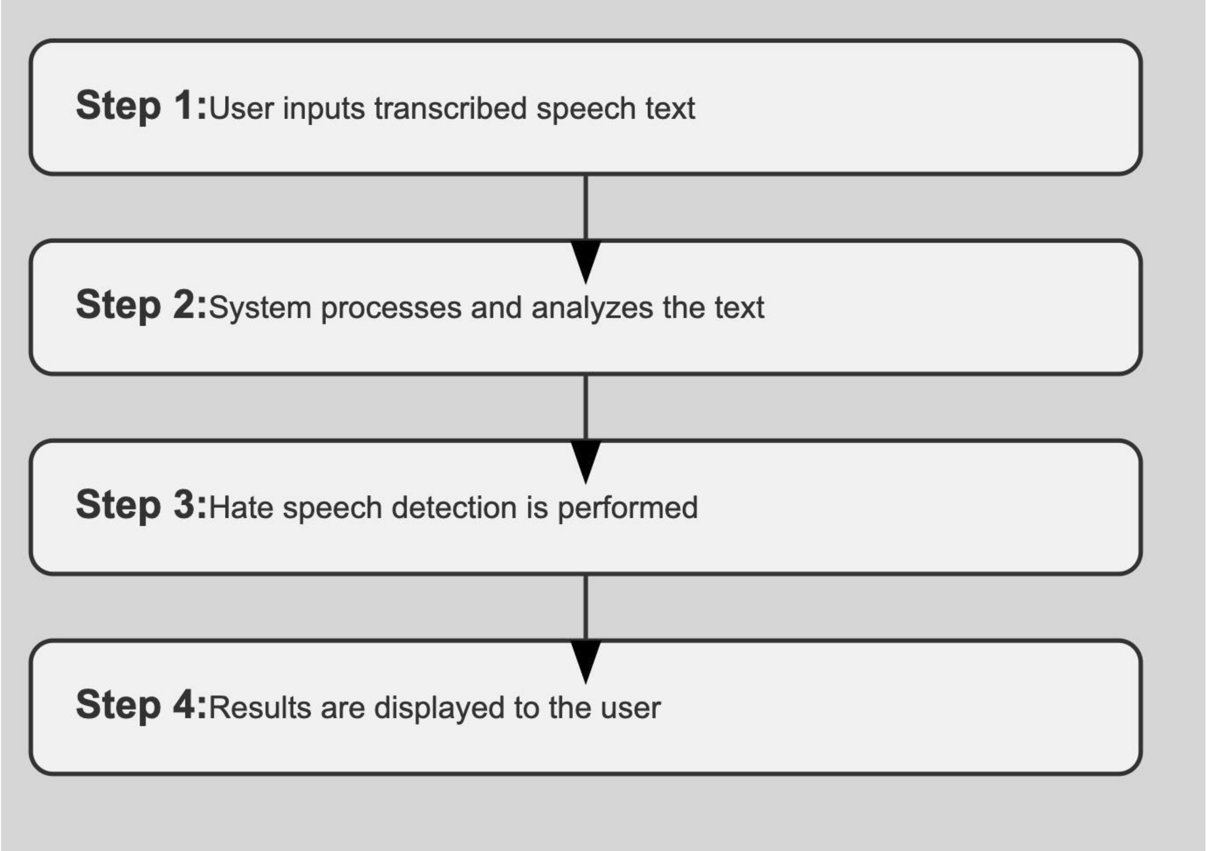
Step 1:User inputs transcribed speech text
Step 2:System processes and analyzes the text
Step 3:Hate speech detection is performed
Step 4:Results are displayed to the user

Figure 1: Simplified Speech Text Hate Speech Detection Workflow

The workflow diagram illustrates the key components and data flow of the proposed system:

1. User Input: The user enters or uploads transcribed speech text.
2. Text Preprocessing: The system prepares the text for analysis, handling speechspecific characteristics.
3. Hate Speech Detection: The trained model analyzes the text for hate speech.
4. Result Display: The system presents its findings to the user, highlighting any detected hate speech.