Project 3 Report

Adrian Hoang

March 28, 2025

1 Overview

This project is divided into two main components. The first component is the Speech-to-Text Conversion, where a YouTube video is converted into an audio file and then transcribed using a speech-to-text model from Hugging Face's Transformers library. The second component is Defamation Detection, which involves fine-tuning an open-source language model on a custom dataset containing defamatory and non-defamatory statements. The fine-tuned model is then applied to analyze each sentence in the transcript.

2 Design Decisions

For the Speech-to-Text Conversion, the process began by downloading the video using `youtube-dl` and subsequently extracting the audio using `ffmpeg`. The audio was converted into a `.wav` file with a sampling rate of 16 kHz and a single audio channel to match the requirements of the chosen ASR model. The OpenAI Whisper-base model was selected due to its robust performance with longer audio segments. Although the initial transcription was generated automatically, minor inaccuracies were later found. However they were not major enough to warrant creating a new, manual transcription.

The Defamation Detection component required the creation of a custom dataset. Due to the small size of the dataset, we simply encoded the statements into a list of python dictionaries. Ten defamatory and ten non-defamatory statements were compiled, ensuring that their tone and style mirrored the courtroom dialogue of the transcript. This tailoring of the dataset to the domain helped create noticeably better testing results of our model compared to using generic defamatory/non-defamatory sentences. The small dataset size meant that training did not take long and required minimal computational resources. The `distilbert-base-uncased` model was chosen as the starting point for its efficiency and strong performance in text classification tasks. The model was fine-tuned using Hugging Face's `Trainer` API, and the training process was tailored to accommodate the limited dataset size. As expected, due to the small number of training samples, the fine-tuned model did not achieve a high level of accuracy upon examination. For inference, the transcript was segmented into individual sentences using NLTK's sentence tokenizer, and each sentence was classified by the fine-tuned model as either defamatory or non-defamatory.

3 HPC Settings

The project was executed on an HPC instance configured with 4 CPU cores (with 6 GB of memory per core) and 1 GPU core. Additionally, due to the small size of the dataset, fine tuning did not require much CPU/GPU resources. No additional API keys (such as those for Weights & Biases) were necessary during the training process.

4 Substitute Transcription

While the initial automatic transcription was generally acceptable, it contained minor errors. However, we decided these errors were not substantial enough to warrant generating our own transcription.