A screenshot of a cell phone

Description automatically generated

Models & Architecture

There were many steps and models explored in the creation of Too Long Didn't Watch, along with interesting challenges. As a team, we generally expected product review videos to be organized and relatively easy to "process" through NLP libraries. However, two areas required significant attention.

First, many videos are "auto-captioned" which means YouTube automatically adds captions but no punctuation. Video posters can add their own captions and include punctuation, but captioning a 20-min video can be tedious. The lack of punctuation provided a challenge when trying to identify sentences and associated product features with descriptors. Second, many product review videos are conversational in nature, complete with filler words (um's/uh's/you know/like) and frequent jumping back and forth between topics and product features.

Obtaining the Text

During the early stages of our project, one of the first tasks that needed to be explored was how to extract the audio from a review video and convert it to text.

We initially built our own function that extracted the audio track and then converted the audio track to text. However, for ease of implementation, we opted to use AWS Transcribe to perform speech to text. AWS Transcribe also has the benefit of attempting to add punctuation to the output, which our model wasn’t capable of doing.

A second method of acquiring text from a video is to pull the video caption directly from YouTube. We implemented a pipeline that would accept a URL and then pull the captions for the video directly from YouTube. This method proved to be a lot quicker and provided pretty good transcripts of the videos. However, the captions provided by YouTube do not always contain punctuation.

After identifying the method to extract the video voice-over into text, we needed to figure out a way to extract keywords and sentences.