Replication of YouTube's User-Generated Content Summarization and Topic Extraction Model

Tommy Shang Wu, Jake McFarland, Abiel Jeong-Joon Kim

Abstract

The objective of our project is to recreate our own variant of YouTube's comment-section AI generator. The principle idea is to define a model that can take swaths of natural language text (such as YouTube comments) and output key overlapping topic in addition to high level summaries of the overall content of said topics. The scope of the application is primarily specialized for native YouTube comments. The scope of the technologies used is focused on transformer models as the architecture base.

1 Introduction

Summarization of content is a vital part of natural language processing (NLP). It is something that can be used in many scenarios, for everyday use and for large projects. Using AI as a tool for summaries has become more common with the growth of AI, and new developments push the field further and further.

2 Problem Statement and Motivation

YouTube released its generative comment summarizer and topic extraction model as a way to dynamically inform users on general viewership thoughts. One of the key features of YouTube's experimental AI is the ability to take swaths of user comments, identify key overlapping topics, and then generate summaries for aforementioned topics. This particular feature will be the focal point of our term project and relates to NLP tasks in the space of supervised deep-learning and constructing a seq2seq, generative architecture.

With respect to what aspects of the project we will work on, our group plans to implement both the topic extraction and generative summarizer system. This will require us to formalize a comprehensive methodology that combines fundamental NLP techniques with deep learning architectures.

We chose this project as a means to gain both broad and in-depth experience in NLP development. Constructing this system will require not just an understanding of deep neural architectures but also insight into defining interconnected NLP systems, applying embedding libraries, requirements engineering in the context of NLP, and exercising software development methodologies.

3 Related Work

The YouTube comment topics is a relatively new addition to the YouTube mobile app exclusively. It appears on large comment sections, with a summary of what many users are commenting about, including an example of each topic (Google, 2025b).

Microsoft's Bing has a "Deep Search" function which utilizes GPT-4 and RAG techniques to help refine search terms by understanding search intents, transforming basic queries into detailed, context-rich search requests that retrieve deeper and more targeted results from across the web.(Microsoft Corporation, 2023). Google also have similar AI search offerings(Google, 2025a). Similarly, Discord, a mass instant messaging platforms is also testing a new 'Summaries' feature which summarizes chats into topics and summarize(Discord, 2025).

4 Approach and Experimental Setup

4.1 High-Level Overview

Our project is decoupled into 2 primary subsystems, namely the topic-extraction stage followed by the generative summarizer system. At a high level, the system architecture will leverage rudimentary NLP techniques and statistical operations for the topic extraction model which should then iteratively pipeline into our deep learning summarizer model.

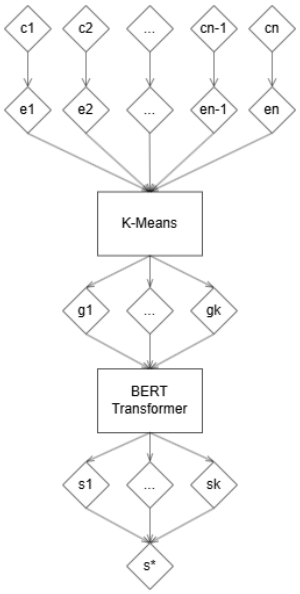

Figure 1: High Level System Pipeline

Figure 2: K-Means Topic Extraction Diagram

4.2 Topic Extraction Subsystem

As shown in Figure 2, the topic-extraction stage will be implemented using k-means clustering of weighted-average embedding vectors. More specifically, given a set of comments c1, c2, ..., cn a set of corresponding embedding vectors will be constructed e1, e2, ..., en; Each embedding vector, ei, is defined as the weighted average embedding of all word-vectors in ci. Mapping out these weighted-average vectors in multidimensional space will allow us to capture viewership semantic relations where k-means clustering (Figure 2) will be applied to identify overlapping or shared topics, g1, ..., gk. In this sense, each cluster represents a corresponding topic. Once clusters have been identified, we iteratively input each cluster set to the summarizer model where deep learning techniques are required.

4.3 Summarizer Model

The summarizer is formalized as a transformed-based seq2seq generative model that will be trained to take concatenated swaths of natural language text to generate a compressed and comprehensive summary. A few key pilot hyper-parameters we've chosen is the softmax activation, cross-entropy loss function, and the mean intersection over union evaluation metric (mIoU) for capturing type 1 and type 2 errors explicitly. In terms of system pipelining, once the transformer is trained, the deep model will iteratively take each cluster from the topic extraction phase and generate corresponding summaries s1, ..., sk. To clarify, member vectors of a given cluster will likely be concatenated or further compressed before being fed into the transformer.

We will prototype various configurations of the summarizer model with the intent of model cross-comparison using the mIoU evaluation metric. Likely, we will try BERT, BART, and T5 then pick the best one. Different hyper-parameter configurations will also be tested, choosing variations of loss functions, regularization, and dropout values. Once the transformer has processed all clusters, an ultimate executive summary, s*, of the combined clusters will finally be produced as a concluding paragraph.

4.4 Datasets

The replication will require multiple datasets, including one for generating summaries for word embeddings. For this purpose, a dataset provided by prithivMLmods hosted on HuggingFace is planned to be used(PrithivMLmods, 2025). It contains 98.4 thousand entries of large articles that are condensed to one-line headlines.

5 Timeline and Work Breakdown

The work for this project is intended to be split equally among all members of the project. As the project develops, the workload will likely shift to favor strengths in the group.

By the first milestone date, we intend to have the topic extraction subsystem working in a satisfactory manner. This includes the embedding vectors, as well as the k-mean clustering. The remainder of the project, the summarizer model, is to be completed after the milestone date has passed.

References

Discord. 2025. In-channel conversation summaries. Accessed: 2025-02-24.

Google. 2025a. Ai overviews. Accessed: 2025-02-24.

Google. 2025b. Explore comment topics. Accessed: 2025-02-24.

Microsoft Corporation. 2023. Introducing deep search. Accessed: 2025-02-24.

PrithivMLmods. 2025. Context-based-chat-summary-plus. Accessed: 2025-02-24.