A YouTube summarize and sentiment analysis tool

GitHub Link: https://github.com/KavyaBalasubramaniam/NLP-Text-Summarization-and-sentiment-analyser

Subbarao Sanka
subbaraosanka@my.unt.edu

Kavya Balasubramaniam
kavyabalasubramaniam@my.unt.edu

Anirudh Chowdary Ravipati
anirudhchowdaryravipati@my.unt.edu

Niveditha Sree Pendli
nivedithasreependli@my.unt.edu

***Abstract*—Sentiment analysis is the field of study related to users' emotions, opinions, evaluations, and emotions. The revolution around social media sites is attracting users to video-sharing sites like YouTube. This paper gives brief techniques to analyze options about a particular video posted by a user.**

Keywords: Sentiment Analysis, Reviews, YouTube, Text summarization

Introduction:

Text Summarization condenses a given piece of text into reduced version but at the same time retaining the key information. Manually performing this task is a tedious and time consuming activity.

Also trying to provide a brief content in a shorter period of time is always a attractive objective for the audience. Especially social medias always tend to observe higher visits to sites and platforms that provide a planned quick content delivery. Hence this Project work tends to optimize the information delivery from YouTube videos.

Additionally, we also aim to perform sentiment analysis on the summarized content for audiences who would like to perform classification tasks like customer segmentation and other review based analysis.

Goals and Objectives

I. Motivation

Problem statement:

Social media is becoming more and more popular since it is simple to use. YouTube is likely the most well-known of them all. YouTube is one application where most people spend a significant time. We tend to watch hour-long videos and try to understand the important ideas from them. Let’s think of a tool which will make our job easier. This tool can grab the key ideas from the video and present them in a text format. This is called the Summarization of the video which is the condensed version of the video.

The project deals with a specific set of YouTube videos i.e., Review videos of products, and implements NLP Text Summarization and Sentimental analysis to achieve the project goal.

II. Significance

The number of uploaded videos has been growing over the years, and researchers have been trying to do many analyses based on cultural and humanities-based approaches.

There are millions of creators making content every day and uploading them to YouTube and many of them include brands and advertisements in them which can potentially influence a lot of people and analyzing those details will provide a lot of insight into Brand promotion culture.

YouTube videos are multimodal and have several aspects that can be analyzed: visual imagery, metadata about the video(details of who made it, duration of the video, and descriptions of the videos), soundtracks used in the videos, recorded sounds, and transcripts. One way to analyze and go-over videos are through their transcripts.

Focusing on the video transcripts opens the door for textual analysis tools that can be used to search for word correlations, frequently used words or phrases, and topics. We can scrape the transcripts out of the videos and use them to analyze and perform Sentiment analysis. This can help us identify how consumers perceive a product, service, or brand in the video via sentiment analysis.

To evaluate whether data is good, negative, or neutral, natural language processing techniques like sentiment analysis are applied. Businesses frequently do sentiment analysis on textual data to track the perception of their brands and products in customer reviews and to better understand their target market. We can use the sentiment analysis on the video transcripts to better understand the way the brands can be better displayed.

III. Objectives

The Objectives of the Project are as follows:

- Provide summarization of video data from YouTube via transcripts
- Perform Sentiment analysis on the summarized data and provide analytical insights with visualizations.

Background of the study:

Sentiment analysis is a method for learning what use rs think and feel about a service or a product. One of the mo st widely used video-sharing websites YouTube receives millions of views daily.

The Project aims to scrape a set of data from the YouTube platform and extract the transcripts from it. There have been numerous academic attempts with two classes (Positive or Negative), three classes (two with neutral), or multiple classes (happy, sad, fear, surprise, and anger). Nevertheless, selecting the most accurate model can be difficult. We must clean the data by eliminating extraneous stop words that don't add anything to the sentiment analysis. As a result, efforts had been made to identify the polarity using sentiment analysis of YouTube comments. The final product will be a model that may demonstrate the effectiveness of a branding strategy or commercial and how viewers reacted to it.

IV. Features

- Speed up browsing of a large collection of video data by aggregating and summarizing YouTube content and achieving efficient access and representation.
- Identifying how consumers perceive a product, service, or brand in the video via sentiment analysis.

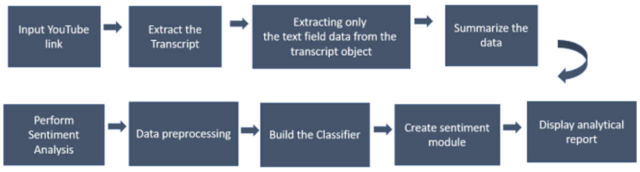


Fig1: Process Pipeline

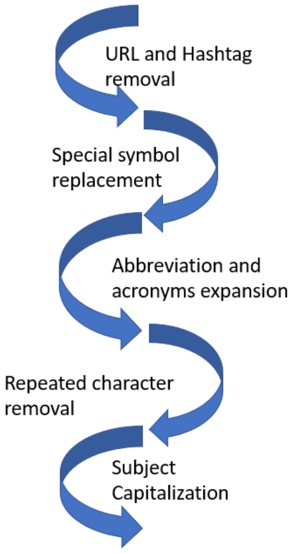

URL and Hashtag removal
Special symbol replacement
Abbreviation and acronyms expansion
Repeated character removal
Subject Capitalization

Fig 2: Data cleaning and preprocessing

Increment 1:

V. Related work

1. Sentiment classification using Supervised learning(Articles attached in the reference)

Pang et al's[1] study was the first to use supervised learning and the bag-of-words model to categorize movie reviews as positive or negative. They demonstrated that the basic machine-learning models might perform better than the baselines created by humans. SVM's outperformed naive Bayes in their results. In their discussion, they addressed the most crucial point that the bag-of-words will perform poorly when a small number of words represent the true overall sentiment.

In the case of movie reviews, a reviewer can make n number of comments about the film. It might be difficult for a bag of words to classify them. So, Gammon[2] classified such noisy customer data by using support vectors. It shows that machine learning classifiers can be used to classify the texts that humans would find difficult to work on by using feature extraction and reduction.

2. Sentiment analysis on YouTube comments:

A bachelor's thesis on the subject of estimating YouTube videos and proportions on its comments was written by Hyberg & laaacs[3]. They have used a straightforward formula for prediction based on the proportion of comments that supervised learning classifies, whether it's positive or negative. Also, their training is limited to classifying only positive and negative remarks. They also implemented the SVM, logistic regression, and multinomial naive Bayes classifiers.

3. Encoder-Decoder Deep Learning Models for Text Summarization:

Urvash Khandelwal[4] has used the RNN-based encoder-decoder for text summarization. The model in Rush et.al. is an encoder-decoder model where the encoder is a convolution network, and the decoder is a feedforward neural network language model.

VI. Dataset

The dataset we have used is a list of review videos from YouTube. The scope of the Project has constraints:

1. The Video must have transcripts available

2. Multilingual data not applicable

We have taken a video from YouTube and extracted the transcript from the video by using the URL with the help of youtube_transcript_api and provide the summarized text as output.

The summarized text from this step goes as the input for the sentiment Analysis.

Then we pre-process the data by applying text normalization techniques like lemmatization which converts any word form into its root word.

We also perform cleaning by removing the punctuations, convert them to lowercase so that the model becomes faster and more efficient.

VII. Detailed Design of Features:

1. Speed up browsing of a large collection of video data by aggregating and summarizing YouTube content and achieve efficient access and representation.
2. We are using transcripts from the YouTubeTransciptsApi, which helps to performs faster and efficient aggregation and summarization.
3. For summarization, we perform tokenization, lemmatization, remove unnecessary characters(punctuations), stop words and extract features before passing this into sentiment analysis model.
4. In sentiment analysis model we perform the analysis over the summarized data which gets classified into features. We will be using classification models like Random Forest, Naïve Bayes.
5. Sentiment analysis helps in identifying how consumers perceive a product, service, or brand in the video.

VIII. Hardware/Software Details

- Python Version-min 3.9.0
- Any Python Compiler

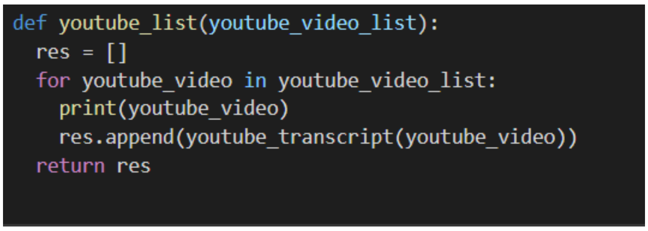
External Dependencies :

- transformers
- youtube_transcript_api-Built-in API for extracting Transcript
- flask-ngrok-To run the apps on localhost
- flask-cors

IX. Analysis

The content on YouTube generally consumes more than the actual time required to understand the information provided. They have irrelevant delivery of content like an introduction about the speaker, promotions, and advertisements which usually are out of context.

In the Project model built here, we make use of built-In API like the 'youtube_transcript_api' which helps to directly extract the auto-generated captions from the YouTube videos.

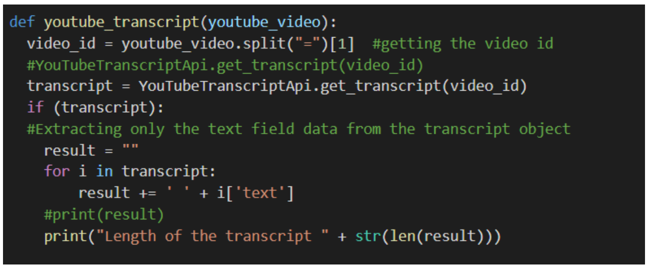


```
def youtube_list(youtube_video_list):
  res = []
  for youtube_video in youtube_video_list:
    print(youtube_video)
    res.append(youtube_transcript(youtube_video))
  return res
```

Tokenization:

To get the video ID we split the URL based on '=' and retrieve the Video ID .For Instance

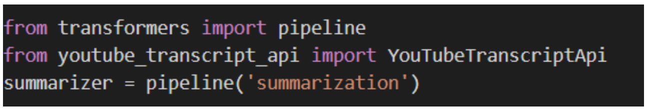
https://www.youtube.com/watch?v=7ThK5UT3BDw



```
def youtube_transcript(youtube_video):
  video_id = youtube_video.split("=")[1]  #getting the video id
  #YouTubeTranscriptApi.get_transcript(video_id)
  transcript = YouTubeTranscriptApi.get_transcript(video_id)
  if (transcript):
  #Extracting only the text field data from the transcript object
    result = ""
    for i in transcript:
        result += ' ' + i['text']
    #print(result)
    print("Length of the transcript " + str(len(result)))
```

Pre-processing

Any Machine learning project is inefficient without pre-processing or data cleaning.



```
from transformers import pipeline
from youtube_transcript_api import YouTubeTranscriptApi
summarizer = pipeline('summarization')
```

X. Implementation

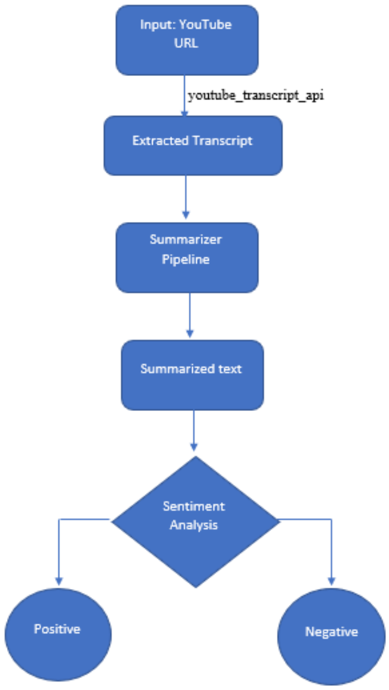


Fig2: Implementation flow chart

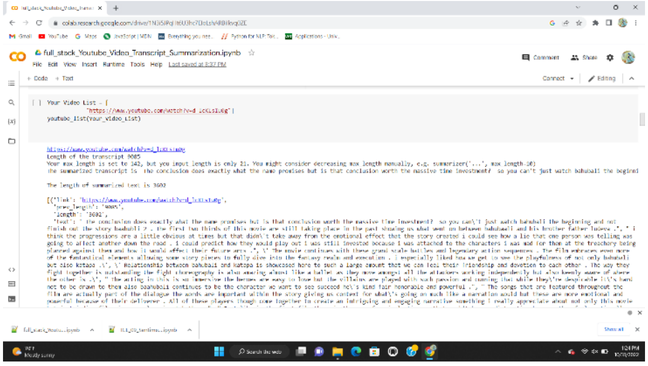
Input is a list of review videos on YouTube which are stored in the array Your_Video_List[].

Transformers module help in transforming data along the pipeline from origin to destination into clean and reduced feature set. We make use of the readymade API youtube_transcript_api to extract the transcripts of the mentioned YouTube videos.

Res[] array stores the transcript, and the YouTube-list() method returns the detailed transcript of the video. We then tokenize and get the length of the transcript. The model considers 1000 lines and summarizes the data by utilizing the Summarization pipeline which is a built-in module used for summarizing texts.

XI. Preliminary results

In increment 1, we have completed the text summarization segment. The primary task is to give a YouTube URL as input and the information contained in it is summarized in a text format using transcripts. Here are some screenshots of our results.

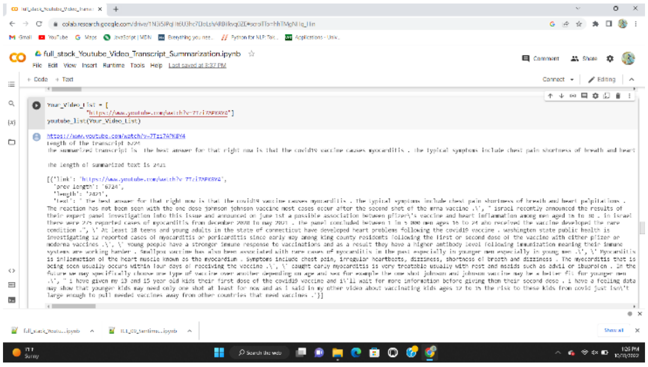


```
Your Video List = [
            "https://www.youtube.com/watch?v=d_1cKLsIu8g"]
youtube_list(Your_Video_List)
```

```
https://www.youtube.com/watch?v=d_1cKLsIu8g
Length of the transcript 9085
Your max length is set to 142, but you input length is only 21. You might consider decreasing max length manually, e.g. summarizer('...', max length=10)
the summarized transcript is  the conclusion does exactly what the name promises but is that conclusion worth the massive time investment?  so you can't just watch bahubali the beginni

The length of summarized text is 3602

[{'link': 'https://www.youtube.com/watch?v=d_1cKLsIu8g',
  'prev_length': '9085',
  'length': '3602',
  'text': " the conclusion does exactly what the name promises but is that conclusion worth the massive time investment?  so you can\'t just watch bahubali the beginning and not finish out the story baahubli 2 . the first two thirds of this movie are still taking place in the past showing us what went on between bahubaali and his brother father ludeva .", " i think the progressions are a little obvious at times but that didn\'t take away from the emotional effect that the story created i could see how a lie that one person was telling was going to affect another down the road . i could predict how they would play out i was still invested because i was attached to the characters i was mad for them at the treachery being planned against them and how it would effect their future arcs .", \' The movie continues with these grand scale battles and legendary action sequences . The film embraces even more of the fantastical elements allowing some story pieces to fully dive into the fantasy realm and execution . i especially liked how we get to see the playfulness of not only bahubali but also katapa .\', \' Relationship between bahubali and katapa is showcased here to such a large amount that we can feel their friendship and devotion to each other . The way they fight together is outstanding the fight choreography is also amazing almost like a ballet as they move amongst all the attackers working independently but also keenly aware of where the other is .\', " The acting in this is so immersive the heroes are easy to love but the villains are played with such passion and cunning that while they\'re despicable it\'s hard not to be drawn to them also baahubali continues to be the character we want to see succeed he\'s kind fair honorable and powerful .", " The songs that are featured throughout the film are actually part of the dialogue the words are important within the story giving us context for what\'s going on much like a narration would but these are more emotional and powerful because of their deliverer . All of these players though come together to create an intriguing and engaging narrative something i really appreciate about not only this movie
```



full_stack_Youtube_Video_Transcript_Summarization.ipynb
File Edit View Insert Runtime Tools Help Last saved at 8:37 PM
Comment
Share
+ Code
+ Text
Connect
Editing

```
Your_Video_List = [
            "https://www.youtube.com/watch?v=7Tzi7APK8Y4"]
youtube_list(Your_Video_List)
```

```
https://www.youtube.com/watch?v=7Tzi7APK8Y4
length of the transcript 6724
the summarized transcript is  the best answer for that right now is that the covid19 vaccine causes myocarditis . the typical symptoms include chest pain shortness of breath and heart

the length of summarized text is 2421

[{'link': 'https://www.youtube.com/watch?v=7Tzi7APK8Y4',
  'prev length': '6724',
  'length': '2421',
  'text': " The best answer for that right now is that the covid19 vaccine causes myocarditis . The typical symptoms include chest pain shortness of breath and heart palpitations .
The reaction has not been seen with the one dose johnson johnson vaccine most cases occur after the second shot of the mrna vaccine .\', " israel recently announced the results of
their expert panel investigation into this issue and announced on june 1st a possible association between pfizer\'s vaccine and heart inflammation among men aged 16 to 30 . in israel
there were 275 reported cases of myocarditis from december 2020 to may 2021 . the panel concluded between 1 in 5 000 men ages 16 to 24 who received the vaccine developed the rare
condition .", \' At least 18 teens and young adults in the state of connecticut have developed heart problems following the covid19 vaccine . washington state public health is
investigating 12 reported cases of myocarditis or pericarditis since early may among king county residents following the first or second dose of the vaccine with either pfizer or
moderna vaccines .\', \' Young people have a stronger immune response to vaccinations and as a result they have a higher antibody level following immunization meaning their immune
systems are working harder . Smallpox vaccine has also been associated with rare cases of myocarditis in the past especially in younger men especially in young men .\', \' Myocarditis
is inflammation of the heart muscle known as the myocardium . Symptoms include chest pain, irregular heartbeats, dizziness, shortness of breath and dizziness . The myocarditis that is
being seen usually occurs within four days of receiving the vaccine .\', \' caught early myocarditis is very treatable usually with rest and nsaids such as advil or ibuprofen . In the
future we may specifically choose one type of vaccine over another depending on age and sex for example the one shot johnson and johnson vaccine may be a better fit for younger men
.\', " i have given my 13 and 15 year old kids their first dose of the covid19 vaccine and i\'ll wait for more information before giving them their second dose . i have a feeling data
may show that younger kids may need only one shot at least for now and as I said in my other video about vaccinating kids ages 12 to 15 the risk to these kids from covid just isn\'t
large enough to pull needed vaccines away from other countries that need vaccines ."}]
```

Now that we have our summarized video data ready, we are ready to perform the sentimental analysis of the obtained results which is the second task of our project. The sentimental analysis will give a complete understanding of whether the video is positive or negative which is the final goal of the project.

XII. Project Management

Name	**Worked on**	**Percentage Contributed**
Kavya Balasubramaniam	Implementation of text summarization	30%
Niveditha Sree Pendli	Analyzing results	20%
Subbarao Sanka	Data pre-processing	30%
Anirudh Chowdary Ravipati	Obtaining datasets	20%

WORK COMPLETE

Name	**Work**	**Percentage of Contribution**
Kavya Balasubramaniam	Implementation of sentiment Analysis	30%
Niveditha Sree Pendli	Analyzing results	20%
Subbarao Sanka	Validation	30%
Anirudh Chowdary Ravipati	Compare different classification algorithms for sentiment analysis	20%

WORK TO BE COMPLETED

Issues and Concerns:

- Measuring the accuracy of the text summarization
- Topic Identification and interpretation
- Sarcastic video content
- Multilingual Data

Challenges of Sentiment Analysis

- When words have more than one meaning (Polysemy)
- Giving polarity scores to words
- Negation Detection

XIII. References

[1] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. “Thumbs Up? Sentiment Classification Using Machine Learning Techniques”. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). Association for Computational Linguistics, 2002, pp. 79–86. doi: 10.3115/1118693.1118704.

[2] Michael Gamon. “Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis”. In: Proceedings of the 20th International Conference on Computational Linguistics. COLING ’04. Association for Computational Linguistics, Jan. 2004. doi: 10.3115/1220355.1220476.

[3] Martin Hyberg and Teodor Isaacs. “Predicting like-ratio on YouTube videos using sentiment analysis on comments”. MA thesis. KTH Royal Institute of Technology, 2018.

[4] https://cs224d.stanford.edu/reports/urvashik.pdf