YouTube Movie Review

Text Extraction, Summarization and Sentiment Analysis

CSCE 5290

Natural Language Processing

Project 12

Sanjib Paudel

Premsai Ganesh Charugundla

Rasagna Vadde

Sarath Durisala

Github Link: https://github.com/krison44/CSCE-5290-project

Motivation:

With the accessibility of huge data, people are being fed too much information these days and it makes it hard to make a right decision. Being a movie lover, I like to watch movies whenever I have free time, but I end up getting disappointed most of the time. I even spent hours trying to figure out a good movie sometimes. Review from YouTube helps, but that also costs me a lot of time. And I am not a guy who likes to listen to every word that critics say nonstop for around ten minutes. So sometimes I check IMDB rating and decide whether to watch it or not. However, 30-40 percent of the time, Critics analysis doesn't really reflect my tastes. So, I am one soul with a lot of disappointment to select the right movie for me. So, in order to solve this problem, our team has decided to extract the information from a YouTube video, summarize the text in an abridged version and run the sentiment analysis based on the summarized text. People won't be bored reading sweet and short texts and analyzing how good or bad the movie is based on percentage.

Objective:

People are very impatient these days and they want to decide in no time. We have an abundance of data, and that data is of no use if not used properly. So, we wanted to utilize the data and feature engineers, so that we could use it in the most effective way. For this purpose, we decided to get the summary and rating of the movie based on reviews from different critics on YouTube. Currently, our objective for this project is to get the result from one video, but in future, it can be extended by traversing multiple videos and predicting the result based on all of them at once. This way, the result will be more accurate per movie, and it will save even more time to analyze the movie. The main focus is to use nouns, verbs and adjectives as descriptive terms to determine the polarity of the review. Reviews fall into two different categories, positive and negative. (Singh Brar & Sharma, 2018).

Significance:

Nowadays, people are more interested in looking and feeling rather than reading texts or watching boring videos to gather information. With the minimum words and graphs, people try to ingest more information. This is where this project plays a vital role. We are trying to give people enough information about movies in just a small paragraph with a visual graph and a movie rating. On top of that, we don't have such tools available on the market to predict movies by machine learning yet. We only have IMDB and rotten-tomatoes ratings, which are made by movie critics, but if we can take the review from normal people from YouTube, we can get the prediction based on non-movie background personalities as well and this will cover the interest of lots of people.

Features:

This project contains the following:

1) Text Extraction

In order to extract the text, we will be using the python library called *youtube-trascript-api*. For this we will need an input box where video id will be provided and the response will give us the array of objects with text, start and duration. We will iterate and concatenate all text in order to perform text extraction.

2) Text Summarization

For text summarization, we will be performing normalization followed by tokenization, word vector, similarity matrix, scoring and finally, summary of raw text.

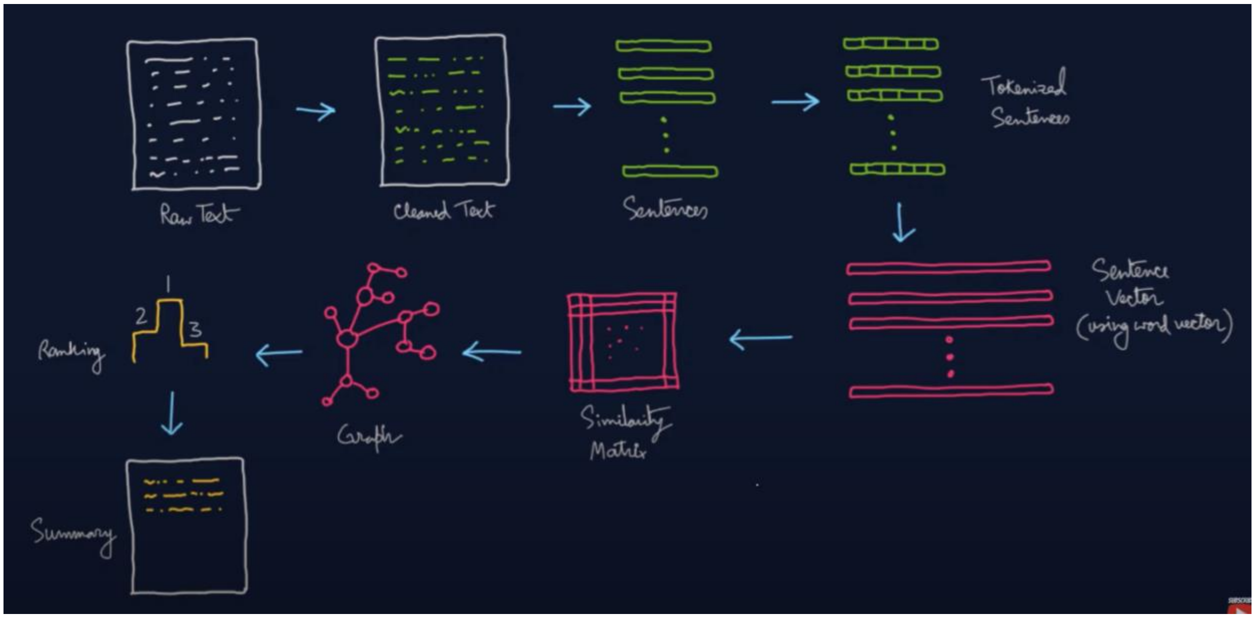

Figure 1: *Text Summarization and Keyword Extraction*, https://www.youtube.com/watch?v=XO97Uon83Os. Accessed Sept 27, 2022

3) Training

For training, we will be using IMDB Dataset. It has 50k movie reviews with positive and negative values, and it is available on Kaggle (https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews). We are going to split data into 9:1 and train the modal with a neural network and validate the result with 10 percent data.

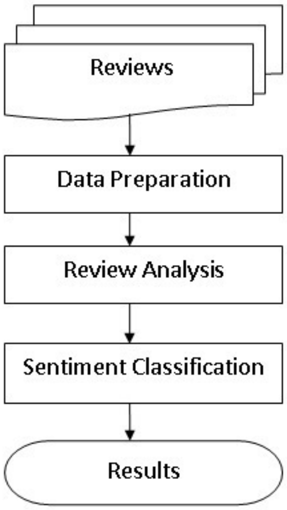

Figure 2: Analysis of different approaches to Sentence-Level Sentiment Classification, *https://www.researchgate.net/publication/286272892_Analysis_of_different_approaches_to_Sentence-Level_Sentiment_Classification*. Accessed September 29

4) Testing

   We will test the result with summarized text and give the output with positive or negative feedback with percentage. For example: *The movie is 90% good.*

REFERENCES

Singh Brar , G., & Sharma, A. (2018). Sentiment Analysis of Movie Review Using Supervised Machine Learning Techniques. Bathinda, Punjab, India. Retrieved September 2022, from https://www.ripublication.com/ijaer18/ijaerv13n16_53.pdf