FINAL REPORT

Topic-YouTube Sentiment Analysis

INT-404 AI

By-

|  |  |  |
| --- | --- | --- |
| Sumeet Bansal | 11802638 | K18KK - B-59 |

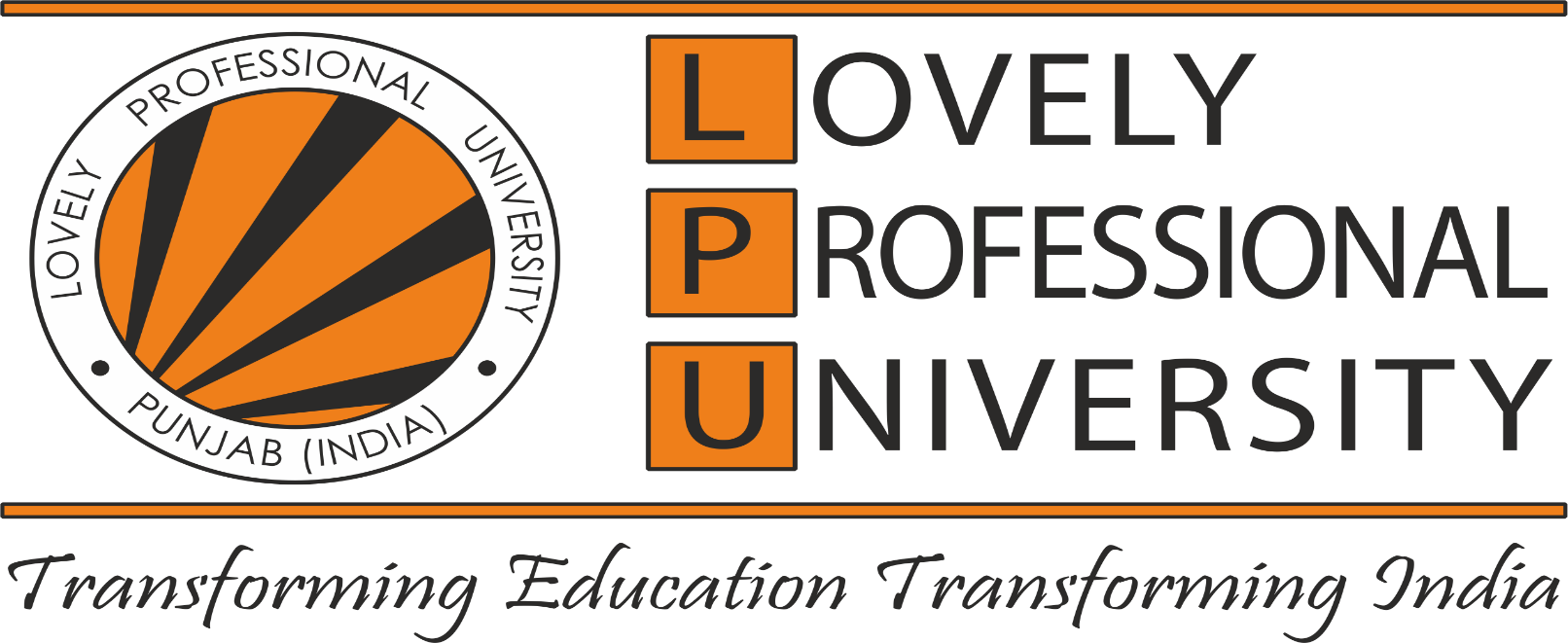
Program:Btech Cse(Hons.)

School:School of Computer Science and Engineering.

**Faculty: Mr. Sagar Pande**

Lovely Professional University

Jalandhar, Punjab, India.



**CONTENT:-**

1. Abstract
2. Introduction
3. Related Work
4. Implementation
5. Result (screenshot)
6. Important Libraries Used
7. Team Responsibilities
8. References

**Abstract**

In this project we use a supervised ML based approach to detect the sentiment of user comments in YouTube. Polarity detection in social media content is challenging not only because of the existing limitations in current sentiment dictionaries but also due to the informal linguistic styles used by users. To analyse sentiment the project programs have been created to perform various tasks. First twitter data set was fetched. Then the data was cleaned. After cleaning a ML model was trained on the data set.The final program takes YouTube video id as the only parameter, the program then fetches all comments based on popularity and public comments then are stored. The comments then are ran through the model which predicts if comments are positive or negative and keeps count. After all comments are analysed a pie chart is plotted depicting positive and negative comments. The model uses Multinomial Naïve Byes Classifier.

**Introduction**

The rapidly-increasing popularity of social media sites such as Facebook , Twitter and YouTube is primarily due to the ease of use and simplicity of these systems for the creation, collaboration and sharing of resources (images, video) even from non-technical users. For video sharing, YouTube is the most popular site on the Web. According to a recent study, YouTube accounts for 20% of Web traffic and 10% of total Internet traffic. YouTube provides many social mechanisms to judge user opinion and views about a video by means of voting, rating, favourites, sharing and negative comments, etc. This context information is useful in studying user and community behaviour and perspective. Analysis of user comments provides a useful data source for many applications such as comment filtering, personal recommendation, and user profiling to name a few. This project analyses the comments to give user a better understanding of the video and its contents before even watching the video. This helps user save time and also allows the user to watch only the helpful videos. Just like how a human would analyse a few comments before watching a video this program does it for the user.

**Related Work**

Sentiment analysis is one of the basic task done in Natural Language Processing**.** Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

There are various projects and works which have already been done on this topic.

There are various twitter sentiment analysis works that have been done. Which fetch data using twitter API on a specific topic, and analyse the sentiment of general public towards that topic. This gives us valuable insights on the topic. Some of the twitter sentiment analysis projects are used in below mention intension: -

* **Business:**In marketing field companies use it to develop their strategies, to understand customers’ feelings towards products or brand, how people respond to their campaigns or product launches and why consumers don’t buy some  
  products.
* **Politics:**In political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well!
* **Public Actions:**Sentiment analysis also is used to monitor and analyse social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

Discussing about YouTube sentiment analysis several research has been undertaken on different aspects of YouTube video features. Among them comments are one of the important one to make a decision (comment rating, topic categories etc.,) about the particular video . These comments are also used to annotate the video object . Comments also reflect the user’s behaviour and could use to find the troll users . Moreover, by analysing the sentiment of comments it is possible to find the users positivity or negativity about video. Based on comments, researchers categorize the videos in several category . Furthermore, for improving video retrieval process. Altingovde et al. Proposed a method based on basic feature and social feature. Lehner et al. Work on YouTube video comments, like, dislikes for showing that user’s perception (like/dislike) are influenced by valuable comments. These two methods worked to find the popularity of video using various features so that it could help to retrieve the useful video. Although these two proposed approaches showed impressive work for video retrieval process but they used like/dislike and views. Sometimes which may lead to inaccurate result. On the contrary, only a large amount of comments are analysed instead of others features (like/views etc.) for finding relevant video which might be useful for YouTube users.

Several research has been undertaken on different aspects

of YouTube video features [5]. Among them comments are

one of the important one to make a decision (comment rating,

topic categories etc.,) about the particular video [6]. These

comments are also used to annotate the video object [7, 9].

Comments also reflect the user’s behavior and could use to

find the troll users [1]. Moreover, by analyzing the sentiment

of comments it is possible to find the users positivity or

negativity about video [10]. Based on comments, researchers

categorize the videos in several category [11, 12].

Furthermore, for improving video retrieval process

Altingovde et al. proposed a method based on basic feature

and social feature [3]. Lehner et al. Work on YouTube video

comments, like, dislikes for showing that user’s perception

(like/dislike) are influenced by valuable comments [4]. These

two methods worked to find the popularity of video using

various features so that it could help to retrieve the useful

video. Although these two proposed approaches [3, 4]

showed impressive work for video retrieval process but they

used like/dislike and views. Sometimes which may lead to

inaccurate result. On the contrary, we only analyze a large

amount of comments instead of others features (like/views

etc.) for finding relevant video which might be useful for

YouTube users.

**Implementation**

The project was implemented in various steps. Each step was very crucial for the project.

At first the task of finding a proper dataset for the model training was tackled. A pre tagged data set was needed and also the data set to be used must be big too. This problem was solved after the Twitter data set with 1.6 million tweets with 800k positive and 800k negative tweets was found. As sentiment was already analysed on these tweets this was the perfect data set for the project.

Second step was to clean the data set and extract only the columns that were helpful and all other columns were removed. Then the rest of the data was cleaned using the steps of text cleaning for sentiment analysis like tokenization was done, stop wards removed, stemming was also performed and URL and links were removed.

The cleaned data was the used to train the model. The algorithm used was Multinomial Naïve Bayes Classifier. This algorithm is one of the most widely used for sentiment analysis purpose. Spam filtering also used this algorithm. Due to computational limitation only a part of data set was used for training and testing. After training the model and count vectoriser were pickled for further use.

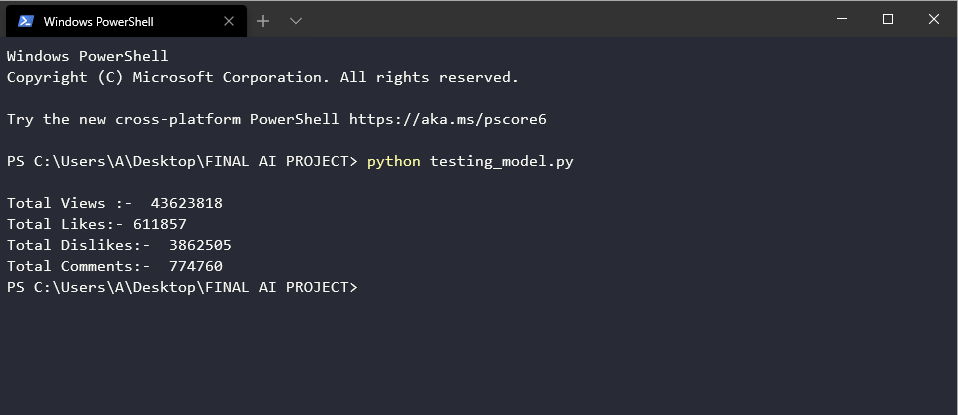
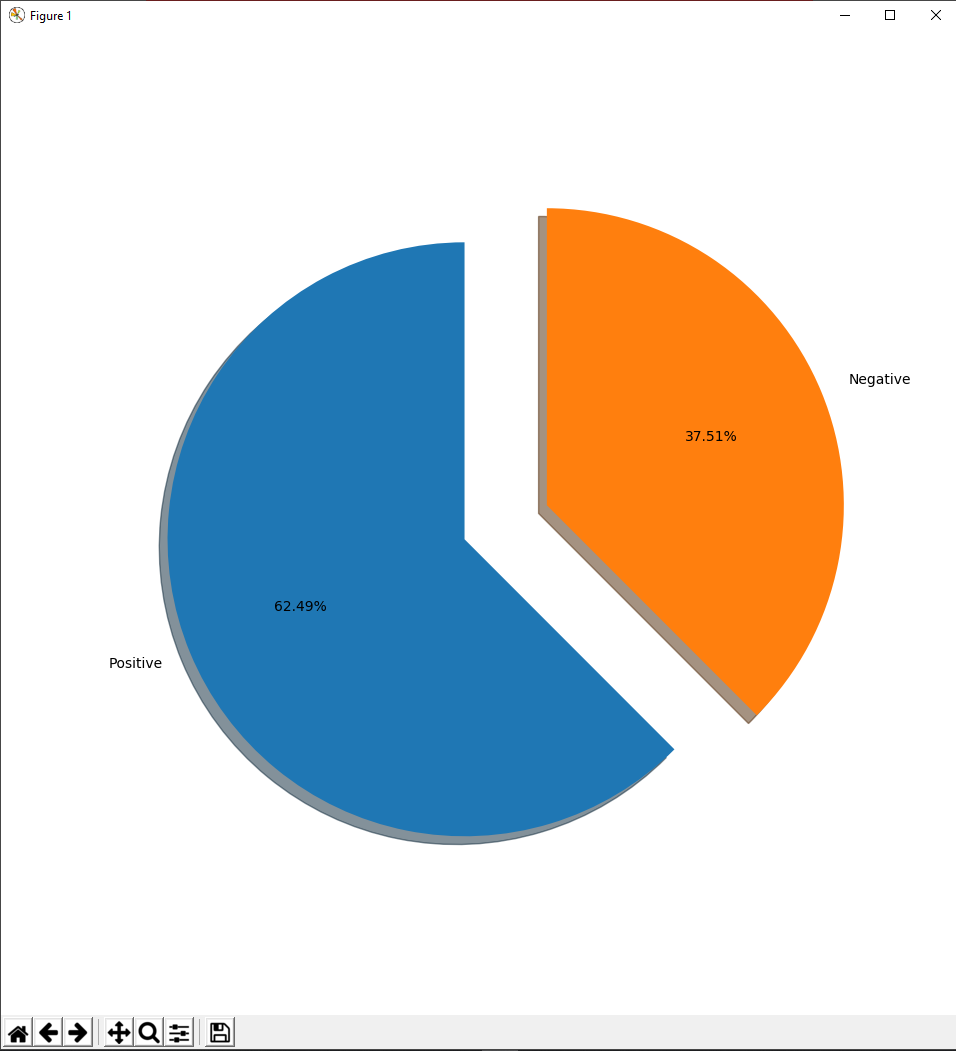
In the last stage using the YouTube AVI V3 and given video id comments were fetched and the model was used to analyse the sentiment. As a result a pie chart depicting percentage of positive and negative comments is displayed along with the stats of the video like views, like count, dislike count, total comments etc.

**Result**

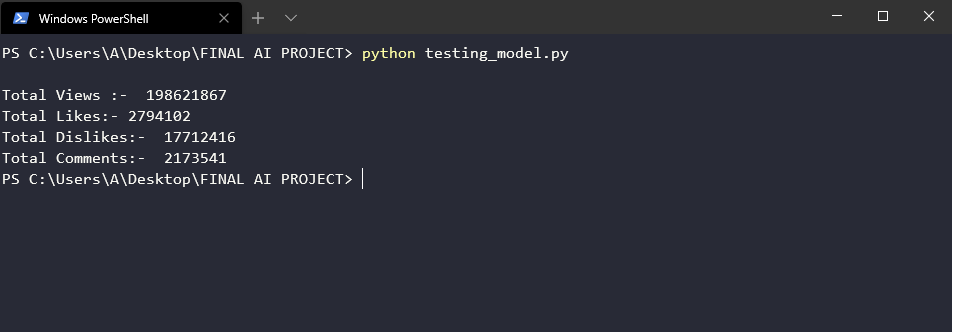
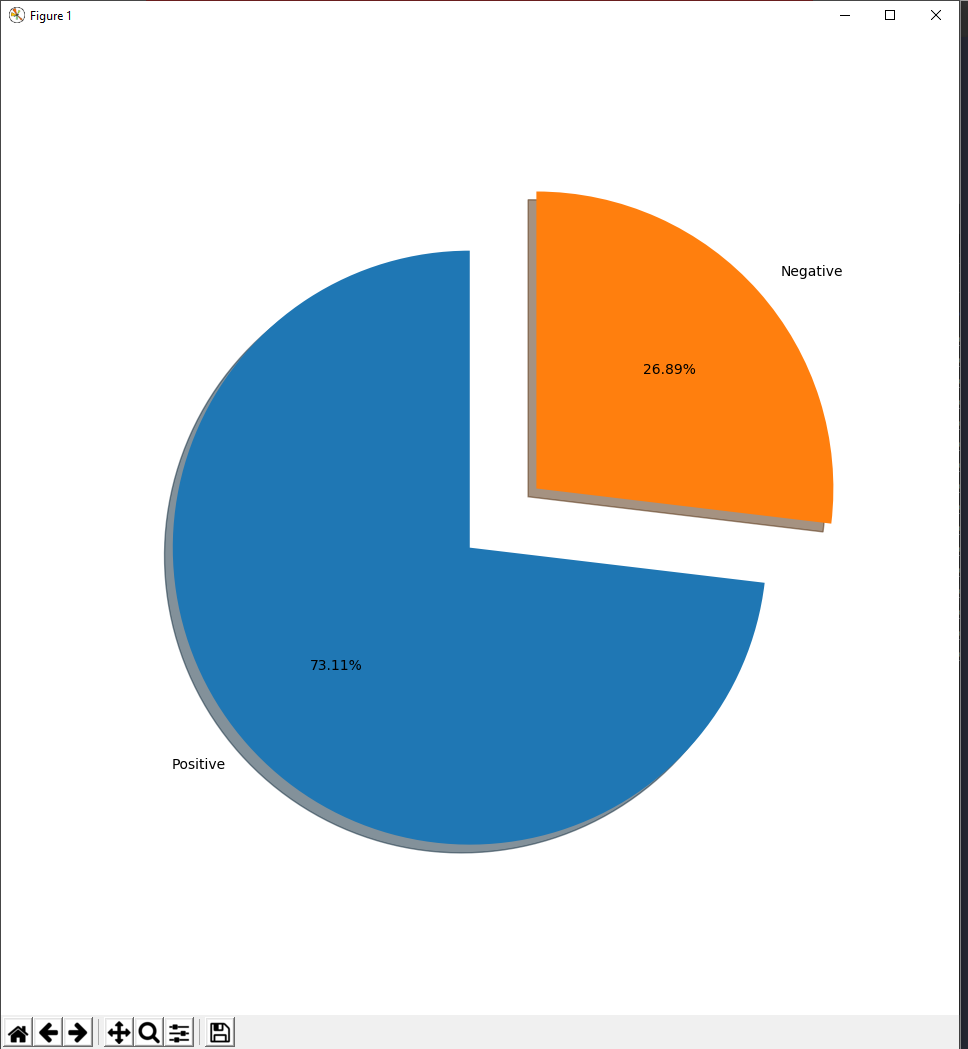
Here are some analysed YouTube videos:-

Pie Chart & Information

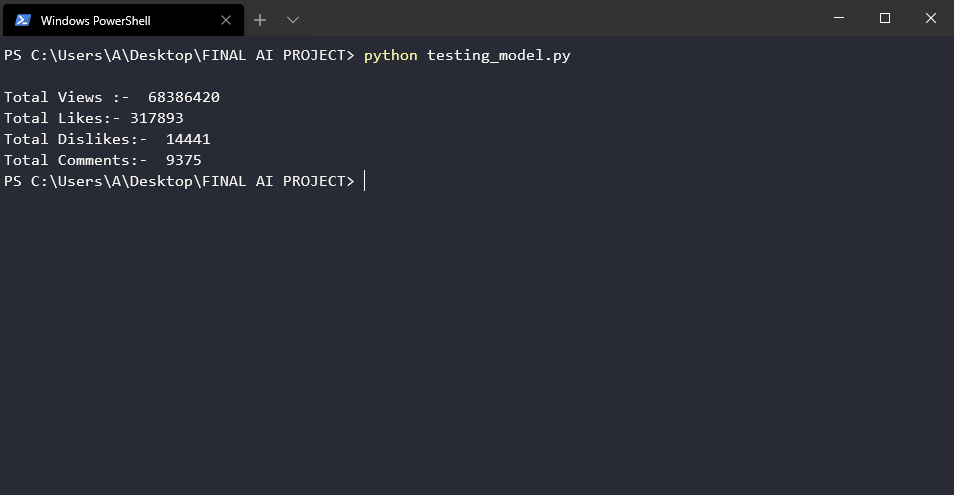
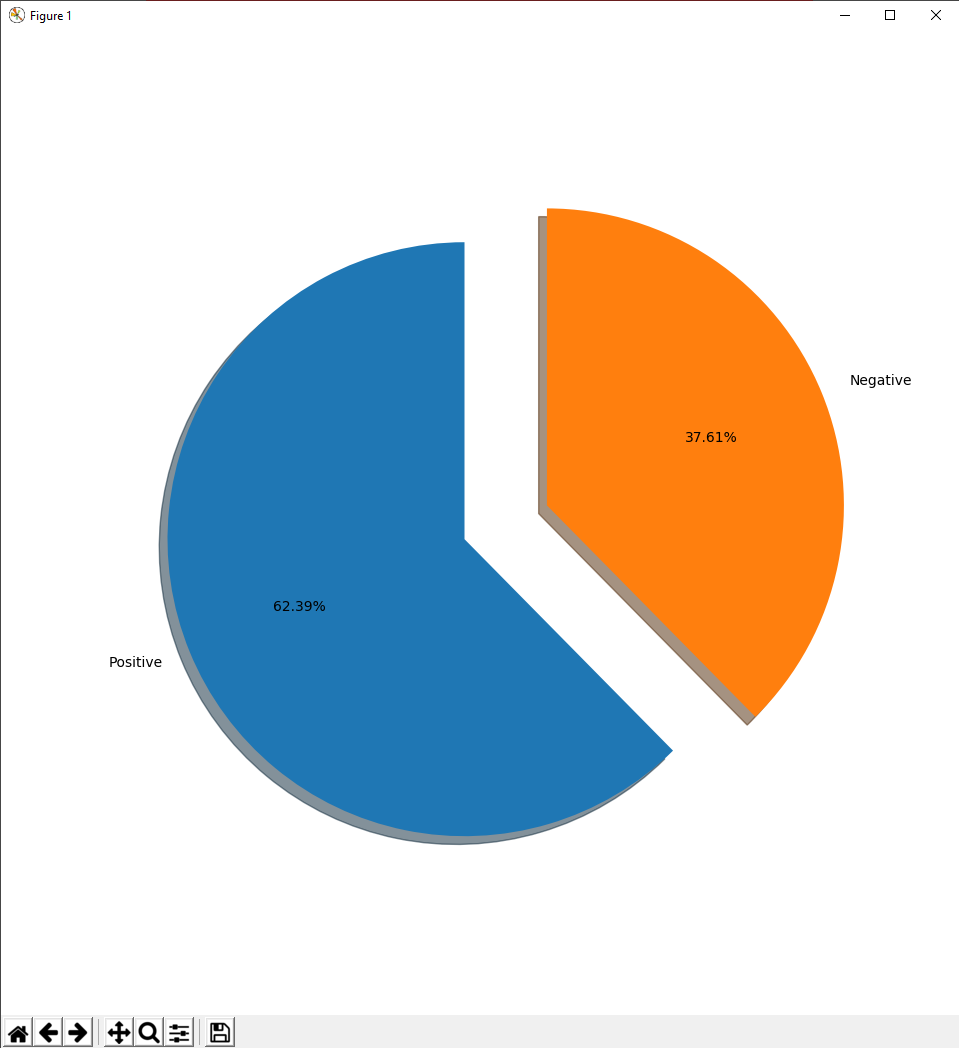
1. Video Id=" EeF3UTkCoxY "



1. Video Id ="YbJOTdZBX1g"



1. Video Id ="1VrWaED18\_g "



**Important Libraries Used**

1. Pandas: - It provides extremely streamlined forms of data representation. This helps to analyse and understand data better. Pandas saves a lot of time by importing large amounts of data very fast. We can use Pandas to perform various tasks like filtering data according to certain conditions, or segmenting and segregating the data according to preference, etc.
2. NLTK- (Natural Language Tool Kit)- NLTK is an essential library supports tasks such as classification, stemming, tagging, parsing, semantic reasoning, and tokenization in Python. It’s basically is the main tool for natural language processing and machine learning.
3. String - cleaning and structuring data
4. Re (Regular Expression)- A RegEx, or Regular Expression, is a sequence of characters that forms a search pattern, hence this was very essential in cleaning the data.
5. Textblob- TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. In this project it was used for cross validation and sentiment analysis in early stages.
6. Matplotlib- It was used to represent the output of the model in a pie chart so it is easier for the user to read the analysis.
7. Sklearn- It was the most crucial library used as the model training, testing, count vecterization and many more steps were achieved.
8. CSV,JSON,PICKEL- There are the file handling libraries used to read and write data and model into and from files.

**Team Responsibilities**

This was an individual project hence all the work that is being done is being an individual.

The tasks include:-

* Collecting data
* Cleaning data
* Training and tuning the model
* Using the YouTube API to collect data

Analysing and verifying data

**References**

Learning the YouTube API

* <https://python.gotrained.com/youtube-api-extracting-comments/>
* Google documentation for YouTube API V3
* <https://backlinko.com/youtube-ranking-factors>
* <https://github.com/youtube/api-samples/blob/master/python/comment_threads.py>

Learning about sentiment analysis

* https://www.learndatasci.com/tutorials/predicting-reddit-news-sentiment-naive-bayes-text-classifiers/
* https://towardsdatascience.com/sentiment-analysis-of-tweets-using-multinomial-naive-bayes-1009ed24276b
* https://streamhacker.com/2010/05/10/text-classification-sentiment-analysis-naive-bayes-classifier/
* <https://textblob.readthedocs.io/en/dev/classifiers.html#classifying-text>
* <https://pythonprogramming.net/naive-bayes-classifier-nltk-tutorial/>
* <https://github.com/JahedulKaderForman/YouTube-Sentiment-Analysis/blob/master/CommentSentiment/comment_extract.py>

Some Helpful YouTube Videos

* <https://www.youtube.com/watch?v=vz_xuxYS2PM>
* <https://www.youtube.com/watch?v=0kPRaYSgblM>
* <https://www.youtube.com/watch?v=Zt83JnjD8zg>
* <https://www.youtube.com/watch?v=l3dZ6ZNFjo0>