Analyze your Video

By
Prasanth Kumar J - 20186080
Venkatesh M - 20186066
Shiva Prasad T - 20186078

Project Submitted to
MSIT, IIIT HYDERABAD

Approved: A. Manasa
Assistant Mentor, MSIT
IIIT Hyderabad

Date: 03-06-2019

Statement of Confidentiality:

This document is submitted in the requirement for the degree of MSIT in IIIT Hyderabad. This is the product of our own labor except where indicated in the text. The report may be freely copied and distributed provided the source is acknowledged.

Acknowledgments:

We would like to specifically thank the following people:
Manasa(Assistant Mentor): For being a supportive guide for our project, that has provided constant support and reassurance throughout the entire project.

All the knowledgeable contributors:

At the stack overflow and the tutorial points where we found some of the free libraries that the project makes use of. The usage of these libraries and individual acknowledgments are included within in the docstring.

Table of Contents

1. Abstract:

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. Identifying and categorizing the opinion of text, especially determine whether the video is positive, negative or neutral. Sentiment analysis is also known as opinion mining or emotion AI.

2. Introduction:

In this project, the main theme is to retrieve the YouTube comments from the given URL and analyze using the TextBlob, Logistic regression and Naive Bias algorithms.

In general sentiment analysis also refers to opinion mining, which is a submachine learning task where we want to determine which is the general sentiment of a given document. Using machine learning techniques and natural language processing we can extract the subjective information of a document and try to classify it according to its polarity such as positive, neutral or negative.

The project was motivated by a thought of trending topic machine learning. As sentiment analysis is based on machine learning concepts it will be great to learn and present it practically using this project. We can use Machine Learning in Finance, Medicine, almost everywhere. That's why I decided to do my project around Machine Learning.

3. Project Prerequisites

1. Programming Language was chosen: Python
2. Automation Tool: Selenium

4. Goals of the project

The goals that are kept for a better project are as follows

- ➢ To automatically scrap the comments and store in CSV based on the given URL
- ➢ Analyze the video based on the comments stored in the CSV file.
- ➢ Train the algorithms based on the dataset which contain 25000 comments where first half are positive and rest are negative.
- ➢ Test the CSV file with comments for accuracy.
- ➢ Integrate the backend code with frontend using the flask.
- ➢ Display the result based on the result produced by TextBlob, Logistic Regression, and Naive Bayes.

5. Coding and implementation

In the project, the coding part is done in the high-level language python. As a start, the main theme is to scrap the comments from YouTube video given. To Scrap the comments we used an automation tool named selenium. As an input, the selenium webdriver will be waiting to get the link from the end-user. If the link is valid then using XPath the pointer is moved to the comments section. In the comments section based on the total number of authors who had commented for the respective video is counted. Now using loop which will iterate for the range of authors and retrieve the text for that author and store it in a list.

Now in the list, we had all the comments as a part of storing the retrieved comments to the CSV file we will iterate the list and using regex we have loaded the comments that are only in the English language.

The second part of the coding is to get the CSV file with the comments. The loaded CSV file is iterated and applied the stopwords function where it will remove the stopwords for the respective comments. Parallel converted to lowercase for better result.

The additional one is we have trained the Logistic and Naive Bayes algorithm based on the dataset with 25000 comments. In those 25000 comments

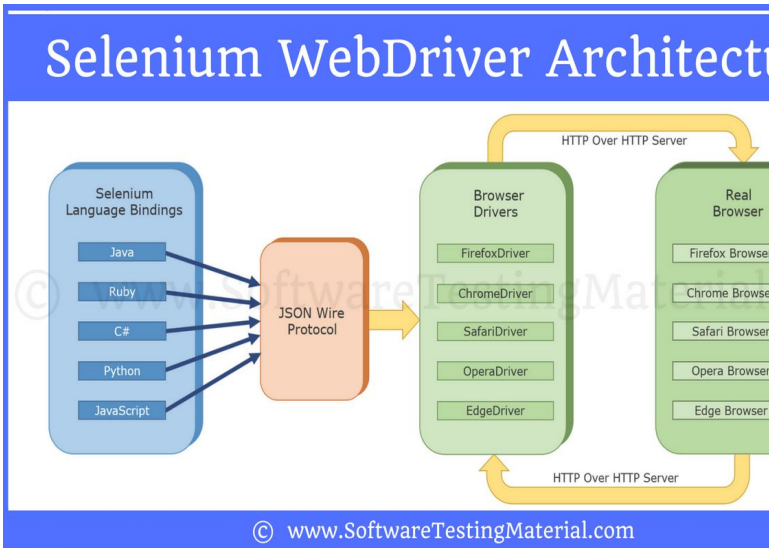
first half will be positive and the rest are negative comments. Once the training is done the CSV file with filtered comments is sent to models and analyzed.
The final part is to integrate the backend python with HTML. To integrate we have used the flask web framework. The result is displayed based on the models and TextBlob.

6. Functionalities:

The functionality of Selenium, TextBlob, Flask, Naive Bayes, Logistic Regression are as follows

6.a. Selenium Tool:

Selenium is an open source web automation tool. Selenium is popular and demanding it is an open source tool freely available on the internet, No project cost involved, No license required and the main important thing is it supports almost 13 different software languages like Java, Python and many more.
ChromeDriver is a separate executable that WebDriver uses to control Chrome. It is maintained by the Chromium team with help from WebDriver contributors.
It is a collection of open source APIs which are used to automate the testing of a web application. Description: Selenium WebDriver tool is used to automate web application testing to verify that it works as expected. It supports many browsers such as Firefox, Chrome, IE, and Safari.



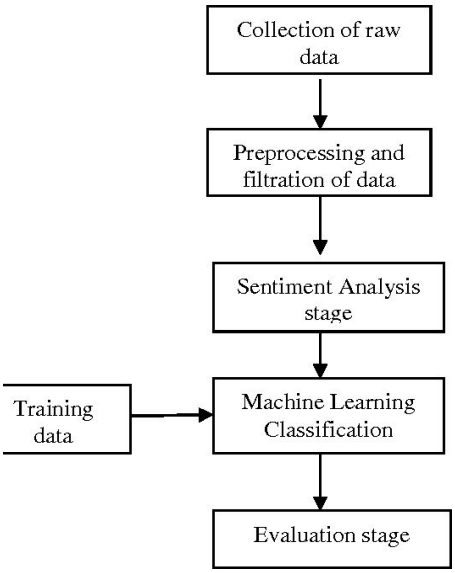
6.b. TextBlob:

TextBlob is a Python 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.

Some features of TextBlob are:

- ➢ Noun phrase extraction
- ➢ Part-of-speech tagging
- ➢ Sentiment analysis
- ➢ Classification (Naive Bayes, Decision Tree)
- ➢ Language translation and detection powered by Google Translate
- ➢ Tokenization (splitting text into words and sentences)
- ➢ Word and phrase frequencies
- ➢ Parsing
- ➢ n-grams
- ➢ Word inflection (pluralization and singularization) and lemmatization

- ➢ Spelling correction
- ➢ Add new models or languages through extensions
- ➢ WordNet integration



6.c. Flask Framework:

Flask is a microframework for Python based on Werkzeug, Jinja 2 and good intentions. It is a third-party Python library used for developing web applications.

It is classified as a microframework because it does not require particular tools or libraries.[3] It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools. Extensions are updated far more regularly than the core Flask program.

Some applications that use the Flask framework include Pinterest, LinkedIn, and the community web page for Flask itself.

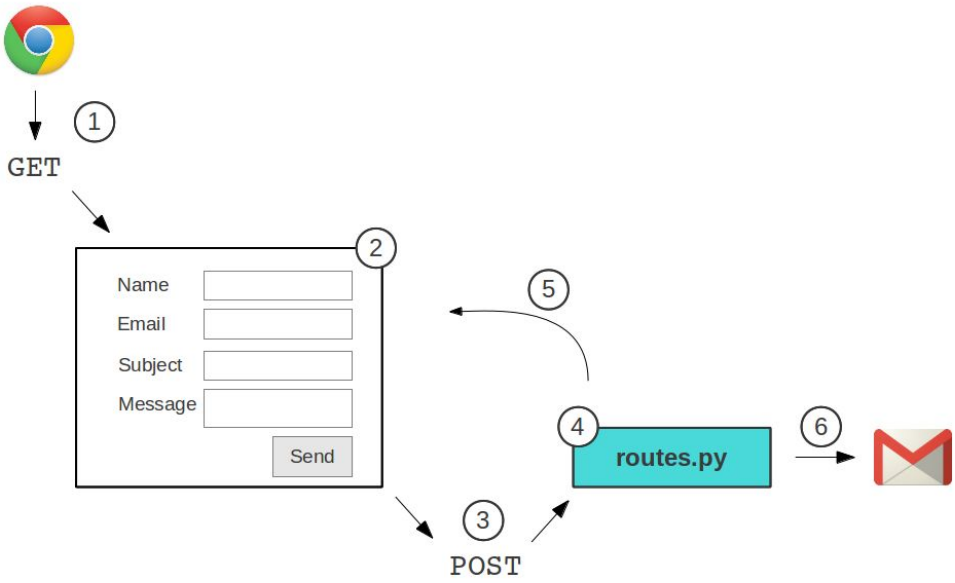


fig: Sample flow of flask with python file

6.d. Naive Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

The fundamental Naive Bayes assumption is that each feature makes an:

- independent
- equal

The assumptions made by Naive Bayes are not generally correct in real-world situations. In fact, the independence assumption is never correct but often works well in practice.

Other popular Naive Bayes classifiers are:

Multinomial Naive Bayes: Feature vectors represent the frequencies with which certain events have been generated by a multinomial distribution. This is the event model typically used for document classification.

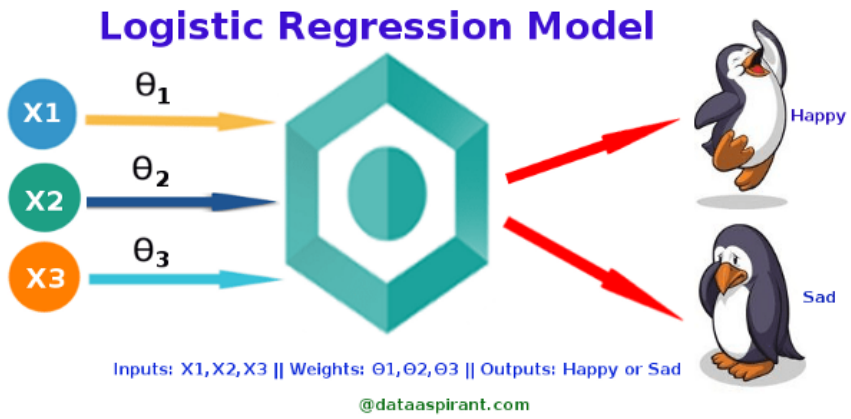
Bernoulli Naive Bayes: In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks, where binary term occurrence(i.e. a word occurs in a document or not) features are used rather than term frequencies(i.e. frequency of a word in the document).

6.e. Logistic Regression:

Logistic Regression was used in the biological sciences in the early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

For example,
To predict whether an email is spam (1) or (0)
Whether the tumor is malignant (1) or not (0)
Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequences in real time.

From this example, it can be inferred that linear regression is not suitable for the classification problem. Linear regression is unbounded, and this brings logistic regression into the picture. Their value strictly ranges from 0 to 1



Algorithms & Technologies

Algorithms used in the project are

- ➢ TextBlob
- ➢ Logistic Regression
- ➢ Naive Bayes.

Technologies are

- ➢ Python
- ➢ Data Analytics
- ➢ Web Technologies

Tools used are

- ➢ Sublime Text IDE
- ➢ Visual Studio Code IDE
- ➢ Jupyter Notebook

➢ Chrome Driver

Overview of Project:

The main theme of the project is to analyze the video based on the user input, scrap the comments and analyze the comments based TextBlob, Logistic Regression and Naive bias models.

The first process is to request the URL from the end_user where the program will test the URL. If the given URL is a valid one then the URL is sent to selenium driver where it will launch the URL to the browser.

As a second part, the selenium tool will start retrieving based on the author-texts.It will retrieve the comment the text under author text and neglects the replies. The data is filtered based on the isEnglish() method which will work based ASCII values. The filtered comments are stored in a list and then sent to the CSV file.

In the final part, the CSV file with retrieved comments is sent to TextBlob, Logistic Regression, and Navie bias models.

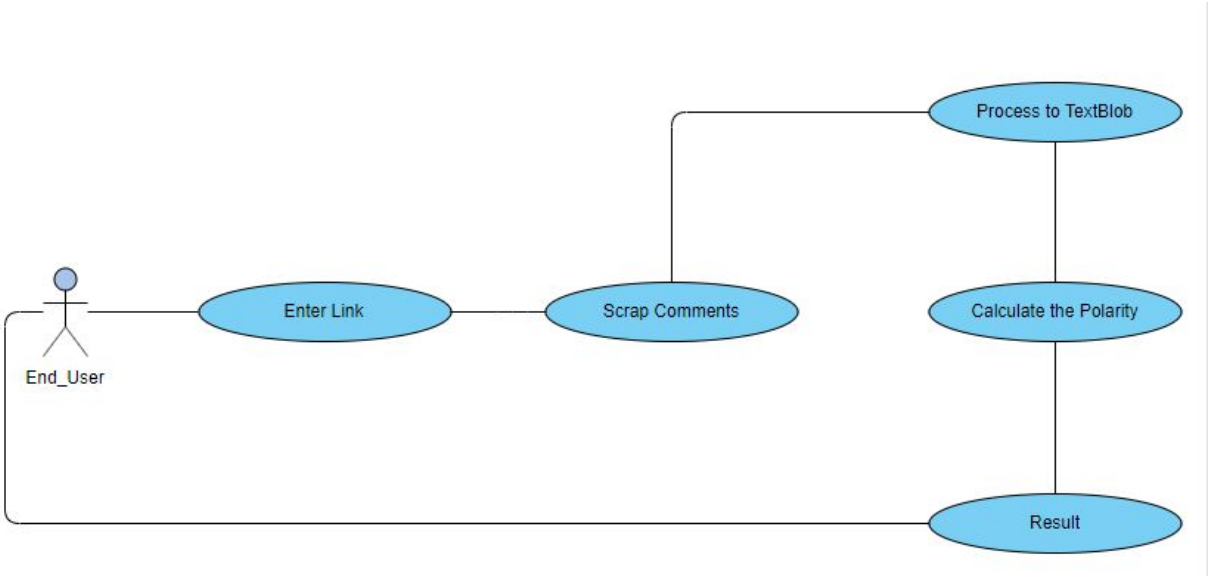
- **TextBlob** will calculate the polarity for each and every comment. If the final result of the polarity is greater than 0 then the video then it will return positive or else it will return negative.
- **Logistic Regression** and **Naive Bayes** algorithms are trained based on the training dataset of 25000 positive and negative comments respectively. Later the retrieved comments dataset is tested for accuracy.

Finally, the result is displayed based on the models and TextBlob result.

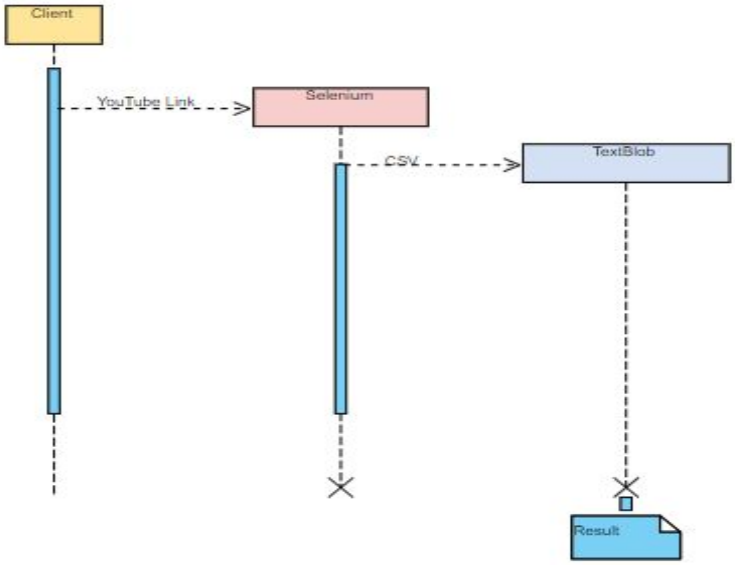
UML diagrams:

The unified modeling language allows the software engineer to express an analysis model using the modeling notation that is governed by a set of syntactic-semantic and pragmatic rules.

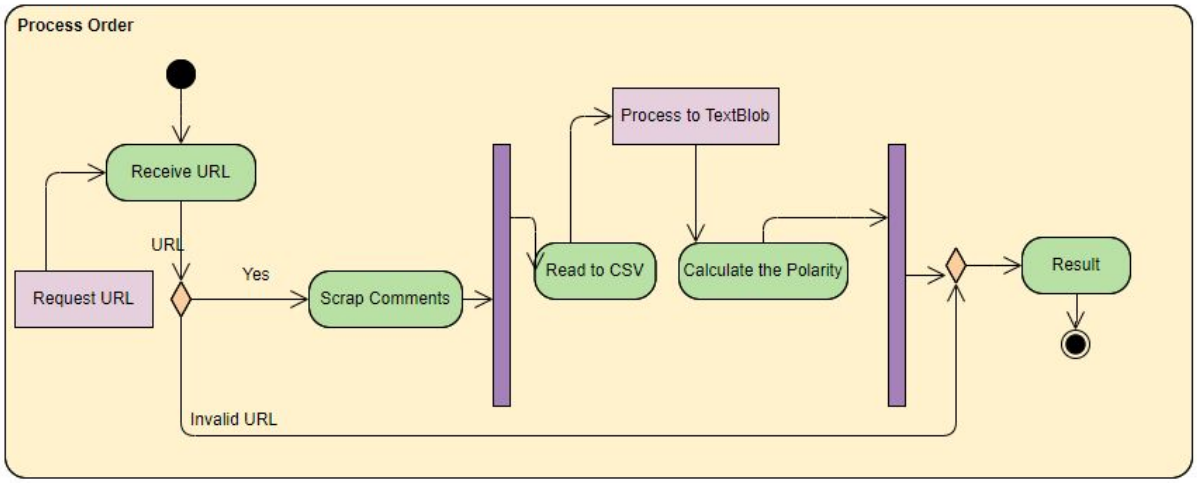
Use Case Diagram:



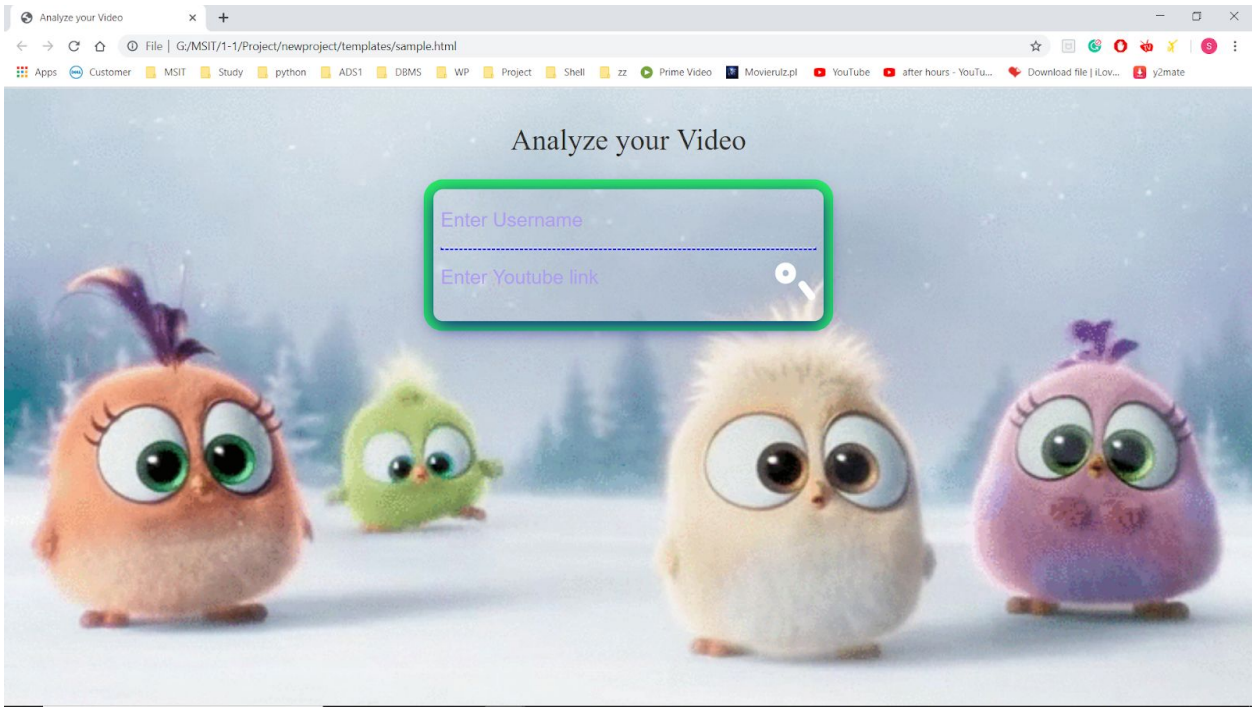
Sequence Diagram:



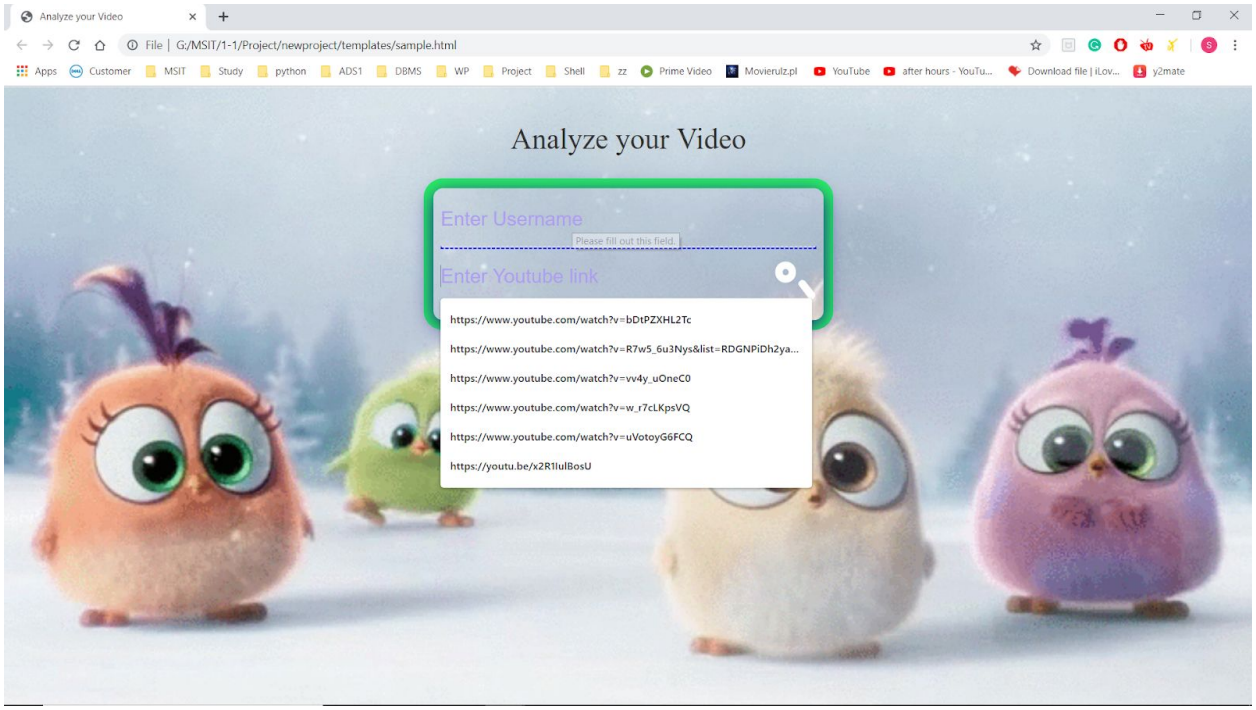
Activity Diagram:



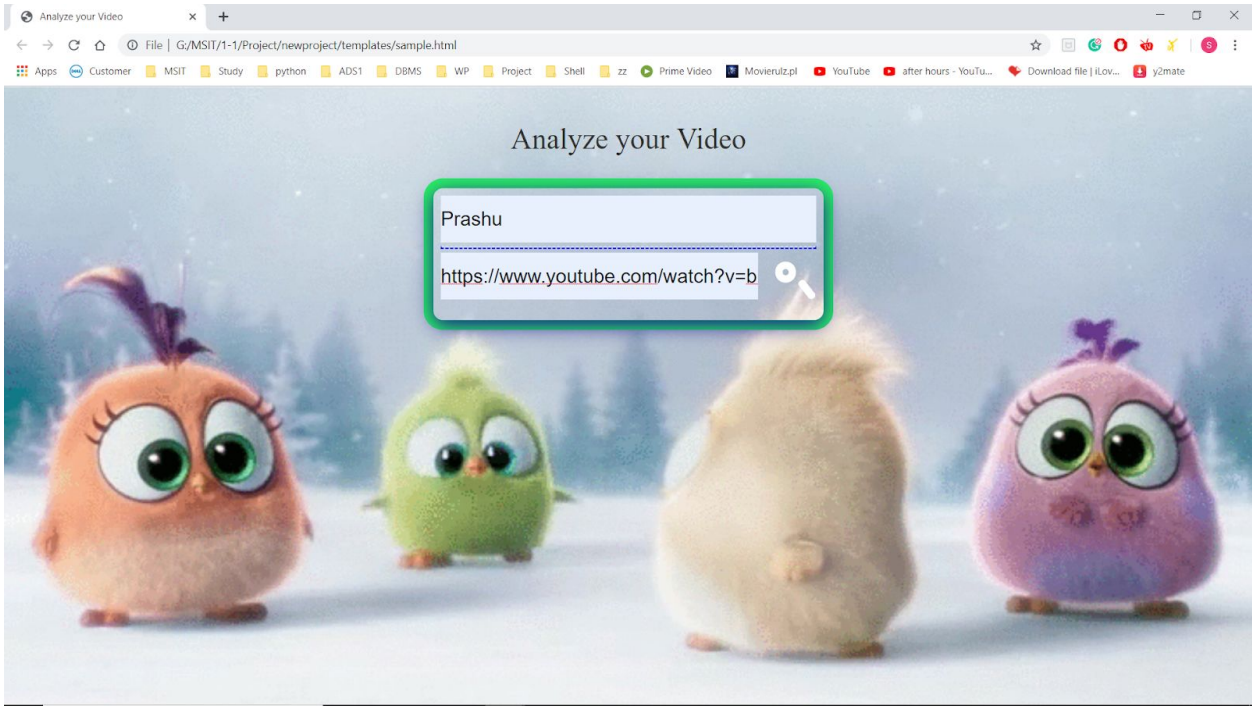
Project Screenshots



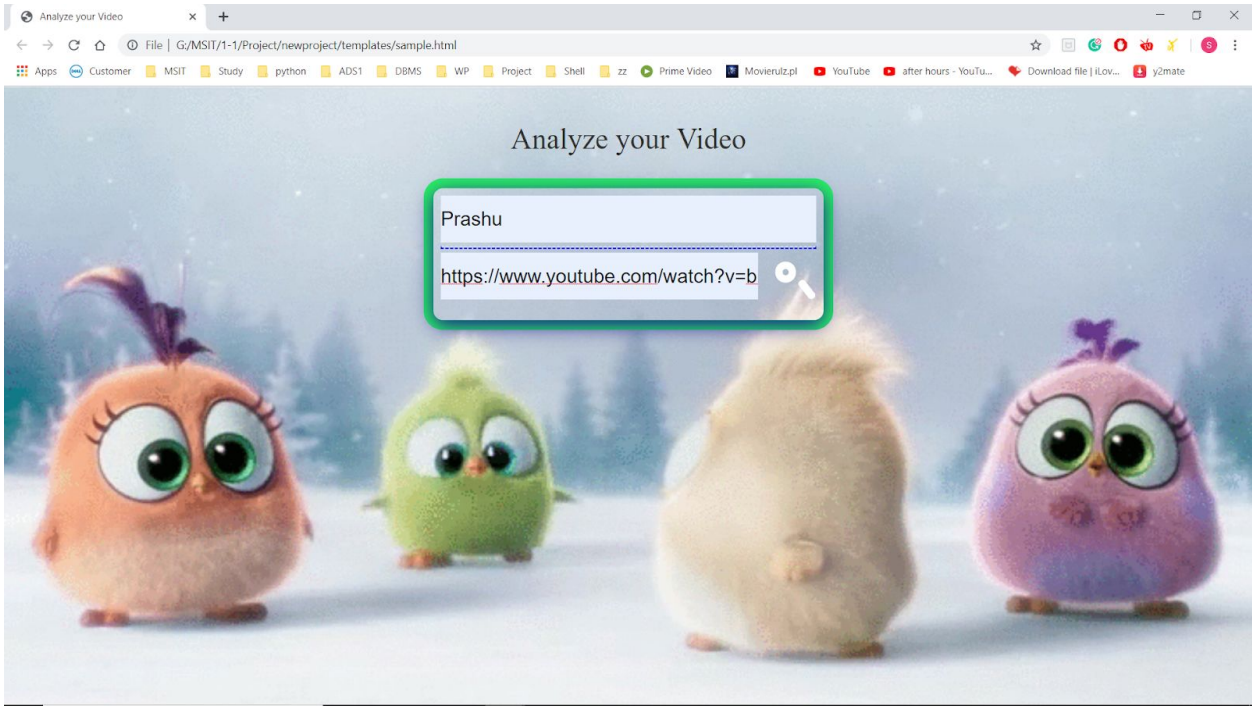
10.a: Home page of the project. Enter the video link to analyze based on the comments.



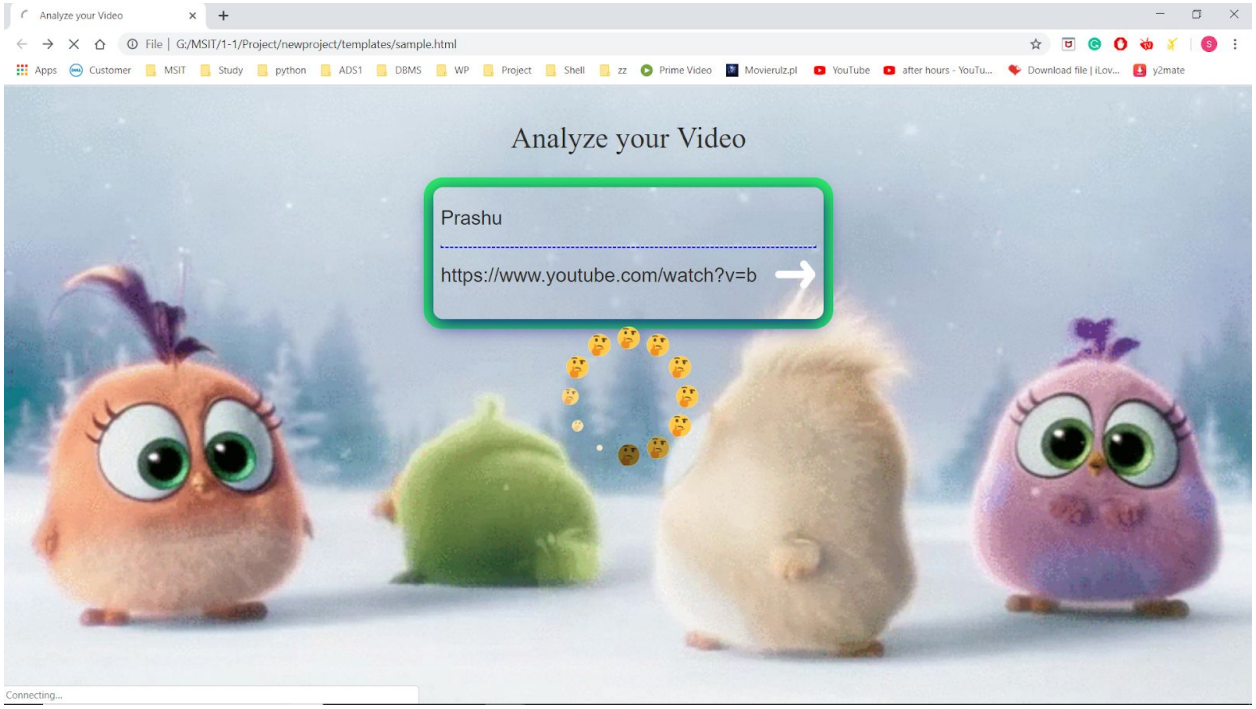
10.b: In the above figure, there are recent searches for the video analyzed.



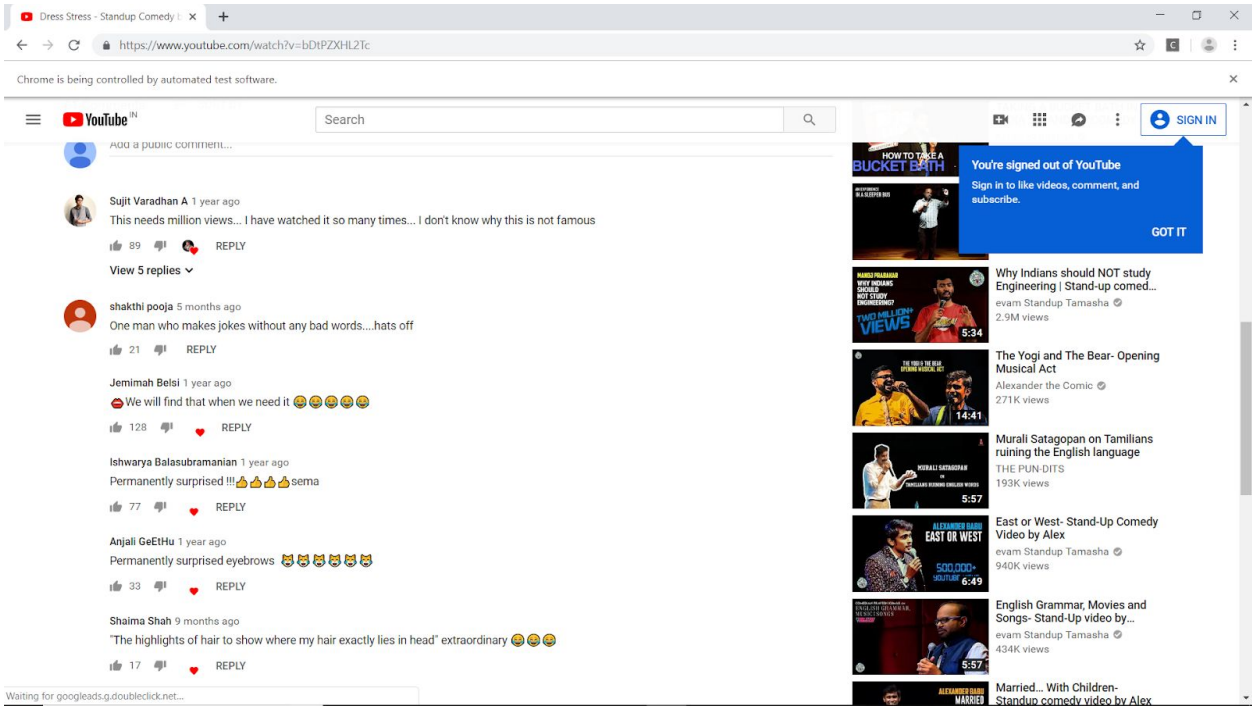
10.c: Enter the Youtube link in the Search bar



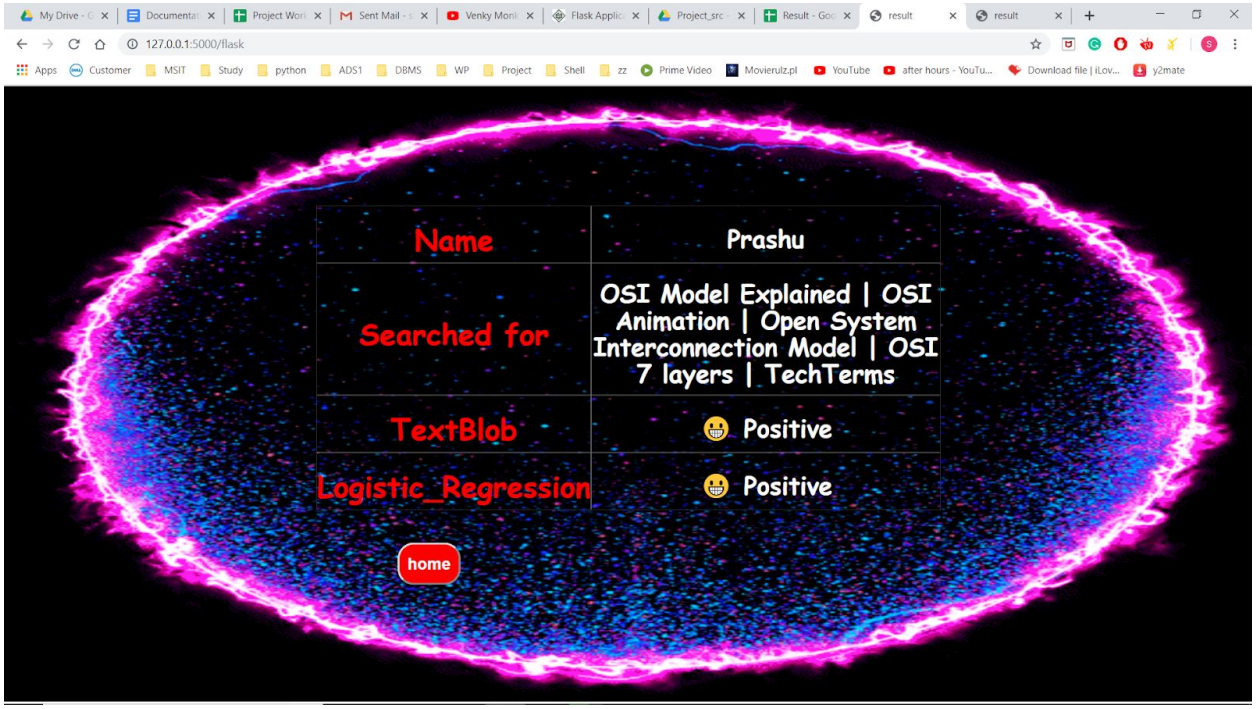
10.d: In the above figure, End_user entered the YouTube link in the text field.



10.e: process loading is mentioned using gif

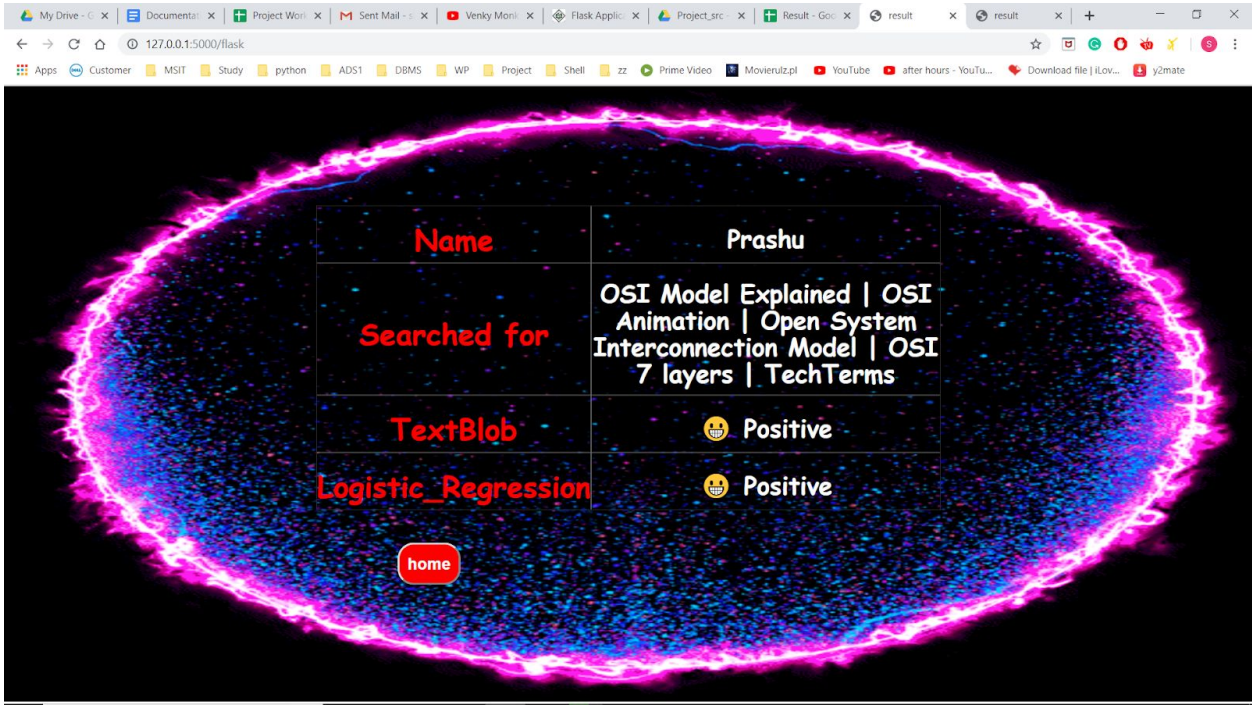


10.f: In the above figure, the selenium tool is scrapping the comments based on user input.

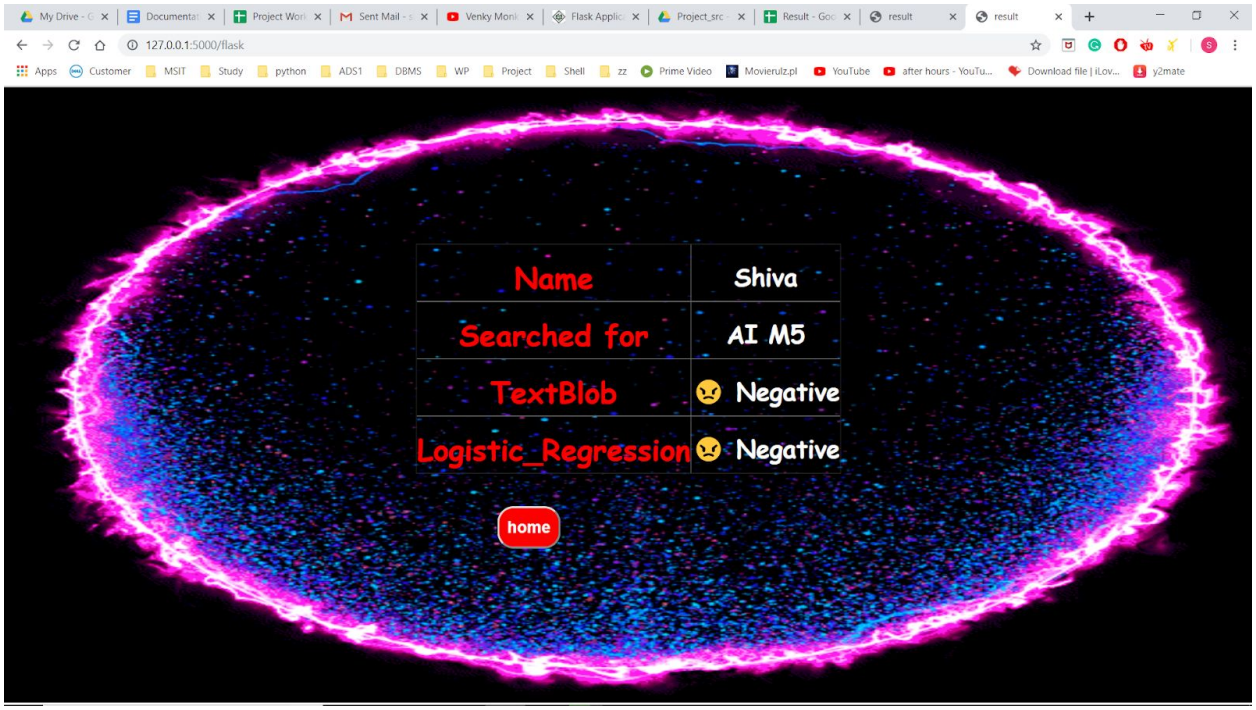


10.g: In the above figure, the result is represented based on the TextBlob and Logistic Regression.

Output



This Screen represents Positive Sentiment for the given URL by the End User based on Textblob and Logistic Regression



This Screen represents Positive Sentiment for the given URL by the End User based on Textblob

Conclusions

Nowadays, sentiment analysis or opinion mining is a hot topic in machine learning. We are still far to detect the sentiments of s corpus of texts very accurately because of the complexity in the English language and even more if we consider other languages such as Chinese.

In this project, we tried to show the basic way of analyzing the video based on the comments into a positive or negative category using TextBlob and Logistic Regression as a baseline. We could further improve our classifier by trying to extract more features from the replies to the comments, likes/dislikes for a particular comment.

References

1. https://www.seleniumhq.org/projects/webdriver/
2. https://stackoverflow.com
3. https://www.tutorialspoint.com/flask/
4. https://www.w3schools.com/html/