YOUTUBE CASE STUDY USING NLP

MINIPROJECT REPORT

Submitted by

DEVANSH JAISWAL [RA2011047010133],

RAHI JAIN [RA2011047010084],

HARISH CHOUDHARY [RA2011047010139],

GAURAV CHATURVEDI [RA2011047010106],

AAYUSH SAXENA [RA2011047010126]

Under the guidance of

Dr. MAHESHWARI A

(Guide Affiliation)

Assistant Professor

Department of Computational Intelligence

FACULTY OF ENGINEERING AND TECHNOLOGY

SCHOOL OF COMPUTING

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Kattankulathur, Kancheepuram

MAY 2023

SRM INSTITUTEOF SCIENCEANDTECHNOLOGY

(UnderSection3ofUGCAct,1956)

BONAFIDE CERTIFICATE

Certified that **18AIC305T_ Inferential Statistics and Predictive Analytics** titled **"YOUTUBE CASE STUDY USING NLP"** is the bonafide work of **"DEVANSH JAISWAL [RA2011047010133], RAHI JAIN [RA2011047010084], HARISH CHOUDHARY [RA2011047010139], GAURAV CHATURVEDI [RA2011047010106], AAYUSH SAXENA [RA20110470126]** who carried out the minor project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. MAHESHWARI A

GUIDE

Assistant Professor

SIGNATURE

Dr. Annie Uthra

HEAD OF THE DEPARTMENT

Professor

Dept.of Computational

2

–YouTube Case Study Using NLP

– -

1. Abstract

Sentiment analysis or opinion mining is the field of study related to analyzing opinions, sentiments, evaluations, attitudes, and emotions of users that they express on social media and other online resources. The revolution of social media sites has also attracted users to video-sharing sites, such as YouTube. Online users express their opinions or sentiments on the videos that they watch on such sites. This project presents a brief survey of techniques to analyze opinions posted by users about a particular video, the category with the maximum engagement, and the channel with the largest number of trending videos.

Corporate companies are using social media for improving their businesses, the data mining and analysis are very important these days. This paper deals with the analysis of YouTube Data. The analysis is done using user features such as Views, Comments, Likes, and Dislikes. Analysis can be performed using NLTK and other Machine learning models and python libraries like Textblob, Wordcloud,
Pandas, Matplotlib, seaborn, and Plotly libraries to classify the YouTube Data and obtain useful information.

Keywords:
Opinion Mining, Sentiment Analysis, Social Media, Social Networking, User Reviews, Video Sharing, Trending
Videos, Maximum Engagement,and YouTube.

2. Introduction

Opinion mining or comments toward attitude evaluation, an individual entity, are usually called sentiment. Everyone is free to give opinions related to the present opinions on youtube. Hence people have free will to express their opinion regarding the performance. Due to the rise of many critics that appear in a short amount of time, there is a need to conduct an analysis of opinion mining. The process of searching or tracing the natural language to find patterns or moods of society against certain products, people, or topics is called Sentiment Analysis. Sentiment analysis is also often referred to as the opinion of mining.

YouTube has a list of trending videos that is updated constantly. Analyzing these trending videos can give content creators greater perspective and knowledge for increasing the popularity and brand of their channels. Companies and businesses using social media and digital platforms can also use this analysis to boost their growth in business by publishing videos or sponsoring appropriate channels at the right time.

The following are the primary objective of this study:

- Performing Sentiment Analysis using Textblob.
- Wordcloud Analysis of Data.
- Perform Emoji Analysis.
- Which category has the maximum likes?
- Whether the Audience is engaged or not.&
- Which channel has the largest number of trending videos?

3. Libraries

3.1 TextBlob:

TextBlob is a python library for Natural Language Processing (NLP).TextBlob actively used Natural Language ToolKit (NLTK) to achieve its tasks. NLTK is a library that gives easy access to many lexical resources and allows users to work with categorization, classification, and many other tasks. TextBlob is a simple library that supports complex analysis and operations on textual data.

For lexicon-based approaches, a sentiment is defined by its semantic orientation and the intensity of each word in the sentence. This requires a pre-defined dictionary classifying negative and positive words. Generally, a text message will be represented by a bag of words. After assigning individual scores to all the words, the final sentiment is calculated by some pooling operation like taking an average of all the sentiments.

TextBlob returns the polarity and subjectivity of a sentence. Polarity lies between [-1,1], -1 defines a negative sentiment and 1 defines a positive sentiment. Negative words reverse the polarity. TextBlob has semantic labels that help with fine-grained analysis.

Import Textblob

```
from textblob import TextBlob
```

3.2 WordCloud:

A word cloud is a data visualization tool for depicting words in an image. Each word's size varies and represents the word's weightage. The larger the size, the higher its importance. In most cases, the frequency of occurrence of a word is used as its weightage

Word cloud has various applications in the data science field. It is normally used during the exploratory data analysis phase of an NLP task. It gives a quick overview of the frequently occurring term in a large text corpus and also gives an indication of a term's usage relative to other terms in the corpus. It is applied to sentiment analysis tasks where a quick summary of the overall sentiment can be visualized. It is also used for topic modeling tasks where the topic terms can be quickly spotted in the word cloud.

Word cloud gives a quick summary of the text corpus from which it is created. Looking at the below word cloud it is easy to identify that the text corpus is about using reinforcement learning, in particular, the deep q-network method on a stock dataset.

4. Data Description

The Dataset that we have used is obtained from the internet. We will analyze USA trending videos Comments. Then we also have a dataset of Youtube video comments from different countries like Canada, India, Great Britain, France, Japan, Korea, and Russia.

```
In [47]: path=r"C:\Users\91842\OneDrive\Desktop\Youtube Analysis\Youtube_Case_Study\additional_data"

In [48]: files=os.listdir(path)
         files

Out[48]: ['CAvideos.csv',
          'CA_category_id.json',
          'DEvideos.csv',
          'DE_category_id.json',
          'FRvideos.csv',
          'FR_category_id.json',
          'GBvideos.csv',
          'GB_category_id.json',
          'INvideos.csv',
          'IN_category_id.json',
          'JPvideos.csv',
          'JP_category_id.json',
          'KRvideos.csv',
          'KR_category_id.json',
          'MXvideos.csv',
          'MX_category_id.json',
          'RUvideos.csv',
          'RU_category_id.json',
          'USvideos.csv',
          'US_category_id.json']
```

Dataset shape

```
In [55]: full_df.shape

Out[55]: (375942, 17)
```

5. Results

5.1 Performing Sentiment Analysis

Polarity lies between [-1,1], -1 defines a negative sentiment and 1 defines a positive sentiment. Negative words reverse the polarity. TextBlob has semantic labels that help with fine-grained analysis.

```
In [16]: comments['polarity']=polarity

In [17]: comments.tail(14)
```

Out[17]:

	video_id	comment_text	likes	replies	polarity
691386	EoejGgUNmVU	❤	0	0	0.000000
691387	EoejGgUNmVU	Замечательно	0	0	0.000000
691388	EoejGgUNmVU	Best song ever 🤘	0	0	1.000000
691389	EoejGgUNmVU	excellent performance	0	0	1.000000
691390	EoejGgUNmVU	Zajebistaa...jak zawsze,live super.	0	0	0.333333
691391	EoejGgUNmVU	Love you	0	0	0.500000
691392	EoejGgUNmVU	"L O S T O N Y O U"	1	0	0.000000
691393	EoejGgUNmVU	<3	0	0	1.000000
691394	EoejGgUNmVU	Amazing!!!!!💚🌹🎷🔥	2	0	0.000000
691395	EoejGgUNmVU	Лучшая	1	0	0.000000
691396	EoejGgUNmVU	qu'est ce que j'aimerais que tu viennes à Roan...	0	0	0.000000
691397	EoejGgUNmVU	Von a mexico! ❤ te amo LP	0	0	0.000000
691398	EoejGgUNmVU	Isliği yeter...	0	0	0.000000
691399	EoejGgUNmVU	Kocham tą piosenkę❤❤❤byłam zakochana po uszy ...	0	0	0.000000

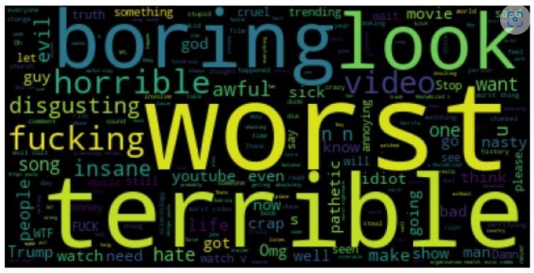
5.2 Frequently Used Words

Each word's size varies and represents the word's weightage. The larger the size, the higher its importance. In most cases, the frequency of occurrence of a word is used as its weightage.

5.2.1 Negative Words

```
Wordcld2=WordCloud(stopwords=set(STOPWORDS)).generate(total_comm2)
plt.figure(figsize=(15,5))
plt.imshow(Wordcld2)
plt.axis('off')
```

```
(-0.5, 399.5, 199.5, -0.5)
```



5.2.2 Positive Words

```
Wordcld1=WordCloud(stopwords=set(STOPWORDS)).generate(total_comm)
plt.figure(figsize=(15,5))
plt.imshow(Wordcld1)
plt.axis('off')
```

```
(-0.5, 399.5, 199.5, -0.5)
```

5.3 Emoji Analysis

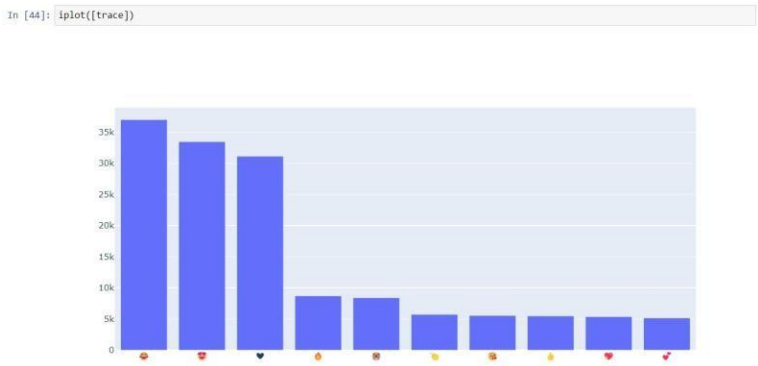
All the comments were labeled with their sentiment polarities, and 294549 emojis were used.

```
len(emoji_list)
```

```
294549
```

The top 10 emoji are

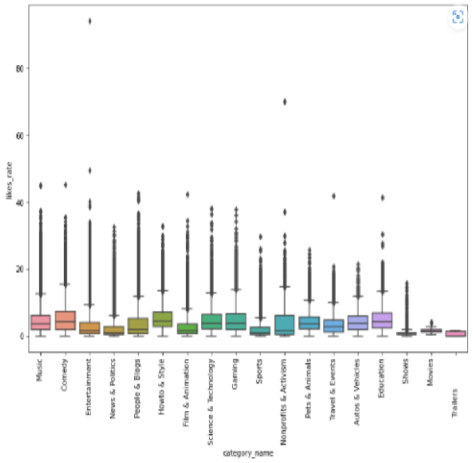
- '😂', 36987,
- '😍', 33453,
- '❤', 31119,
- '🔥', 8694, • '😭', 8398,
- '🅰', 5719,
- '😘', 5545, • '👍', 5476, • '💔', 5359,
- '💕', 5147



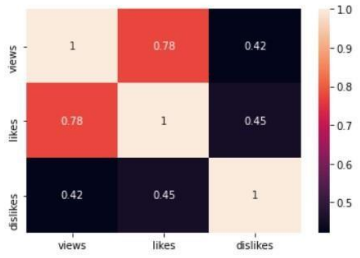
Above Graph Fig. Show the frequently used emoji.

5.4 Most Viewed and Liked Category

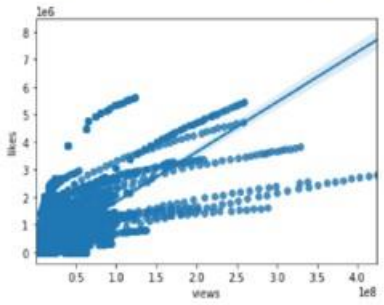
Below are the statistics of popular categories on YouTube. As you can see, the three most popular content categories on YouTube are Entertainment, Music, and
People & Blogs.




```
]: sns.heatmap(full_df[['views','likes','dislikes']].corr(),annot=True)
]: <AxesSubplot:>
```




```
In [73]: sns.regplot(data=full_df,x='views',y='likes')
Out[73]: <AxesSubplot:xlabel='views', ylabel='likes'>
```

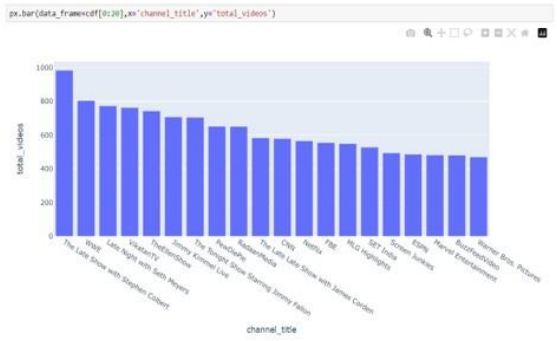


The above Graph shows the relation between likes and views.

5.5 Analysing trending videos

YouTube is a widely used and famous online video platform in the world today. YouTube has a list of trending videos that is updated constantly. Analysing these trending videos can give content creators greater perspective and knowledge for increasing the popularity .

Here we did an analysis of the channel with the most trending videos.



6. Conclusion

In this paper, we perform Sentiment Analysis using Textblob, Wordcloud Analysis of the Dataset, and Perform Emoji Analysis also, then we find which category has the maximum number of likes. Whether the Audience is engaged or not & Which channel has the largest number of trending videos?

7. References

https://www.mukpublications.com/resources/i jcic%20v13-2-16%3D%3Ddeshak%20revise d.pdf

https://www.irjet.net/archives/V7/i8/IRJET-V7 I8732.pdf

https://www.twinword.com/blog/features-of-to p-250-youtube-channels/

https://www.researchgate.net/publication/215 709959_Case_Study_Research_Methodolog y

https://www.researchgate.net/publication/266

137891_A_CASE_STUDY_IN_USING_YOU

TUBE_AND_FACEBOOK_AS_SOCIAL_ME DIA_TOOLS_IN_ENHANCING_STUDENT_

CENTERED_LEARNING_AND_ENGAGEM

ENT

https://www.researchgate.net/publication/284

724207_Sentiment_Analysis_on_YouTube_A_Brief_Survey

https://www.academia.edu/75267532/Sentim ental_Analysis_of_YouTube_Videos

https://www.irjet.net/archives/V7/i12/IRJET-V 7I12374.pdf

https://ritikasingh95.github.io/Documents/Pu blications/YOUTUBE%20COMMENTS%20S ENTIMENT%20ANALYSIS.pdf

https://www.semanticscholar.org/paper/Senti ment-Analysis-on-YouTube%3A-A-Brief-Surv ey-Asghar-Ahmad/6c49a4d193e61387340b9 b6448d9d2295b55b814