ISSN No:-2456-2165

A Sentiment Analysis Case Study to Understand How a Youtuber can Derive Decision Insights from Comments

Chandreyi Chowdhury
School of Advanced Sciences
Vellore Institute of Technology
Vellore, India

BaibhavPathy
School of Electrical Engineering
Vellore Institute of Technology
Vellore, India

Abstract:- YouTube is considered the biggest platform for content creators to share their content with the world. Usually, a YouTuber aims to give his/her viewers the best content possible by going through the comments of their past videos. On average, the comments can go up to 10 thousand; hence, it becomes practically impossible to go through every comment and get an idea of what the viewers want or expect.

Our work provides a model based on Python that extracts the comments of a YouTube video which then becomes our dataset. A Machine Learning techniqueknown as Sentiment Analysis (Classification Model) is applied to the dataset extracted to provide the YouTuber with a better understanding of the distribution of the sentiment of his/ her viewers, which in turn helps them get an idea of the thoughts of the viewers and also what the viewers expect from their future videos.

***Keywords**:- YouTube, Sentiment Analysis, Classification, Decision Insights, Case Study.*

I. INTRODUCTION

Whether one's material is intended for the entire public or is tailored to a community, age group, etc., social media platforms are thought to be the simplest and fastest method to simultaneously reach millions of individuals. Through social media, one may communicate with someone who is thousands of miles away. Non-textual information, such videos, photos, and animations, is shared on many websites using systems that let people leave comments on individual items. With millions of videos submitted by its users and billions of comments for each one, YouTube is the most well-known of these programmes for sharing material in the form of videos [4]. These clips include material that might significantly harm a person's or a company's reputation. By counting the likes and dislikes of a video, it is easy to determine its reputation. It is good material if there are significantly more likes than dislikes, while typically terrible content has many more dislikes than likes[1].

There are, however, a few approaches to quantify reputation. This highlights how crucial it is to automatically extract thoughts and views shared on social media[5].In order to ascertain how YouTube viewers feel about video content, sentiment analysis may be used to text data from the comments area of a video. The most effective option for deciphering each comment's significance is a text-mining strategy. For YouTube viewers to assess the significance of the uploaded video based on user opinion comments, the categorization of positive and negative content becomes extremely crucial.[7].

This study focuses on utilising several classifiers from Python's Scikit-learn module to do sentiment analysis using a machine learning method on two case study datasets scraped from YouTube. Functions from the Python libraries Selenium and BeautifulSoup are used in the text mining portion. On the scraped data, stemming, lemmatization, and text pre-processing are carried out utilising a variety of Natural Language Toolkit methods.

II. RELATED WORKS

Numerous studies have explored the use of various sentiment analysis approaches, including machine learning, Naive Bayes, Lexicon-based techniques, and mBERT. Abbi Nizar Muhammad, Saiful Bukhori, and Priza Pandunata developed a sentiment analysis strategy by combining Naive Bayes-Support Vector Machine (NBSVM) with a binary classification approach, achieving 91% accuracy [7]. Sudhanshu Ranjan, Dheeraj Mekala, and Jingbo Shang applied the mBERT approach to code-switching data, improving the performance across multiple datasets [20]. Tanvi Mehta and Ganesh Deshmukh compared various methods such as linear regression, SVM, decision trees, random forests, and artificial neural networks, and found decision trees and ANN had the lowest root mean square errors [23]. Ritika Singh and Ayushka Tiwari used six machine learning techniques, including Gaussian Naive Bayes, SVM, logistic regression, decision trees, KNN, and random forest, to develop a sentiment analysis system, evaluating accuracy and F-score [19].

Hanif Bhuiyan, Rajon Bardhan, Jinat Ara, and M. Rashedul Islam studied retrieving YouTube videos using sentiment analysis on user comments. Their approach utilized natural language processing (NLP) and sentiment analysis to find relevant and well-liked videos on YouTube that match the search criteria [2]. Martin Wöllmer, Felix Weninger, Tobias Knaup, BjörnSchuller, and Congkai Sun aimed to analyze voice sentiment in internet cinema review recordings, taking into account the speaker's positive valence content and speech-based emotional audio elements [3]. FazalMasudKundi, AfsanaMarwat, Shakeel Ahmad, and Muhammad Zubair Asghar employed sentiment lexicons, such as word net and senti-word-net, to detect the polarity of feeling in their research paper [4]. Alessandro Moschitti, Aliaksei Severyn, Barbara Plank, Agata Rotondi, and Olga Uryupina discussed the "SenTube" dataset, which comprises user-generated comments on YouTube videos that have been catalogued for informativeness and sentiment polarizability, to create classifiers for several critical NLP tasks [5].

S. Nawaz, M. Rafiq, and M. Rizwan proposed a unique approach for calculating the suggestions of efficacy of a YouTube video content by using the Google API to determine the total sentiment text analytics of comments and responses [9]. Philipp A. Toussaint, Sebastian Lins, Maximilian Renner, Ali Sunyaev, and Scott Thiebes used Bing (binary), National Research Council Canada (NRC) emotion, and 9-level sentiment analysis to determine user behaviors toward videos about DTC genetic testing, as expressed in the comment section of their research paper [13]. However, there is still no specific method designed for YouTubers/Influencers to understand the sentiment distribution of their video's comments to make informed decisions based on user feedback. This work presents an initial attempt to automatically extract or scrape the comments from YouTube, apply text processing techniques, and fit them into a machine learning model to provide a comprehensive study of how content creators can leverage sentiment analysis to their advantage and understand the users' overall viewpoint.

III. METHODOLOGY

This section elaborates on the step–by–step procedure we have used in order to arrive at the final model. The procedure being proposed operates through a sequence of four steps, which are presented visually in Figure 1. The initial step is to select the video one wants to perform sentiment analysis onextraction code so that the comments can be downloaded and stored as a CSV (Comma Separated Values) file. Then using Python packages like Selenium and Beautiful Soup, one can extract the comments. Furthermore, one can perform Sentiment Analysis using our proposed model to get almost accurate classifications on those comments. The accuracy metric also has to be decided, and the validation score needs to be over ninety per cent to be called almost accurate. The steps are described more elaborately in the following sub-section.

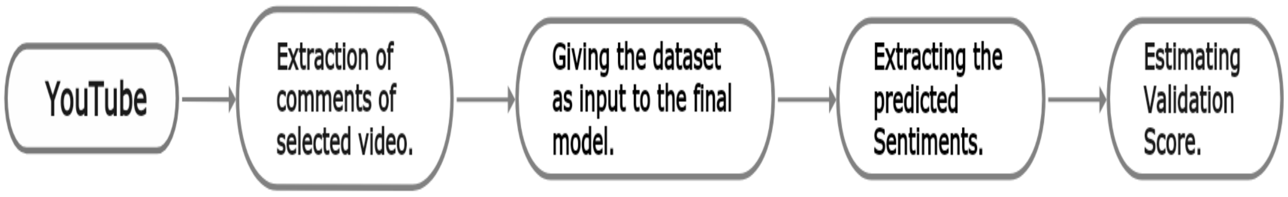
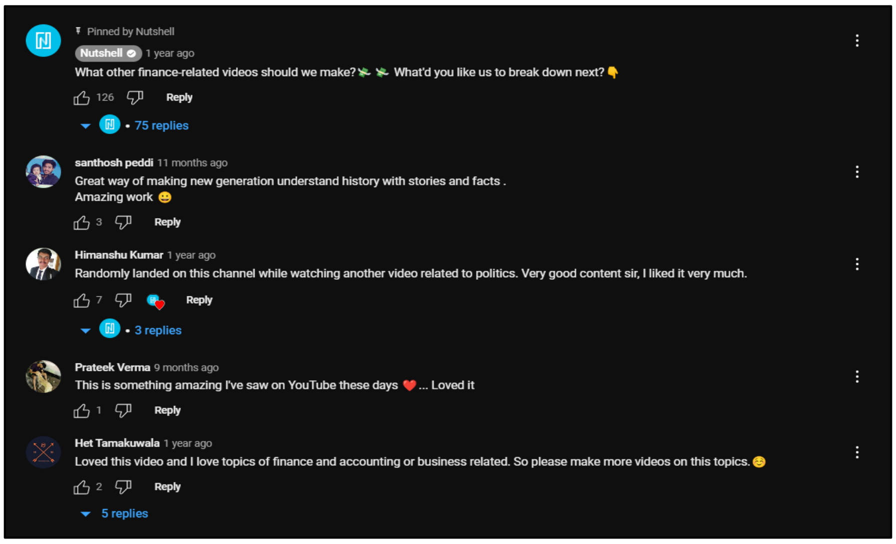

YouTube
Extraction of comments of selected video.
Giving the dataset as input to the final model.
Extracting the predicted Sentiments.
Estimating Validation Score.

Fig. 1: Sentiment Analysis Procedure

A. Comments Extraction

Comments extraction is a vital process in our research that involves extracting comments from selected YouTube videos and storing them in a CSV file. We utilize two web-scraping libraries in Python, Selenium and BeautifulSoup, in our model for this task. Selenium is used to automate web browser interaction through Python scripts, while BeautifulSoup simplifies the process of extracting data from websites by providing Python-based idioms for parsing HTML or XML code. Our model code, which carries out the comments extraction process, is available online, along with Figures 1, 2, and 3, which illustrate the steps involved in the process and some relevant code snippets. However, we note that we only extract a random subset of comments from each video, as scraping the entire comments section violates YouTube's policies.


Pinned by Nutshell
Nutshell 1 year ago
What other finance-related videos should we make? What'd you like us to break down next?
126 Reply
75 replies

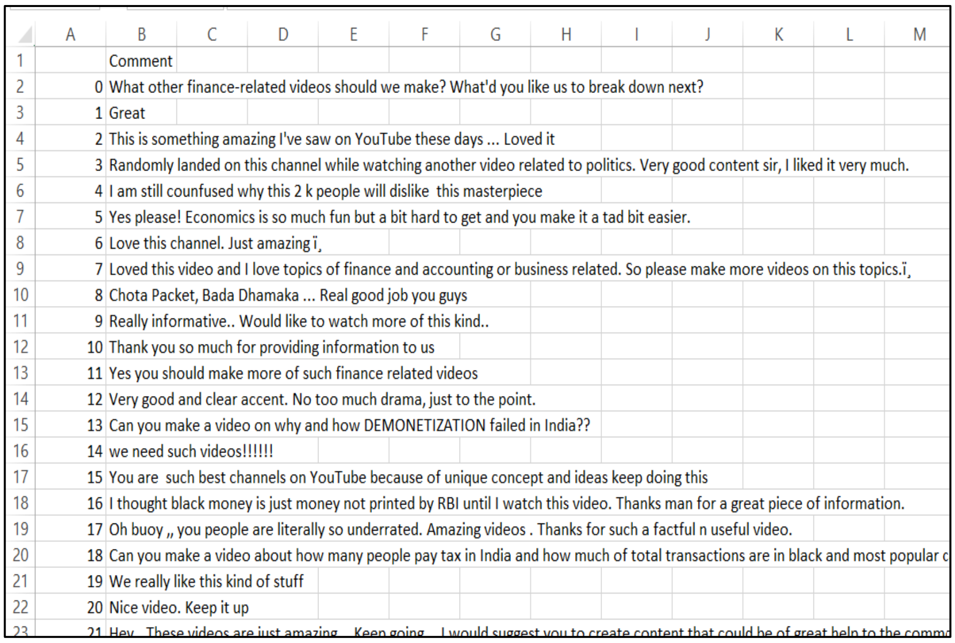
santhosh peddi 11 months ago
Great way of making new generation understand history with stories and facts .
Amazing work
3 Reply

Himanshu Kumar 1 year ago
Randomly landed on this channel while watching another video related to politics. Very good content sir, I liked it very much.
7 Reply
3 replies

Prateek Verma 9 months ago
This is something amazing I've saw on YouTube these days ❤️ ... Loved it
1 Reply

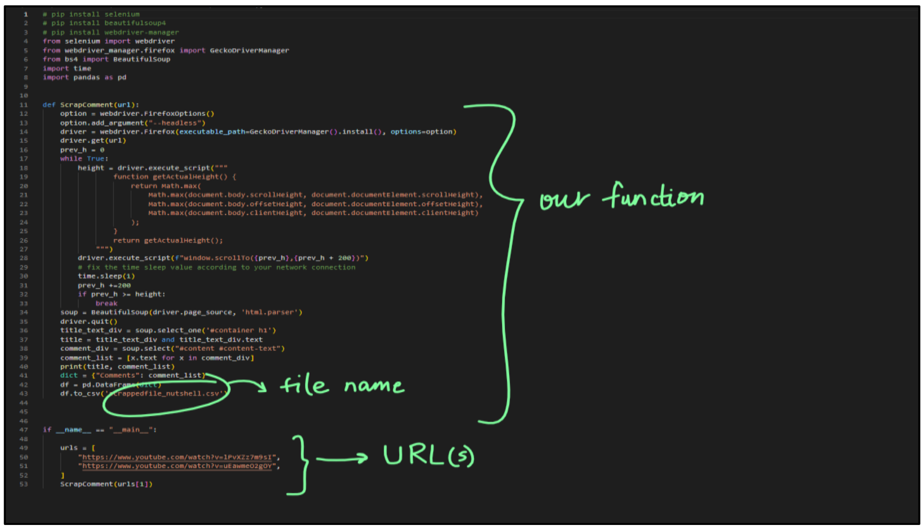
Het Tamakuwala 1 year ago
Loved this video and I love topics of finance and accounting or business related. So please make more videos on this topics.
2 Reply
5 replies

Fig. 2: User Comments on YouTube



	A	B
1		Comment
2	0	What other finance-related videos should we make? What'd you like us to break down next?
3	1	Great
4	2	This is something amazing I've saw on YouTube these days ... Loved it
5	3	Randomly landed on this channel while watching another video related to politics. Very good content sir, I liked it very much.
6	4	I am still counfused why this 2 k people will dislike this masterpiece
7	5	Yes please! Economics is so much fun but a bit hard to get and you make it a tad bit easier.
8	6	Love this channel. Just amazingï¸
9	7	Loved this video and I love topics of finance and accounting or business related. So please make more videos on this topics.ï¸
10	8	Chota Packet, Bada Dhamaka ... Real good job you guys
11	9	Really informative.. Would like to watch more of this kind..
12	10	Thank you so much for providing information to us
13	11	Yes you should make more of such finance related videos
14	12	Very good and clear accent. No too much drama, just to the point.
15	13	Can you make a video on why and how DEMONETIZATION failed in India??
16	14	we need such videos!!!!!!
17	15	You are such best channels on YouTube because of unique concept and ideas keep doing this
18	16	I thought black money is just money not printed by RBI until I watch this video. Thanks man for a great piece of information.
19	17	Oh buoy ,, you people are literally so underrated. Amazing videos . Thanks for such a factful n useful video.
20	18	Can you make a video about how many people pay tax in India and how much of total transactions are in black and most popular c
21	19	We really like this kind of stuff
22	20	Nice video. Keep it up
23	21	Hey.. These videos are just amazing... Keep going... I would suggest you to create content that could be of great help to the comm

Fig. 3: Scrapped Comments



```python
# pip install selenium
# pip install beautifulsoup4
# pip install webdriver-manager
from selenium import webdriver
from webdriver_manager.firefox import GeckoDriverManager
from bs4 import BeautifulSoup
import time
import pandas as pd


def ScrapComment(url):
    option = webdriver.FirefoxOptions()
    option.add_argument("--headless")
    driver = webdriver.Firefox(executable_path=GeckoDriverManager().install(), options=option)
    driver.get(url)
    prev_h = 0
    while True:
        height = driver.execute_script("""
                function getActualHeight() {
                    return Math.max(
                        Math.max(document.body.scrollHeight, document.documentElement.scrollHeight),
                        Math.max(document.body.offsetHeight, document.documentElement.offsetHeight),
                        Math.max(document.body.clientHeight, document.documentElement.clientHeight)
                    );
                }
                return getActualHeight();
            """)
        driver.execute_script(f"window.scrollTo({prev_h},{prev_h + 200})")
        # fix the time sleep value according to your network connection
        time.sleep(1)
        prev_h +=200
        if prev_h >= height:
            break
    soup = BeautifulSoup(driver.page_source, 'html.parser')
    driver.quit()
    title_text_div = soup.select_one('#container h1')
    title = title_text_div and title_text_div.text
    comment_div = soup.select("#content #content-text")
    comment_list = [x.text for x in comment_div]
    print(title, comment_list)
    dict = {"Comments": comment_list}
    df = pd.DataFrame(dict)
    df.to_csv('scrappedfile_nutshell.csv')


if __name__ == "__main__":

    urls = [
        "https://www.youtube.com/watch?v=lPvXZz7m9sI",
        "https://www.youtube.com/watch?v=uEawmeO2gOY",
    ]
    ScrapComment(urls[1])
```

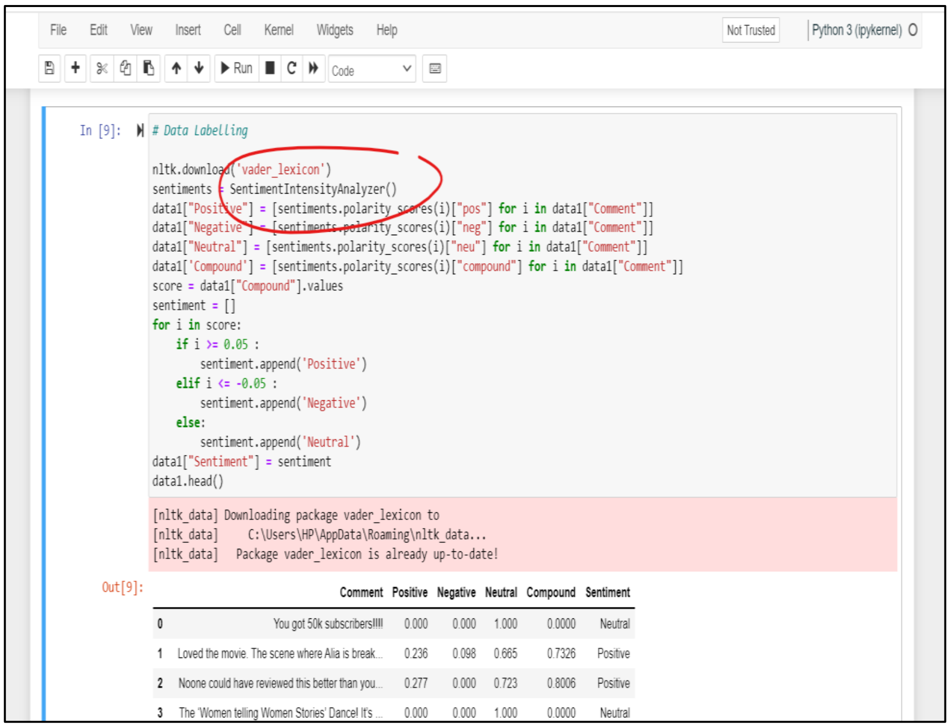
Fig. 4: Extraction Code Snip

B. *Text Preprocessing and Model Building*

Once the comments were extracted, we imported the CSV file into Python and performed the necessary pre-processing steps. Non-English comments and symbolic comments were removed from the dataset, although emoji-based comments could have been valuable in conveying users' perspectives, they were not included due to the limitations of our model's scope and context. We transformed the dataset to classify sentiments as Positive, Negative, or Neutral.

To perform text sentiment analysis, we used the VADER (Valence Aware Dictionary for Sentiment Reasoning) model, which considers both the polarity (positive/negative) and intensity (strong) of emotions. This model can be applied to unlabelled text data immediately and is included in the NLTK package. VADER uses a dictionary that assigns sentiment scores to lexical data, allowing us to determine the intensity of emotion in a text by summing up the intensity of each word. For example, words like "love," "enjoy," "glad," and "like" all convey positive feelings. VADER also understands the underlying meanings of these words, such as the negative connotation of "did not love" and how capitalization and punctuation can emphasize words, such as "ENJOY."

The Sentiment Intensity Analyzer() function of VADER takes a string as input and returns a dictionary containing scores for each category of negative, neutral, positive, and compound (normalized by the above scores). The implementation of this function is shown in Figure 4.



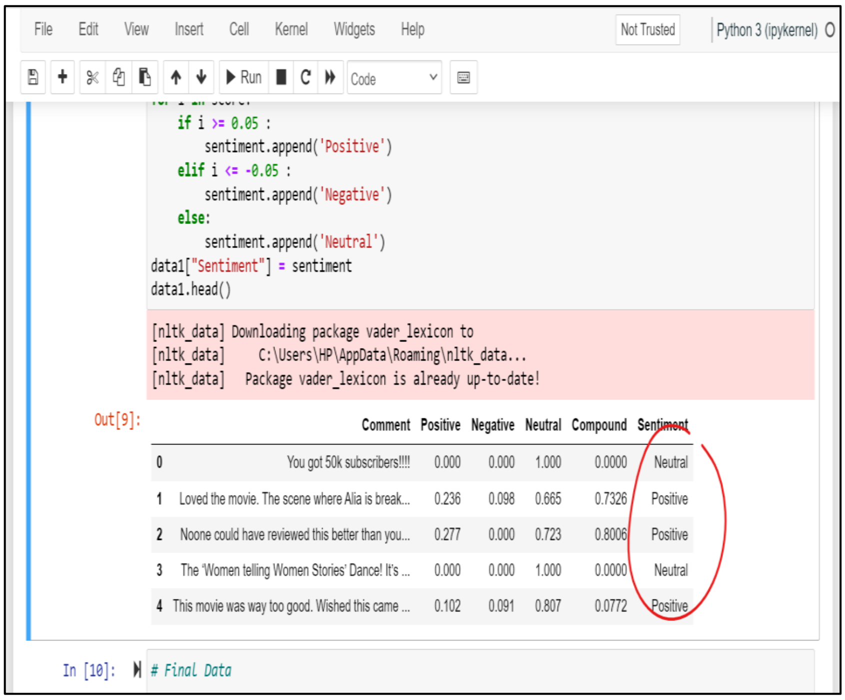
```
# Data Labelling

nltk.download('vader_lexicon')
sentiments = SentimentIntensityAnalyzer()
data1["Positive"] = [sentiments.polarity_scores(i)["pos"] for i in data1["Comment"]]
data1["Negative"] = [sentiments.polarity_scores(i)["neg"] for i in data1["Comment"]]
data1["Neutral"] = [sentiments.polarity_scores(i)["neu"] for i in data1["Comment"]]
data1['Compound'] = [sentiments.polarity_scores(i)["compound"] for i in data1["Comment"]]
score = data1["Compound"].values
sentiment = []
for i in score:
    if i >= 0.05 :
        sentiment.append('Positive')
    elif i <= -0.05 :
        sentiment.append('Negative')
    else:
        sentiment.append('Neutral')
data1["Sentiment"] = sentiment
data1.head()
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\HP\AppData\Roaming\nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

	Comment	Positive	Negative	Neutral	Compound	Sentiment
0	You got 50k subscribers!!!!	0.000	0.000	1.000	0.0000	Neutral
1	Loved the movie. The scene where Alia is break...	0.236	0.098	0.665	0.7326	Positive
2	Noone could have reviewed this better than you...	0.277	0.000	0.723	0.8006	Positive
3	The 'Women telling Women Stories' Dance! It's ...	0.000	0.000	1.000	0.0000	Neutral

Fig. 4: Sentiment Intensity Analyser() Function Usage

Finally, we get a complete analysis of every review as positive, negative, or neutral. Figure 5 shows the result quite evidently.



```
    if i >= 0.05 :
        sentiment.append('Positive')
    elif i <= -0.05 :
        sentiment.append('Negative')
    else:
        sentiment.append('Neutral')
data1["Sentiment"] = sentiment
data1.head()
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\HP\AppData\Roaming\nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

	Comment	Positive	Negative	Neutral	Compound	Sentiment
0	You got 50k subscribers!!!!	0.000	0.000	1.000	0.0000	Neutral
1	Loved the movie. The scene where Alia is break...	0.236	0.098	0.665	0.7326	Positive
2	Noone could have reviewed this better than you...	0.277	0.000	0.723	0.8006	Positive
3	The 'Women telling Women Stories' Dance! It's ...	0.000	0.000	1.000	0.0000	Neutral
4	This movie was way too good. Wished this came ...	0.102	0.091	0.807	0.0772	Positive

```
# Final Data
```

Fig. 5: Sentiments Classification

In the Sentiments Classification stage, text normalization techniques of stemming and lemmatization are applied. These techniques are widely used by natural language processing experts to prepare text, words, and documents for processing. Stemming is the process of generating morphological variations of a root or base word to help improve search accuracy when looking up information in the text. Stemming algorithms, also known as stemmers, are frequently used for this purpose. Our model employs three standard stemming methods: PorterStemmer(), LancasterStemmer(), and SnowballStemmer(). On the other hand, lemmatization applies a morphological analysis to words by considering the entire Lexicon of a language and reducing words to their base form. Lemmatization does not classify sentences; instead, it analyzes the context of a word to determine its part of speech. In our model, the standard Word Net Lemmatizer() method is used for lemmatization.

- Upon completing the necessary data transformations, we employed a range of machine learning models to perform sentiment analysis on our dataset. Specifically, we utilized the Gaussian Naïve Bayes, Logistic Regression, AdaBoost Classifier, Random Forest Classifier, K-Nearest Neighbors (K-NN), and Decision Tree Classifier algorithms.
- The Gaussian Naïve Bayes Classifier was our first choice due to its simplicity and effectiveness in classification tasks. It is capable of producing quick, accurate predictions by assuming that the features are independent of one another, which makes it ideal for high-dimensional datasets.
- Our second model, Logistic Regression, is a widely used supervised learning algorithm that predicts categorical outcomes based on a predefined set of independent variables. It is often used in machine learning due to its simplicity, interpretability, and ability to handle linearly separable data.
- The AdaBoost Classifier, our third model, employs a boosting technique called the AdaBoost algorithm to create an ensemble model. It operates by redistributing weights to each instance, with higher weights assigned to misclassified instances, to produce a series of learners that gradually improve on one another. This approach effectively reduces bias and variance in the predictions.
- We also employed the Random Forest Classifier, which uses a large number of decision trees to improve the accuracy of predictions. It avoids overfitting by training on different subsets of the input dataset and combining their results to obtain the final prediction.
- The K-NN algorithm, our fifth model, relies on the assumption that new and existing cases are comparable, and places the new instance in the category most similar to existing ones. It is commonly used for classification tasks but can also be used for regression.
- Finally, our last model was the Decision Tree Classifier, which creates a tree structure that represents the features of a dataset, decision-making branches, and classification results at each leaf node. It is a widely used algorithm in machine learning for classification and regression problems.

We will discuss the application of each of these models in detail in the case studies section (Section IV).

C. Validation Steps

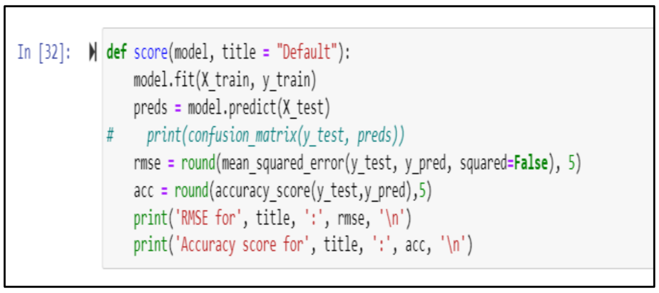
The process of ensuring that a model truly serves its intended function is known as model validation. This usually entails verifying that the model is accurate in the circumstances of its intended application. Validation of a model is done using an accuracy metric. Several accuracy metrics include accuracy score, logarithmic loss, confusion matrix, root mean squared error, the area under the curve, etc. Our model used accuracy score, and root mean squared error as our accuracy metrics. The basic intuitive notion of the formula for accuracy score is:

$$Accuracy = \frac{number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \qquad (1)$$

The method that gives the accuracy score in Python is accuracy_score(), found in the package sklearn. Moreover, for the root mean squared error, the intuitive formula is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted\ value - Actual\ Value)^2}{N}} \qquad (2)$$

The method that gives the root mean squared error in Python is mean_squared_error(), found in the package sklearn. We used a scoring function to give out the accuracy score, and the root mean squared error value for any model inserted. Figure 6 shows the body of the function used.



```
In [32]:  def score(model, title = "Default"):
              model.fit(X_train, y_train)
              preds = model.predict(X_test)
          #     print(confusion_matrix(y_test, preds))
              rmse = round(mean_squared_error(y_test, y_pred, squared=False), 5)
              acc = round(accuracy_score(y_test,y_pred),5)
              print('RMSE for', title, ':', rmse, '\n')
              print('Accuracy score for', title, ':', acc, '\n')
```

Fig. 6: Scoring Function Used In Our Model

D. Final Model

The final model is selected based on the accuracy metric's scores. Whichever model gives the highest score is chosen as the final model. The selection and the scores will vary for every dataset and hence cannot be predicted beforehand. So, the final model selection will be discussed further in the case studies section (Section IV).

IV. CASE STUDIES

A case study of two YouTube videos is presented to understand how a YouTuber can derive decision insights from the comments on his/her videos to improve the content or to know what his/her viewers want.Below are the two enlisted analyses and inferences of the sentiment distributions of the respective videos.

A. Movie Reviews by Sucharita -> Darlings (the 5th of August, 2022)

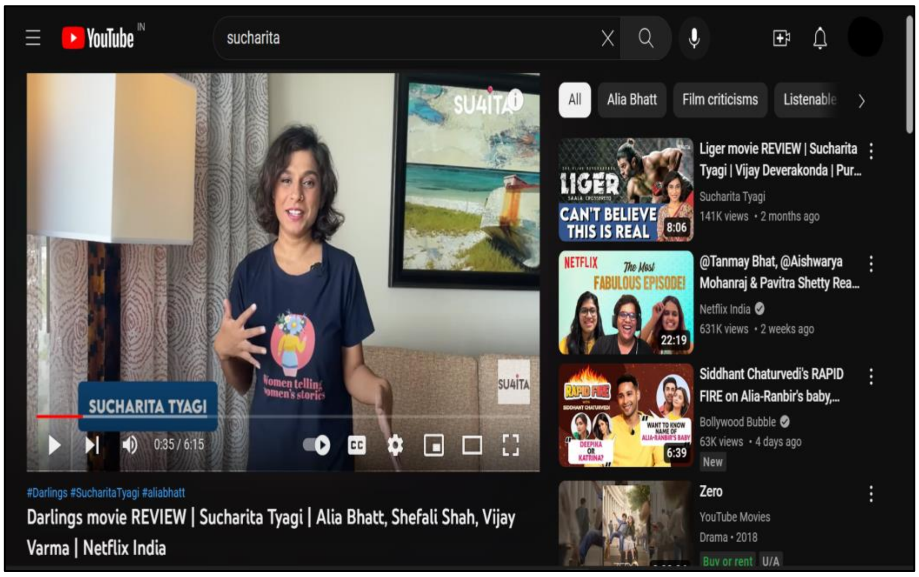

YouTube
sucharita
All
Alia Bhatt
Film criticisms
Listenable
Liger movie REVIEW | Sucharita Tyagi | Vijay Deverakonda | Pur...
Sucharita Tyagi
141K views • 2 months ago
@Tanmay Bhat, @Aishwarya Mohanraj & Pavitra Shetty Rea...
Netflix India
631K views • 2 weeks ago
Siddhant Chaturvedi's RAPID FIRE on Alia-Ranbir's baby,...
Bollywood Bubble
63K views • 4 days ago
New
Zero
YouTube Movies
Drama • 2018
Buy or rent
SUCHARITA TYAGI
0:35 / 6:15
#Darlings #SucharitaTyagi #aliabhatt
Darlings movie REVIEW | Sucharita Tyagi | Alia Bhatt, Shefali Shah, Vijay Varma | Netflix India

Fig. 8: Video 1 Snip

We chose our first case study for the account "Sucharita" on YouTube, shown in Figure 8. She is a full-time social media influencer whose content mainly concerns reviewing recent movies, series and online stream-able entertainment. We chose her because, being an influencer, she has to create content now and then and keep her audience engaged. It is highly beneficial for her to analyse and see the distribution of her viewers' opinions. What they want more of, what they detest, and so forth.For her account, we proceeded to analyse the video for the movie review: "Darlings", which is among the most viewed videos on her YouTube channel. The video was published on YouTube through the internet on the 5th of August, 2022. There has been a collection of 193 comments on this video over the three months, and we have scrapped 119 comments to work.The exact caption for the video is: “Darlings movie REVIEW | Sucharita Tyagi | Alia Bhatt, Shefali Shah, Vijay Varma | Netflix India”.

Table 1: First Case Study Video Details

Account	Sucharita Tyagi
Subscribers	62.5K
Video Caption	Darlings movie REVIEW | Sucharita Tyagi | Alia Bhatt, Shefali Shah, Vijay Varma | Netflix India
Published Date	The 5th of August, 2022
Video URL	https://www.youtube.com/watch?v=lPvXZz7m9sI&t=2s
Total Comments	193 (as of the 2nd of November 2022)
Scrapped Comments	119

Proceeding towards the model description and analysis of this dataset, we found that the processed data had 20 negative classes, 63 positive classes and 36 neutral classes. Figure 7 shows the same distribution result we got in the code output.

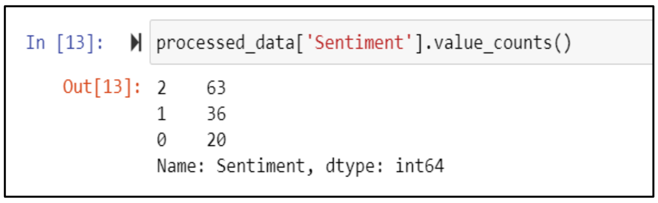

```
In [13]:  processed_data['Sentiment'].value_counts()

Out[13]: 2    63
         1    36
         0    20
         Name: Sentiment, dtype: int64
```

Fig. 7: Processed Data Distribution For Video 1

Note that 0 denotes the 'negative' class, 1 denotes the 'neutral' class, and 2 denotes the 'positive' class in the Figure. After this step, we unsample the minority classes, the 'neutral' and 'negative' classes. Upsampling is a technique used to balance imbalanced datasets. It involves generating new data points for the minority class to increase its representation in the dataset. This results in a more balanced distribution of data between different classes.The model does not tend to favour the class with most members, thanks to this equalisation process and concatenates the new dataframes to the majority class. So finally, our data at this stage contains 205 negative classes, 205 neutral classes and 63 positive classes.

Applying several machine learning classifiers on the data by training the models with 70 per cent of the data and saving the rest 30 per cent for testing, we get the best accuracy score of 0.9577 from the Gaussian Naïve Bayes Classifier. Thus, we conclude that the distribution we initially got is almost correct. It means that approximately 53 per cent of her viewers have a favourable opinion, 30 per cent have a neutral opinion, and the rest 17 per cent have a negative opinion about her video. Figure 9 shows the same diagrammatically.

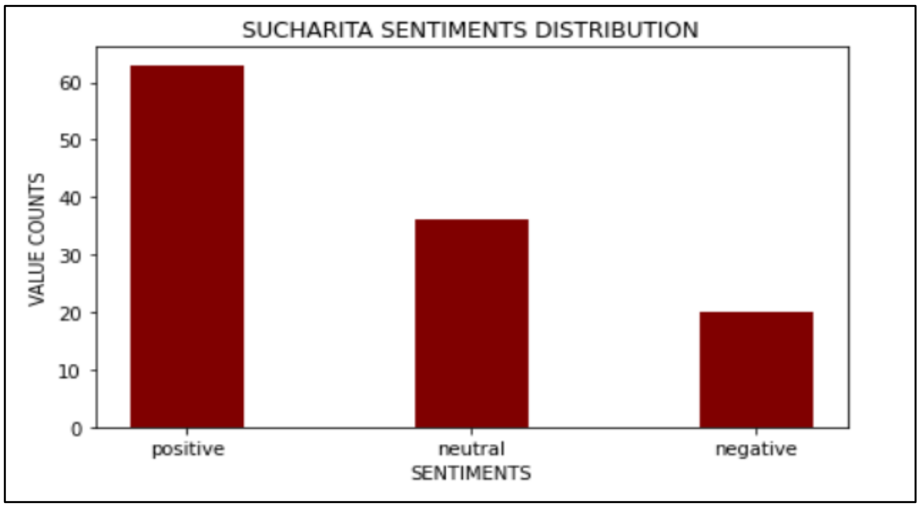


Fig. 9: Video 1 Sentiments Distribution

B. Nutshell -> Black Money (the 21st of May, 2021)

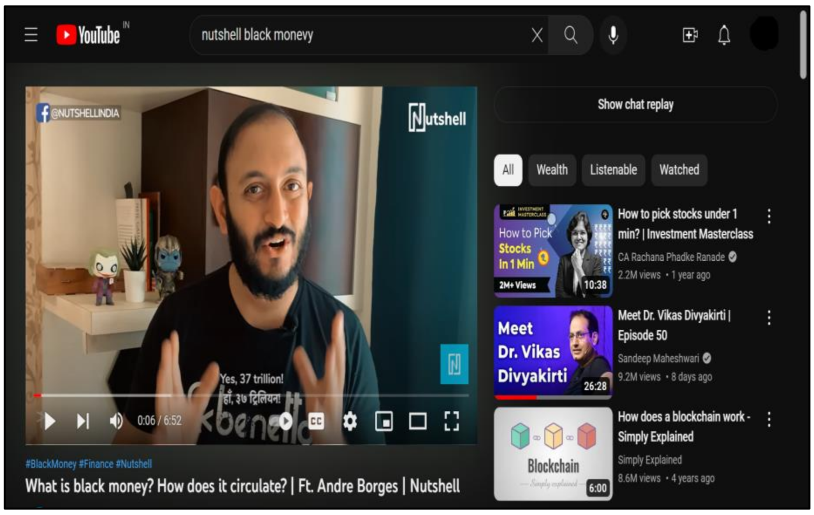


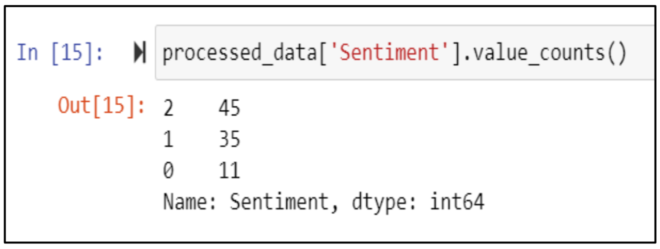
Fig. 10: Video 2 Snip

Secondly, we proceeded to analyse and draw inferences from one of the channel's " Nutshell " videos. It is shown in Figure 10. This channel's main aim is to bring our long lost and un-digged history stories or awareness videos about current affairs in a relatable way, revealing new fact survey videos etc. It can be called an encyclopaedia of knowledge of current and past news. We again chose a top-viewed video on their channel, which was about "Black Money". The exact caption of the video is "What is black money? How does it circulate? | Ft. Andre Borges | Nutshell".The video was published on YouTube through the internet on the 21st of May, 2021. There are 228 comments on the video for the seventeen months, and our algorithm has scrapped 93 comments to work upon.

Table 2: Second Case Study Video Details

Account	**Nutshell**
Subscribers	243K
Video Caption	What is black money? How does it circulate? | Ft. Andre Borges | Nutshell
Published Date	The 21st of May, 2021
Video URL	https://www.youtube.com/watch?v=uEawmeO2gOY&t=7s
Total Comments	228 (as of the 2nd of November 2022)
Scrapped Comments	93

For this video, proceeding towards the model description and analysis, we found that the processed data had 11 negative classes, 45 positive classes and 35 neutral classes. Figure 11 shows the same distribution result we got in the code output.



```
In [15]:   processed_data['Sentiment'].value_counts()

Out[15]: 2    45
         1    35
         0    11
         Name: Sentiment, dtype: int64
```

Fig. 11: Processed Data Distribution for Video 2

Again, note that 0 denotes the 'negative' class, 1 denotes the 'neutral' class, and 2 denotes the 'positive' class in the Figure. Unsampling the minority classes, which are the classes for ‘neutral’ and ‘negative’ here (again), we get a dataset containing 205 negative classes, 205 neutral classes and 45 positive classes.

Finally, on this dataset, applying several machine learning classifiers by similarly training the models with 70 per cent of the data and saving the rest 30 per cent for testing, we get the best accuracy score of 0.9781 from the Gaussian Naïve Bayes Classifier. Hence, we can also concludethat the distribution we initially got is almost correct. So, it means that approximately 48 per cent of its viewers have a favourable opinion, 37 per cent have a neutral opinion and the rest 15 per cent have a negative opinion about the video. Figure 12 shows the same diagrammatically.

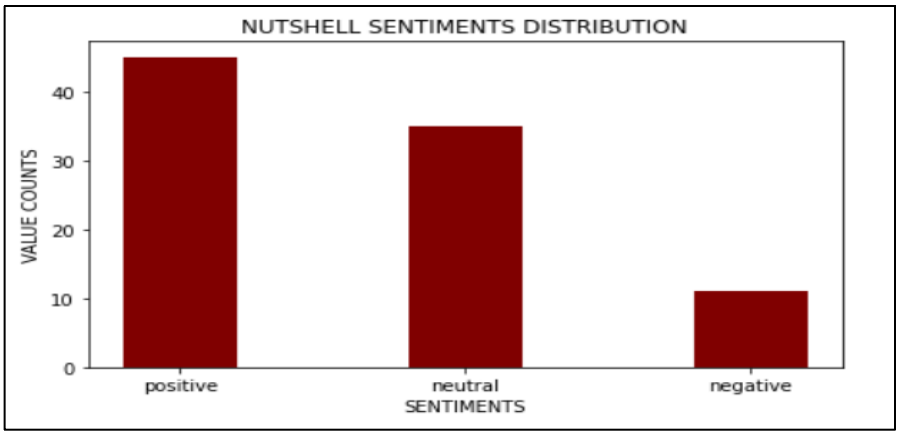


Fig. 12: Video 2 Sentiments Distribution

V. CONCLUSION

After performing these two case studies, we can say that most of the comments were on the positive side of both videos. Moreover, the respective viewers liked the contents of the two videos and want more of the same kind in future. The ratio of negative comments was meagre, i.e., less than 20 per cent for both videos, which is an ideal scenario for the content creators. These creators or YouTubers want the least number of negative comments, ideally. Hence, through this project, we have concluded that by performing sentiment analysis on the comments of one's videos, they can understand their viewers' opinions as a gist. And then further, if they want, they can go through the positive, negative and the neutral opinions separately according to their purpose. When they would like to thank the viewers for their appreciation, they can reply the positive comments. They can go through the negative comments when they want to improve their content. Furthermore, when they want suggestions for other content creation, they can go through the neutral comments, which generally contain suggestions from the viewers.

REFERENCES

[1.] Alexandre Ashade Lassance Cunha(B), Melissa Carvalho Costa, and Marco Aur'elio C. Pacheco "Sentiment Analysis of YouTube Video Comments Using Deep Neural Networks" the 24th of May 2019.

[2.] Hanif Bhuiyan; Jinat Ara, Rajon Bardhan, Md. Rashedul Islam "Retrieving YouTube video by sentiment analysis on user comment" the 14th of September 2017.

[3.] Martin Wöllmer,Felix Weninger,Tobias Knaup,Björn Schuller,Congkai Sun “YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context" the 27th of March 2013.

[4.] Muhammad Zubair Asghar, Shakeel Ahmad, Afsana Marwat, Fazal Masud Kundi “Sentiment Analysis on YouTube: A Brief Survey”the 30th of November 2015.

[5.] Olga Uryupina, Barbara Plank, Aliaksei Severyn, Agata Rotondi, Alessandro Moschitti “SenTube: A Corpus for Sentiment Analysis on YouTube Social Media” May 2014.

[6.] Rawan Fahad Alhujaili, "Sentiment Analysis for Youtube Videos with user comments" the 12th of April 2021.

[7.] Abbi Nizar Muhammad,Saiful Bukhori,Priza Pandunata "Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes" the 5th of December 2019.

[8.] Annamaria Porreca, Francesca Scozzari&Marta Di Nicola "Using text mining and sentiment analysis to analyse YouTube Italian videos concerning vaccination" the 19th of February 2020.

[9.] S. Nawaz, M. Rizwan and M. Rafiq "Recommendation Of Effectiveness Of YouTube Video Contents By Qualitative Sentiment Analysis Of Its Comments And Replies" December 2019.

[10.] Shanta RangaswamyShubham Ghosh,Srishti Jha,Soodamani Ramalingam"Metadata extraction and classification of YouTube videos using sentiment analysis" the 16th of January 2017.

[11.] Lakshmish Kaushik, Abhijeet Sangwan, John H. L. Hansen"Metadata extraction and classification of YouTube videos using sentiment analysis" the 9th of January 2014.

[12.] Amar Krishna, Joseph Zambreno, Sandeep Krishnan"Polarity Trend Analysis of Public Sentiment on YouTube" the 1st of January 2014.

[13.] Philipp A Toussaint Maximilian Renner,Sebas, Scott Thiebes Ali Sunyaev"Direct-to-Consumer Genetic Testing on Social Media: Topic Modeling and Sentiment Analysis of YouTube Users' Comments" the 15th of September 2022.

[14.] Rahul Pradhan "Extracting Sentiments from YouTube Comments" the 10th of February 2022.

[15.] F Peng, A McCallum "Accurate Information Extraction from Research Papers using Conditional Random Fields" January 2006.

[16.] Mike Thelwali "Social media analytics for YouTube comments: potential and limitations" September 2017.

[17.] Rhitabrat Pokharel Dixit Bhatta "Classifying YouTube Comments Based on Sentiment and Type of Sentence" the 31st of October 2021.

[18.] M. Viny Christanti, Walda1, Tri Sutrisno, "Comments Scraping Application For Review Youtube Content" 2019.

[19.] Ritika Singh "Youtube comment analysis" May 2021.

[20.] Sudhansu Ranjan, Dheeraj Mekala, Jingbo, Shang "Progressive Sentiment Analysis for Code-Switched Text Data" the 25th of October 2022.

[21.] Arpit Khare, Amisha Gangwar, Sudhakar Singh, Shiv Prakash "Sentiment Analysis and Sarcasm Detection of Indian General Election Tweets" the 3rd of January 2022.

[22.] Rhitabrat Pokharel Dixit Bhatta "Classifying YouTube Comments Based on Sentiment and Type of Sentence" the 31st of October 2021.

[23.] Tanvi Mehta, Ganesh Deshmukh "YouTube Ad View Sentiment Analysis using Deep Learning and Machine Learning" the 11th of May 2022.

[24.] https://www.youtube.com/watch?v=lPvXZz7m9sI&t=2s

[25.] https://www.youtube.com/watch?v=uEawmeO2gOY&t=7s