Measurement and Sentiment Analysis of YouTube Video Comments

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xinyu Sui

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Haiyang Wang

November 2022

Acknowledgements

First, I would like to thank my advisor, Dr. Haiyang Wang for helping me determine the topic of my thesis, guiding me throughout the research process and helping me through my whole graduate study. Next, I would like to thank Dr. Yongcheng Qi and Dr. Eleazar Leal. I learned the statistical methods from Dr. Qi's course and machine learning and database from Dr. Leal's course, which built the foundation of my research. Thanks also to my committee members for your tremendous effort in reviewing my thesis, making valuable suggestions, and participating in my thesis defense. Finally, I would like to thank the CS department, the Mathematics & Statistics department, the faculties, stuff and classmates, you have really helped me learn and improve a lot both in academic and my life.

Dedication

To my family, thank you for your love, support and company along my journey, especially to my mother, who always supports all my decisions without any hesitation. To Sun, thank you for your love, tolerance and accompanying me no matter how far we are. To my friends, you are treasure in my life, thank you for giving me laugh and support even during my hardest time. Finally I dedicate it to my hometown, where I was born and grew up, I miss my hometown and my family all the time.

Abstract

According to the latest video consumption statistics in 2022, 92.7 percent of global Internet users worldwide visits online video-sharing platforms, such as YouTube and TikTok, every week. These users share their videos and exchange image/text comments to establish crucial social network interactions. Based on the existing research, users' likes and comments are evidence commonly used to quantify the popularity of videos and social media creators. However, it remains largely unclear if the sentiment of comments, e.g., negative comments, will also affect the video or video creators' popularity.

In this thesis, we take initial steps to explore YouTube video comments via sentiment analysis. We present an in-depth measurement study of commenting and user's comment behaviors on a sample of more than 7 million comments on 4 million YouTube videos. Our measurement indicates that Music and Gaming videos attract more feedback and are more likely to be affected by the sentiment of comments. To better understand this, we utilize three popular machine learning models and two deep learning models to analyze the sentiment of video comments. Unlike Twitter and Facebook-based research, our study proves that negative comments do not significantly impact the popularity of YouTube videos. This means the online video-sharing platforms are more robust against unhealthy comments or rumors.

Contents

List of Tables

List of Figures

1 Introduction

The rise of social networking websites has brought about a lot of famous social media companies, such as Meta, Instagram, YouTube and LinkedIn. Many social scenes in daily life have been transferred to the online world, where people can make friends, listen to music, play games, and apply for jobs. Social networking websites have reshaped people's daily lives, redefined popular trends, and changed the way people surf the Internet. For example, according to data statistics [1], the number of YouTube users has reached 1.3 billion. In the past two years, due to the outbreak of the COVID-19, people were forced to live in isolation at home. Eight out of every 10 adults used YouTube. There is no doubt that YouTube is the No.1 in the video field.

With the development of information interaction forms, it has become a very common behavior to leave messages in the comment area of social media. YouTube users express their emotions or views through comments, interact with video creators or other users, and as a result, create a sense of connection with others. For video creators, the comment area is an important source of inspiration. They can mine interesting UGC (user generated content) in the comment area and create more popular works. For social media platforms, the comment function does not only meet the needs of users but also increases their sense of participation, which is helpful to increase the users' stickiness. However, at the same time, some studies showed that if the comment area is filled with negative comments, other users, especially video creators, will feel excluded, which could lead to some creators delete user comments,

[1] https://fortunelords.com/youtube-statistics/

close the comment area or even leave the platform. Although diversity and openness may be the original intention and charm of the Internet platform, it seems that setting up comment barriers and clearing the comment area have become the consensus of social media.

The purpose of this research project is to discuss whether too many negative comments will affect the popularity of the video and thus affect the income of the video creators. To be able to answer this question, we used the YouTube API and crawled three datasets, namely 100 YouTubers, 400,000 videos, and 700,000 comments. Also, we conducted statistical analysis on the YouTuber dataset and video dataset. After that, we utilized three popular machine learning models, namely SVM, NB, RF, and two deep learning models, LSTM and CNN to analyze the sentiment of comments. All these five models performed well and got over 80% accuracy. The correlation coefficient of positive ratio and like ratio was less than 0.4, indicating a weak correlation relationship. Moreover, the correlation coefficient between the positive ratio and the number of views was less than 0.2. We can conclude that negative comments did not affect the popularity of the videos. When we calculated it using video categories, we found that different video categories presented different characteristics. The correlation for Entertainment was 0.4, while the correlation for Science & Technology was only 0.2. We assumed that Science & Technology comments are more neutral and objective, and that's why sentiment analysis cannot effectively mine user opinions.

This research paper includes five chapters. Chapter 1 is Introduction. Chapter 2 briefly introduces the YouTube website, the statistics theory, the programming library, and the algorithms. Chapter 3 focuses on how we collect the YouTube data includes channel list, video data and video comments, and the measurement of observations. Chapter 4 presents our analysis results for the user comments. Finally, Chapter 5 discusses the specific conclusions of our research.

2 Background

2.1 Online Video Sharing

The progress of video compression technology makes the application of online video hosting and video streaming possible. The world's first online video hosting platform is ShareYourWorld, which was founded in 1997 and allows users to upload videos in different formats. However, the bandwidth technology was limited at that time, and the transcoding technology of video was not yet mature. When the growth of user scale brought more video content, ShareYourWorld finally closed its service in 2001 because of serious budget problems. After that, Pandora TV, an online video hosting platform established in South Korea, officially began its service in October 2004. Although it also provides video hosting services for users, Pandora TV provides users with unlimited storage space for uploading videos. And it is the first online video platform in the world to add advertisements to videos, so as to ensure its continuous operation.

Founded in 2005, YouTube is the first video streaming platform. Users can upload their videos in various formats to YouTube platform. Based on YouTube's own video transcoding technology, these user generated content can be streamed anywhere on the Internet. By 2006, YouTube had accounted for more than half of the online video market share, far surpassing its competitors in the same period. Such excellent performance contributed to Google's sky high purchase of YouTube in 2006. Then YouTube found a perfect advertising form to achieve a balance between user experi-

ence and advertising effect. After years of exploration, 5 seconds can skip the plug-in advertisement, which can appear in the beginning, middle, or end of the video, providing the possibility to increase the density of advertisements. Users can decide whether to finish watching the advertisement or skip it according to their preferences. Five seconds is a very short time, which has little impact on users' movie viewing experience. This kind of advertisement benefits advertisers, users and platforms. YouTube's revenue is close to $20 billion by 2020, accounting for about 13% of Google's total advertising revenue. However, YouTube's growth rate was faster than that of Google's other major advertising sources, with advertising revenue of an increase of 49% year-on-year.

2.1.1 The Development of YouTube

The development of Web 2.0 technology gave birth to the new generation of social networking websites. The user is no longer just a information receiver, but a producer of website content, and can communicate and interact with other users. Therefore, social networking sites are also called user generated content (UGC) sites. For example, Facebook, Twitter and YouTube are well known.

YouTube was founded in 2005. After 16 years of development, YouTube has already become a part of people's daily life. According to the statistics of Internet traffic[1] , YouTube has become the second most visited website after Google, ahead of Facebook, Wikipedia and Twitter. YouTube has more than 30 million visitors every day. On average, about 300 hours of videos are uploaded to the YouTube website every minute, and more than 5 billion videos are viewed by users every day. Especially during the COVID-19, more and more people choose to become full-time or part-time YouTubers to make money by making videos. More excellent video content

[1]https://www.semrush.com/website/top/

will continue to improve the user experience and enhance user stickiness. In the video field, YouTube stands out as a leader in the Internet video platform.

YouTube has designed many functions to facilitate users to interact with others. On the one hand, users can directly comment on the video to express their views. On the other hand, users can also subscribe to the channel of YouTuber or send private messages to YouTuber. The website also provides sharing, rating and favoriting functions. For YouTuber, the website allows content providers to easily upload videos they have made, and provides the tag function to extract keywords for easy retrieval by users. The above functions can record user behaviors, constitute the analysis data source of this article, and can directly crawl the metadata provided by the website through the YouTube API. The attributes of the video itself, such as publish time, video length, video category and country, can also be obtained directly through the YouTube API, which is convenient for us to compare different short video websites.

2.2 Statistical Analysis Methods

2.2.1 Pearson Correlation

Pearson correlation coefficient is usually used to measure the linear correlation between variables (Benesty et al. 2009), and its formula is as follows:

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

where numerator refers to the covariance of variable X and variable Y, which reflects the correlation of two random variables. Specifically, if a variable increases or decreases with another variable, the covariance of the two variables is positive, and vice versa.

According to the formula, the Pearson correlation coefficient is obtained by dividing the covariance by the standard deviation of the two variables. Although the covariance can reflect the correlation of the two random variables, when the covariance is greater than 0, it means the two are positively correlated, and when the covariance is less than 0, it means the two are negatively correlated, the covariance value can not well measure the correlation of the two random variables. Pearson correlation coefficient divides the standard deviation of two random variables on the basis of covariance, thus limiting Pearson's value between -1 and 1. Specifically, when the linear relationship between the two variables is enhanced, the correlation coefficient tends to 1 or -1. When one variable increases and the other variable also increases, it indicates that there is a positive correlation between them, and the correlation coefficient is greater than 0. If one variable increases while the other decreases, it indicates that there is a negative correlation between them, and the correlation coefficient is less than 0. If the correlation coefficient is equal to 0, there is no linear correlation between them.

2.2.2 Standard Deviation

The standard deviation is the square root of the arithmetic mean of the square of the difference between each individual in the sample and its mean, which reflects the dispersion of a dataset. The greater the value, the more discrete, that is, the greater the difference between individuals (Wan et al. 2014). The formula is as follows:

$$std = \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

where μ represents the mean value of the data, and:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

2.2.3 Mean Squared Error

Mean squared error reflects the relationship between the estimated value and the real value of the data, and can be used to measure the average squared difference between the estimated value and the real value. Mean squared error is the average of the sum of squares of the difference between the data and the actual value, that is, the average of the sum of squares of errors (Wang and Bovik 2009). It is formulated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where Y_i is the actual value, and $\hat{Y}_i$ is the predicted value. When the predicted value is completely consistent with the real value, it is equal to 0, that is, the perfect model; the greater the MSE value, the greater the error, that is, the bad model.

2.3 Programming Library

We used Scikit-learn library for implementing machine learning algorithms and Keras library for implementing deep learning algorithms. Scikit-learn is an open source Python library, we can access its interface to implement all kinds of machine learning algorithms such as Support Vector Machine, Random Forest, and Naive Bayes, etc; it also provides some data preprocessing methods, evaluation methods for machine learning (Bisong 2019). Keras is an open source Python library which

uses TensorFlow as its backend, we can access its interface to build the deep learning networks such as Recurrent Neural Network, Long short-term memory, and Convolutional Neural Network, etc (Gulli and Pal 2017). We also used the NumPy and Pandas library for data processing and Matplotlib library for data visualization.

2.4 Supervised Machine Learning Algorithms

2.4.1 Support Vector Machine

Sentiment analysis in this paper needs to divide comments into positive comments and negative comments, which belongs to a binary classification problem. Support Vector Machine (SVM) is a binary classification model, which can be constructed by training set, so as to divide the new data into two categories. Moreover, a comment often contains multiple words, so it belongs to high-dimensional problems, and SVM can also solve high-dimensional problems. So SVM model can meet our needs.

Support Vector Machine (SVM) is a supervised classification algorithm. An ordinary SVM is just a straight line in two dimensions, which is used to perfectly divide two linearly separable categories. It is the most perfect straight line among countless separable lines, which is the same distance from the points of the two classes. The SVM architecture is as Figure 2.1. The SVM selects a hyperplane that seperates the two classes with maximum margin (X. Zhang 2017). However, if the two classes can not be separated in the dimension, then you can use the kernel function, it can map the data to an additional dimension, that is to say, we can project the data from lower dimension to higher dimension, then we can separate the classes (Noble 2006).

The SVM model is widely used for sentiment analysis of user comments on social media platforms. Li used the SVM model to mine consumer opinions on the

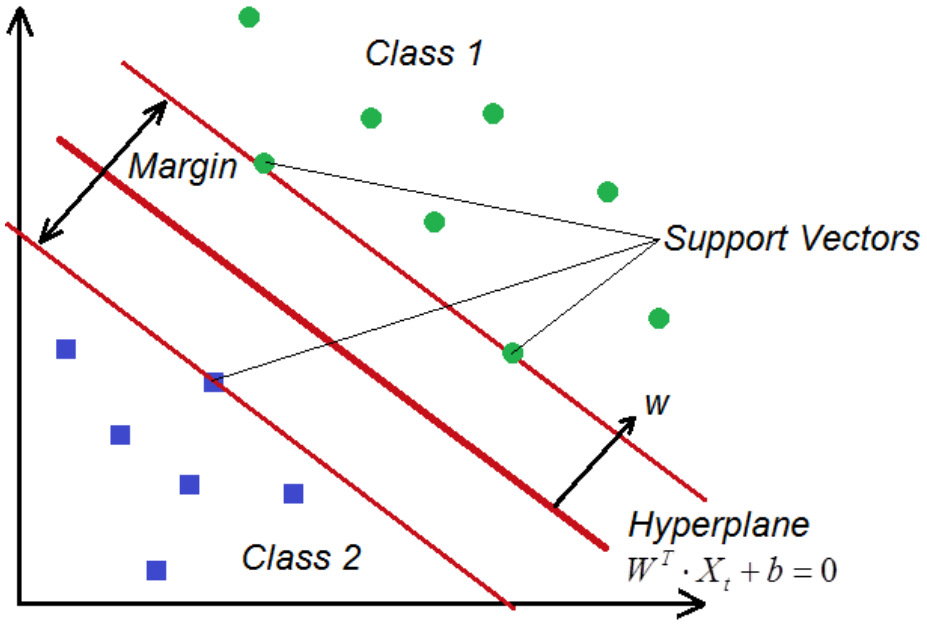

Figure 2.1: Support Vector Machine Architecture(Gorguluarslan and Choi 2013)

Microblogs website, effectively discovering market intelligence to support market decision makers (Y.-M. Li and T.-Y. Li 2013). Eleonora D'Andrea and Pietro Ducange demonstrated that SVM classification results yielded the highest accuracy when monitoring user attitudes towards vaccination using Twitter comments (D'Andrea et al. 2019).

2.4.2 Random Forest

The SVM algorithm is difficult to implement for large-scale training samples, and the calculation will consume a lot of memory and computing time. While Random Forest is easy to implement, have low computational overhead, and show strong performance in many real-world tasks. It can handle very high-dimensional data without feature selection.

Random Forest (RF) is a classification and regression method that integrates many decision trees into a forest and it is used to predict the final result. RF builds bagging

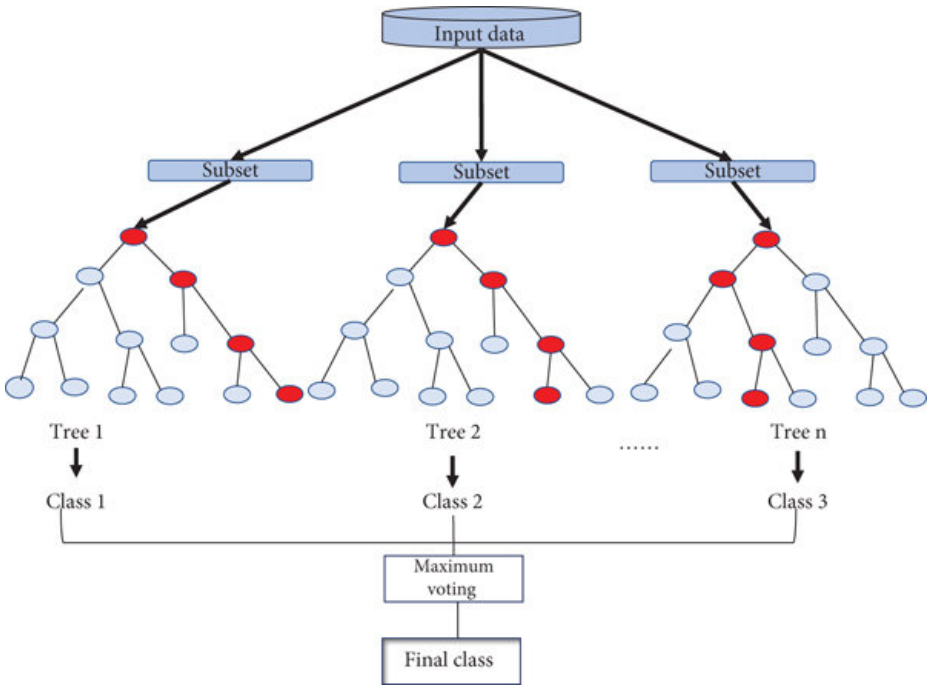

Figure 2.2: Random Forest Architecture(Allehaibi, Y. Khan, and S. a. Khan 2021)

integration with decision tree, but add a random selection of features during the decision tree training process (Breiman 2001). The RF architecture is as Figure 2.2. And RF gets the final results based on all the decision trees result, for example by majority voting or averaging all the decision trees results (Oshiro, Perez, and Baranauskas 2012). Gupte found that there are many experimental results for opinion mining showed good accuracy in different data sets when do the classification problem (Gupte et al. 2014). Karthika and Murugeswari demonstrated that both the RF model and the SVM model can be used to classify user reviews (positive, negative, neutral) on the online shopping website flipkart.com, but the RF model performed better with 97% accuracy (Karthika, Murugeswari, and Manoranjithem 2019).

2.4.3 Naive Bayes

Naive Bayes classifier is another widely used algorithm based on Bayes' theorem, which can classify data (Pang, Lee, and Vaithyanathan 2002). Naive Bayes model combines prior probability and posterior probability, and assumes that features are independent of each other. Feature independence means that the probabilities of different words in the text are not affected by each other, that is, the occurrence of one word will not affect the occurrence of another word. Although this condition is often not true in the real world, Naive Bayes classifier still shows high accuracy. Moreover, the algorithm is very simple, insensitive to missing data, and has good robustness. Kang and Yoo used SVM algorithm and NB algorithm when doing sentiment analysis on restaurant reviews, and the improved NB algorithm, which improved the recall and precision of the performance (Kang, Yoo, and Han 2012). The formula is:

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}$$

where $P(Y|X)$ is posterior probability, $P(Y)$ is prior probability.

2.5 Supervised Deep Learning Algorithms

2.5.1 Long short-term memory

Recurrent Neural Network (RNN) is a kind of neural network used to process sequence data. For example, RNN can solve the meaning of a word having different meanings in different context (Medsker and Jain 2001). Long short-term memory (LSTM) is a special RNN, which is mainly used to solve the gradient vanishing and gradient explosion in the process of long sequence training. That is to say, LSTM can

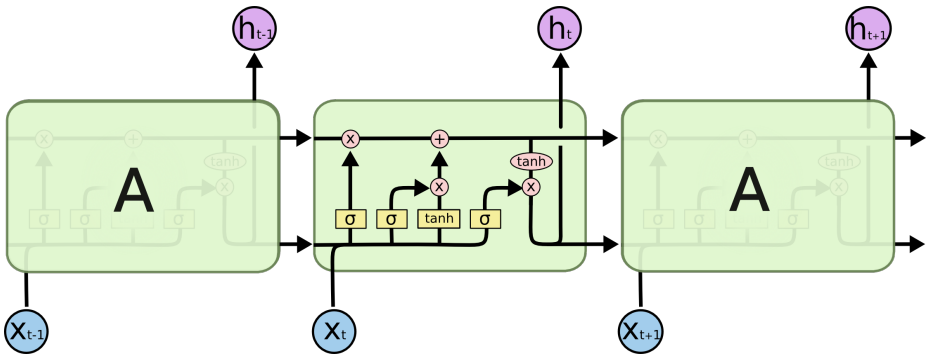

Figure 2.3: LSTM Architecture(Olah 2015)

perform better in longer sequences (Sherstinsky 2020). The LSTM architecture is as Figure 2.3. Basically there are three gates in LSTM, the "forget gate" decides what information should be deleted, that is to say not all information can be used in the next layer; the "memory gate" decides what new information should be added to the cell; the "output gate" decides what value should be outputed (Olah 2015). Many papers have used LSTM for sentiment analysis, for example Li and Qian's paper has conducted several experiments show that the LSTM can produce good accuracy and recall rate on comments sentiment analysis (D. Li and Qian 2016). And Dr.Murthy used the LSTM to do sentiment analysis on raw text and achieved good success on sentiment classification (Murthy et al. 2020).

2.5.2 Convolutional Neural Network

Convolutional Neural Network (CNN) is a kind of neural network. The model structure of CNN is relatively simpler than that of LSTM, and training is more time-saving than LSTM. The structure of CNN can be divided into 3 layers. The CNN basic architecture is as Figure 2.4. The first layer is Convolutional Layer which is mainly used to extract feature. The second layer is Max Pooling Layer which is used to do downsampling. The third layer is Fully Connected Layer which is used to do the

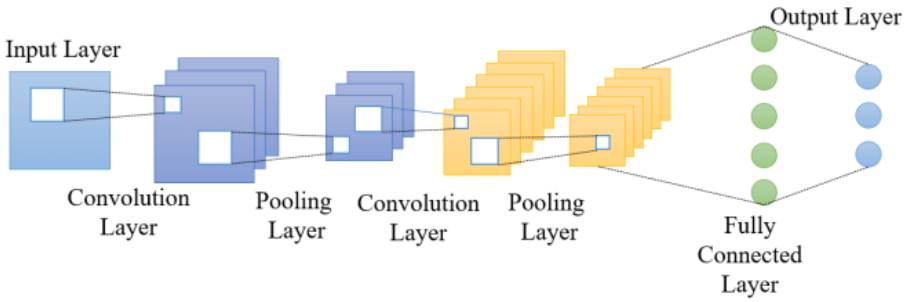

Figure 2.4: CNN Basic Architecture(Gu et al. 2019)

classification (O'Shea and Nash 2015). CNN is mainly used for image classification and voice recognition, while CNN can also work good for sentiment analysis since we can extract text data piece by piece (Liao et al. 2017). CNN is being used to extract opinions by Ishaq, he implemented CNN to do the sentiment analysis and get a better result than the state-of-the-art techniques (Ishaq, S. Asghar, and Gillani 2020).

2.6 Related Work

2.6.1 Comment Related Studies

Compared with the traditional media, social media allows people to easily share opinions, experiences and views with each other. Antonis and Samuel compared the performance of users on traditional news websites and social media platforms, and found that people who used social platforms to access news were more willing to comment or share news (Kalogeropoulos et al. 2017). As the most popular video sharing platform, YouTube has gathered many users. Users can express their views on video content, express their emotions, and communicate with each other, which constitutes a large amount of text data that can be used for sentiment analysis. At present, the research on YouTube comments is mainly divided into three categories: Event Classification, which analyzes video titles and descriptions, and classifies videos

into music, movies, sports, etc; The second is to detect the Sentiment Polarity, that is, to divide user comments into positive comments or negative comments; The third type is Comment Prediction, which predicts the polarity of the video through all the comments and to get a high polarity video (M. Z. Asghar et al. 2015).

Paying attention to the comment area can bring direct benefits to the YouTubers. First, video creators can get more traffic on the platform. According to YouTube Search Engine Optimization (SEO) algorithm, several metrics are used to determine the ranking of videos, including completion rate, like rate, comment rate, click-through rate, audience retention, etc. Therefore, increasing the comment rate helps creators get more traffic on the platform. Second, creators can get creative inspiration in the comment area. User comments are often an intuitive reflection of user needs. When making videos, by checking comments, counting the topics and hot words with many responses, finding out user interests, it helps to enhance the appeal of the video and inspire resonance. Hanif and Jinat's research validates that sentiment analysis of user comments on YouTube can help users retrieve the most relevant and popular videos (Bhuiyan et al. 2017).

It has been confirmed by many studies that comments influence the opinions and judgments of others. For example, investors' perceptions of negative news are influenced by the comments of others, and if the majority of comments are positive, investors will not be extremely pessimistic about bad news (Trinkle, Crossler, and Bélanger 2015). However, the comment area has both advantages and disadvantages. One study shows that negative comments can make video creators feel excluded and even quit social interaction (Lutz and Schneider 2021). This also explains why some YouTubers close video comments.

To be sure, negative comments can frustrate video creators. The more negative comments a video creator receives, the harder it is for them to have fun. Our research

interest is whether the ratio of positive and negative comments and the number of positive and negative comments affect the popularity and ranking of videos. Previous studies have shown that the ratio of positive and negative comments(positive/positive + negative) has nothing to do with the like ratio (like/like + dislike) (Lorentz and Singh 2021). Lisette and Sonja analyzed 355 brand posts from 11 international brands and found that the rate of positive comments on brand posts was positively correlated with the number of likes (De Vries, Gensler, and Leeflang 2012). Our research comprehensively examines the popularity and ranking of videos, including views, likes, comments, and the number of positive comments, like ratio, positive ratio. In addition, we also analyzed different video categories to explore the different performances between them.

2.6.2 Sentiment Analysis

Sentiment analysis is the process of using algorithms to analyse people's emotions, attitudes and opinions expressed in the text. Now we are in the explosive growth of social media era, which has produced a huge amount of text data to express people's emotions and opinions (Liu and L. Zhang 2012). So it is necessary to use computer to do the sentiment analysis which will save a lot of time.

There are mainly three methods for the sentiment analysis. The first one is lexicon based approach, we can indicate positive, negative or neutral sentiment(or polarity) with a sentiment score to each word in the lexicon, then to calculate the sentence sentiment score (Kiritchenko, Zhu, and Mohammad 2014). The second one is machine learning based approach, we can process the text to extract the feature, then use machine learning algorithms to train our model and predict the new sentence polarity. And there are also some deep learning algorithms have applied into the sentiment

analysis tasks. The third one is the combination of machine learning and lexicon-based methods (Wankhade, Rao, and Kulkarni 2022).

It is very useful to do the sentiment analysis on social media user comments. By doing sentiments analysis on users' comments, we can evaluate the YouTube video quality popularity and relevancy (Bhuiyan et al. 2017). We can also do the sentiment analysis on the particular video to understand people's opinion on some policies, like Italian people's opinion change about vaccination policy (Porreca, Scozzari, and Di Nicola 2020), etc.

3 Data Collection and Measurement

3.1 Data Gathering

In order to analyze the statistical properties of YouTube video websites, this paper uses the YouTube API[1] to crawl the relevant data of YouTube videos and channels. First, we get the top 100 YouTubers from each country, such as the United States, China, France, Germany, Japan, India, Russia, etc., and get a total of 1,000 YouTubers, and then sort it according to the number of subscriptions, filter out the top 100, such that get our YouTuber channel table. The second step is to get all the videos published by each YouTuber through the YouTube API, and a total of about 500,000 videos are collected. The third step is to crawl all user comments under the video through each video url. Because the average number of video comments is more than 5,000, we have no way to process the comments of all videos. So when crawling comments, we only randomly selected 2,000 videos with dislike numbers greater than 0 in the United States, and then collect the corresponding comments of these videos. In the end, we collected three data sets in total, namely channel table, video table, and comment table, contains 100 channels, 500,000 videos, 700,000 comments respectively.

[1] https://developers.google.com/youtube/v3/quickstart/python

3.1.1 Crawl the Channel List

The Influencer ranking website provides influencer rankings on internet platforms such as YouTube, TikTok, and Instagram. By directly crawling and analyzing relevant web pages, the channel list can be easily obtained, thereby helping us to directly use the YouTube API to get more needed data. BeautifulSoup, webdriver, requests and re are packages that are often used by scraping web pages. webdriver is used to drive chrome browser, request to get all the information of the web page, BeautifulSoup and re can parse the coding information of the web page, so as to crawl the required data. The python crawler code is shown below.

```
1  #Crawler_Get_ChannelID
2  from bs4 import BeautifulSoup
3  from apiclient.discovery import build
4  from selenium import webdriver
5  import requests
6  import re
7  import csv
8
9  driver = webdriver.Chrome()
10
11 def getChannelID(page_url):
12     driver.get(page_url)
13     html = driver.page_source
14     page_soup = BeautifulSoup(html, 'html.parser')
15
16     links = page_soup.find_all("a", {"class": "link clearfix"})
17
18     data = []
19     for link in links:
20         href = link['href']
21         ID = re.match('/youtube/channel/(.*)', href)
22         #print(ID)
23         if ID != None :
24             data.append(ID.group(1))
25
26     return data
27
28 new_page = 'https://www.noxinfluencer.com/youtube-channel-rank/top-100-
       all-all-youtuber-sorted-by-subs-weekly'
29 channel_ids = getChannelID(new_page)
```

```
{'kind': 'youtube#channelListResponse',
 'etag': 'laHNr-AWSmkkeDhLTtZLFXWJHWs',
 'pageInfo': {'totalResults': 1, 'resultsPerPage': 5},
 'items': [{'kind': 'youtube#channel',
   'etag': 'O7-cAicaigwG6OEnr3COocXgKz4',
   'id': 'UCYLNGLIzMhRTi6ZOLjAPSmw',
   'snippet': {'title': 'Shakira',
    'description': "Welcome to Shakira's Official YouTube Channel. Bienvenidos al Canal Oficial de Shakira en Yout
ube",
    'customUrl': 'shakira',
    'publishedAt': '2005-10-16T09:08:20Z',
    'thumbnails': {'default': {'url': 'https://yt3.ggpht.com/fTXj82s0-YJSxrpnJHDgZezMmqPeioRkAvLbOdCAGvqD_byC_OCd_
fizik25P1NogEkWKH2Rww=s88-c-k-c0x00ffffff-no-nd-rj',
      'width': 88,
      'height': 88},
     'medium': {'url': 'https://yt3.ggpht.com/fTXj82s0-YJSxrpnJHDgZezMmqPeioRkAvLbOdCAGvqD_byC_OCd_fizik25P1NogEkW
KH2Rww=s240-c-k-c0x00ffffff-no-nd-rj',
      'width': 240,
      'height': 240},
     'high': {'url': 'https://yt3.ggpht.com/fTXj82s0-YJSxrpnJHDgZezMmqPeioRkAvLbOdCAGvqD_byC_OCd_fizik25P1NogEkWKH
2Rww=s800-c-k-c0x00ffffff-no-nd-rj',
      'width': 800,
      'height': 800}},
    'localized': {'title': 'Shakira',
     'description': "Welcome to Shakira's Official YouTube Channel. Bienvenidos al Canal Oficial de Shakira en You
tube"}},
   'contentDetails': {'relatedPlaylists': {'likes': '',
     'uploads': 'UUYLNGLIzMhRTi6ZOLjAPSmw'}},
   'statistics': {'viewCount': '23051088389',
    'subscriberCount': '36300000',
    'hiddenSubscriberCount': False,
    'videoCount': '276'}}]}
```


Figure 3.1: Json File of Channels

3.1.2 Collect Channel and Video Information

Once we have the channel id, it is the most efficient way to get channel information directly through the YouTube API. Use the api key to send a request to google, and we will get the json file as shown in Figure 3.1. It contains all channel information officially provided by YouTube. Our channel dataset selects 6 attributes, namely Channel name, Channel id, Subscribers, Views, Total videos and Videolist id. The obtained channel table and video table are shown in Figure 3.2 and Figure 3.3. And the Python script is shown below.

```
#get_channel_data

import requests, sys, time, os, argparse

ID = 'UCYLNGLIzMhRTi6ZOLjAPSmw'
request_url = f"https://www.googleapis.com/youtube/v3/channels?part=
    snippet%2CcontentDetails%2Cstatistics&id={ID}&key={api_key}"
request = requests.get(request_url)
request.json()
```

```
def getChannelData(items):
    channel_data = []
    for data in items:
```

	Youtuber	Channel_id	Subscribers	Views	Total_videos	Videolist_id	category	category_id	country
0	T-Series	UCq-Fj5jknLsUf-MWSy4_brA	218000000	1.934730e+11	17013	UUq-Fj5jknLsUf-MWSy4_brA	Music	10	India
1	SET India	UCpEhnqL0y41EpW2TvWAHD7Q	135000000	1.197710e+11	86100	UUpEhnqL0y41EpW2TvWAHD7Q	Shows	24	India
2	PewDiePie	UC-lHJZR3Gqxm24_Vd_AJ5Yw	111000000	2.837783e+10	4485	UU-lHJZR3Gqxm24_Vd_AJ5Yw	Gaming	24	United States
3	MrBeast	UCX6OQ3DkcsbYNE6H8uQQuVA	96900000	1.600769e+10	722	UUX6OQ3DkcsbYNE6H8uQQuVA	Entertainment	24	United States
4	✿ Kids Diana Show	UCk8GzjMOrta8yxDcKfylJYw	96200000	7.606463e+10	993	UUk8GzjMOrta8yxDcKfylJYw	Entertainment	24	Ukraine
...	...	...	...	...	...	...	...	...	...
95	Troom Troom	UCWwqHwqLSrdWMgp5DZG5Dzg	23100000	9.712553e+09	1945	UUWwqHwqLSrdWMgp5DZG5Dzg	Howto & Style	26	United States
96	Mister Max	UC_8PAD0Qmi6_gpe77S1Atgg	22900000	1.454633e+10	1038	UU_8PAD0Qmi6_gpe77S1Atgg	Entertainment	24	Great-Britain
97	Sesame Street	UCoookXUzPciGrEZEXmh4Jjg	22900000	2.086400e+10	3395	UUoookXUzPciGrEZEXmh4Jjg	Entertainment	24	United States
98	ABP NEWS HINDI	UCmphdqZNmqL72WJ2uyiNw5w	22900000	6.621893e+09	138837	UUmphdqZNmqL72WJ2uyiNw5w	People & Blogs	22	India
99	Sandeep Maheshwari	UCBqFKDipsnzvJdt6UT0lMIg	22900000	1.710219e+09	492	UUBqFKDipsnzvJdt6UT0lMIg	People & Blogs	22	India

100 rows × 10 columns

Figure 3.2: Top 100 Channels Table

```
        if bool(data):
            element = dict(Channel_name = data['snippet']['title'],
                           Channel_id = data['id'],
                           Subscribers = data['statistics']['
                               subscriberCount'],
                           Views = data['statistics']['viewCount'],
                           Total_videos = data['statistics']['videoCount'],
                           Videolist_id = data['contentDetails']['
                               relatedPlaylists']['uploads'])

            channel_data.append(element)

        else:
            continue

    return channel_data
```

Regarding scraping video information, we use similar code, and also crawled through YouTube API. The scraper code is also shown below. For each video, we collected 10 attributes, namely Video name, Video id, PublishedAt, ChannelId, CategoryId, Video length, Views, LikeCount, CommentCount and DislikeCount.

```
#Crawler_Get_VideoInformation
import requests, sys, time, os, argparse
from googleapiclient.discovery import build
import pandas as pd
import seaborn as sns
import numpy as np
```

```

df = pd.read_csv("Video_IDs.csv")
```

```
def getVideoData(items):
    video_data = []
    for data in items:
        if data:
            try:
                dislikeCount = data['statistics']['dislikeCount']
            except:
                dislikeCount = data.get('dislikeCount', 0)

            try:
                likeCount = data['statistics']['likeCount']
            except:
                likeCount = data.get('likeCount', 0)

            try:
                commentCount = data['statistics']['commentCount']
            except:
                commentCount = data.get('commentCount', 0)


            element = dict(Video_name = data['snippet']['title'],
                           Video_id = data['id'],
                           PublishedAt = data['snippet']['publishedAt'],
                           ChannelId = data['snippet']['channelId'],
                           CategoryId = data['snippet']['categoryId'],
                           Video_length = data['contentDetails']['duration'
                               ],
                           Views = data['statistics']['viewCount'],
                           LikeCount = likeCount,
                           CommentCount = commentCount,
                           DislikeCount = dislikeCount)

            video_data.append(element)
    return video_data
```

```
import requests, sys, time, os, argparse

video_data = []
for ID in df:
    #print(ID)
    request_url = f"https://www.googleapis.com/youtube/v3/videos?part=
        statistics, snippet, contentDetails&id={ID}&key={api_key}"
    #print(request_url)
    request = requests.get(request_url)
    page = request.json()
    items = page.get('items', [])
```

	Video_name	Video_id	PublishedAt	ChannelId	CategoryId	Video_length	Views	LikeCount	CommentCount	DislikeCou
0	Saiyaan Ji ▶ Yo Yo Honey Singh, Neha Kakkar|Nu...	VEkj2sanAeU	2021-01-27T07:31:05Z	UCq-Fj5jknLsUf-MWSy4_brA	10	PT3M40S	527333120	6074559	289823	15404
1	Pachtaoge (Female Version) | Nora Fatehi | Ase...	olLLDnTdGFg	2020-08-14T05:31:04Z	UCq-Fj5jknLsUf-MWSy4_brA	10	PT2M55S	36633166	744438	44878	13167
2	Billo Tu Agg Official Video | Singhsta Feat. Y...	Q3ge_F5ei38	2020-08-17T04:29:20Z	UCq-Fj5jknLsUf-MWSy4_brA	10	PT3M52S	62869113	1415176	126697	11377
3	Baby Girl | Guru Randhawa Dhvani Bhanushali | ...	pLhNdJNwGC8	2020-10-01T05:30:12Z	UCq-Fj5jknLsUf-MWSy4_brA	10	PT3M43S	416880726	3725519	210264	10989
4	Dil Ko Maine Di Kasam Video | Amaal M Ft.Ariji...	UsMRgnTcchY	2020-08-10T05:30:49Z	UCq-Fj5jknLsUf-MWSy4_brA	10	PT3M38S	99951238	1614398	278724	10404
...	...	...	...	...	...	...	...	...	...	

Figure 3.3: Videos Information

	videoID	comment
0	tXZ0Qq_Nbc0	Love JLo! Love how she sang this song...diffe...
1	tXZ0Qq_Nbc0	She can really rock out!!!!! I can't get h...
2	tXZ0Qq_Nbc0	Queen of performance and always outdoing herse...
3	tXZ0Qq_Nbc0	I personally think JLo is a brilliant performe...
4	tXZ0Qq_Nbc0	When she speaks I want to cry, she awakes so m...
...	...	...

Figure 3.4: Video Comments

```
11      element = getVideoData(items)
12
13      video_data +=element
```

3.1.3 Crawl Video Comments

Before crawling video comments, we first filter the videos, and select those videos published in the United States with dislikes >0, so that likes/(likes + dislikes) can be calculated. At the same time, we ensure that most of the comments are in English, so as to simplify the word processing process. Using 2,000 randomly selected video ids, we finally get 700,000 video comments. The dataset is shown in Figure 3.4.

3.2 Data Cleaning

As we all know, it is necessary to preprocess the YouTube comments before doing the sentiment analysis. Firstly, according to the language characteristics, in most cases, valuable comments from different users will not be completely the same. If the comments from different users are completely repeated, then these comments are generally meaningless. Obviously, only the earliest of these comments are meaningful, that is, only the first one works. Some comments are very similar, but there are differences in the use of same words. If the comments with similar words are deleted, they will be deleted by mistake. As there is a lot of useful information in similar comments, it is obviously inappropriate to remove such comments. Therefore, we only focus on deleting the completely repeated comments to ensure that the useful comments information is retained. We used the drop_duplicates() methods in pandas, the Python data analysis library, to drop exactly the same comments.

Then, we preprocess the comment sentences crawled from Youtube so that they can be easily learned by various classifiers. Since there are many URL links and various random symbols in user comments on social media, we use regular expressions to remove them. Because we only need the alphanumeric character to do the sentiment analysis, so this will not affect our results. Next because there are a lot of stop words in English usually include adverbs, prepositions and interjections, such as "the", "is" and "and" etc, those words have no sentiment meaning, we need to delete them too. The python code is shown below.

```
1 import nltk
2 import re
3
4 def preprocess(data):
5     nltk.download('stopwords')
6     stop_words = set(stopwords.words('english'))
7
8     data = re.sub(r'((https?://[\S]+))', '', data)
```

```
9       data = re.sub(r'<.*?>', '', data)
10      data = re.sub(r'[^a-zA-Z0-9\s]', '', data)
11      data = data.apply(lambda x:[w for w in x.split() if w not in
            stop_words])
12
13      return data
```

3.3 Measurement Observations

In this subsection, we have made statistical analysis on the YouTuber dataset, the video dataset and comment dataset respectively, so as to have a basic understanding of the YouTube website.

3.3.1 YouTuber Statistics

The YouTuber dataset selects the top 100 channels of the YouTube website based on the number of channel subscriptions, and captures 9 channel-related attributes, namely 'Youtuber', 'subscribers', 'views', 'total videos', 'category', 'start year', "Country", "Category ID", and "Channel ID".

The distribution diagram drawn according to YouTuber's category is shown in Figure 3.5. As can be seen from Figure 3.5, the top 100 Youtubers mainly focus on music and entertainment, accounting for more than half of users' attention. Education ranks third, while sports, News&Politics, Style, Trailers and Science rank last, indicating that YouTube users pay less attention to these topics.

For numerical attributes, such as 'subscribers', 'views', 'total Videos', and 'Start Year', we calculated their mean and variance, as shown in Figure 3.6, and drew their boxplots respectively, see Figure 3.8. According to the principle of boxplot, the purple rectangle part gathers 50% of the data. Through observation, it is found that the distribution of subscribers and views are very similar, However, total videos data is

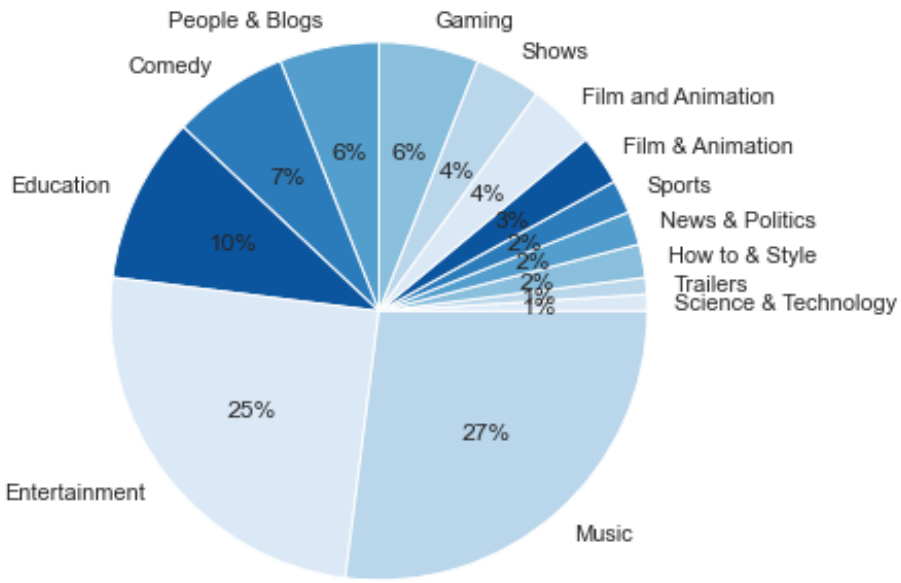

Figure 3.5: Category distribution

	count	mean	std	min	25%	50%	75%	max
Subscribers	100.00	40,560,000.00	26,693,531.92	22,900,000.00	26,075,000.00	31,900,000.00	44,425,000.00	218,000,000.00
Views	100.00	22,472,279,619.52	24,345,975,375.97	1,710,219,330.00	11,570,651,274.00	17,544,202,405.00	23,758,644,375.75	193,473,000,000.00
Total_videos	100.00	16,444.78	42,189.36	46.00	522.00	2,238.50	6,714.75	232,143.00
Start_year	100.00	2,011.33	3.48	2,005.00	2,008.00	2,012.00	2,014.00	2,017.00

Figure 3.6: Numerical attributes analysis

too concentrated near the mean value. The distribution of year is very scattered and has no outliers.

Next, we did pairplot and heatmap to compare the correlation between different attributes. The scatter diagram in Figure 3.9 shows the correlation between each two attributes. Generally, if the scatter points are arranged in a straight line, it indicates that there is obvious correlation between the two attributes. The graph in the second row and the first column has obvious linear distribution characteristics, indicating that there is a strong correlation between views and subscribers. The figures in the third and fourth rows do not show obvious linear relationship, indicating that total videos has a very weak linear correlation with other attributes. The same is true for

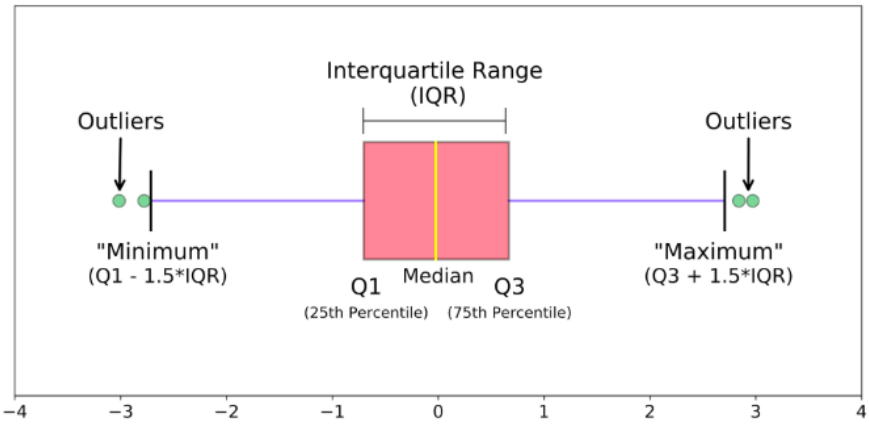

Figure 3.7: Boxplot

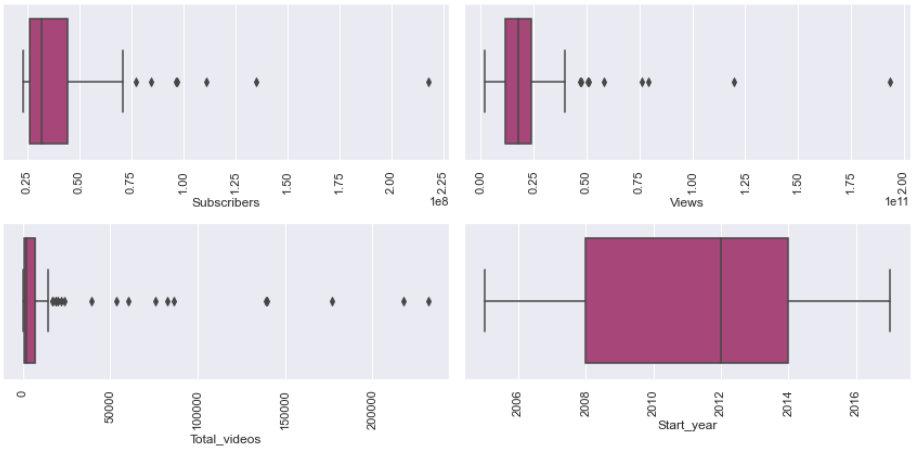

Figure 3.8: Boxplot of numerical attributes

start year.

The heatmap confirms our observations. In Figure 3.10, the closer the color is to red, the stronger the correlation of attributes. It can be seen that there is a high correlation between views and subscribers. After calculation, the Pearson correlation reaches 0.84. while the Pearson value between other variables is close to 0. It can be concluded that there is no linear correlation between them. The results of heatmap are consistent with those of pairplot.

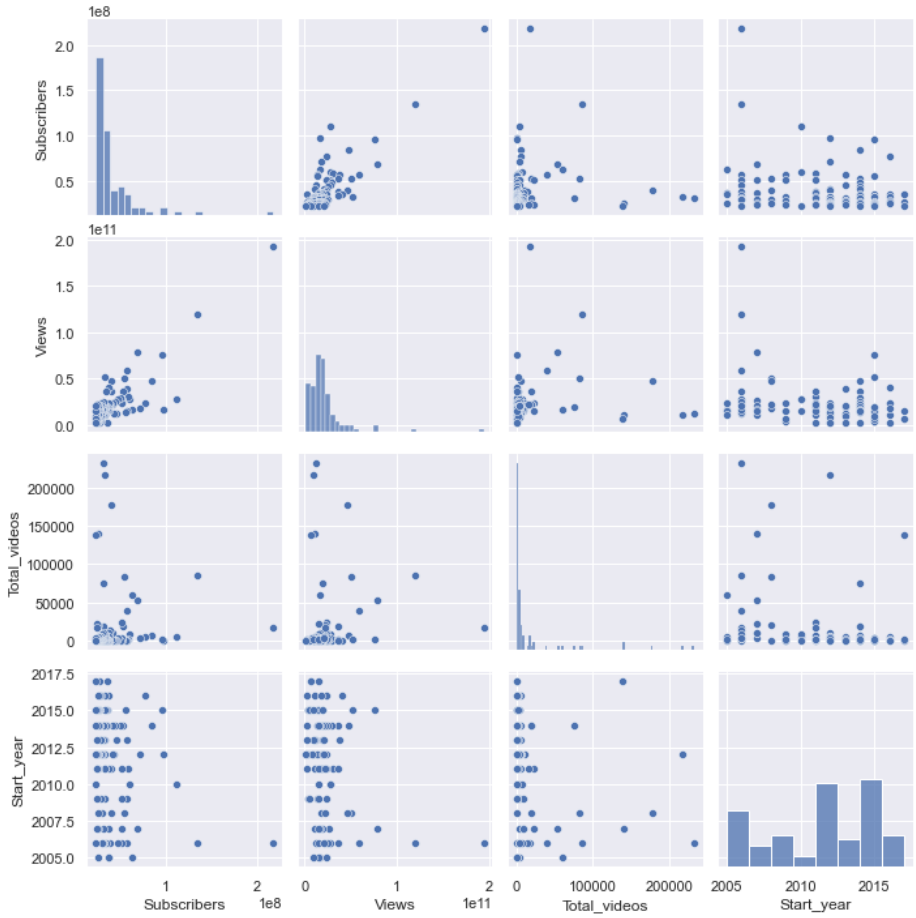

Figure 3.9: Correlations between numerical columns

3.3.2 Video Statistics

For video data, we focused on analyzing its seven attributes, namely category, publish year, video length, view count, likes, dislikes and comment count.

As can be seen from Figure 3.11, about 50% of the views belongs to Music, followed by Film&Animation and Entertainment, Science received the least amount of views. This indicates that the main purpose of YouTube users using the website is for music and entertainment.

Next, we count the publish year of videos and get the distribution as shown in Figure 3.12. It can be seen that although the number of videos released in 2011

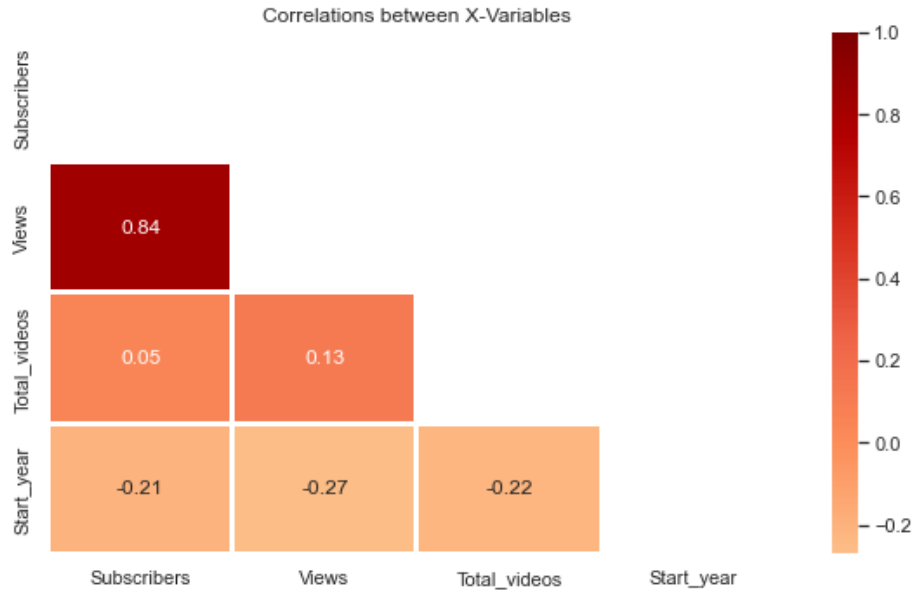

Figure 3.10: Heatmap of correlations

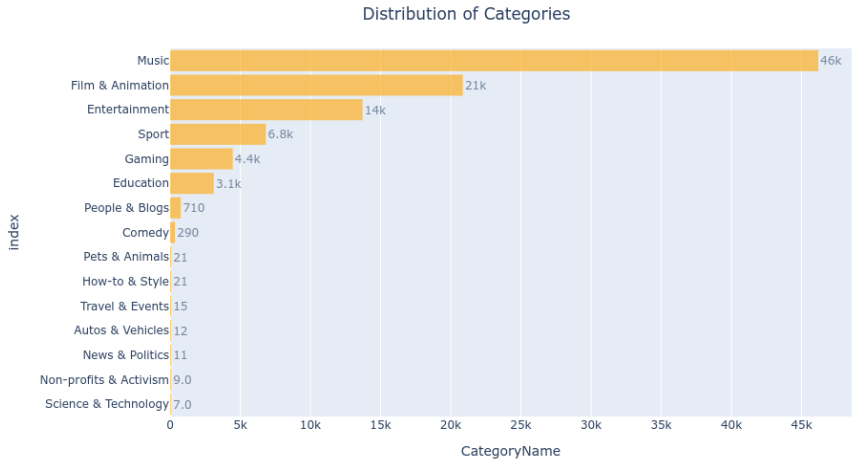

Figure 3.11: Video category distribution

and 2019 increased substantially compared to the previous year, the total number of videos released continued to grow over time. It can be predicted that the popularity of the YouTube website will continue to increase.

For video length, we can intuitively see from the box diagram and probability density diagram in Figure 3.13 that most video lengths are between 200 seconds and

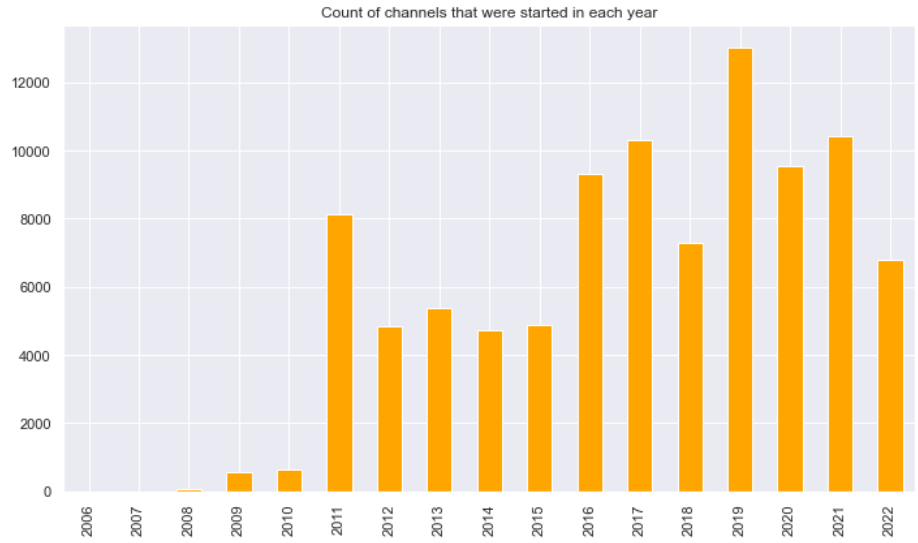

Figure 3.12: Video year distribution

300 seconds. Therefore, YouTube is a short video based platform. Moreover, the video length has an obvious long tail distribution, that is, the duration of a few videos on the website is very long, which can reach more than 1,000 seconds.

Next, we analyze views, like count, dislike count and comment count. As can be seen in Figure 3.14 and Figure 3.15, the four attributes have similar distributions. And Figure 3.16 shows that most of the data are concentrated on the left side of the median value, with a long tail distribution on the right and many outliers on the right.

Figure 3.17 shows the results of Figure 3.16 more intuitively. The diagonal bar chart shows the distribution of each attribute. It can be seen that most videos can only get few views, likes, dislikes and comments, and only a few videos can get more attention.

The calculation results of Pearson correlation are shown in Figure 3.22. There is correlation between each two attributes, indicating the consistency of user behavior. A very interesting finding is that comment and likes have a strong association, and the

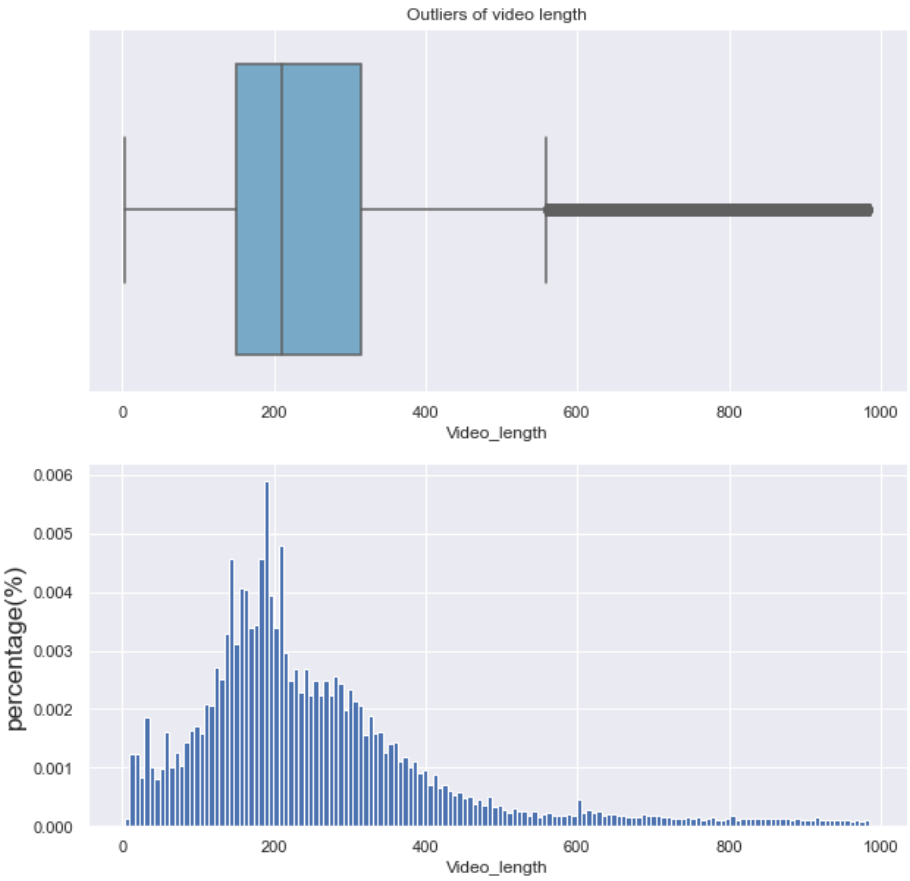

Figure 3.13: Distribution of video length

Pearson value reaches 0.82. Moreover, likes and views also have a strong association, and the correlation coefficient is 0.76. If a video has a high number of likes, its views will be correspondingly high. Because the high number of likes indicates that the video is popular with other users. And more views have a strong impact on platform revenue because more people watch advertisements. Therefore, the YouTube platform will give higher rankings to videos with high likes. Moreover, studies have shown that more positive comments correspond to more likes, and the increase of inferior comments will lead to more dislikes(Schultes, Dorner, and Lehner 2013).

We are interested in whether the YouTube platform has this quality. And in order to reduce malicious negative comments and help videos get more views, YouTube hid

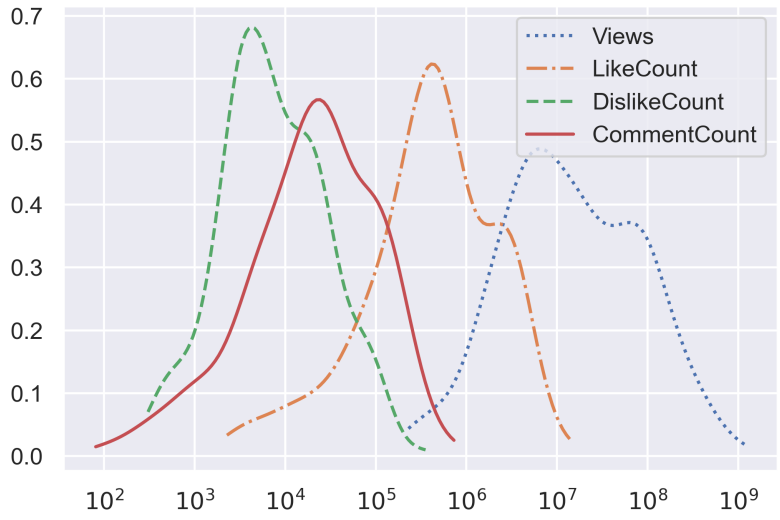

Figure 3.14: PDF

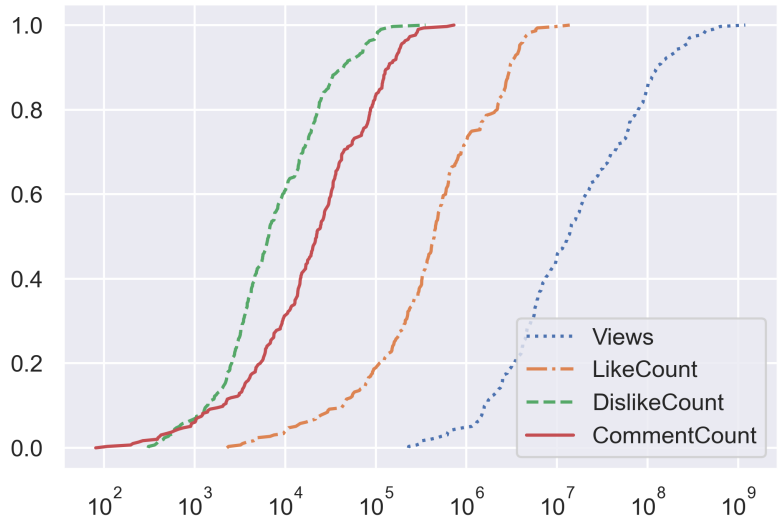

Figure 3.15: CDF

the number of dislikes in 2022. Therefore, if the relationship between the positive and negative of comments and like-dislike can be proved, it can be used to predict the number of dislikes. And users could save their time by taking a look at the comments.

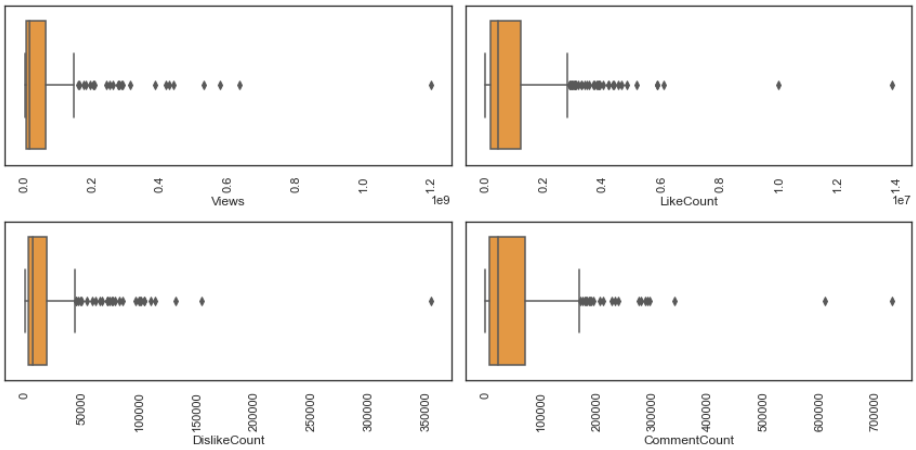

Figure 3.16: Boxplot of Views, likeCount, dislikeCount and commentCount

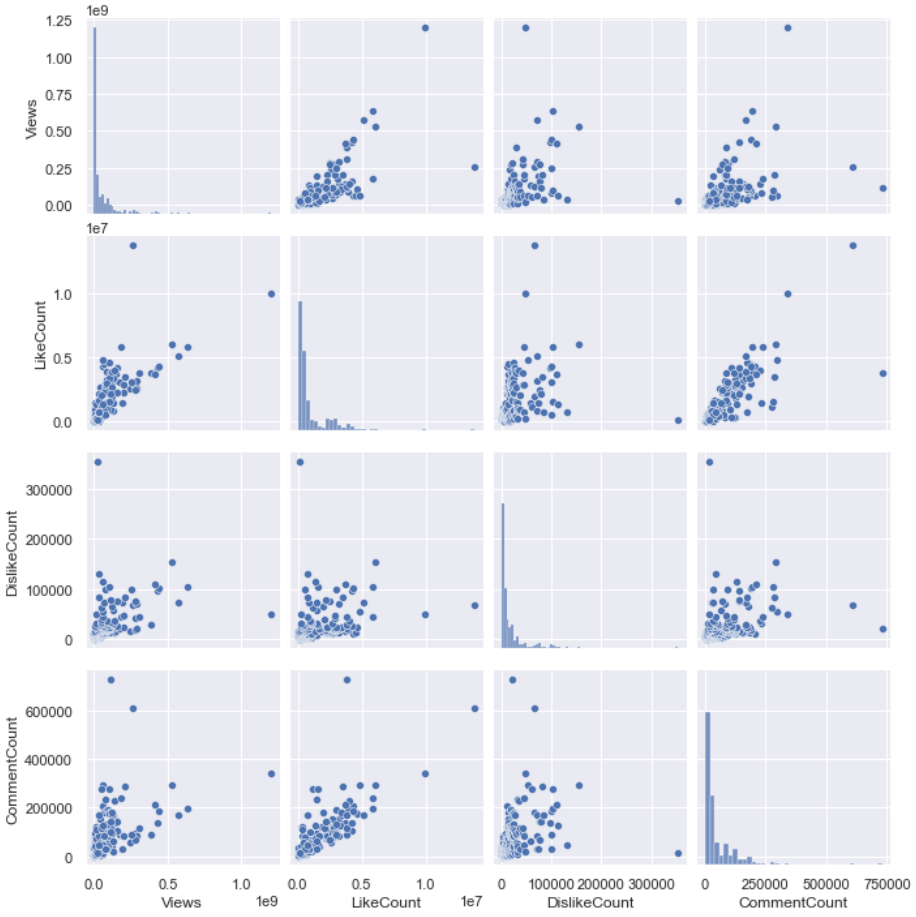

Figure 3.17: Pairplot of Views, likeCount, dislikeCount and commentCount

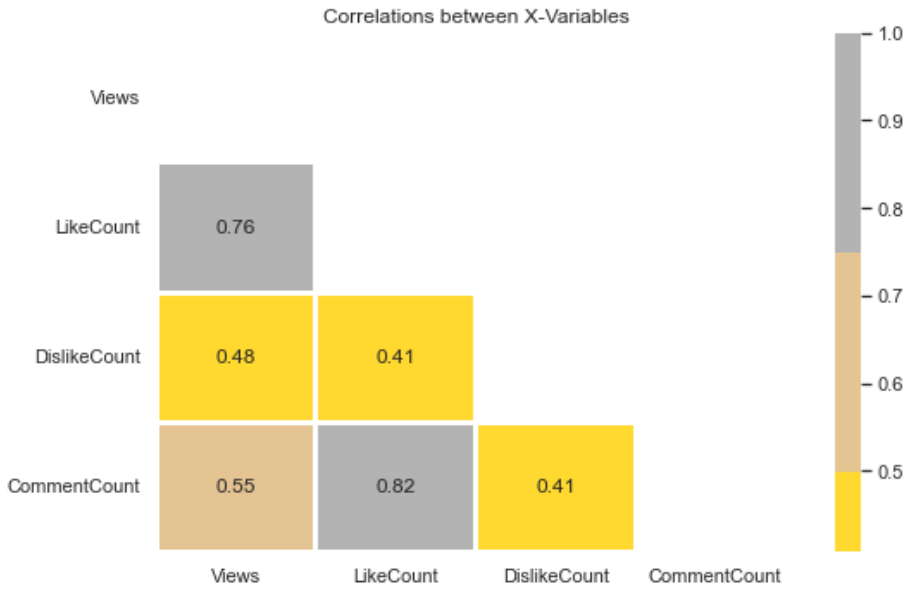

Figure 3.18: Heatmap of correlation

3.3.3 YouTube Comments Statistics

Sorting the comment data by year and category yields the distribution shown in the following figures. As can be seen from Figure 3.19, the number of comments shows an increasing trend year by year, and peaked in 2019, followed by a downward trend. However, after calculating the average of comments, it is found that except for 2018, the average number of comments is high, and the average number of comments in other years shows a steady trend.

When the comments are calculated by video category, it is found that Entertainment has the largest number of comments, followed by Music and Gaming, and these three categories have far more comments than other categories. While the peak of the average number of comments occurs in Comedy, the average number of comments for Music is relatively low.

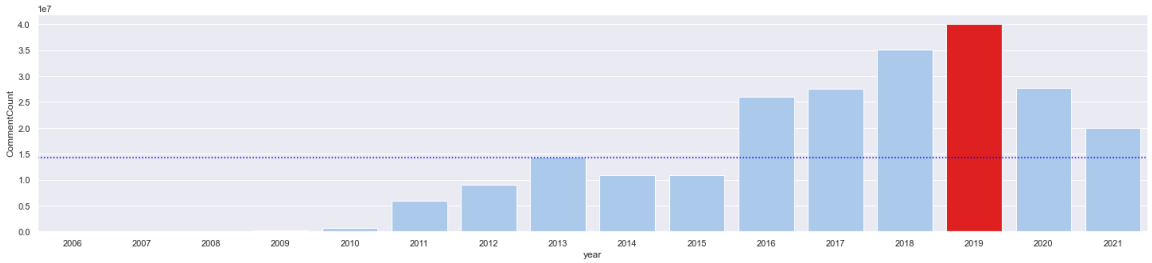

Figure 3.19: Comments by Year

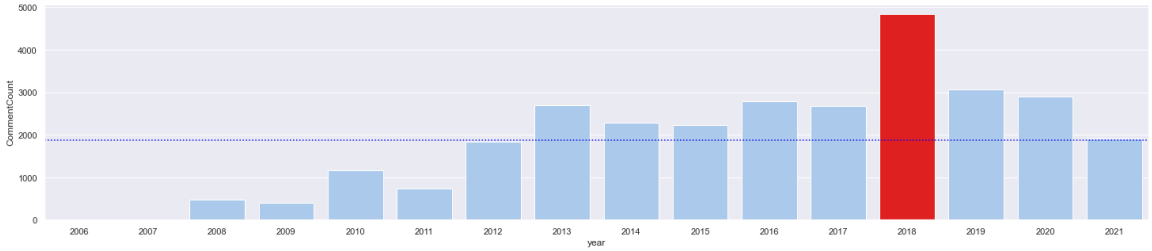

Figure 3.20: Average comments by Year

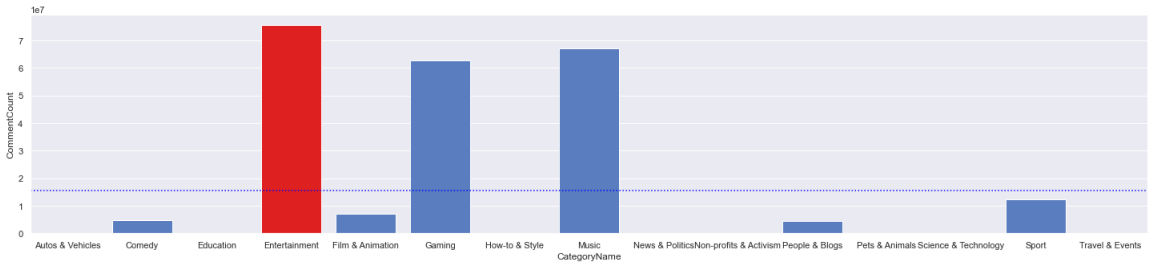

Figure 3.21: Comments by Category

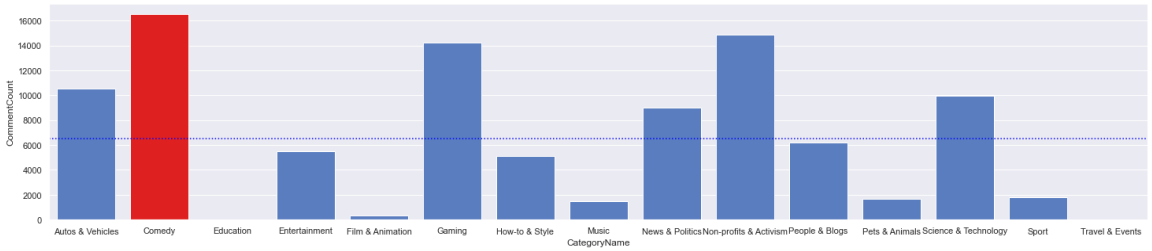

Figure 3.22: Average Comments by Category

4 Sentiment Analysis and Correlations on YouTube Comments

4.1 Models Configuration and Training

Since we're going to analyze YouTube comments, we need to train our classifiers using comments with positive or negative labels. The dataset we used to train our different classifiers are the IMDB dataset and the Reddit dataset. There are 50,000 highly polar movie reviews in the IMDB dataset with positve or negative labels(Maas et al. 2011), and there are 37,000 comments with its sentimental label in Reddit dataset(Gowda et al. 2019). We applied the Scikit-learn machine learning package train_test_split() method to randomly split our data, and we took 80% of the data as the training set and the remaining 20% of the data as the test set.

Before inputed our data to train our model, we needed to do the word embeddings, and we needed to use a corresponding number to replace each word in the sentence and transform each comment sentence into number sequence. We did this using CountVectorizer() from Scikit-learn Python machine learning package or Tokenizer() and pad_sequeces() from Keras Python deep learning package. After that, we used the sequences of numbers to train our models to do the sentiment analysis task.

4.1.1 Machine Learning Classifiers

We implemented the Support Vector Machine, Random Forest and Naive Bayes using Scikit-learn package methods. For the Support Vector Machine, we used LinearSVC method from Scikit-learn package, and set the loss to the square of the hinge loss and let it run 1,000 iterations. For the Random Forest, we set the depth of the tree to 20. For the Naive Bayes, we used the MultinomialNB method from Scikit-learn package and let our models to learn the class prior probabilities. The code is as follows.

```
1  from sklearn.svm import LinearSVC
2  from sklearn.ensemble import RandomForestClassifier
3  from sklearn.naive_bayes import MultinomialNB
4
5  def SVM():
6      LinearSVC(loss='squared_hinge', tol=tol=1e-4, max_iter=1000)
7
8  def RandomForest():
9      RandomForestClassifier(n_estimators=100, max_depth=20, random_state
           =0)
10
11 def NB():
12     MultinomialNB(alpha=1.0, fit_prior=True)
```

4.1.2 Deep Learning Classifiers

For the deep learning classifier, we implemented the LSTM and CNN using Keras as the frontend and TensorFlow as the backend. There were 4 layers of our LSTM model and total 33,148,200 parameters in our LSTM model need to be trained. Also,there were 10 layers of CNN model and total 62,095,200 parameters need to be trained.

```
1  from tensorflow.keras.models import Sequential
2  from tensorflow.keras.layers import Embedding, LSTM, Dense, Dropout
3  from tensorflow.keras.layers import Flatten, Conv1D, MaxPooling1D
4
```

```
dim = 300
def LSTM(totalWords):
    model = Sequential()
    model.add(Embedding(totalWords, dim, input_length = 150))
    model.add(LSTM(128))
    model.add(Dense(32))
    model.add(Dense(1, activation='sigmoid'))
    model.compile(optimizer = 'adam', loss = 'binary_crossentropy',
        metrics = ['accuracy'])

def CNN(totalWords):
    model = Sequential()
    model.add(Embedding(totalWords, dim, input_length=1000))
    model.add(Conv1D(filters=128, kernel_size=4, padding='same',
        activation='relu'))
    model.add(MaxPooling1D(pool_size=2))
    model.add(Conv1D(filters=64, kernel_size=4, padding='same',
        activation='relu'))
    model.add(MaxPooling1D(pool_size=2))
    model.add(Conv1D(filters=32, kernel_size=4, padding='same',
        activation='relu'))
    model.add(MaxPooling1D(pool_size=2))
    model.add(Flatten())
    model.add(Dense(256, activation='relu'))
    model.add(Dense(1, activation='sigmoid'))
    model.compile(loss='binary_crossentropy', optimizer='adam', metrics
        =['accuracy'])
```

The accuracy of different classifiers is as shown in Figure 4.1. The accuracy of CNN is the highest, which is 86.13%, SVM's accuracy is in the second place which is 84.28%, LSTM's accuracy is the third one, which is 83.46%, followed by Naive Bayes which is 81.37%, and the accuracy of Random Forest is in the last place which is 79.97%.

4.2 Results and Correlations

In this section, we used different machine learning models such as: Support Vector Machine, Random Forest, Naive Bayes and different deep learning models such as Long short-term memory, Convolutional Neural Network to do the sentiment analysis on 700,000 comments from 2,000 videos. These 2,000 videos belong to different

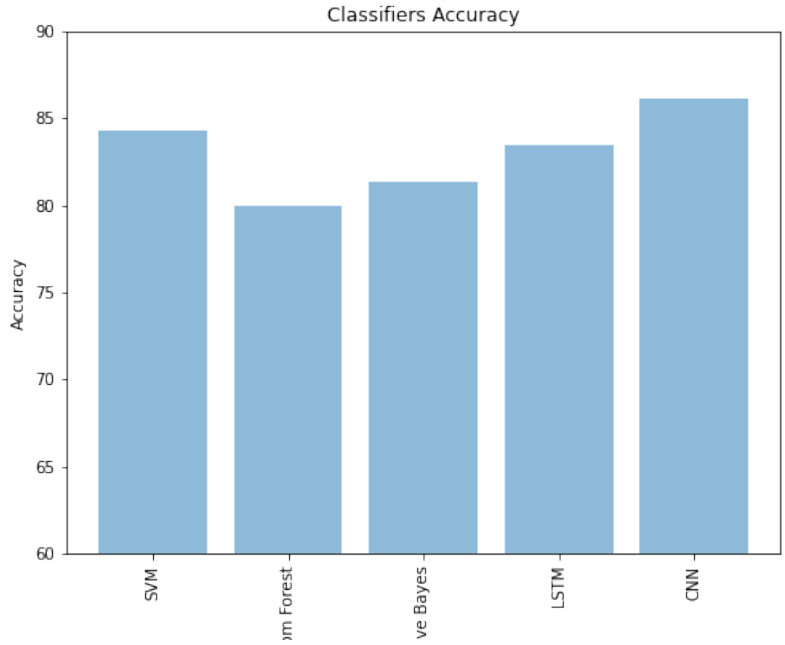

Figure 4.1: The Accuracy of Different Classifiers for Sentiment Analysis

YouTube video categories such as Film & Animation, Entertainment, News & Politics, Science & Technology etc. First we did the sentiment analysis on all the comments, then we chose 3 categories to further analysis the comments. We applied the PR as the abbreviation of positive ratio (PR = positive comments / positive comments + negative comments), and LR refers to like ratio (LR = likes / likes + dislikes), and VN refers to view number, and PN refers to positive number in this section.

4.2.1 Support Vector Machine

We trained our Support Vector Machine model on 50,000 comments, and got an accuracy of 84.28%. Then we used our trained model to do the sentiment analysis of our crawled comments, the results are as Figure 4.2 and Table 4.1. We can see that our predicted positive ratio has some correlation with the actual like ratio since the Pearson correlation value is 0.429, and the MSE is 0.07 which means the predicted

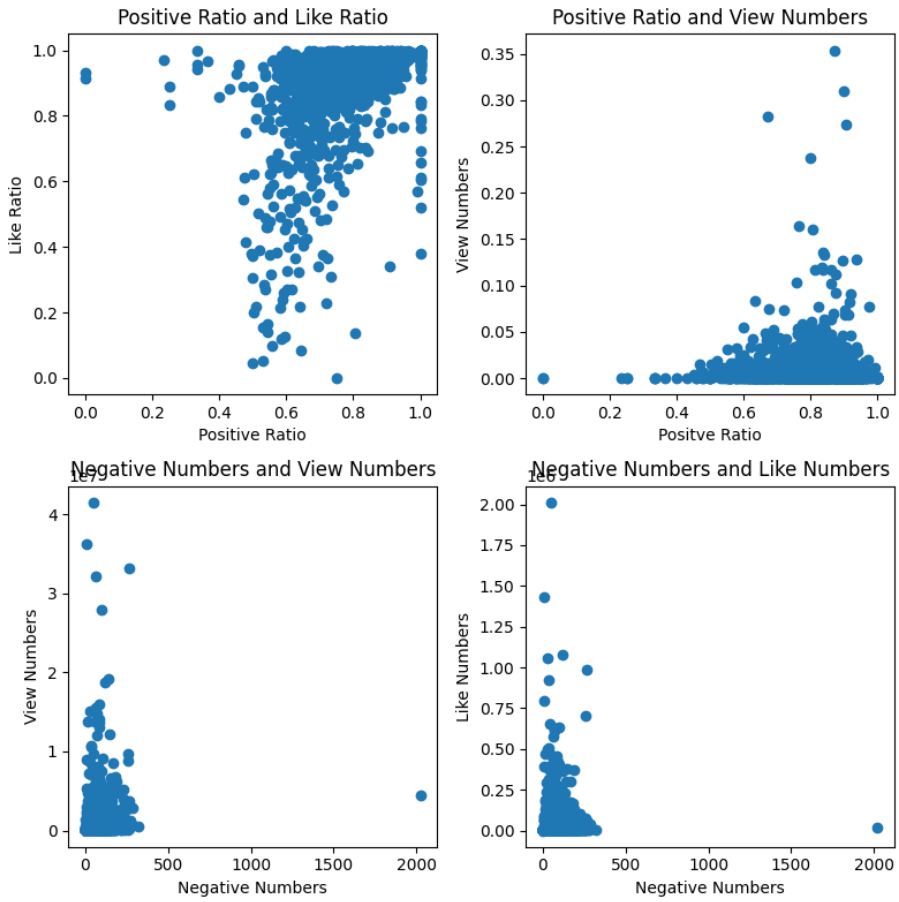

Figure 4.2: Correlation between predicted positive comments with other factors using Support Vector Machines

positive ratio is close to the actual like ratio. That is to say, we can use SVM to predict the actual like ratio from user comments. But we can also get that the predict positive ratio has weak correlation with the view numbers since the absolute Pearson correlation value is -0.004 and the MSE is 0.488. The same as the relationship of the predict positive number, actual view number and actual like number since the Pearson value is less than 0.2, we conclude that there is no linear relationship between them. The Pearson correlation value is the highest using SVM model among all the 5 models.

Category	Pearson correlation	Mean squared error	Standard error
$PR\&LR$	0.429	0.070	0.133
$PR\&VN$	-0.004	0.488	0.123
$NN\&VN$	0.188	NA	NA
$NN\&LN$	0.111	NA	NA

Table 4.1: Predicted positive comments statistics using Support Vector Machines

Category	Pearson correlation	Mean squared error	Standard error
$PR\&LR$	0.314	0.132	0.114
$PR\&VN$	0.003	0.721	0.076
$NN\&VN$	0.173	NA	NA
$NN\&LN$	0.102	NA	NA

Table 4.2: Predicted positive comments statistics using Random Forest

4.2.2 Random Forest

We trained our Random Forest model on 50,000 comments, and get an accuracy of 79.97%. Then we used our trained model to do the sentiment analysis of our crawled comments, the results are as Figure 4.3 and Table 4.2. We can see that our predicted positive ratio has weak correlation with the actual like ratio since the Pearson correlation value is 0.314 and the MSE is 0.132 close to 0, which means the predicted positive ratio is close to the actual like ratio. And we can also get that the predict positive ratio has no linear relationship with the view numbers since the Pearson correlation value is 0.003 and MSE is 0.721. The same as the relationship of the predict positive number, actual view number and actual like number since the Pearson value is less than 0.2. As a result, there is weak correlation between them.

4.2.3 Naive Bayes

The Naive Bayes model got an accuracy of 81.37%. We used our trained model to do the sentiment analysis of our crawled comments. The results are as Figure 4.4

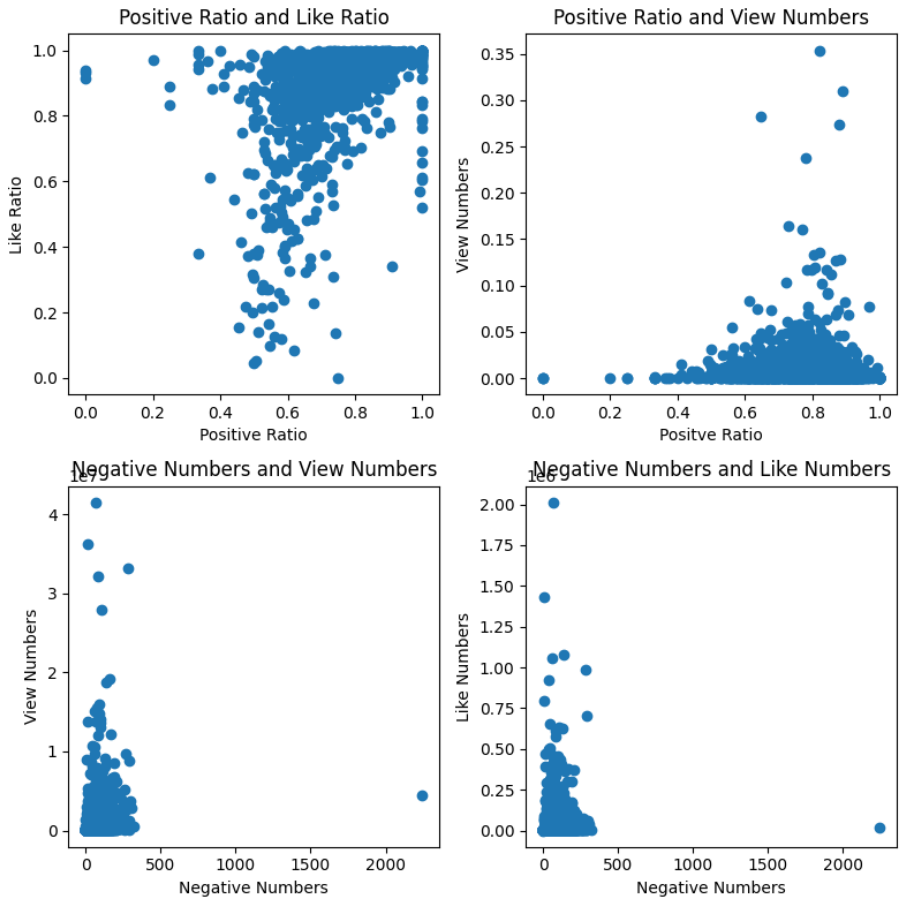

Figure 4.3: Correlation between predicted positive comments with other factors using Random Forest

and Table 4.3. We can see that our predicted positive ratio has some correlation with the actual like ratio since the Pearson correlation value is 0.276 and MSE is 0.206. The Pearson correlation is smaller than SVM's maybe because the accuracy is smaller than SVM's. And we can also get that the predict positive ratio has weak correlation with the view numbers since the Pearson correlation value is -0.006. The same as the relationship of the predict positive number, actual view number and actual like number since the Pearson value is less than 0.2, we can think there is weak linear correlation between them. The Pearson correlation of PR & LR is the lowest the the MSE is the highest using Naive Bayes in the 5 models.

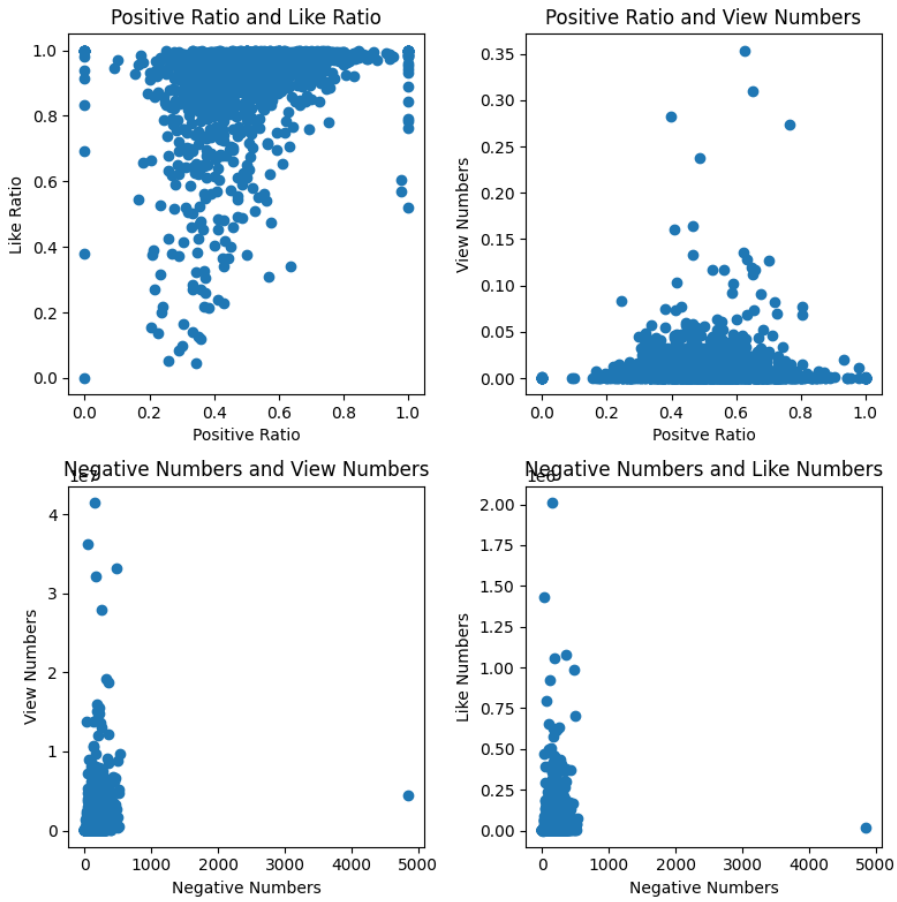

Figure 4.4: Correlation between predicted positive comments with other factors using Naive Bayes

Category	Pearson correlation	Mean squared error	Standard error
$PR\&LR$	0.276	0.206	0.167
$PR\&VN$	-0.006	0.273	0.151
$NN\&VN$	0.158	NA	NA
$NN\&LN$	0.084	NA	NA

Table 4.3: Predicted positive comments statistics using Naive Bayes

4.2.4 Long Short-term Memory

LSTM model was trained on 75,000 comments, and reached an accuracy of 83.46%. Then we used our trained model to do the sentiment analysis of our crawled comments,

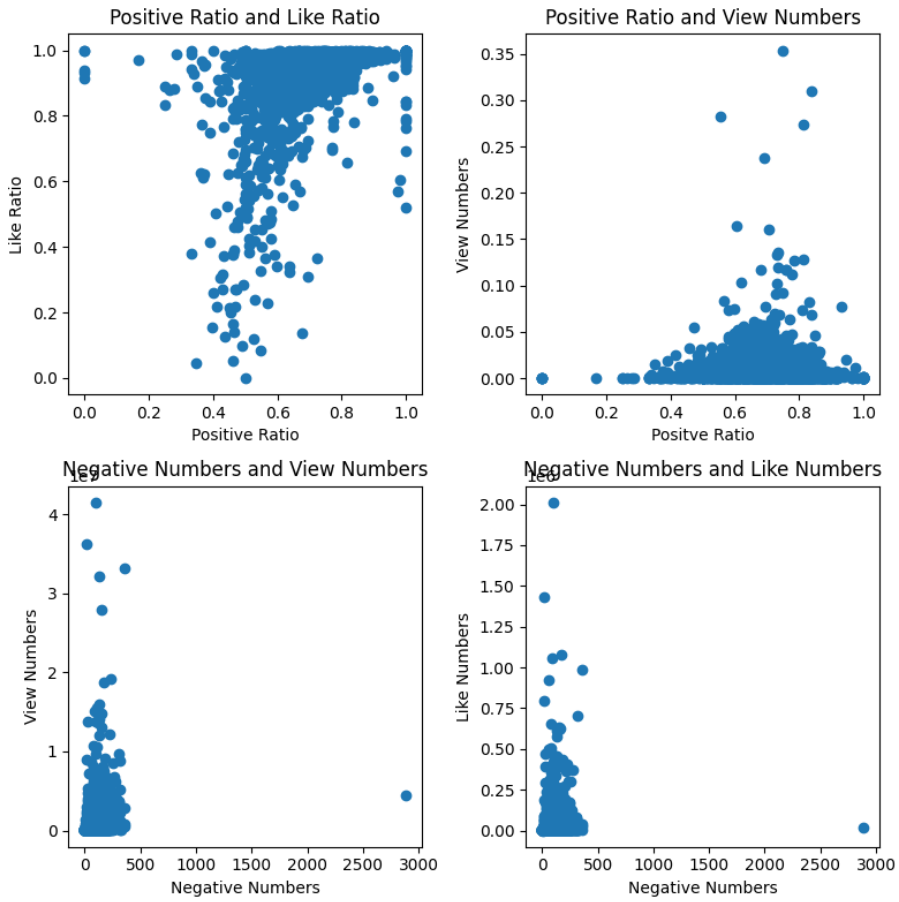

Figure 4.5: Correlation between predicted positive comments with other factors using LSTM

the results are as Figure 4.5 and Table 4.4. We get that our predicted positive ratio has some correlation with the actual like ratio as the Pearson correlation value is 0.355 and MSE is 0.039 which means the predicted positive ratio is close to the actual like ratio. And we can also get that the predict positive ratio has weak correlation with the view numbers since the absolute Pearson correlation value is 0.024. The same as the relationship of the predict positive number, actual view number and actual like number since the Pearson value is less than 0.2, we can think there is no linear relationship between them.

Category	Pearson correlation	Mean squared error	Standard error
$PR\&LR$	0.355	0.039	0.136
$PR\&VN$	0.024	0.227	0.089
$NN\&VN$	0.171	NA	NA
$NN\&LN$	0.105	NA	NA

Table 4.4: Predicted positive comments statistics using LSTM

Category	Pearson correlation	Mean squared error	Standard error
$PR\&LR$	0.372	0.073	0.159
$PR\&VN$	0.049	0.509	0.127
$NN\&VN$	0.183	NA	NA
$NN\&LN$	0.117	NA	NA

Table 4.5: Predicted positive comments statistics using CNN

4.2.5 Convolutional Neural Network

We trained our CNN model on 75,000 comments, and got an accuracy of 86.13%. Then we used our trained model to do the sentiment analysis of our crawled comments, the results are as Figure 4.6 and Table 4.5. We can see that our predicted positive ratio has some correlation with the actual like ratio since the Pearson correlation value is 0.372 and MSE is 0.073 which means the predicted positive ratio is close to the actual like ratio. And we can also get that the predict positive ratio has weak correlation with the view numbers since the absolute Pearson correlation value is 0.049. The same as the relationship of the predict positive number, actual view number and actual like number since the Pearson value is less than 0.2, we can think there is no linear relationship between them.

4.2.6 Video Category

In this part, we further conducted relevant experiments to verify whether positive comments ratio is related to the video like ratio and video view numbers, and

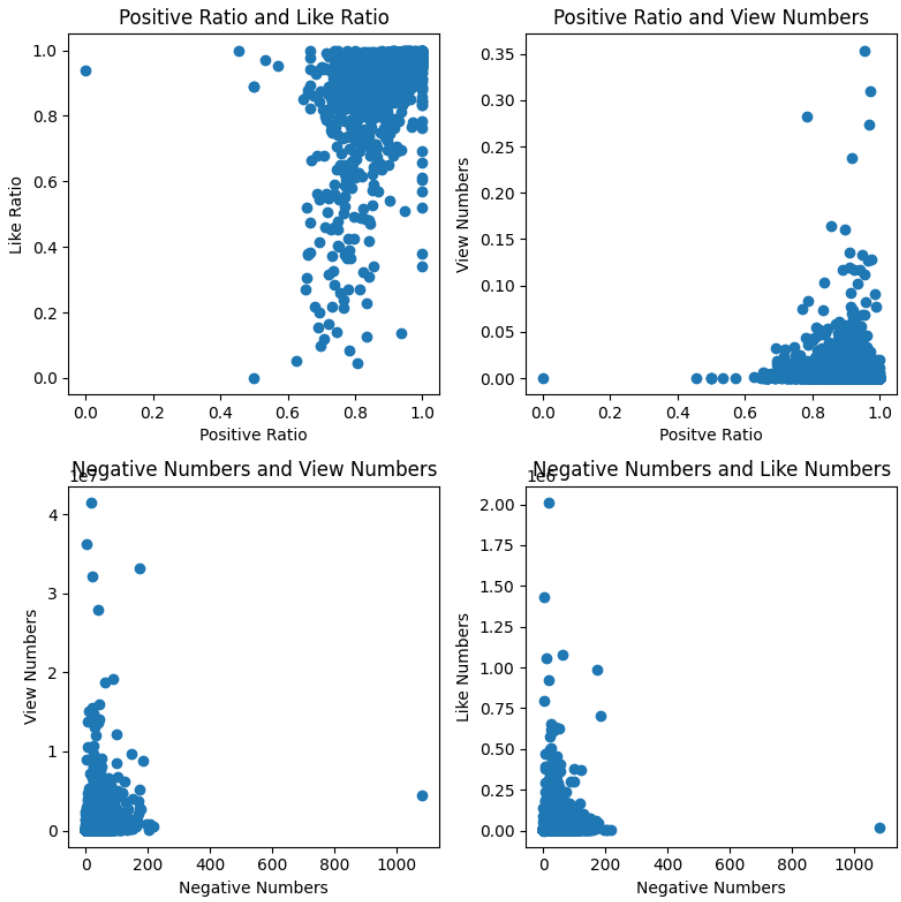

Figure 4.6: Correlation between predicted positive comments with other factors using CNN

whether different video category has different results. So we selected 3 different video categories from YouTube and used the 5 classifiers respectively to check Pearson correlation between positive comment ratio, actual like ratio and view numbers. The experiment results is Table 4.6, Table 4.7, Table 4.8. We can see that the Pearson correlation of the News & Politics category is the highest, and Pearson correlation value of Entertainment category is the second, and that of Science & Technology category is the lowest. This maybe because the user comments of the News & Politics and Entertainment usually have strong emotion polarity, and the models are easier to do sentiment analysis. However, the user comments of Science & Technology category

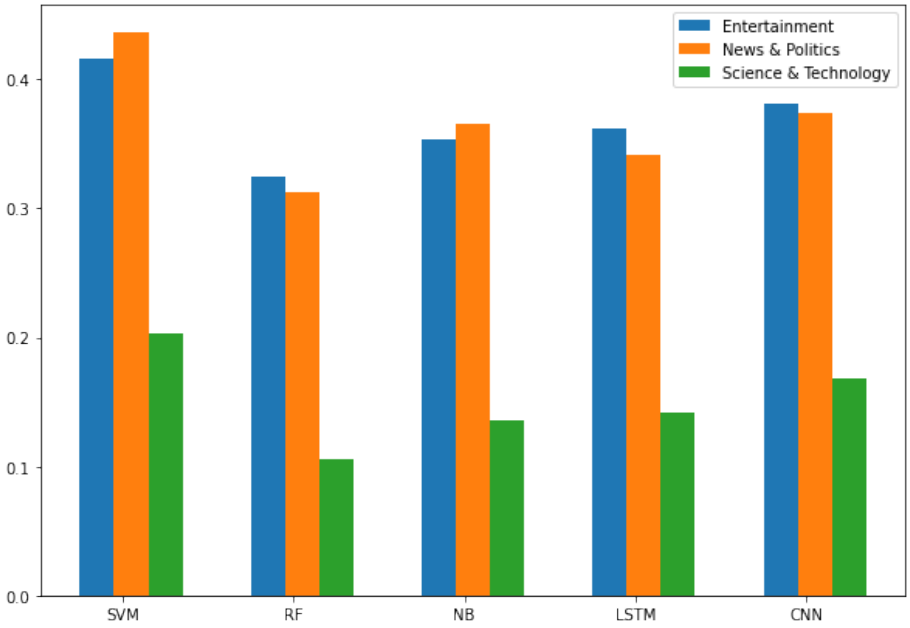

Figure 4.7: Correlation of the three video categories

Classifiers	PR & LR Pearson correlation	PR & VN Pearson correlation
SVM	0.416	-0.070
RF	0.324	0.001
NB	0.354	-0.013
$LSTM$	0.362	0.011
CNN	0.381	0.129

Table 4.6: Entertainment Pearson Correlation

are often more rational, neutral and objective, and the models are not easy to do the sentiment analysis. But in a nutshell, we can get the result that the predicted positive comments ratio has a weak correlation with user actual like ratio.

Classifiers	PR & LR Pearson correlation	PR & VN Pearson correlation
SVM	0.436	0.109
RF	0.312	0.011
NB	0.366	-0.101
$LSTM$	0.342	0.021
CNN	0.374	-0.009

Table 4.7: News & Politics Pearson Correlation

Classifiers	PR & LR Pearson correlation	PR & VN Pearson correlation
SVM	0.203	0.021
RF	0.106	0.003
NB	0.136	-0.153
$LSTM$	0.142	0.010
CNN	0.168	-0.029

Table 4.8: Science & Technology Pearson Correlation

5 Conclusions

This paper deeply and systematically studies the characteristics of YouTube, crawls the 100 most popular YouTuber data and the corresponding 400,000 video data and 700,000 comment data, analyzes their statistical characteristics, and uses the model of machine learning and deep learning to analyze the sentiment of video comments, so as to obtain the emotional polarity of comments. Our research results have constructive advice on the website and YouTubers.

First, among the 100 most popular channels, more than half belong to Music and Entertainment, while the proportion of Science is the smallest. It can be inferred that the best choice to become a popular YouTuber is to create music and entertainment related videos and avoid Style, Trailers and Science. Second, the average number of videos exceeds 16,000, but the median value is only 2,238, and the 75% quantile is 6,714. This shows that most YouTuber upload only a few thousand videos. Moreover, the correlation coefficient between the number of videos and the views of the channel and subscribers is close to 0, indicating that the number of videos has weak linear relationship with the popularity of YouTuber. So to become a popular YouTuber depends on video quality, not video number. The year of channel establishment are evenly distributed, and the Pearson correlation coefficient between the year of establishment and views and subscribers is around 0.2. It can be inferred that whether the channel is popular has nothing to do with the time of establishment. Therefore, new YouTuber has no disadvantages. As long as they can make popular videos, their channel can still become the top of YouTube website. The views of the channel

are only highly correlated with subscribers, and the Pearson correlation coefficient reaches 0.84. Therefore, for advertisers, finding YouTuber with the largest number of subscriptions is enough to ensure more viewing of advertisements and better promote products.

For the video dataset, we have also obtained many interesting conclusions through our observation. The number of videos published on YouTube shows an increasing trend, indicating that the more users accumulated, the more content generated, the stronger the advantages of the website, and the smaller the living space of competitors. In addition, according to the statistics of views by video category, it is found that Music contributes about 50% of the views, indicating that users mostly use the YouTube website to listen to music, and YouTube can increase the proportion of videos about music recommended to users. Film&Animation ranks second, accounting for 22% of the views, but only 4% of the channels is in this category, indicating that the demand for Film&Animation is far greater than the supply. Considering that the production of film and animation requires a lot of professional people and equipment input, for YouTuber, the input-output ratio for the production of this category of videos is too low. Therefore, it seems that traditional film and animation companies will not face strong competition from short video websites in the short term. Regarding the video length, statistics show that 50% of the video length is between 3 mins and 5 mins, which is consistent with the length of a song. Compared with another popular short video application TikTok, the average video duration is 47 seconds. Therefore, YouTube and TikTok do not seem to have a strong competitive relationship.

When users click on a video, they usually rate it or make comments. Therefore, we put the four numerical data, views, likes, dislikes and comments together for analysis. The heatmap shows that there is a correlation between every two attributes,

especially the strong correlation between likes and comments. The Pearson correlation coefficient reaches 0.82.

At the end of 2021, YouTube announced that it would not disclose the number of dislikes to the public. Therefore, we intend to analyze the emotional polarity of user comments to get the positive ratio of comments, and then compare it with the like ratio of video, so as to see whether we can infer the number of dislikes through user comments in the future. For sentiment analysis, we selected three popular machine learning models, namely SVM, NB, RF, and two deep learning models, LSTM and CNN. All five models performed well, basically all got 80% accuracy. The correlation coefficients calculated by the five models ranged from 0.3 to 0.4, indicating a weak correlation between the positive ratio and the like ratio of video comments. Moreover, the correlation coefficient between the positive ratio and the number of views is less than 0.2, indicating that there is no linear correlation between them. Therefore, it can be inferred that the number of views and likes does not seem to be affected by the positive comments ratio. However, the data will show different characteristics when subdivided according to the video category. For example, the correlation coefficient between News & Politics and Entertainment can reach 0.4, while the correlation coefficient of Science & Technology is only 0.2. this phenomenon can be explained that user comments of the News & Politics and Entertainment normally have strong emotion polarity, and the models are easy to do sentiment analysis. However, the Science & Technology category users comments are often more neutral and objective.

For YouTubers troubled by negative reviews, our conclusions show that negative reviews do not actually affect a video's popularity and ranking. Hopefully this conclusion will ease their sense of exclusion.

References

Allehaibi, Khalid, Yaser Khan, and Sher afzal Khan (Sept. 2021). "ITAGPred: A Two-Level Prediction Model for Identification of Angiogenesis and Tumor Angiogenesis Biomarkers". In: *Applied Bionics and Biomechanics* 2021, pp. 1–15. DOI: 10.1155/2021/2803147 (cit. on p. 10).

Asghar, Muhammad Zubair et al. (2015). "Sentiment analysis on youtube: A brief survey". In: *arXiv preprint arXiv:1511.09142* (cit. on p. 14).

Benesty, Jacob et al. (2009). "Pearson correlation coefficient". In: *Noise reduction in speech processing*. Springer, pp. 1–4 (cit. on p. 5).

Bhuiyan, Hanif et al. (2017). "Retrieving YouTube video by sentiment analysis on user comment". In: *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, pp. 474–478 (cit. on pp. 14, 16).

Bisong, Ekaba (2019). "Introduction to Scikit-learn". In: *Building machine learning and deep learning models on Google cloud platform*. Springer, pp. 215–229 (cit. on p. 7).

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32 (cit. on p. 10).

D'Andrea, Eleonora et al. (2019). "Monitoring the public opinion about the vaccination topic from tweets analysis". In: *Expert Systems with Applications* 116, pp. 209–226 (cit. on p. 9).

De Vries, Lisette, Sonja Gensler, and Peter SH Leeflang (2012). "Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing". In: *Journal of interactive marketing* 26.2, pp. 83–91 (cit. on p. 15).

Gorguluarslan, Recep and Seung-Kyum Choi (Aug. 2013). "Predicting Reliability of Structural Systems Using Classification Method". In: vol. 2. DOI: 10.1115/DETC2013-13323 (cit. on p. 9).

Gowda, Charan et al. (2019). *Twitter and Reddit Sentimental analysis Dataset*. DOI: 10.34740/KAGGLE/DS/429085. URL: https://www.kaggle.com/ds/429085 (cit. on p. 35).

Gu, Hao et al. (Aug. 2019). "Blind Channel Identification Aided Generalized Automatic Modulation Recognition Based on Deep Learning". In: *IEEE Access* PP, pp. 1–1. DOI: 10.1109/ACCESS.2019.2934354 (cit. on p. 13).

Gulli, Antonio and Sujit Pal (2017). *Deep learning with Keras*. Packt Publishing Ltd (cit. on p. 8).

Gupte, Amit et al. (2014). "Comparative study of classification algorithms used in sentiment analysis". In: *International Journal of Computer Science and Information Technologies* 5.5, pp. 6261–6264 (cit. on p. 10).

Ishaq, Adnan, Sohail Asghar, and Saira Andleeb Gillani (2020). "Aspect-Based Sentiment Analysis Using a Hybridized Approach Based on CNN and GA". In: *IEEE Access* 8, pp. 135499–135512. DOI: 10.1109/ACCESS.2020.3011802 (cit. on p. 13).

Kalogeropoulos, Antonis et al. (2017). "Who shares and comments on news?: A cross-national comparative analysis of online and social media participation". In: *Social media+ society* 3.4, p. 2056305117735754 (cit. on p. 13).

Kang, Hanhoon, Seong Joon Yoo, and Dongil Han (2012). "Senti-lexicon and improved Naıve Bayes algorithms for sentiment analysis of restaurant reviews". In: *Expert Systems with Applications* 39.5, pp. 6000–6010 (cit. on p. 11).

Karthika, P, R Murugeswari, and R Manoranjithem (2019). "Sentiment analysis of social media network using random forest algorithm". In: *2019 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS)*. IEEE, pp. 1–5 (cit. on p. 10).

Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M Mohammad (2014). "Sentiment analysis of short informal texts". In: *Journal of Artificial Intelligence Research* 50, pp. 723–762 (cit. on p. 15).

Li, Dan and Jiang Qian (2016). "Text sentiment analysis based on long short-term memory". In: *2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI)*, pp. 471–475. DOI: 10.1109/CCI.2016.7778967 (cit. on p. 12).

Li, Yung-Ming and Tsung-Ying Li (2013). "Deriving market intelligence from microblogs". In: *Decision Support Systems* 55.1, pp. 206–217 (cit. on p. 9).

Liao, Shiyang et al. (2017). "CNN for situations understanding based on sentiment analysis of twitter data". In: *Procedia computer science* 111, pp. 376–381 (cit. on p. 13).

Liu, Bing and Lei Zhang (2012). "A Survey of Opinion Mining and Sentiment Analysis". In: *Mining Text Data*. Ed. by Charu C. Aggarwal and ChengXiang Zhai. Boston, MA: Springer US, pp. 415–463. ISBN: 978-1-4614-3223-4. DOI: 10.1007/978-1-4614-3223-4_13. URL: https://doi.org/10.1007/978-1-4614-3223-4_13 (cit. on p. 15).

Lorentz, Isac and Gurjiwan Singh (2021). *Sentiment Analysis on Youtube Comments to Predict Youtube Video Like Proportions* (cit. on p. 15).

Lutz, Sarah and Frank M Schneider (2021). "Is receiving dislikes in social media still better than being ignored? The effects of ostracism and rejection on need threat

and coping responses online". In: *Media Psychology* 24.6, pp. 741–765 (cit. on p. 14).

Maas, Andrew L. et al. (June 2011). "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 142–150. URL: http://www.aclweb.org/anthology/P11-1015 (cit. on p. 35).

Medsker, Larry R and LC Jain (2001). "Recurrent neural networks". In: *Design and Applications* 5, pp. 64–67 (cit. on p. 11).

Murthy, GSN et al. (2020). "Text based sentiment analysis using LSTM". In: *Int. J. Eng. Res. Tech. Res* 9.05 (cit. on p. 12).

Noble, William S (2006). "What is a support vector machine?" In: *Nature biotechnology* 24.12, pp. 1565–1567 (cit. on p. 8).

O'Shea, Keiron and Ryan Nash (2015). "An introduction to convolutional neural networks". In: *arXiv preprint arXiv:1511.08458* (cit. on p. 13).

Olah, Christopher (2015). *Understanding LSTM Networks*. URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/ (cit. on p. 12).

Oshiro, Thais Mayumi, Pedro Santoro Perez, and José Augusto Baranauskas (2012). "How many trees in a random forest?" In: *International workshop on machine learning and data mining in pattern recognition*. Springer, pp. 154–168 (cit. on p. 10).

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (2002). "Thumbs up? Sentiment classification using machine learning techniques". In: *arXiv preprint cs/0205070* (cit. on p. 11).

Porreca, Annamaria, Francesca Scozzari, and Marta Di Nicola (2020). "Using text mining and sentiment analysis to analyse YouTube Italian videos concerning vaccination". In: *BMC Public Health* 20.1, pp. 1–9 (cit. on p. 16).

Schultes, Peter, Verena Dorner, and Franz Lehner (2013). "Leave a comment! An in-depth analysis of user comments on YouTube". In: (cit. on p. 30).

Sherstinsky, Alex (2020). "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network". In: *Physica D: Nonlinear Phenomena* 404, p. 132306 (cit. on p. 12).

Trinkle, Brad S, Robert E Crossler, and France Bélanger (2015). "Voluntary disclosures via social media and the role of comments". In: *Journal of Information Systems* 29.3, pp. 101–121 (cit. on p. 14).

Wan, Xiang et al. (2014). "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range". In: *BMC medical research methodology* 14.1, pp. 1–13 (cit. on p. 6).

Wang, Zhou and Alan C Bovik (2009). "Mean squared error: Love it or leave it? A new look at signal fidelity measures". In: *IEEE signal processing magazine* 26.1, pp. 98–117 (cit. on p. 7).

Wankhade, Mayur, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni (2022). "A survey on sentiment analysis methods, applications, and challenges". In: *Artificial Intelligence Review*, pp. 1–50 (cit. on p. 16).

Zhang, Xinhua (2017). "Support Vector Machines". In: *Encyclopedia of Machine Learning and Data Mining*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, pp. 1214–1220. ISBN: 978-1-4899-7687-1. DOI: 10.1007/978-1-4899-7687-1_810. URL: https://doi.org/10.1007/978-1-4899-7687-1_810 (cit. on p. 8).