EBA5002 Business Analytics Practice

Report for YouTube Video Analysis for BigFame company

Group16_Report.pdf

16 November 2022

Team: Respect
Qiaoling Chen: e0983208
Lyu Xinyu: e0983260
Xie Siteng: e0983376
Yashna Gogineni: e0983387
Zhao Rudan: e0983390

Table of Contents

BIGFAME

Executive Summary

Purpose of Report

Big Fame is a multi-channel network (MCN). It is a company which can help vloggers become popular in the short video industry. It includes ng the selection of web stars, incubation, content development, technical support of UGC platform, IP development, and other complicated work. MCN companies are faced with lots of challenges. This project can deal with the problems of lack of talent vloggers, creative ideas, KOL erosion, decentralized investment that BigFame co. faced with.

The purpose of this report are:

- Make BigFame co. as a powerful differentiator by targeting different groups of viewers and making precise video making decisions.
- Discover the video popularity trends and make well preparation for the hot category next period

Methods

Extracting the business insights from the TikTok datasets through statistical analysis. Making time series analysis and text analysis to meet our business and technical objectives.

Key Finding & Conclusions

Video popularity of each category shows a different performance next period. Entertainment video shows an upward trend, technology video shows a stable trend but with fluctuations, health video shows a stable trend without seasonal variation, science topic will keep fluctuating with a slight decrease.

By doing sentiment analysis, building clustering model and topic model. We define 5 groups of viewers with different characteristics. They show strong interest in game, music, learning, technical and life areas with specific personalities respectively.

Suggestions for Improving

- Contract anchors who can attract an audience interested in gaming, knowledge, music, finance and life according to the company's audience targeting
- According to different viewers clusters and popular video categories, enlighten the creators to create special video(eg. Entertainment video: more eating, animal, makeup contents)
- Pay more attention on KOL’s training. Provide them more technical and investment support
- Save costs by releasing other small anchors in other areas from their contracts
- By shrinking the business lines, improve other popular categories performance.
- By marketing and IP building to increase the reputation

1. Introduction & Background

Currently, the user is spending more and more time on Currently, the user base is spending more and more time on short-form video platforms. As the main production site of UGC, short video platforms are playing an increasingly important role in business socialization. The number of users of 'live and short-form video' application platforms has reached 888 million, accounting for 94.5% of the total number of Internet users in China. And the sales revenue of various flat broadcast e-commerce has reached trillions in 2020 (Guo, 2021). Meanwhile, the The growth rate of MCN institutions remains stable in 2021, and the number of MCN institutions is expected to exceed 40,000 in 2022 and 60,000 in 2025 （iiMedia Report).

MCN company is a new product in the emergence of the UGC video industry. MCN companies work with video platforms to sign up video creators, aiming at turning video flow's bonuses into profit. At the same time, they provide video creators with content editing, creator training, digital rights management, selling, audience discovery, and advertising (YouTube, n.d.). MCNs help creators find advertising sponsors and provide key resources for creating videos (e.g., video shooting locations, equipment, copyright, etc.). In return, they gain benefits from a percentage of the revenue from these channels (The Wrapbook Team, 2022).

2. Problem Statement and Objectives

BigFame is an MCN company， as more and more agencies of the same type enter this market and compete for the limited number of KOL, the company has received a significant loss (TopKLOut, 2020, p. xx) 。

Due to the lackness of ability in managing all the outstanding KOLs with limited resources and providing them with technical and content ideas support, resulted in many outstanding KOLs feeling unappreciated and choosing to hop to other MCN companies from BigFame.

At the same time, BigFame's KOLs created a wide variety of videos, but did not follow the popularity trend of the times. The unpredictable popularity trend became one of the biggest problems for the company's content creation department.

Advertisers' high demand for qualified video for advertisements producing leads to the difficulty of making profit from cooperation with advertisers.(BaiZhun, 2022) 。

In order to create a clear distinction from other companies, BigFame must develop a differentiation strategy to attract more talented KOLs. Help contracted video creators to create more distinctive videos to attract a specific group of viewers. And sign up more suitable advertisers to increase revenue.

Table 2-1 Business & Technical Objectives

Business Objectives	Technical Objectives
Make agile, flexible, and resilient operations based on full range of clean data, providing business insights and risk management within one month ● Make data-driven tactics can win more trust from KOL and brands ● Find the business opportunity more quickly to make contract with brands	● Preparing data including integration within 3 days ● Generating tags for each video based on the title and comments by doing text analysis within 1 week ● Building an interactive dashboard to find potential business insights within 3 days
Looking at video popularity trends and make well preparation for the hot category next period ● Catch the opportunity in the rapid changing era so the MCN company can help the KOL attract more followers ● Customizing development plan for each KOL and cultivate new vloggers in a particular field	● Building mathematical model such as AHP/ TOPSIS to calculate videos' popular score within 3 days ● Building time series model to predict the future popularity trend for all categories within 1 week
Become a much more strategic generator of value and potentially also a powerful differentiator ● Analysing viewers sentiment and grouping them, can help MCN guide vloggers to generate videos with more specific style, then attract target viewers. ● An opportunity to provide customer interests-oriented video generation and make the company become a leader in a special space.	● Classifying the sentiment based on the text within 5 days ● Building machine learning model such as clustering based on sentiment analysis to identify different viewers group within 5 days

Once achieved these goals, BigFame is able to build a successful differentiation strategy. Realized optimizing the team of signed video creators by reducing the type of video categories. For KOLs who already have a large amount of fans, the company can use its resources wisely to provide them with better services and training. For video creators with a medium to small number of followers, the content creation department might help them create more innovative video content after capturing popular trends. In this way, BigFame's contracted KOLs are transformed from popularity trend followers into popularity tread leaders.

In addition, having a better understanding of the video creators we have already signed up with can help them to match with more suitable advertising content and get a better connection with advertisers. Thus creating more profit to the company.

Such a trend-focused company can attract more KOLs with the same video viewership to join. And the differentiation strategy can streamline BigFame's comprehensive business lines, reduce cost investment and increase profits. In addition, the various sentiment analysis models, clustering models and category models built by the Respect project team provide good templates for prediction and segment capture thereafter. It helps the company to adjust tactics conveniently at any time.

3. Data Understanding & Data Preprocessing

3.1. Basic Information

- **Data sources:** The data source used for this project is from Kaggle website. The links are attached in reference.
- **Data Quantity:** The dataset 'YouTube Statistics' has two files which is video-stats.csv and comments. csv. *video-stats.csv* has 1846 rows with 7 columns about the basic information of each video. *comments.csv* has 18247 rows contains the comment details and sentiments which are corresponded to the video-stats dataset.
- **Data Quality:** The data in this table is relatively clean, but the text was encoded by different way.
- **Data Types:** The raw data contains unstructured data like comments and structured data like Published At, Likes, Views e.g.

3.2. Data Dictionary

The following table is part of the data dictionary, full details are given in Appendix 1.

Table 3-1 Part of Data Dictionary

Field Name	Data Type	Field Length	Description	Null Value Acceptation	Example
Video ID	Integer	11	Every video has its own unique ID	N	wAZZ-UWGVHI
Title	String	256	The video's title describes the main content of the video	N	Apple Pay
Published At	Date	256	This basically describes when the video was released	N	2022/8/23
Keyword	Character	10	The keyword means the field that this video belongings to	N	tech
Likes	Integer	256	The number of viewers who like this video	Y	3407
Views	Integer	256	The number of times the video has been viewed	Y	135612

Comments	Integer	256	The number of comments on this video	Y	672
Comment Detail	String	256	The viewers' comment content on the video	Y	This video is good
Comment Likes	Integer	256	The number of people who like this comment	Y	95
Comment Sentiment	Integer	256	0: dislike 1: neutral 2: like	N	1

3.3. Data Preparation

Data preprocessing and text preprocessing helps achieve the technical objectives about building dashboard and analytic models.

3.3.1 Data Pre-processing - Cleaning & Feature Extraction

For the two datasets, we used the REMOVE DUPLICATES function to remove them and used the DATA VALIDATION function to remove the inconsistencies of data which exceed the range of the variable. And some extreme outliers of each variable are removed based on this step.

This dataset has three columns that are suitable as dependent variables of a video's feature: views, likes, comments. It is necessary to use these to assign weights to calculate the popularity degree of each video for time series analysis.

We defined the popularity score as "pop_score". It is generated by TOPSIS Method based on Entropy Weight:

Firstly, we standardized the data of the three indicators (views, likes, comments counts) to form a matrix.

```python
# Def: Standardize the matrix
def Y_ij(data1):
    for i in data1.columns:
        for j in range(n+1):
            if i == str(f'X{j}negative'):  # negative
                data1[i]=(np.max(data1[i])-data1[i])/(np.max(data1[i])-np.min(data1[i]))
            else:   # positive
                data1[i]=(data1[i]-np.min(data1[i]))/(np.max(data1[i])-np.min(data1[i]))
    return data1
```

Then, calculate the information entropy to get the weight for the 3 indicators.

```
# Def: Caculate the entropy
# Build empty matrix
None_ij = [[None] * n for i in range(m)]
def E_j(data2):
    data2 = np.array(data2)
    E = np.array(None_ij)
    for i in range(m):
        for j in range(n):
            if data2[i][j] == 0:
                e_ij = 0.0
            else:
                p_ij = data2[i][j] / data2.sum(axis=0)[j]
                e_ij = (-1 / np.log(m)) * p_ij * np.log(p_ij)
            E[i][j] = e_ij
    E_j=E.sum(axis=0)
    return E_j
```

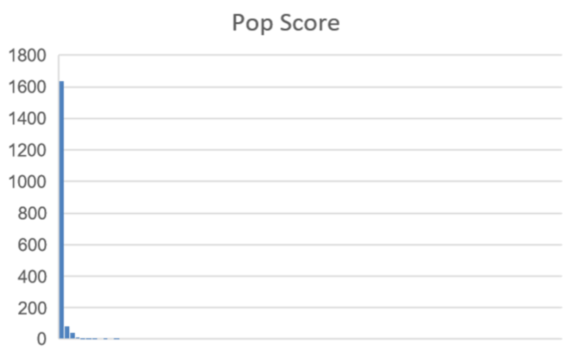
```
Likes       0.301295
Comments    0.294011
Views       0.404695
```

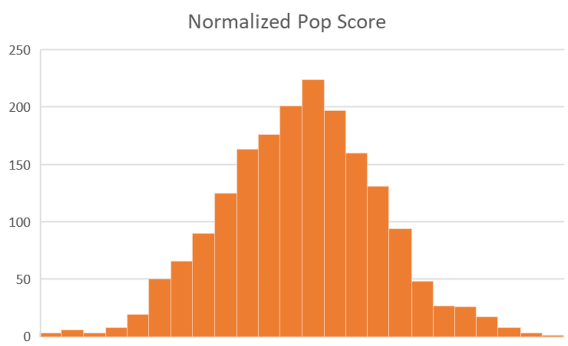
Considering that it is hard to define the degree of the score. Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) can deal with the problem. By using the distance and the weight to generate the popular score of each video.

```
# Calculate the optimum and the worst
Imax_j = Z_ij.max(axis=0)
Imin_j = Z_ij.min(axis=0)
Dmax_ij = np.array(None_ij)
Dmin_ij = np.array(None_ij)
for i in range(m):
    for j in range(n):
        Dmax_ij[i][j] = (Imax_j[j] - Z_ij[i][j]) ** 2
        Dmin_ij[i][j] = (Imin_j[j] - Z_ij[i][j]) ** 2
```

```
# Calculate the pop score
Dmax_i = Dmax_ij.sum(axis=1)**0.5
Dmin_i = Dmin_ij.sum(axis=1)**0.5
S_i = Dmin_i/(Dmin_i + Dmax_i)
S_i= pd.Series(S_i, index=video_data.index, name='pop_score')
```

Finally, we visualized the result. The high skewed histogram should be normalized into normal distribution. We tried LOG Function to do the work. Then the feature "pop_score" can be used in time series analysis.





3.3.2 Text Pre-processing - Cleaning, Lemmatization & Tokenization

"comments.csv" dataset includes 10 detailed comments for each video. To make it more convenient on building text analysis models. Firstly, we replaced the contractions and transform the capital case into lowercase. Then removed the Non-ASCII words, punctuation, stopwords.

```
def replace_contractions(text):
    """Replace contractions in string of text"""
    return contractions.fix(text)


def remove_non_ascii(words):
    """Remove non-ASCII characters from list of tokenized words"""
    new_words = []
    for word in words:
        new_word = unicodedata.normalize('NFKD', word).encode('ascii', 'ignore').decode('utf-8', 'ignore')
        new_words.append(new_word)
    return new_words

def to_lowercase(words):
    """Convert all characters to lowercase from list of tokenized words"""
    new_words = []
    for word in words:
        new_word = word.lower()
        new_words.append(new_word)
    return new_words
```

```
def remove_punctuation(words):
    """Remove punctuation from list of tokenized words"""
    new_words = []
    for word in words:
        new_word = re.sub(r'[^\w\s]', '', word)
        new_word = re.sub(r'(.)\1+', r'\1\1', new_word)
        if new_word != '':
            new_words.append(new_word)
    return new_words

def remove_stopwords(words):
    """Remove stop words from list of tokenized words"""
    new_words = []
    sw = stopwords.words('english')
    for word in words:
        if word not in sw:
            new_words.append(word)
    return new_words
```

Lemmatization and tokenization helped us to build sentiment analysis, clustering models and topic modelling more explainable.

4. Modelling & Evaluation Methodologies

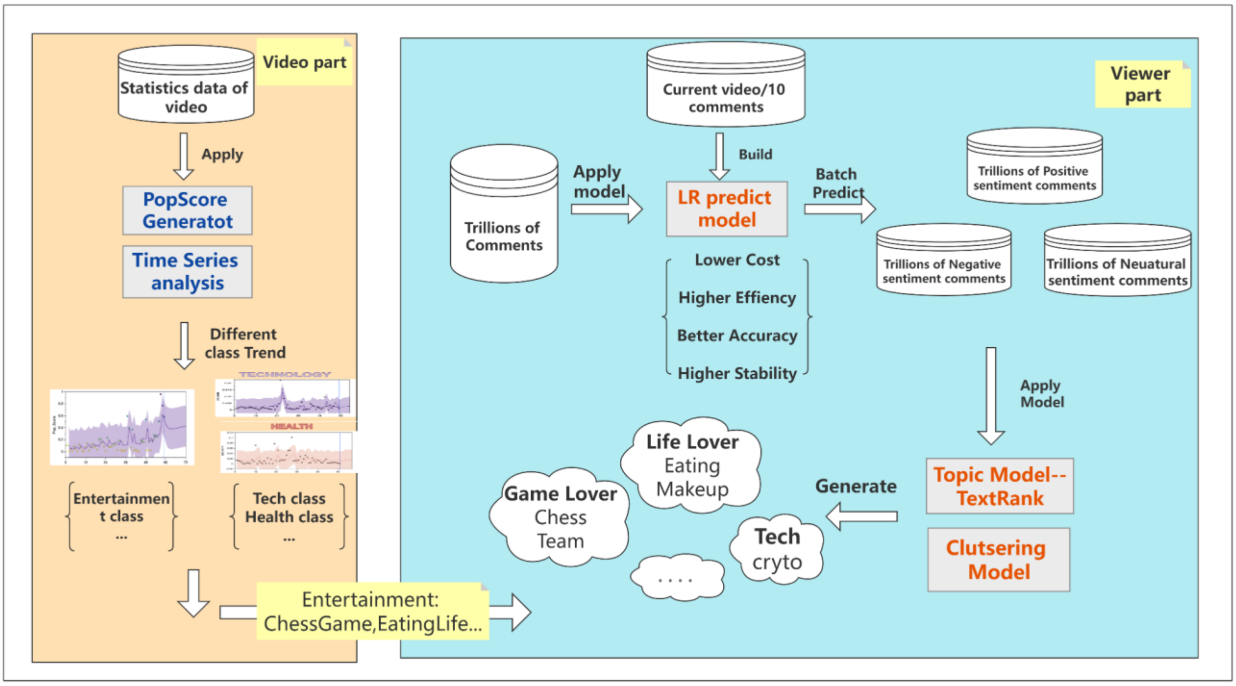


Figure 4-1 Overview of WorkFlow

4.1. Overview Of WorkFlow

Video Trend Prediction

Pop score generator （As shown in data preparation）

Time Series Analysis

Viewer Feature Extraction

Sentiment Analysis Model Selection

Topic Model

Clustering Model

This workflow is separated into two parts, one is to build a time series predictive model to predict pop trend videos, and the other is to extract more insight from viewers’ comments. At the end ,by integrating this two parts, we can give more specific biz suggestions.

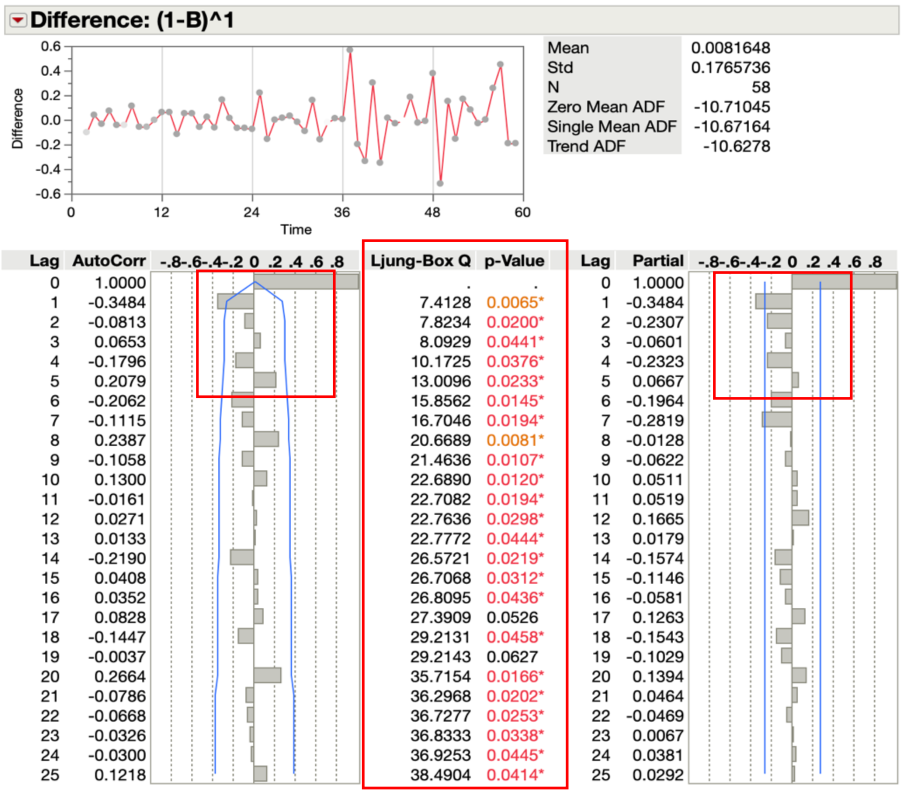
4.2. Video Trend Prediction

For the video part, we are trying to use the time series analysis to figure out the popular trend of some topics, which can combine with the words used frequently in viewers’ comments. So that it will be clearer for our KOL to create videos catering to our target viewers. In time series analysis, we compared different models for each category, Time Series regression with Seasonal Variation, Modeling Seasonal Variation by Dummy Variables, Exponential Smoothing, Holt's Method, Holt-

Winters' Method and ARIMA. As the R Square and AIC, SBC of ARIMA has the best performance of each category, we mainly focus on ARIMA analysis and the result of other models will only be shown by category 'entertainment' in the report, while all the results are in the R code file.

As each keyword's data amount is about 40 to 50, and when it was converted to monthly data, the number decreased to no more than 20, which cannot be used for time series analysis. Therefore, we divided the 41 keywords into five categories and conducted a time series analysis to explore the trend of each category's popularity. Based on each video's keyword, they are divided into entertainment, health, technology, science, and humanities. To make it specific, Science contains Natural science and applied science, focusing on theory. While technology focuses on the product of the application of knowledge, it includes physical objects like machines and intangible tools such as software.

For topic 'entertainment', from the pop score plot and ACF, autocorrelation will drop to zero quickly if the series is a stationary process, so for the entertainment pop score series, as it has an increasing trend, it can be initially determined that the time series is a non-stationary process. We may take one differencing to transform it into a stationary series. But as sometimes it is difficult to figure out whether the series' ACF is die down or cut off, we will try both 0 and 1 differencing. Additionally, based on The Ljung–Box test, we check whether any of a group of autocorrelations of this time series are different from zero, as almost all the P values are less than 0.05, we can determine that this series is not a white noise series. Then according to the diagram of autocorrelation function and partial-autocorrelation function, we can initially decide to try P=0, P=1, Q=0, Q=1, Q=2. And as it's difficult to figure whether a seasonal variation exists in this series, we also tried one seasonal differencing, but the result is not better than taking one regular differencing.



Mean	0.0081648
Std	0.1765736
N	58
Zero Mean ADF	-10.71045
Single Mean ADF	-10.67164
Trend ADF	-10.6278

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.3484	7.4128	0.0065*	1	-0.3484
2	-0.0813	7.8234	0.0200*	2	-0.2307
3	0.0653	8.0929	0.0441*	3	-0.0601
4	-0.1796	10.1725	0.0376*	4	-0.2323
5	0.2079	13.0096	0.0233*	5	0.0667
6	-0.2062	15.8562	0.0145*	6	-0.1964
7	-0.1115	16.7046	0.0194*	7	-0.2819
8	0.2387	20.6689	0.0081*	8	-0.0128
9	-0.1058	21.4636	0.0107*	9	-0.0622
10	0.1300	22.6890	0.0120*	10	0.0511
11	-0.0161	22.7082	0.0194*	11	0.0519
12	0.0271	22.7636	0.0298*	12	0.1665
13	0.0133	22.7772	0.0444*	13	0.0179
14	-0.2190	26.5721	0.0219*	14	-0.1574
15	0.0408	26.7068	0.0312*	15	-0.1146
16	0.0352	26.8095	0.0436*	16	-0.0581
17	0.0828	27.3909	0.0526	17	0.1263
18	-0.1447	29.2131	0.0458*	18	-0.1543
19	-0.0037	29.2143	0.0627	19	-0.1029
20	0.2664	35.7154	0.0166*	20	0.1394
21	-0.0786	36.2968	0.0202*	21	0.0464
22	-0.0668	36.7277	0.0253*	22	-0.0469
23	-0.0326	36.8333	0.0338*	23	0.0067
24	-0.0300	36.9253	0.0445*	24	0.0381
25	0.1218	38.4904	0.0414*	25	0.0292

Figure 4-2 Visualize 'entertainment' Pop Score

Visualize 'entertainment' Pop Score after 1 regular differencing

After trying different variables in seasonal ARIMA, according to the Akaike Information Criterion and Schwartz Bayesian Criterion, the AIC and SBC should be as small as possible. As ARIMA (1, 1, 1) has a very small AIC and although it's SBC is not the smallest, it's R Square has a better performance. Therefore, we found that ARIMA (1,1,1) most adequately represents the data.

Model	*DF*	*Variance*	*AIC*	*SBC*	*RSquare*	*-2LogLH*	*MAPE*	*MAE*
ARIMA(1, 1, 1)	**55**	**0.024106**	**-45.44431**	**-39.26298**	**0.404**	**-51.44431**	**692.7482**	**0.105634**
AR(1)	57	0.026241	-44.93511	-40.78003	0.345	-48.93511	564.0437	0.10742
IMA(1, 1)	56	0.02596	-44.82172	-40.70083	0.362	-48.82172	616.0846	0.108038
ARMA(1, 1)	56	0.025985	-44.41087	-38.17826	0.362	-50.41087	557.495	0.105578
AR(2)	56	0.026079	-44.25493	-38.02232	0.36	-50.25493	556.4575	0.105483
IMA(1, 2)	55	0.02572	-44.20452	-38.02319	0.379	-50.20452	651.6136	0.104614
ARIMA(2, 1, 1)	54	0.024432	-43.93216	-35.69039	0.407	-51.93216	679.9363	0.105215
Seasonal ARIMA(1, 1, 1)(0, 0, 1)12	54	0.024561	-43.45321	-35.21144	0.404	-51.45321	693.1448	0.10542
Seasonal ARIMA(1, 1, 1)(1, 0, 0)12	54	0.024559	-43.45152	-35.20975	0.404	-51.45152	693.0725	0.105457
ARIMA(1, 1, 2)	54	0.026167	-42.35606	-34.11429	0.38	-50.35606	609.5492	0.105772

Figure 4-3 Compare Different ARIMA Model

The residuals left over after fitting the model should be white noise. Therefore, we firstly check the residual values distribute on both sides of the zero. And then check the ACF and PACF, as all the spikes are in blue lines, showing that the residuals do not have significant autocorrelations or partial autocorrelations. Thirdly, according to the Portmanteau test, we can see that the p-value is significantly larger than 0.05 and as such we can state that there is strong evidence for discrete white noise being a good fit to the residuals. Hence, the ARIMA (1,1,1) model is adequate, as expected.

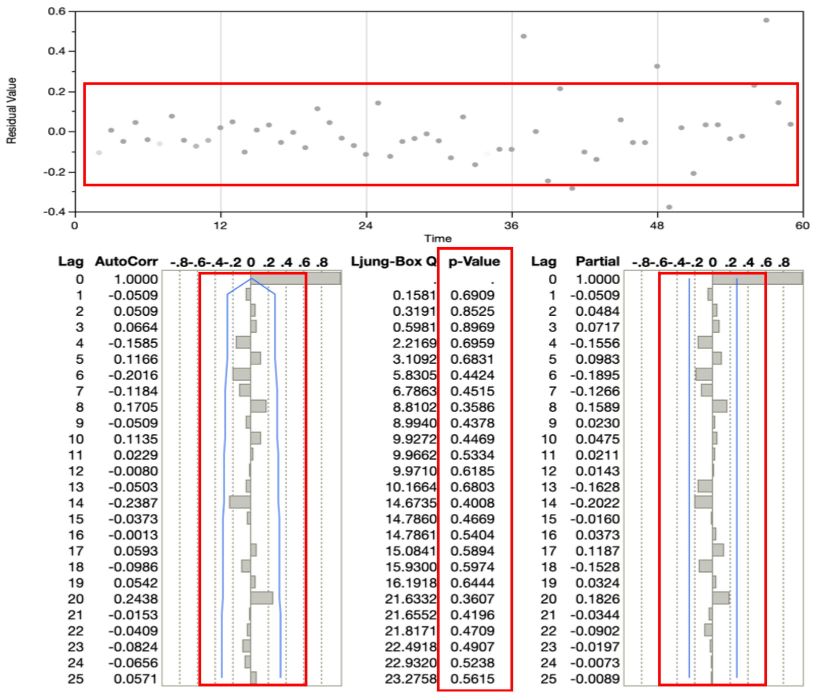
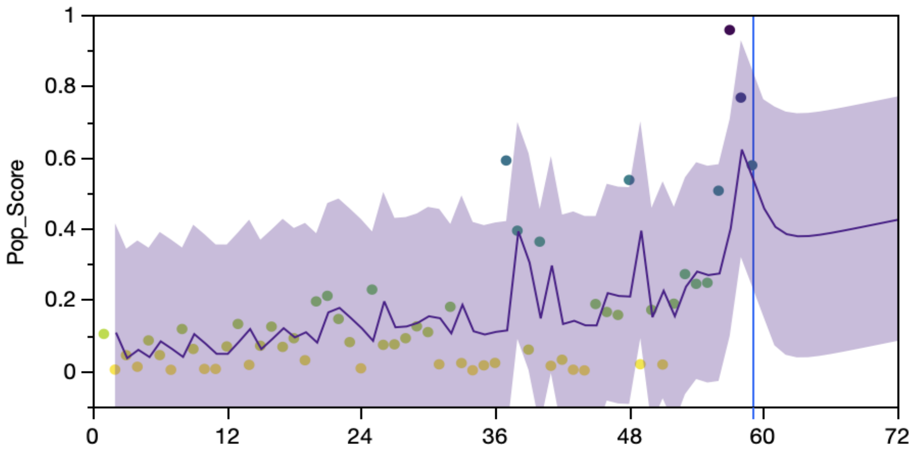


Figure 4-4 Residual of ARIMA (1,1,1)

After the diagnosing residuals, we checked the model is stable and invertible. In general, based on this model, the popularity of the topic of ENTERTAINMENT shows an upward trend, the formular of this model is as fellow. Besides, we also tried other time series models (in appendix) to compare with the ARIMA model, ARIMA (1,1,1) has the best performance. Based on this analysis and prediction, we can predict that the topic of entertainment will have an upward tendance in the next few months, we are suggested to encourage our KOL to pay more attention on this kind topic.



Model: ARIMA(1, 1, 1)

Model Summary

DF	55	Stable	Yes
Sum of Squared Errors	1.32581033	Invertible	Yes
Variance Estimate	0.02410564		
Standard Deviation	0.15525992		
Akaike's 'A' Information Criterion	-45.444309		
Schwarz's Bayesian Criterion	-39.26298		
RSquare	0.40369567		
RSquare Adj	0.38201188		
MAPE	692.74816		
MAE	0.10563359		
-2LogLikelihood	-51.444309		

Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR1	1	0.4617515	0.1225077	3.77	0.0004*	0.00341949	0.006353
MA1	1	1.0000000	0.0735173	13.60	<.0001*		
Intercept	0	0.0063530	0.0020585	3.09	0.0032*		

Figure 4-5 Model Summary of ARIMA (1,1,1)

With the same process and passed the diagnostic checking shown before, we found that topics 'technology' and 'health' are both stationary processes. The result of time series analysis is ARIMA (2,0,1), ARIMA (0,0,2) respectively. Technology's popularity shows a stable trend but with fluctuates in September and October from 2019, which may be inferred as apple's launch in about October every year. Our video creators are encouraged to focus on the technology innovation, especially some popular topics like APPLE. And the 'health' category shows a stable trend without seasonal variation, to attract more followers, KOL has to explore some creative video content.

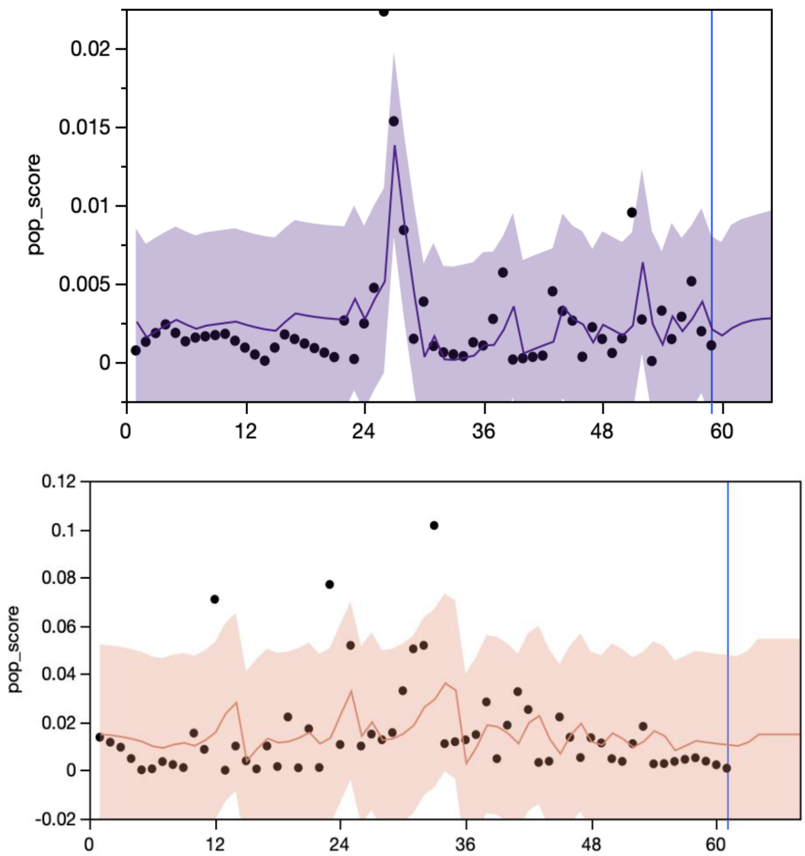


Figure 4-6 Prediction of category TECHNOLOGY and HEALTH

Besides, we also used R to generate training set and testing set, 70% and 30% respectively. And we can find that for 'science' category, the residual scatters on both sides of zero, all sticks are in two blue lines and the distribution of residuals approximate the normal distribution. This model predicts that the popularity of science topic will keep fluctuation with a slightly decrease in the next few months, trend is not easy to catch. But for this model we noticed both AIC and BIC are very small, overfitting may exist in this model, so as a limitation, we may have to normalize the popularity score to make it a better performance.

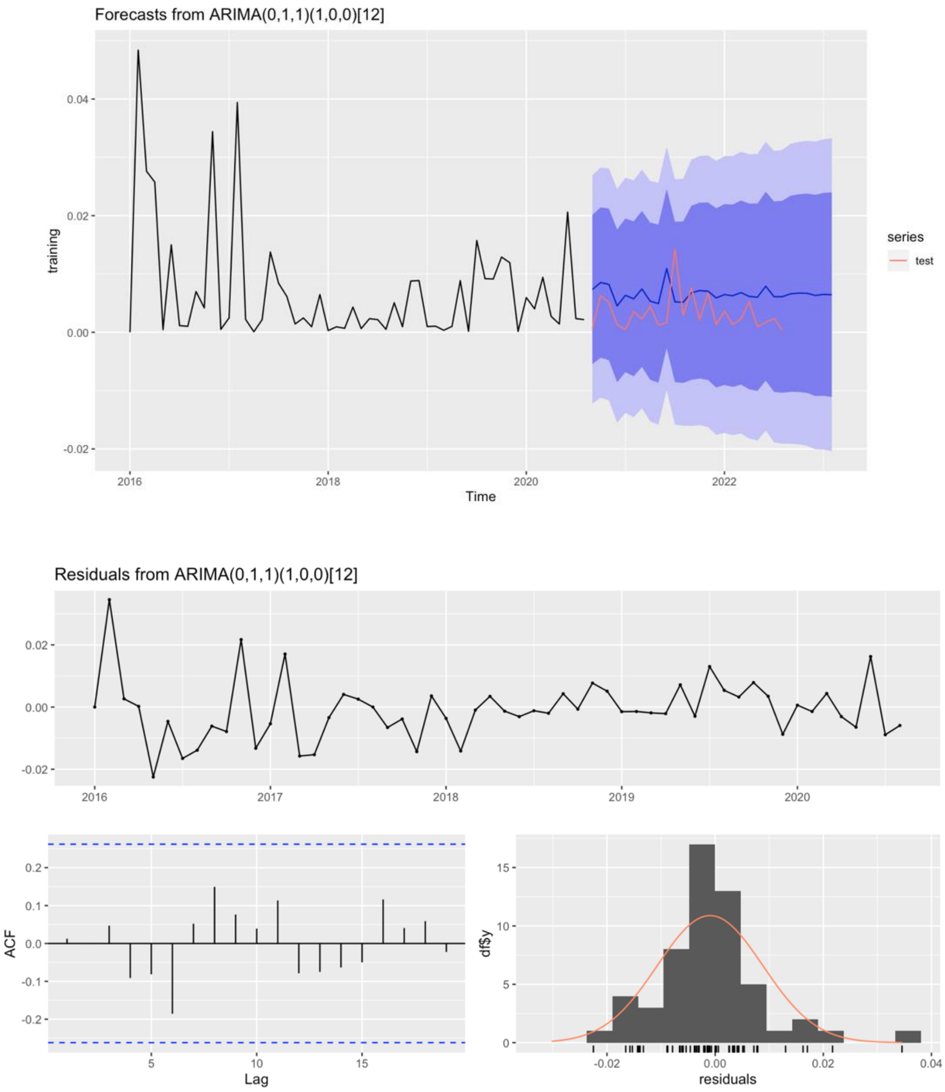


Figure 4-7 Prediction of category SCIENCE

For the humanities and social topic, the result of Randomness Test shows that all the P-Values are larger than 0.05 which means it is a white noise series. To further check this result, we used auto. arima function in R studio this time, and the result shows that pdq all equal to 0 is the best model, so we can confirm that the humanities and social's popularity score is a white noise series. Due to this result, we can not come up with any business strategy, so as one of the limitations, we determine to recategorize the keywords for improvement.

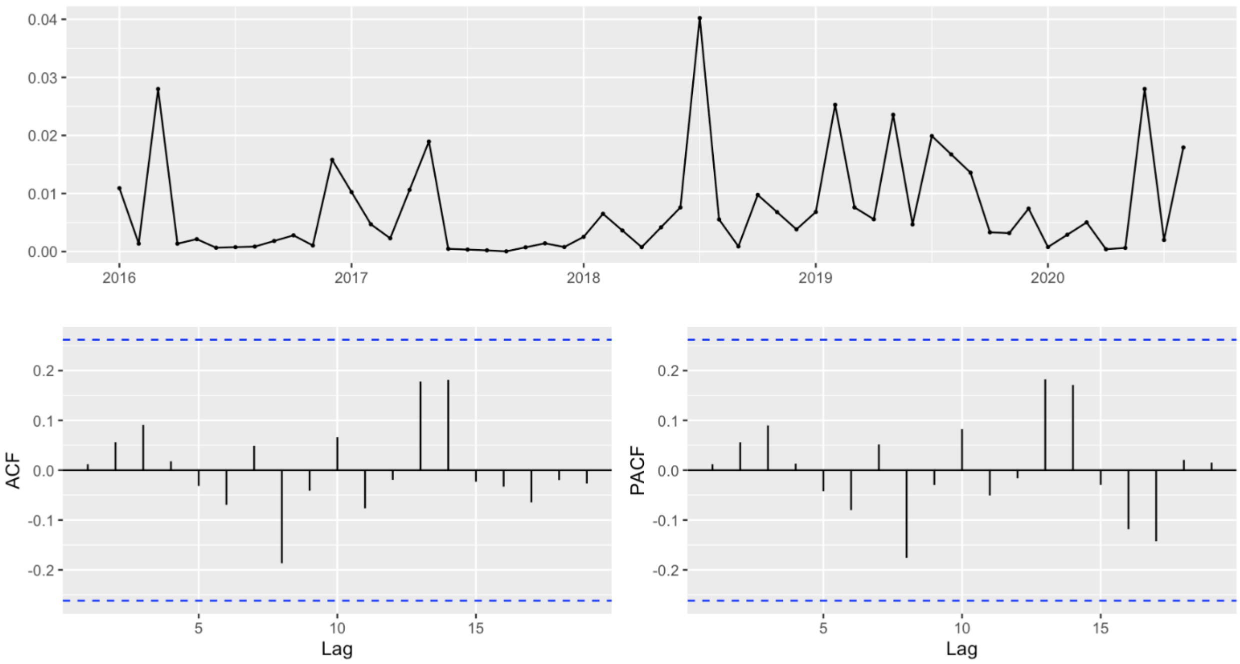


Figure 4-8 Prediction of category HUMANITIES

4.3. Sentiment Analysis Model Selection

Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral. Today, companies have large volumes of text data like emails, customer support chat transcripts, social media comments, and reviews. Sentiment analysis tools can scan this text to automatically determine the author's attitude towards a topic. Companies use the insights from sentiment analysis to improve customer service and increase brand reputation.

We aim to build a Lower Cost Higher Efficiency Better Accuracy Higher Stability Sentiment Analysis model for batch prediction of sentiment orientation on Youtube using a small number of comments. This workflow will take the large number of comments obtained and place them in three different datasets. The next workflow will analyse the characteristics of a particular sentiment based on these three datasets.

This model is the basis for the entire feature extraction of the comments generated by the Viewer.

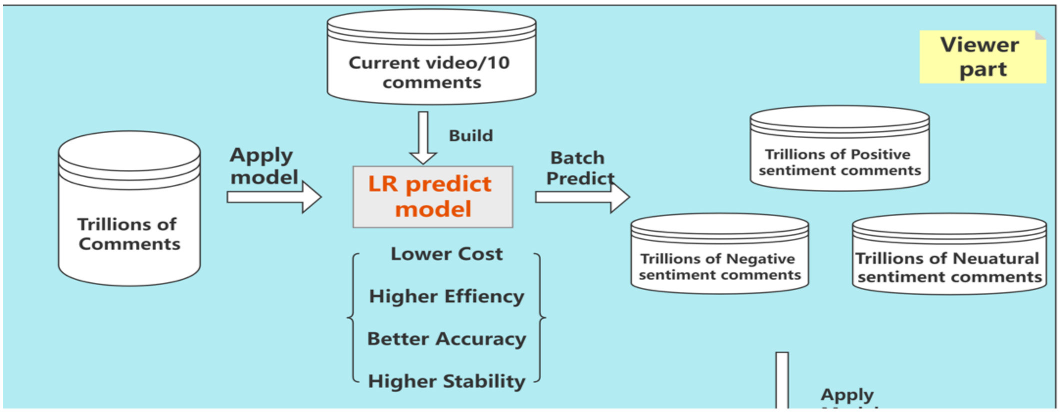


Figure 4-9 Sentiment Analysis WorkFlow

4.3.1 Problem Statement for Sentiment Analysis

The problem statement is to predict the sentiment of the comments i.e., Positive/ Negative/Neutral. The bar chart below shows the distribution of the comments. This is an imbalanced data set.

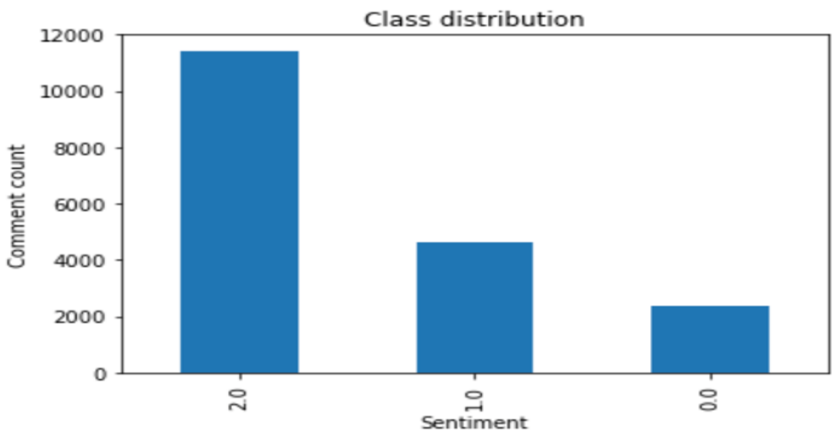


Figure 4-10 Class Distribution

4.3.2 Word Cloud for different Sentiments

The word cloud for Positive, Negative and Neutral sentiment is represented here.

Sentiment 2 / Positive Sentiment.This has words like ‘Thank’, ‘Love’, ‘Great’, ‘much’

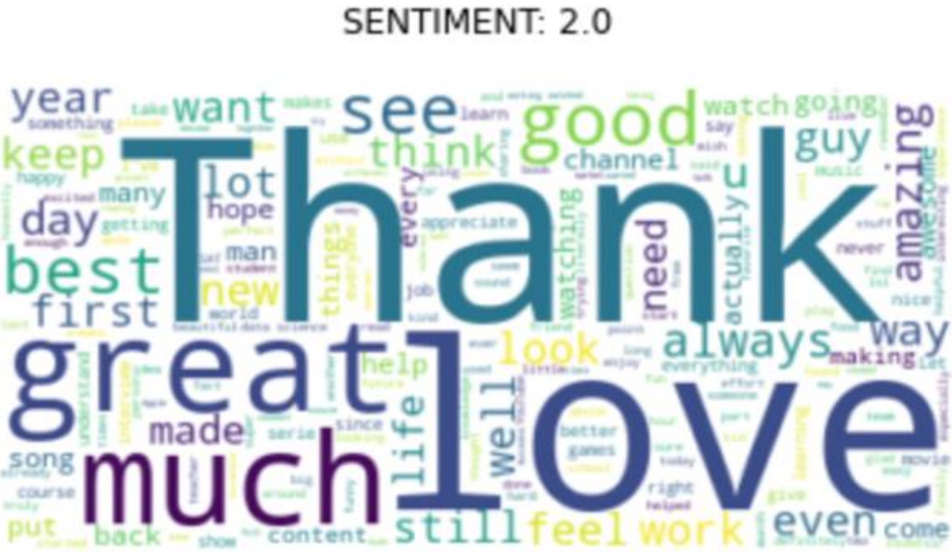


Figure 4-11 Sentiment: 2.0

Sentiment 1 / Neutral Sentiment.his has words like ‘see’, ‘think’

Figure 4-12 Sentiment 1.0

Sentiment 0 / Negative Sentiment.This has words like ‘never’, ’need’, ‘even’

Figure 4-13 Sentiment 0.0

4.3.3 Steps involved in Building the Model

The steps involved in building the Machine Learning model are listed here. The comments are pre-processed. The data is split into training and test set. TF-IDF Vectorizer is used. Stratified K-fold cross validation technique is used in training the model. Class weight method is used to deal with the imbalanced data set. Next, Model selection is done. The model with the highest 'F1-Score' is selected. We don't use accuracy as an evaluation metric. Instead, we use F1-Score which is the harmonic mean of precision and recall since the data set is imbalanced. Hyper-parameter tuning using GridSearchCV is performed. Finally, we predict the sentiment on the test set. The model evaluation is carried out with the help of confusion matrix and classification report.

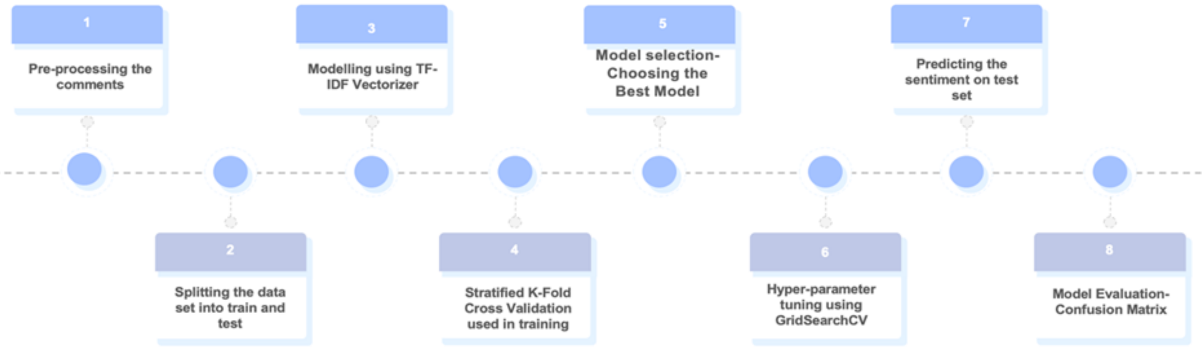


Figure 4-14 Sentiment Analysis Steps

a) Model Selection

Different models such as XLNET, BERT, Support Vector Machine, Logistic Regression, Gaussian Naive Bayes, Random Forest Classifier, Gradient Boosting Classifier, XGB, LGBM, KNN were used to predict the sentiment. The chart below shows the F1-Score w.r.t to the different models used. The evaluation metric used here is 'F1-Score' since the dataset is imbalanced. Deep learning models like XLNET, BERT gave a F1-Score of 0.78. Among the Machine Learning model, Logistic regression model performed the best with a F1-Score of 0.67. The deep learning model has a F1-Score 15% better than Logistic Regression.

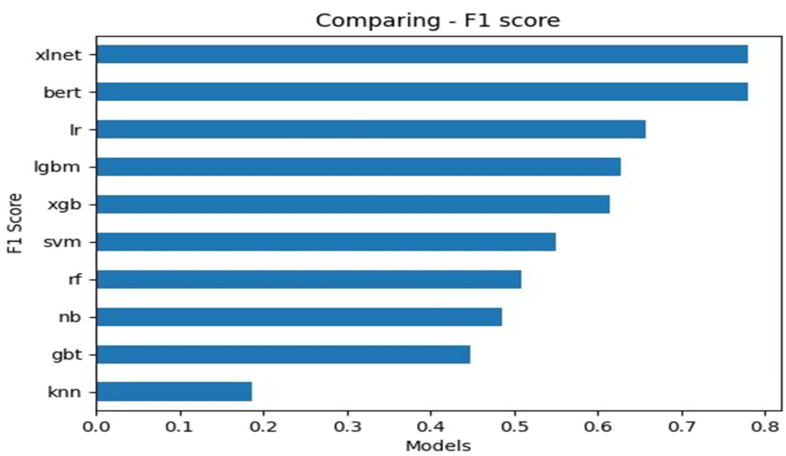
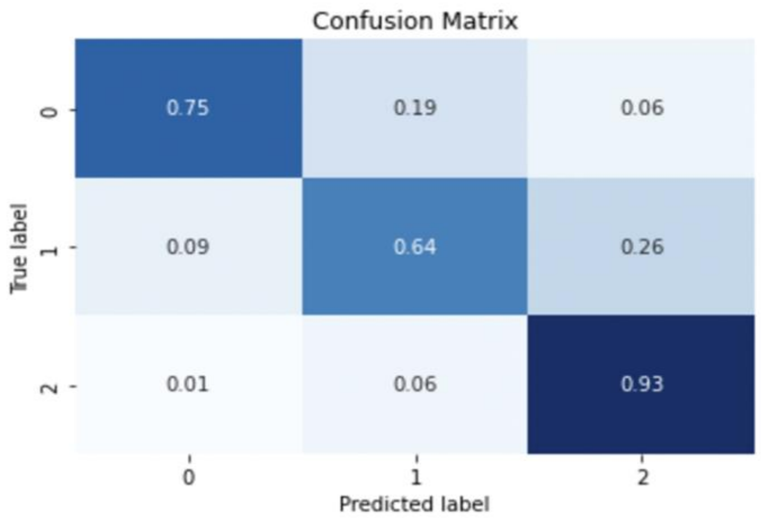


Figure 4-15 F1 Score

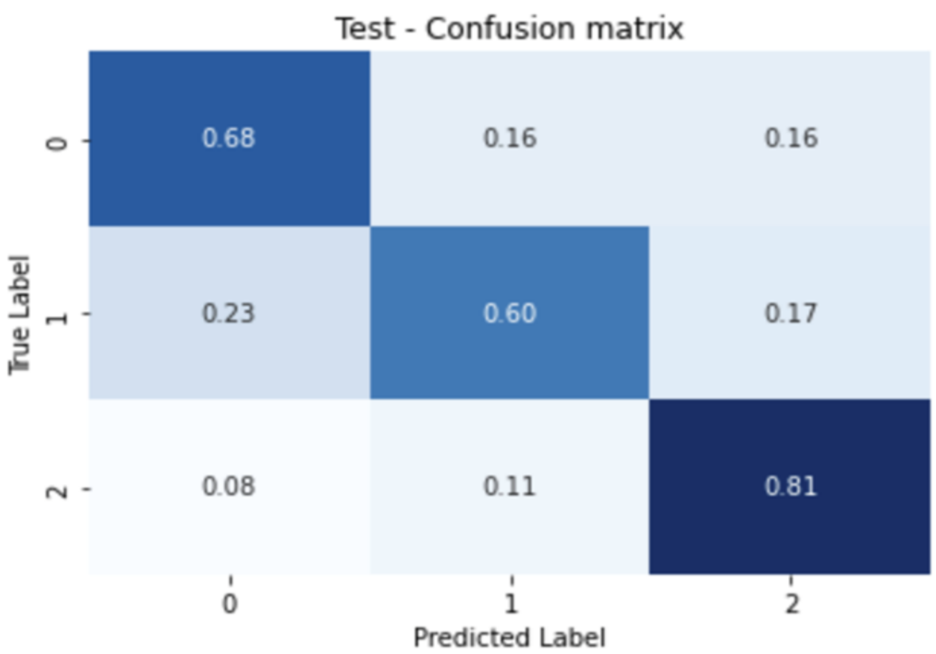
b) Confusion Matrix

The confusion matrix for both models- Deep Learning model and Logistic Regression model are shown below. The diagonal elements represents the correct predictions i.e the actual and predicted values by the model are same. Further analysis is carried out in the subsequently taking in consideration both technical and business aspect.



DEEP LEARNING MODEL

Figure 4-16 Deep Learning



LOGISTIC REGRESSION MODEL

Figure 4-17 LR model

c) Cost Comparison

The deep learning model have an F1-Score of 0.78 while Logistic Regression model has a score of 0.67 i.e 15% higher 'F1-Score' in comparison to Logistic Regression model. But, we need to keep the business aspect in mind. Cost is also an important factor which needs to be considered. Deep learning models need GPU to run which are very expensive in comparison to CPU, which are used for ML models. Keeping this is mind, The Logistic Regression model is a better model to be implemented in this business cause we need to analyse high volume of comments in trillions. We need a model which gives a good evaluation metric at very low cost such as the Logistic Regression model.

Cost= Device_price * time_consuming

	XLNET	LR
Device_price	GPU= $15/h	CPU= $0.15/h
Time_consuming	60s	32ms
Cost	**0.25**	**0.000001**
F1-score(Reward)	**0.78**	**0.67**

Figure 4-18 Cost Analysis

d) Classification Report of Logistic Regression model

The classification report using Logistic Regression model on both the train and test set is shown. The model is performing well on both train and test. Hence, there is no problem of overfitting or underfitting. Thus concluding that Logistic Regression model is used to predict the sentiment of new comments.

```
Classification report on train set
              precision    recall  f1-score   support

         0.0       0.79      0.99      0.88      1753
         1.0       0.75      0.93      0.83      3479
         2.0       0.98      0.84      0.91      8574

    accuracy                           0.88     13806
   macro avg       0.84      0.92      0.87     13806
weighted avg       0.90      0.88      0.88     13806

Classification report on test set
              precision    recall  f1-score   support

         0.0       0.45      0.68      0.54       585
         1.0       0.62      0.60      0.61      1159
         2.0       0.89      0.81      0.84      2858

    accuracy                           0.74      4602
   macro avg       0.65      0.70      0.67      4602
weighted avg       0.76      0.74      0.75      4602

CPU times: user 18 s, sys: 3.73 s, total: 21.7 s
Wall time: 3.58 s
```

Figure 4-19 Training and Testing Result

4.4. Topic Model

Topic modeling is a type of statistical modeling tool which is used to assess what all abstract topics are being discussed in a set of documents. Topic modeling, by its construction solves the problem of creating topic in an unsupervised manner.

This model is used to extract keywords for each comment, which is shown in the figure below, not only extracting the keywords for each comment, but also providing the text data in TOKEN units for the subsequent viewer clustering analysis. Figure below shows the steps of the model building process.

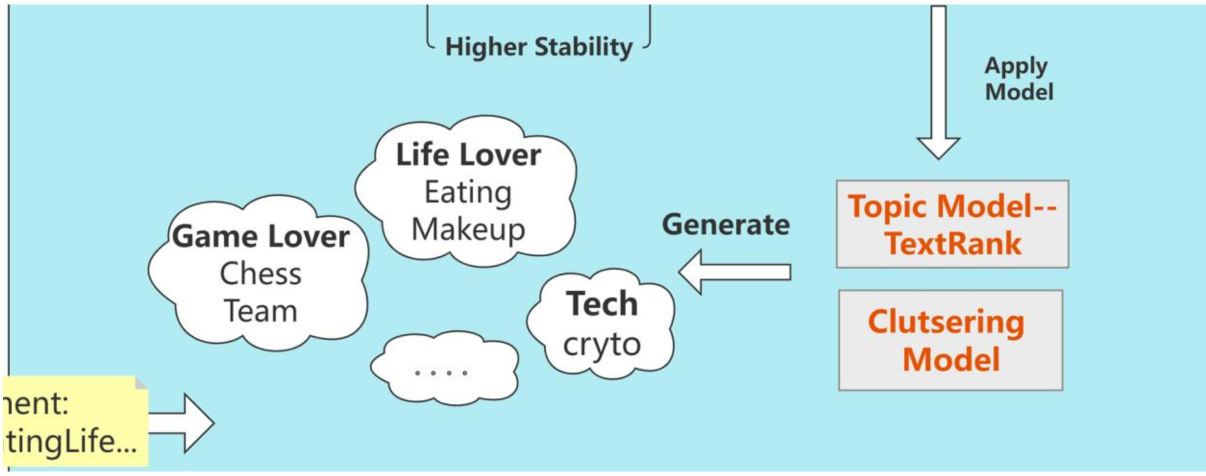


Figure 4-20 Topic Model WorkFlow

4.4.1 Steps & Model Selection

Data from the Postive Sentiment Comments dataset was used and keywords were extracted using TextRank's SUMMA method.

TextRank – is a graph-based ranking model for text processing which can be used in order to find the most relevant sentences in text and also to find keyword

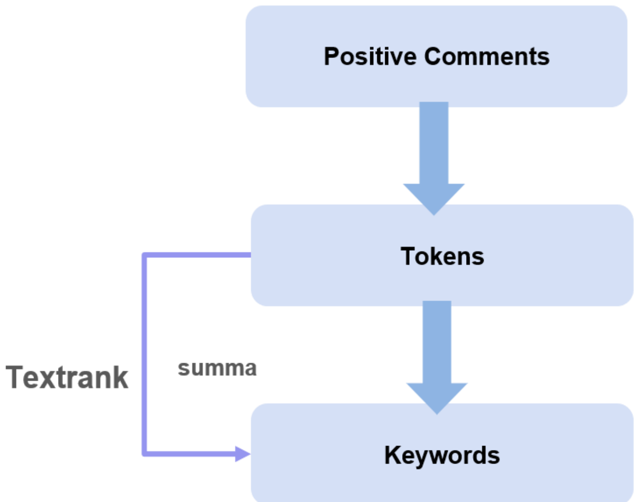


Figure 4-21 TextRank

4.4.2 Results

The following graph shows some of the results of the prediction after the Topic model, you can see that the topics of the 10 comments generated by each video have been extracted

```
df_comment.to_csv('data/comment_keyword_true_tag.csv')
df_comment['Comment_keyword'][0:10]
```

```
Video ID
--ZIOdSbbNU    [eating, eat, eats, fry, feel, smile, man, suf...
--hxd1Cr0qg    [russia, til, n, wo, realize, ukraine, analyza...
--ixiTypG8g    [college, colleges, tuition, regulated, regula...
-64r1hcxtV4    [life, inspiration, inspiring, inspire, health...
-6IgkG5yZfo    [thankful, thank, thanks, time, help, helped, ...
-7hzaGya86g    [good, like, mikey, dinner, food, eating, dump...
-8TnsjDRXUE    [kid, amazing, know, knowing, guy, time, work,...
-9hjdvULDyc    [reading, prediction, th, nd, wa, writing, tes...
-Cr69sGnrk8    [doe, u, fan, thank, thanks, sccoby, support, ...
-D4S6Tpn044                                          [getbonus]
Name: Comment_keyword, dtype: object
```

Figure 4-22 Keyword Generating

4.5. Clustering Model

However, it is difficult for us to draw conclusions with a large number of keywords, so we will make a cluster analysis model. The graph below shows Clustering Model's position in the entire WorkFlow, by receiving the keywords generated by Topic Mode and send them into clustering model to achieve the object of classify the viewer who generate this comments.

K-means is an unsupervised clustering algorithm designed to partition unlabelled data into a certain number (thats the " K") of distinct groupings. In other words, k-means finds observations that share important characteristics and classifies them together into clusters. A good clustering solution is one that finds clusters such that the observations within each cluster are more similar than the clusters themselves.

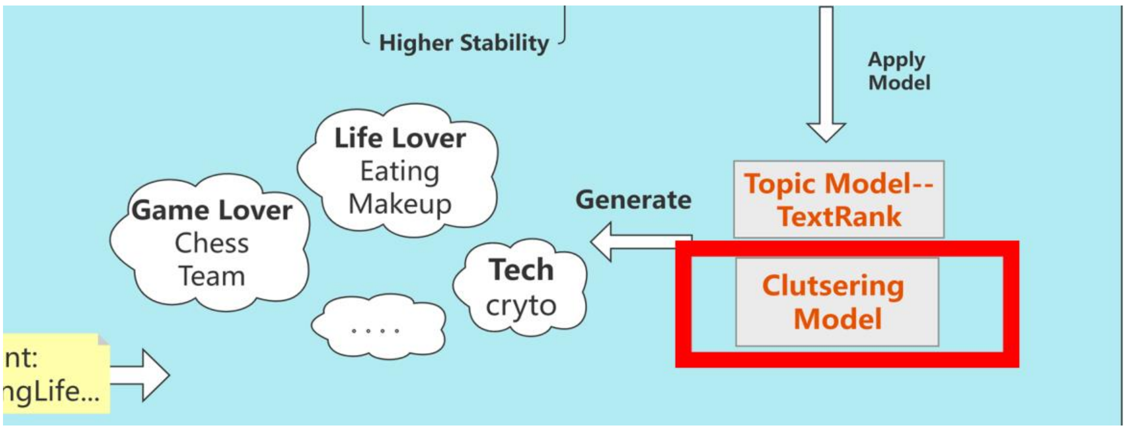


Figure 4-23 Clustering WorkFlow

4.5.1 Find the optimal value of K

The number of clusters (k) is the most important hyperparameter in K-Means clustering. In this report we will cover two such methods:

a) Elbow Method

The elbow method involves finding a metric to evaluate how good a clustering outcome is for various values of K and finding the elbow point. Initially, the quality of clustering improves rapidly when changing the value of K but eventually stabilizes. The elbow point is the point where the relative improvement is not very high anymore. This is shown pictorially in the two graphs below for the metric average within-cluster sum of squares. The elbow point shown below is not obvious, so the Silhouette Method was used to find the best K value

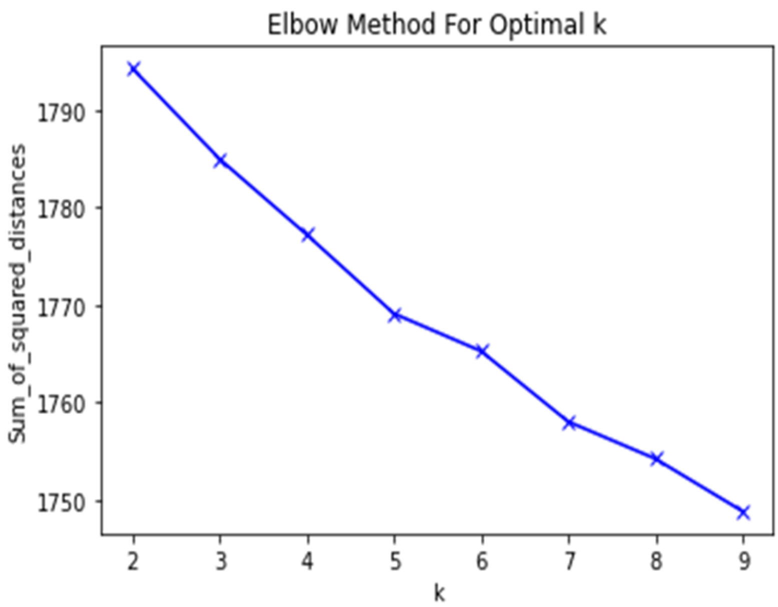


Figure 4-24 Elbow Method

b) Silhouette Method

Better than Elbow Method to find Optimal Clusters

The silhouette Method is also a method to find the optimal number of clusters and interpretation and validation of consistency within clusters of data. The silhouette method computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters. by providing a succinct graphical representation of how well each object has been classified.

As you can see from the graph below, K = 5 is the best K.

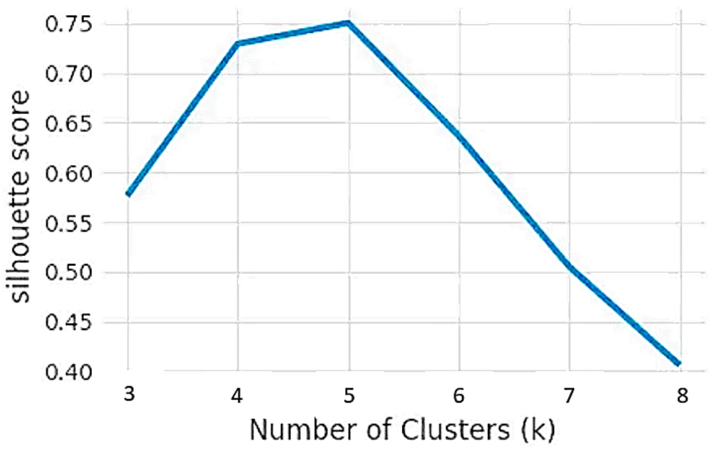


Figure 4-25 Silhouette Method

4.5.2 Clustering Result

The results of our clustering are visualised via WordCloud and are shown below. Assign labels to each clustering category as 1.GAME 2.MUSIC 3.LEARNING 4.TECHNICAL 5.LIFE These five categories are the result of our classification of active users based on their comments.

Figure 4-26 5 Viewers' Clusters

Specifically:

1. Users classified as GAME have a greater interest in videos in the GAME category. Combined with WordCloud, it can be seen that Chess, Magnus, Goal-Oriented,Hit type games have a larger user base and user activity.

Figure 4-27 Cluster 0

2. Users categorised as MUSIC have a greater interest in MUSIC videos, and combined with WordCloud analysis, we can see that MUSIC music in the playlist, para genre has a greater user base and user activity.

Figure 4-28 Cluster 1

3. Users classified as LIFE have a greater interest in videos in the LIFE category, which, when combined with WordCloud analysis, is focused on Animal, food and makeup.

Figure 4-29 Cluster 2

4. Users in the LEARNING category have a greater interest in online learning videos, as most of them are students and have more positive feedback when they gain knowledge in videos, making them the most viscous category of viewer. It is worth noting that the MATH category has a particularly high level of interest, which means that videos in the MATH category have the largest user base of any discipline and could be a focus for the company in the learning area.

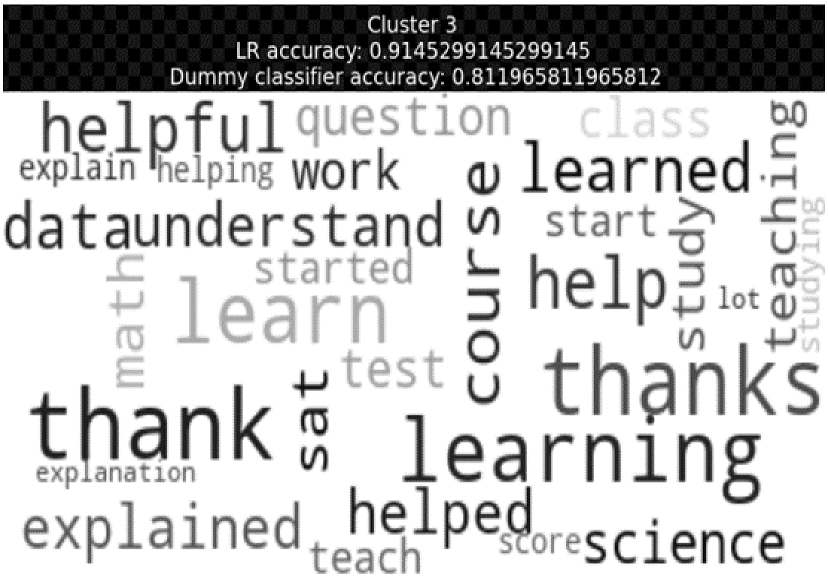


Figure 4-30 Cluster 3

5. Users in the TECHNICAL category have a greater interest in videos in the TECHNICAL category, which combined with WordCloud analysis shows that there is a greater interest in Tech videos about Crypto, which is in line with the current trend in Tech and means that Youtube videos can reflect trends in technology. Keeping up with current events and uploading videos to Youtube is a key focus for the company in the TECH sector.

Figure 4-31 Cluster 4

5. Strategic Planning

As an important strategic planning model which focuses on the enhancement of the basic plan, Issue-Based Model is mainly used by companies with limited resources but diverse problems and wish to go deeper into strategic planning. Issue-Based Model is a more dynamic and fluid conceptual model, because it helps to identify various both internal and external factors and evaluate problems that affect the company's operation and achieve various goals. Besides, reviewing the company's mission, vision, and guiding principles is the beginning stage of this model. This ensures deeper understanding and brainstorming to resolve business issues.

BigFame as an MCN company facing heavy operational problems. BigFame had to make a change in its strategy. In order to more accurately define the business situation in which the company operates. A SWOT analysis was conducted.

Table 5-1 SWOT Analysis

Strengths	Weaknesses
• **Reputation:** Famous MCN company • **Resource:** Long-term and deep cooperation with TikTok video platform • **Host:** Rich resources of hosts • **Organizational Structure:** Well-organized company with clear division of responsibilities in each department • **Technology:** Enough technical support to complete business transformation	• **Business Lines:** Too many resources spread across too many business lines • **Host Management:** Lack of attraction and management of talented video creators • **Content Creation:** Inability to keep track of video trends and lack of creative video creation
Opportunities	**Threats**
• **Industry:** Lack of distinctive MCN agencies, most companies are comprehensive business lines • **Advertising:** Advertisers pay attention to the real impact of advertising on users and emphasize the real and effective ROI • **Investment:** Many investment institutions have started to study MCN companies, trying to analyze the profits that MCNs can bring from multiple perspectives. • **Viewers:** Viewers' viewing habits are changing and they are more demanding on content quality	• **Industry:** High competitive pressure in the market, a large number of MCN agencies established • **Costs:** High labor costs and difficulty in realizing traffic • **Platform:** bottleneck in increasing the number of fans • **Capital:** Capital requires MCN companies to be sustainable, and the profitability of MCN agencies is measured from multiple perspectives. • **Politics:** MCNs in various countries lack policy support

The application of the SWOT Analysis model shows that under the trend of video popularity, BigFame, as a company with a comprehensive business line of channel production, has a strong foundation for business transformation. More and more capital is starting to focus on the development of MCN agencies. If BigFame can successfully implement a differentiation strategy, it will attract significant financing to help grow. has Since it has already cooperated with TikTok and TikTok's big data recommendation level has greatly improved in recent years, BigFame is likely to achieve profit growth.

In addition, the SWOT analysis shows that BigFame has some shortcomings and external threats. The comprehensive line of business consumes a lot of capital costs and human costs, resulting in the company not paying enough attention to the anchors and not being able to give them specific guidance and resources to help them. With the public demanding a more stringent quality of video content, the lack of creativity in BigFame content creation and the difficulty in tracking the trend of content creation in each category in the market became an urgent problem. The lack of management of video content has also created the problem of no way to establish close long-term cooperation with advertisers. The emergence of more and more MCN agencies has further increased competition in the market. The lack of policy support makes the development of MCN agencies even more difficult.

Table 5-2 Current Problems & Strategies

Current Problems	**Strategies**	**Model be used**	**Department followed**	**Timeline**	**Budget ($)**
Host Management: Lack of quality talent	• Contract anchors who can attract an audience interested in gaming, knowledge, music, finance and life according to the company's audience targeting	Text analysis models	Financial, Host Management	3 months	100 million
Content Creation: The content is seriously homogenized. Lack of creative content ideas.	• According to different viewers clusters and popular video categories, enlighten the creators to create special video (eg. Entertainment video: more eating, animal, makeup contents)	Time series analysis; text analysis model	Content	6 months	10 million
Competitive Pressure: KOL erosion. The problem of professionalization and contract	• Pay more attention on KOL's training. Provide them more technical and investment support	Time series analysis; text analysis model	Operation	1 year	No Budget

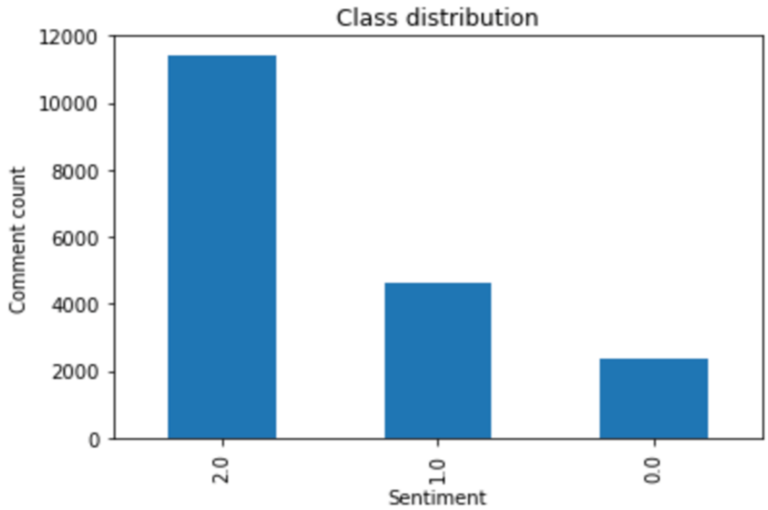
Platform Pressure: More competitors and the reputation get threatened	• Save costs by releasing other small anchors in other areas from their contracts • By shrinking the business lines, improve other popular categories performance. • By marketing and IP building to increase the reputation	Time series analysis	Marketing	6 months	50 million

6. Limitations

For this project, we have 3 limitations:

1. Sentiment analysis have long tail class distribution

In the natural case, the data tend to show the same long-tail distribution as follows. This trend is also seen in various problems in various fields from natural sciences to social sciences, cf. Zipf's Law or what we often refer to as the 28 law. Classification and recognition systems trained directly on long-tail data tend to over-fit the head data and thus ignore the tail categories in their predictions. In this case, the long tail class distribution of sentiment, positive sentiment accounts for the majority of the training samples, which leads to overfitting problems for positive comment predictions and poor performance for negative comment predictions in the LR model trained with this dataset.



2. Some results of ARIMA have extremely low AIC or SBC

For the HUMANITIES topic, the result of the Randomness Test shows that all the P-Values are larger than 0.05 which means it is a white noise series. Due to this result, we can not come up with any business strategy, so as one of the limitations, to address this problem, we determine to recategorize these topics.

For the SCIENCE topic, we noticed both AIC and BIC are very small, overfitting may exist in this model, so as a limitation, we may have to normalize the popularity score to make it a better performance.

3. Not practice massive data into this process to test its stability

The entire WorkFlow currently available can perform well on collected data sets, but is not practiced in a large number of industrial scenarios. However, the system is not optimized for communication and memory capacity within the system when inputting large amounts of text data, which can lead to inefficiencies at runtime.

7. Project Management

7.1. Project Assumptions

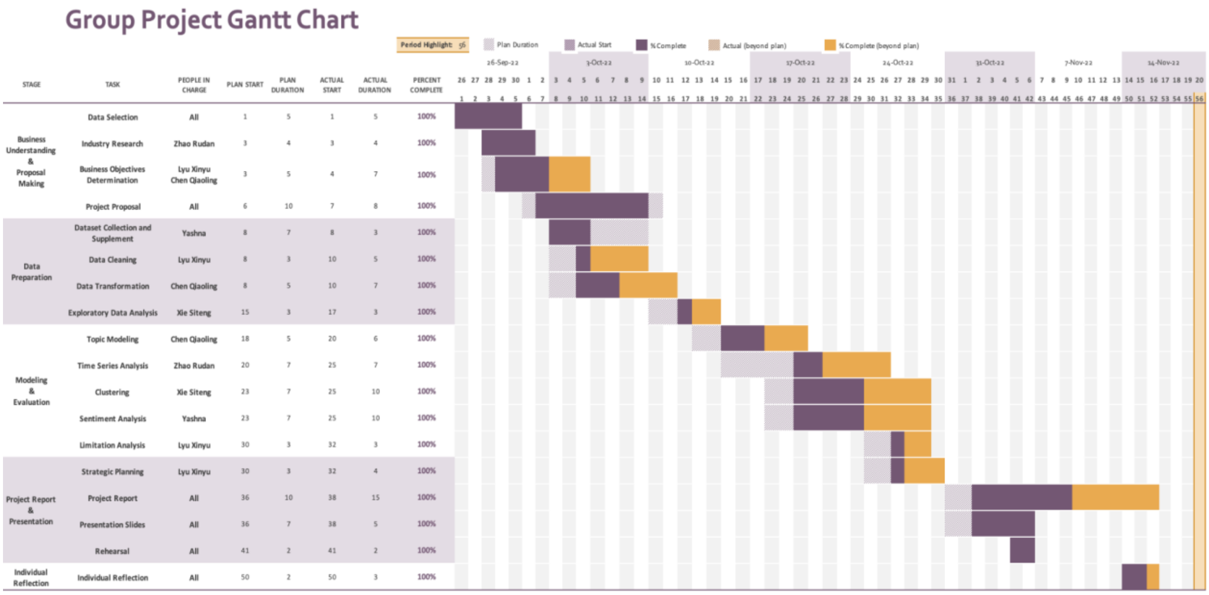
- **Human Resources:**
  Respect's team members are supposed to be sufficient during the project, so that the personnel's shortage leading to the project bankruptcy will not exist. Also, the efficiency of Respect's team members can be guaranteed, and there are no delays in delivery.
- **Equipment:**
  Respect's equipment remains stable during the whole project, taking on the data analysis and visualization work within the scope of the project.
- **Project Scope:**
  The scope of the project will change based on the real situation. But Respect will retain sufficient budget to accommodate these changes.
- **Information Integrity:**
  The Respect's project team will have access to any project-related information from all sides during the project.

7.2. Project Roles and Responsibilities

Table 7-1 Project Roles and Responsibilites

Role	Responsibilities	Participant(s)
Project Sponsor	● Provide necessary funds and resources for the project ● Communicate and express views with other project stakeholders on progress and success factors ● Provide guidance on the current or future value and relevance of the project ● Participate in the review of project implementation requirements ● Sign and approve project closure	Dr. Charles
Project Managers	● Communicate with project sponsor ● Check project quality ● Communicate and coordinate personnel and resources ● Control project schedule	QiaoLing Chen
Project Participants	● According to the project division, complete each task ● Complete project report on time ● Make presentation ● Timely report project progress and problems ● Engage and collaborate with other members as part of an effective team ● Understand the goals and objectives of the project	Respect's Project Team

7.3. Gantt Chart



STAGE	TASK	PEOPLE IN CHARGE	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Business Understanding & Proposal Making	Data Selection	All	1	5	1	5	100%
	Industry Research	Zhao Rudan	3	4	3	4	100%
	Business Objectives Determination	Lyu Xinyu Chen Qiaoling	3	5	4	7	100%
	Project Proposal	All	6	10	7	8	100%
Data Preparation	Dataset Collection and Supplement	Yashna	8	7	8	3	100%
	Data Cleaning	Lyu Xinyu	8	3	10	5	100%
	Data Transformation	Chen Qiaoling	8	5	10	7	100%
	Exploratory Data Analysis	Xie Siteng	15	3	17	3	100%
Modeling & Evaluation	Topic Modeling	Chen Qiaoling	18	5	20	6	100%
	Time Series Analysis	Zhao Rudan	20	7	25	7	100%
	Clustering	Xie Siteng	23	7	25	10	100%
	Sentiment Analysis	Yashna	23	7	25	10	100%
	Limitation Analysis	Lyu Xinyu	30	3	32	3	100%
Project Report & Presentation	Strategic Planning	Lyu Xinyu	30	3	32	4	100%
	Project Report	All	36	10	38	15	100%
	Presentation Slides	All	36	7	38	5	100%
	Rehearsal	All	41	2	41	2	100%
Individual Reflection	Individual Reflection	All	50	2	50	3	100%

Figure 7-1 Project Gantt Chart

7.4. Issue Definition & Solution

1. Time Issue

Team communication time: Meeting at 8:00 AM every Monday to communicate the project progress and ensure that the team can timely understand the project progress.

2. Technique Issue

In case of data analysis technical problems, such as merging the visual created by the different versions of Tableau, language setting, and deliverables submission, we will use Tableau Public to public our data analysis result and share the link with others.

3. Resource Issue

The team will manage the backup of the work product at any time, so as not to lose files due to equipment problems such as the computer breaks down.

If there is a problem with the equipment, the team immediately activates backup equipment to ensure the project is completed.

If there isn't enough time to re-create a new file, the project manager will explain the situation to project sponsor and make a Plan B to finish the task as quickly as possible.

5. Team Management Issue

If a team member is unable to participate in the project immediately, the division of labor will be adjusted, and another member will complete the task in advance. Then this member will complete other tasks of the same workload to make up for the absence. And tasks will be assigned based on task priorities. So that we can ensure balance within the team.

6. Project Change Issue

To solve the problems which occurred during the completion process of tasks, after the assignment of the project's tasks timely. We created a Google shared document to record the scope change. It includes various fields as below:

Table 7-2 Change Management

Requested By	Person who discovers the problem need to be solved or the change need to be made
Related Person	People who will affected by the changing or can help solve the problem
Date	Date the problem discovered
Meeting	Whether need for meetings to solve problems together
Priority	How important is the change? Circle whether it is high, medium or low
Task Name	The name of the task associated with the issue
Problem Description	Problem details
Solution	How the change be made or the problems be solved If not solved, ignore this field
Task/Scope Affected	Task/Scope Affected: What tasks will be impacted to make this change? Estimate how the scope of the project will be impacted by the change.
Recommendations	What others have said about the issue

7.5. Communication Plan

The communications methodology utilizes 2 directions for effective communication:

- Middle-In:
  Monday meeting, occasional email notifications, and social media discussion groups.
  The inclusion of all members, their ideas, and their perceptions of the business in order to make the most informed decisions for the project and the company. Members are able to communicate about their work process and current issues, and coordinate work load.

- Middle-Out:
  Email: Ask problems and make sure our work progress is in line with the requirement according to the project sponsor.
  Report: Project content and specific schedule and submit it by CANVAS.
  PPT presentation: Presentation of project results.

Besides, considering that our team members have different culture background, we need to make sure everyone understand the task details and project objectives effectively and clearly. And we made the following template to record each project team status meeting to track the progress of the project over time.

Table 7-3 Meeting Record

ID	Main Topic	Sponsor	Participant	Method	Output	Time	Place	Meeting Summary
1	Scope Meeting	Project Manager	Respect's Project Team	Face to face (best), Online	Project Scope	November 12th	ZOOM Meeting Room	1. XXXXXXXXXX 2. XXXXXXXXXX

8. Reference (APA 6th)

BaiZhun. (2022, July 11). 2022，The dilemma and breakthrough of MCN commercialization. Sina. Retrieved November 13, 2022, from https://t.cj.sina.com.cn/articles/view/3284109041/c3bf86f100100zb3m

Guo, j. (2021). *Research On The Marketing Innovation Of "Live + Short Video" In The Culture And Tourism Industry In We Media Era*. EDP Sciences. https://doi.org/10.1051/e3sconf/202125103036

Iimedia. (2022, July 20). MCN Industry Development Research Report (2022-2023). Retrieved November 1, 2022, from https://www.iimedia.cn/c400/87027.html

TopKLOut. (2020). 2020 China MCN Industry Development Research. CBNData. https://chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/pdf.dfcfw.com/pdf/H3_AP202005081379467432_1.pdf

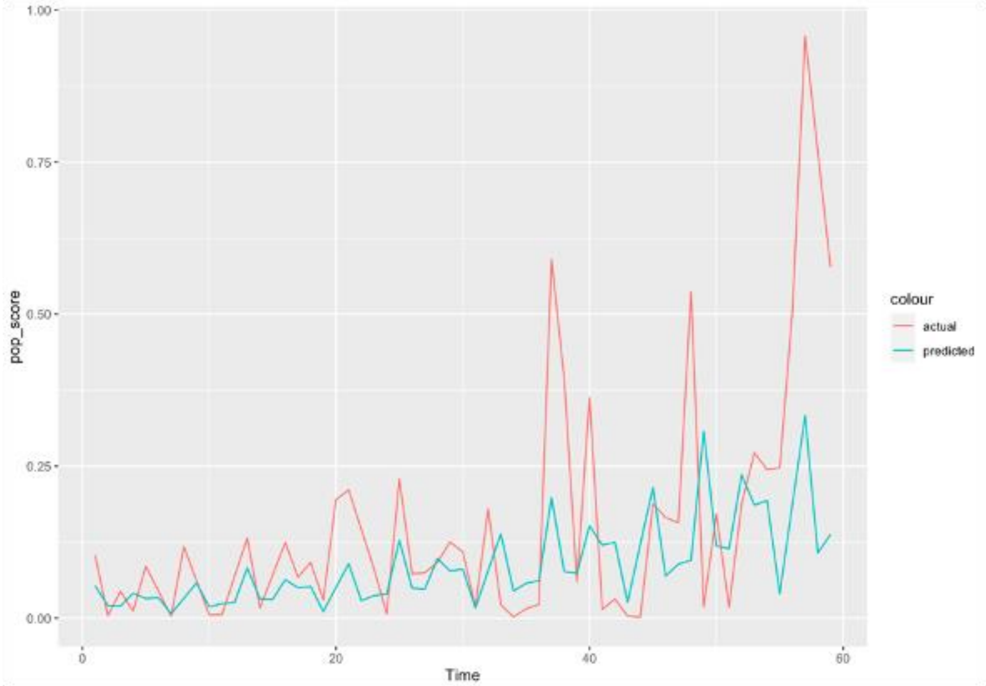
The Wrapbook Team. (2022, March 1). How to create a multi-channel network: Use MCNs to grow your client list. Wrapbook. Retrieved November 16, 2022, from https://www.wrapbook.com/blog/multi-channel-network

YouTube. (n.d.). Multi-channel network (MCN) overview for YouTube creators. https://support.google.com/youtube/answer/2737059

9. Appendix

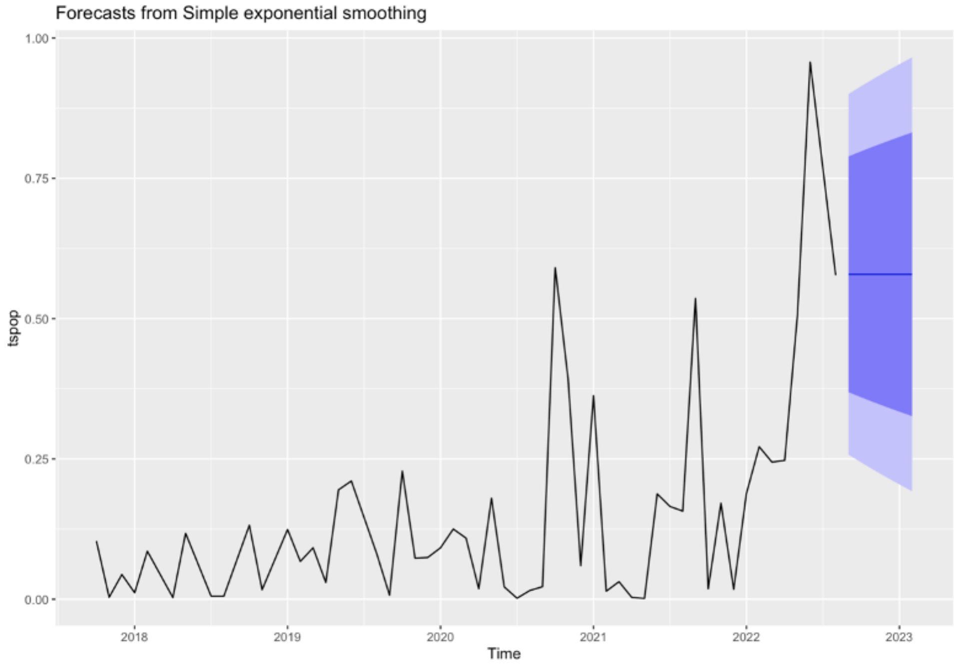
Other Models used in Time Series Analysis Taking ENTERTAINMENT as Example.

- **Time Series with Seasonal Variation**



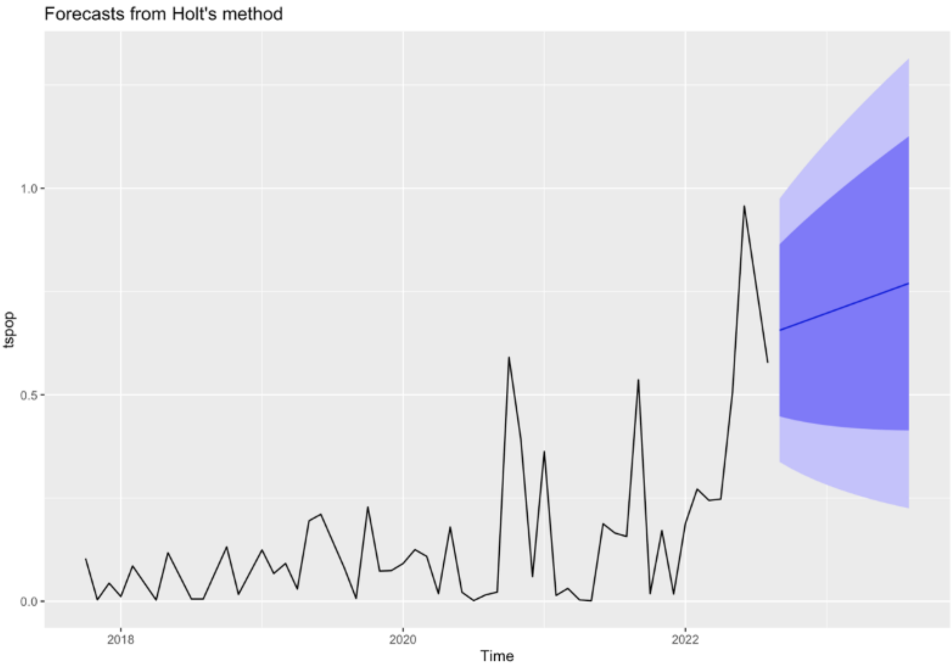

```
Call:
lm(formula = data$transY ~ Time + Month, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4633 -0.9286  0.3930  0.8720  1.9692

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5.22072    0.77343  -6.750 2.16e-08 ***
Time            0.03651    0.01183   3.086  0.00343 **
MonthAugust     1.08325    0.96413   1.124  0.26703
MonthDecember   1.19025    0.96413   1.235  0.22327
MonthFebruary   1.60404    0.96326   1.665  0.10267
MonthJanuary    1.87476    0.96362   1.946  0.05784 .
MonthJuly       0.86879    0.96362   0.902  0.37198
MonthJune       2.03918    0.96326   2.117  0.03970 *
MonthMarch      1.60545    0.96304   1.667  0.10230
MonthMay        1.50289    0.96304   1.561  0.12548
MonthNovember   1.26219    0.96479   1.308  0.19728
MonthOctober    2.25115    0.96558   2.331  0.02417 *
MonthSeptember  1.11478    1.02145   1.091  0.28079
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.523 on 46 degrees of freedom
Multiple R-squared:  0.2715,	Adjusted R-squared:  0.08145
F-statistic: 1.429 on 12 and 46 DF,  p-value: 0.1877
```

- **Simple Exponential Smoothing**



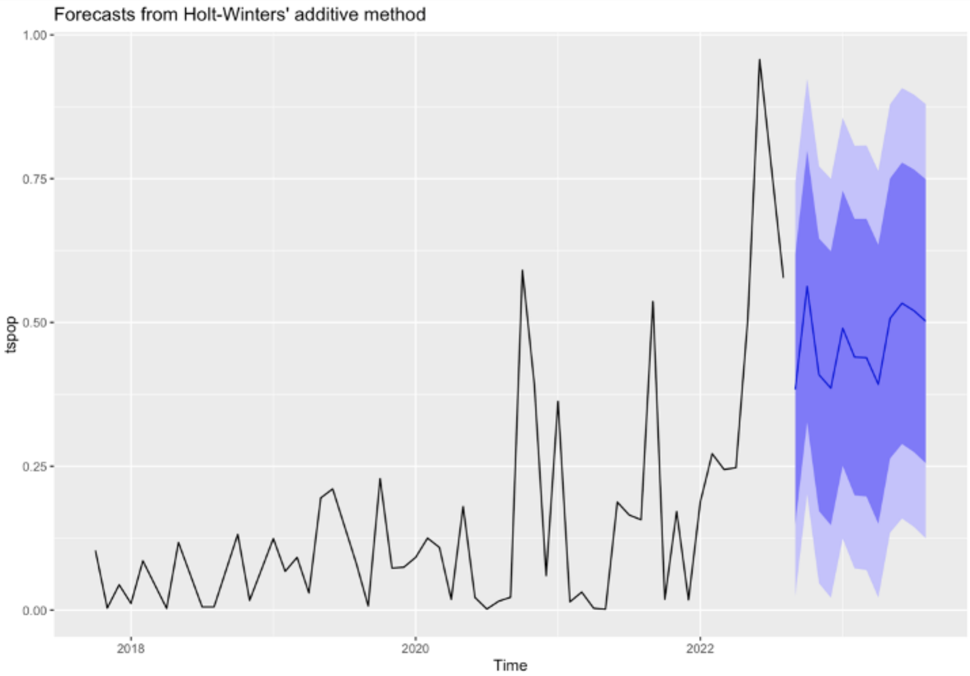

```
Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing

Call:
 ses(y = tspop, h = 6, alpha = 0.3)

  Smoothing parameters:
    alpha = 0.3

  Initial states:
    l = 0.0537

  sigma:  0.1638

     AIC     AICc      BIC
29.07468 29.28897 33.22976

Error measures:
                     ME      RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set 0.02967662 0.1610128 0.103945 -471.2293 515.9934 0.6233589 0.1896714

Forecasts:
         Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
Sep 2022      0.5789843 0.3690492 0.7889194 0.2579163 0.9000523
Oct 2022      0.5789843 0.3598056 0.7981630 0.2437795 0.9141892
Nov 2022      0.5789843 0.3509364 0.8070322 0.2302152 0.9277535
Dec 2022      0.5789843 0.3423995 0.8155692 0.2171590 0.9408096
Jan 2023      0.5789843 0.3341600 0.8238086 0.2045579 0.9534108
Feb 2023      0.5789843 0.3261890 0.8317797 0.1923672 0.9656014
```

- **Holt's Method**




```
Forecast method: Holt's method

Model Information:
Holt's method

Call:
 holt(y = tspop, h = 12)

  Smoothing parameters:
    alpha = 0.4174
    beta  = 1e-04

  Initial states:
    l = 0.0337
    b = 0.0103

  sigma:  0.1626

     AIC     AICc      BIC
32.07511 33.20719 42.46280

Error measures:
                       ME      RMSE       MAE       MPE     MAPE      MASE       ACF1
Training set 0.0001713629 0.1569714 0.1066961 -566.6153 599.7491 0.6398573 0.07683088

Forecasts:
         Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
Sep 2022      0.6560875 0.4477337 0.8644413 0.3374379 0.9747371
Oct 2022      0.6664374 0.4406568 0.8922180 0.3211358 1.0117390
Nov 2022      0.6767873 0.4348241 0.9187506 0.3067365 1.0468382
Dec 2022      0.6871372 0.4300005 0.9442740 0.2938805 1.0803940
Jan 2023      0.6974872 0.4260167 0.9689576 0.2823090 1.1126654
Feb 2023      0.7078371 0.4227461 0.9929280 0.2718281 1.1438460
Mar 2023      0.7181870 0.4200909 1.0162831 0.2622884 1.1740856
Apr 2023      0.7285369 0.4179738 1.0391000 0.2535716 1.2035022
May 2023      0.7388868 0.4163323 1.0614413 0.2455823 1.2321914
Jun 2023      0.7492367 0.4151153 1.0833582 0.2382421 1.2602314
Jul 2023      0.7595867 0.4142801 1.1048933 0.2314858 1.2876875
Aug 2023      0.7699366 0.4137906 1.1260825 0.2252584 1.3146148
```

- **Holt's-Winters' Method**




```
Forecast method: Holt-Winters' additive method

Model Information:
Holt-Winters' additive method

Call:
 hw(y = tspop, h = 12)

  Smoothing parameters:
    alpha = 0.0969
    beta  = 1e-04
    gamma = 1e-04

  Initial states:
    l = -0.02
    b = 0.0073
    s = -0.0401 -0.0018 0.0238 0.044 0.0251 -0.0823
           -0.0286 -0.0203 0.0372 -0.0597 -0.029 0.1317

  sigma:  0.1832

     AIC     AICc      BIC
55.62372 70.55055 90.94186

Error measures:
                      ME      RMSE       MAE       MPE     MAPE      MASE      ACF1
Training set 0.001513034 0.1563721 0.1057409 -682.7215 797.5373 0.6341293 0.3725242

Forecasts:
         Point Forecast     Lo 80     Hi 80      Lo 95     Hi 95
Sep 2022      0.3835158 0.1487758 0.6182559 0.02451197 0.7425197
Oct 2022      0.5625017 0.3266597 0.7983437 0.20181254 0.9231908
Nov 2022      0.4091392 0.1721981 0.6460803 0.04676907 0.7715093
Dec 2022      0.3857127 0.1476753 0.6237502 0.02166590 0.7497596
Jan 2023      0.4899879 0.2508568 0.7291189 0.12426848 0.8557073
Feb 2023      0.4397713 0.1995493 0.6799932 0.07238346 0.8071591
Mar 2023      0.4387715 0.1974613 0.6800818 0.06971939 0.8078237
Apr 2023      0.3923980 0.1500021 0.6347939 0.02168546 0.7631105
May 2023      0.5070929 0.2636139 0.7505718 0.13472397 0.8794618
Jun 2023      0.5333504 0.2887909 0.7779099 0.15932897 0.9073718
Jul 2023      0.5204619 0.2748244 0.7660994 0.14479177 0.8961320
Aug 2023      0.5021359 0.2554229 0.7488490 0.12482093 0.8794510
```