EBA5002: Graduate Certificate in Business Analytics Practice

Prediction of YouTube video trend for BigFame company

13 October 2022
Team: TMID
Qiaoling Chen: e0983208
Lyu Xinyu: e0983260
Xie Siteng: e0983376
Yashna Gogineni: e0983387
Zhao Rudan: e0983390

1. Introduction

Big Fame is a multi-channel network (MCN), which was born within YouTube and its ecosystem which generally aims at profiting by assisting YouTubers in making an income. We provide services to video creators including audience development, content programming which is an important revenue stream for us. Therefore, we need to tap into the insight of popular videos to provide these services to increase the company revenue. With this goal, our BA team is trying to define trending rules, build a time series model, sentiment analysis model and clustering model based on the data sets provided by YouTube to predict trends, extract audience features and provide creative suggestions for video creation as needed.

2. Business Problem & 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

3. Project Scope and Design

a) Project Data

- **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. We've combined two tables into one which contains 18646 rows and 10 columns.
- **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.
- **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

- **Potential challenges:**

1) We have two separate datasets that store the statistics of the video and ten comments for each video. It was a challenge to bring the two data sets together and perform a joint analysis.
2) The comment text data within the dataset contains different languages, and which encoding method to use to extract the tags is a strategic issue to consider.
3) This dataset has three columns that are suitable as dependent variables and it is a question of how to assign weights to these columns to calculate a completely new dependent variable for statistical analysis.

b) Analytical Methods:

Predictive models & techniques	Relevance and Effective	Evaluation
• Do token frequency analysis to find top5 token named tags. • Calculate the popular score for each video based on the number of views, comments, likes.	• Apply these tags and score into datasets as a separated column for comments in each video.	----------
• Constructing an ARIMA model will be used to predict the popular score in the coming year for each category.	• Comparing the predicted popular score, we will have an insight of the coming trend.	RMSE MAPE MAE
• Apply BERT model to do sentiment analysis for comments in each video.	• By looking at the sentiment of the comments on the video, we can give advice on the direction of future video production.	Confusion matrix
• Clustering the tags of videos to find some hidden features of viewers.	• With the outcomes, we can group viewers by recognizing the feature	Davies-Bouldin Index

c) Critical Success Factors

Project:

Clear Project Plans	Creating clear plans that document clear deliverables. Define how we will achieve them and by when will go a long way to securing success.
Specific, Measurable, Attainable Deliverables	Deliverables should always be realistic, measurable against KPIs, and specific.
Professional Software	Choose the right tools to make sure we can see the result in an effective way.
Risk Management	We need to plan for any kind of risk. If there are any emergencies, we need to change our timeline or project scope immediately.
Monitoring & Control	Check the progress and regularly evaluate the results. Communicate with the professor and make sure everything is on the track.

Team:

Clear Collaboration and Communication	With proper communication, many undesirable developments in a project can be prevented or at least detected earlier
Models Familiarity	Everyone in the team should be familiar with all the models we have learned. So that we can choose the most proper model to achieve our business objectives

4. Key Deliverables

a) Dashboard- This is an effective method for data visualization. This helps stakeholders understand and leverage business intelligence to make more informed decisions

b) Word Cloud- The word cloud is developed to understand the most frequently used tags in the comment section. This helps to understand the audience perspective of the video and accordingly taken some actions for the future.

c) Popularity Score Generator- The is a score generated based on number of likes, comments, and views. A popularity score is going to be generated for every video using mathematical model such as AHP/ TOPSIS

d) ARIMA Univariant Times Series model- This 'Autoregressive Integrated Moving Average' machine learning model is useful to predict the future popularity score of each category

e) Classification model for Sentiment Analysis- This supervised machine learning model is useful for predicting the sentiment of the viewers using the existing data collected. The viewer sentiment is classified into negative, neutral and positive

f) Clustering model- This unsupervised machine learning model is built to cluster the tags to help identify similar patterns or groups and thereby achieve some actionable insights.

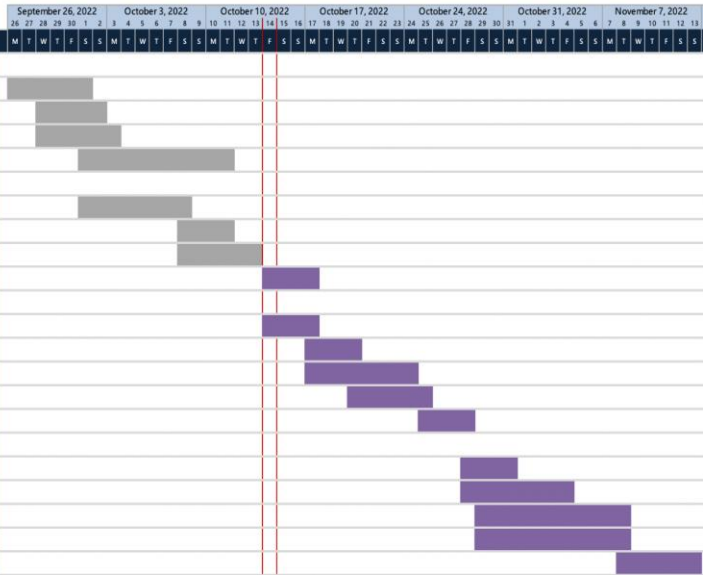
5. Effort Estimates and Timeline

Group 16: Lyu Xinyu, Qiaoling Chen, Xie Siteng, Yashna Gogineni, Zhao Rudan

Start Date: 26-Sep-2022

Week Number: 1

TASK NAME	Person in Charge	PERCENT COMPLETE	START DATE	END DATE	DURATION (Days)
Business Understanding and Proposal Making					
Data Selection	All	100%	26/09/22	01/10/22	5
Industry Research	Xie Siteng	100%	28/09/22	02/10/22	4
Business Objectives Determination	Lyu Xinyu	100%	28/09/22	03/10/22	5
Project Proposal	Qiaoling Chen	100%	01/10/22	11/10/22	10
Data Understanding & Preparation					
Data Collection and Supplement	Yashna Gogineni	100%	01/10/22	08/10/22	7
Data Integration	Zhao Rudan	100%	08/10/22	11/10/22	3
Data Cleaning	Qiaoling Chen	100%	08/10/22	13/10/22	5
Descriptive Analysis	Lyu, Xie	20%	14/10/22	17/10/22	3
Modeling & Evaluation					
Tokenization	Zhao Rudan	0%	14/10/22	17/10/22	3
AHP Model Building	Chen, Lyu	0%	17/10/22	20/10/22	3
ARIMA Model Building	Xie Siteng	0%	17/10/22	24/10/22	7
BERT Model Building	Yashna, Chen	0%	20/10/22	25/10/22	5
Model Evaluation	Zhao, Xie	0%	25/10/22	28/10/22	3
Project Report					
Dashboard and Storyboard	Chen, Lyu	0%	28/10/22	31/10/22	3
Strategic Planning	Yashna, Xie	0%	28/10/22	04/11/22	7
Project Report	Zhao, Lyu	0%	29/10/22	08/11/22	10
Presentation Slides and Rehearsal	All	0%	29/10/22	08/11/22	10
Individual Reflection	All	0%	08/11/22	13/11/22	5



6. References

ADVAY PATIL. (2022, August). *YouTube statistics*. Kaggle. Retrieved October 1, 2022, from https://www.kaggle.com/datasets/advaypatil/youtube-statistics

DMYTRO NIKOLAIEV. (2022, January). *YouTube dislikes dataset*. Kaggle. Retrieved October 1, 2022, from https://www.kaggle.com/datasets/dmitrynikolaev/youtube-dislikes-dataset