Data Analysis in Trending YouTube Videos for Category People and Blog

TUNKU KHAIRI BIN TUNKU HANIZD[1] & MOHAMMAD BIN HOSSIN[2]

[1]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
Corresponding authors: tunkukhairi@gmail.com; hmohamma@unimas.my

ABSTRACT

The study of users' behaviours on YouTube has been an interesting topic in the research world since YouTube started to rapidly grow into one of the largest video sharing platforms on internet. In this project, we aim to study the factors that can possibly give positive impacts for videos in the category "*People and Blog"* to attract the viewers' interest to interact with the videos using data analysis.

Keywords: YouTube, Trending Videos, *Opinion mining*, Exploratory Data Analysis, Sentiment Analysis, data science, content creators

INTRODUCTION

YouTube as a medium of expression provides public statistics for each of uploaded videos which are upload date, number of views, number of likes and dislikes. YouTube also provides comment section for sentimental engagement of the videos. YouTube determines the high engagement videos called "*trending videos"* by indicating the level of popularity of the videos based on high number of views, likes and positive comments for each video. Although YouTube is a platform that content creators could freely upload their videos based on YouTube 's guidelines, it is difficult to have high interactions in a single video. YouTube categorize the videos by which topic is covered in the video. One of the categories is *"People and Blog"*. This category covers videos that is related to people's lifestyle, news about people, promotions, reviews, blogs, and topics that has correlation with people.

YouTube has massive datasets that is categorized as Big Data. According to Alexa report, YouTube has become one of the most preferred digital video platforms that intercept more than 30 million visitors in a day. The largest portion of viewers comes from USA with 16.4% of traffics followed by India (9.2%) and Japan (4.8%) respectively. Because of high traffic and more videos are being added every day, unveiling viewership pattern is considered as a complex process for a normal content creator with no Data Science background. Thus, it is time consuming and potentially can be misleading when interpreting the data. To assist content creators for better insights will also potentially open the opportunity to greatly increase the engagement quality of their videos thus improving their marketing strategies of their videos. This research intends to uncover the hidden pattern by answering various questions about the trending YouTube videos using co-related analysis of Exploratory Data Analysis and Sentiment Analysis.

MATERIALS & METHODS

Specifically, Language used is Python and Jupyter Notebook extension is utilized for code documentation. Library used for this project will be explained according to each method step. The project overview is shown in Figure 1.

Formatted: Indent: Left: 1.27 cm

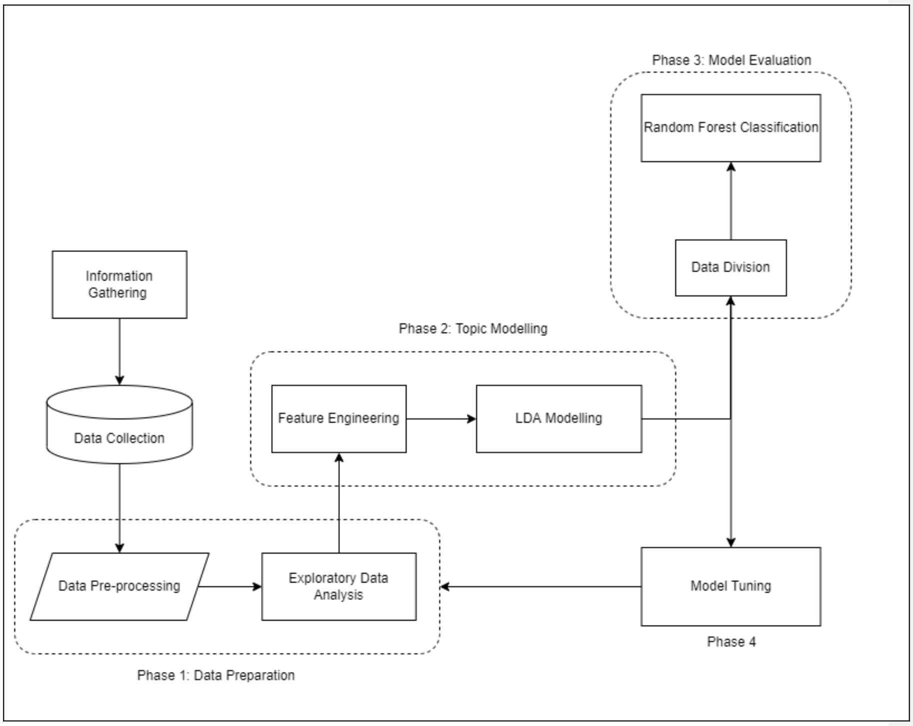

Figure 1. The detailed Project Pipeline.

Data Collection

The dataset used will be a pre-cleaned dataset from Kaggle website, "YouTube Trending Video Dataset (updated daily)" uploaded by Rishav Sharma because of high dimensionality of the dataset with collected data from 11 different countries which are India, US, Great Britain, Germany, Canada, France, Russia, Brazil, Mexico, South Korea, and Japan respectively. The dataset is consisted of up to 200 listed trending videos per day. The dataset is fetched is separated by regions. List of data attribute is as below:

Video ID

Video title

Date published

Channel ID

Channel title

Category ID

Trending date

Video tags

Count of views

Likes

Data Preparation

As an overall pre-processing method, the dataset is inspected to identify missing values. The missing values is then treated. The data preparation process is then divided by numerical columns pre-processing and textual columns pre-processing. For numerical column, the dataset is pre-processed by observing outliers are and treating it to avoid lower accuracy of model training later steps. For textual column, the dataset is pre-processed by removing punctuations and special characters, data segmentation, normalization of out-of-vocabulary words, conversion of lowercase and lemmatization in the column 'title', 'description' and 'tags'. Cleaning dataset will be adjusted according to usability for Exploratory Data Analysis and Topic Modelling on the later steps. Python Libraries involved; Klib, NumPy, Pandas, NLTK and SpaCy.

Exploratory Data Analysis

Exploratory Data Analysis is then used for more comprehensive views of what given data in the dataset is about. By doing so, the best practice of machine learning and roadmap of the project is identified. The purpose of doing Exploratory Data Analysis is to provide better insights from statistical evidence in Exploratory Data Analysis and utilize it to determine the results in findings.

The analysis involved are:

Univariate Analysis

Bivariate Analysis

Time-series Analysis

Univariate analysis is the simplest form of analysis which involves summarization and pattern of only one "Uni" variable in the dataset dimension. This analysis is conducted to discover the meaning of each column of data, whether it is a categorical or continuous, or independent or dependent to other variables in the dataset. Compared to univariate analysis, Bivariate analysis is a correlation analysis that is conducted to gather insights for causal and relationship of two "Bi" variables. Because of the dataset containing timestamps, time series analysis is necessary to discover hidden insights based on time intervals. Time series analysis is a specific approach to analysing a set of data points accumulated over an extended period of time. Time series analysis is done by manipulating the record the data points over a set period of time that is extracted from video published time. Multivariate analysis is the statistical analysis of correlations between several measurements are made on each experimental unit and where the relationship between multivariate measurements and their structure are crucial to understanding the experiment. It is done to find suitable features to feed into machine learning models, or to decide whether the data type should be transformed or not in the feature engineering process.

Feature Extraction & Feature Selection

Feature extraction and feature selection will be conducted to reduce the data dimension and further pull relevant data to use for hypothesis testing. The data types is transformed for better data representation. For topic modelling purpose in later steps, the textual column 'title', 'tags', and 'description' is merged into 'all_text' column. The purpose of merging is to enhance the context of specific video for model training later. Reducing dimension of the data helps to improve the accuracy of the model and reduce the execution time when building and tuning model. Feature Selection include the process of removing stop words, and removing null rows in the dataset, creation of dictionary and data annotation.

Topic Modelling

For topic modelling phase, the chosen analysis for the proposed project is using unsupervised LDA modelling method. LDA modelling process includes building base LDA model, computing perplexity and coherence score. The perplexity and coherence score are then observed and evaluated by assigning each document in the dataset with its predicted topic.

Model Evaluation

For model evaluation process, the dataset with its predicted topic is then divided into training set and testing set with ratio 80% training set and 20% training set. The supervised Random Forest Classification method is used to evaluate the predicted topic in terms of Bag-Of-Word model and TF-IDF model. The evaluation is observed from the value of precision, recall, F-1 score and supports obtained after cross validation process from training and testing sets.

Model Tuning

To increase the performance of LDA model, there are 2 steps of model tuning:

- Building LDA Mallet Model.
- Finding the optimal number of topics, *k.*

Hyperparameter is manipulated to determine the optimal number of topics, *k*. This is done by building many LDA models then plotting computed coherence score graph for range of *k.* Generally, *k* value is determined by model that has the highest coherence score before flattening out.

The project will then repeat the preprocessing method followed by topic modelling phase as part of tuning model process until optimal number of topic *k* is achieved. The optimal *k* value is then used to predict true dominant topic for each of documents.

RESULTS & DISCUSSION

Exploratory Data Analysis

Univariate analysis is conducted based on how videos in the dataset is distributed respectively to view counts, like counts and dislikes count, and tags.

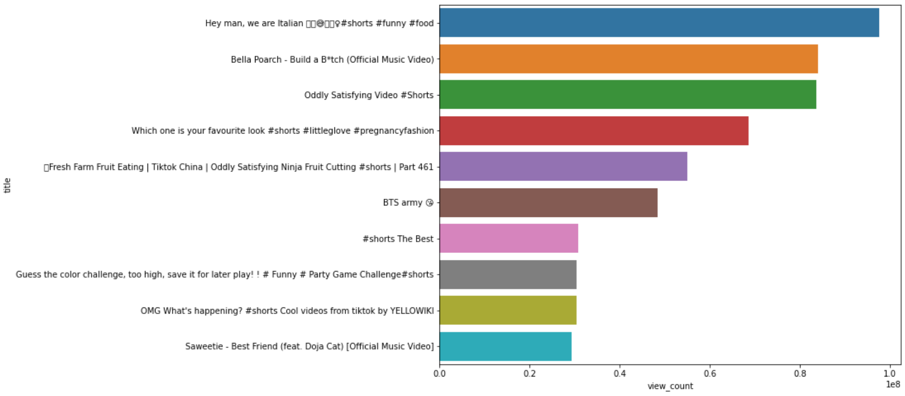

Figure 2. Top 10 Most Viewed Videos in Category *People and Blog*.

Based on Figure 2 above, proves that most viewed videos from category "People and blog" have wide topics posted. Other than that, majority of the top viewed videos consist of hashtag #shorts in count of 7 out of 10 most viewed videos. Roughly, #shorts are used to indicate that the duration of video is short.

```
#shorts              697
#Shorts              121
#viral                58
#funny                47
#POV                  35
#food                 25
#4                    20
#pov                  20
#meme                 19
#trending             16
#minecraft            16
#shorts]              16
#FYP                  16
#Dpeezy2099           16
#1:                   16
#2                    13
#4!                   12
#short                12
#ad                   12
#dowehaveaproblem     11
Name: 0, dtype: int64
```

Figure 3. Most Frequent Hashtags used in Category *People and Blog*.

YouTube has a new feature that enables content creators to upload their videos in a form of short videos. Based on Figure 3 above, the analysis has proved that usage of the new feature is effective to attract viewers. After removing word 'short' along with other stop words, the WorkCloud for tags used in the videos is generated as shown in Figure 4.

Figure 4. *WordCloud* for Most Tags used in Category *People and Blog*.

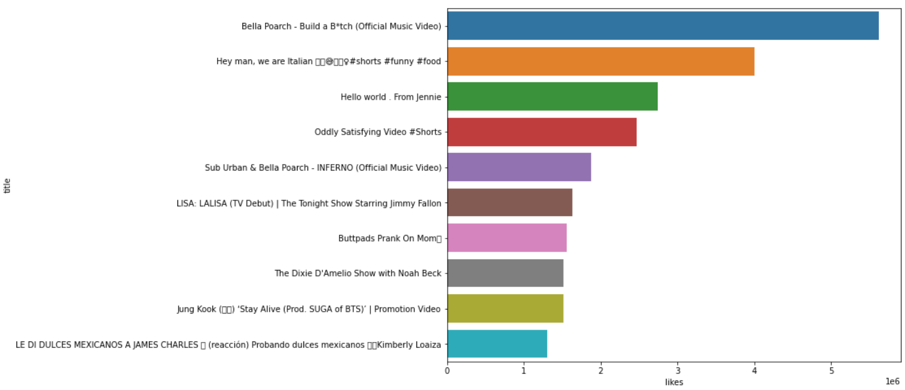

Figure 5. Top 10 Most Liked Videos in Category *People and Blog.*

Based on Figure 5 above, there are slight changes in order of Top 10 videos. This indicates that not every trending videos has positive feedback despite higher views.

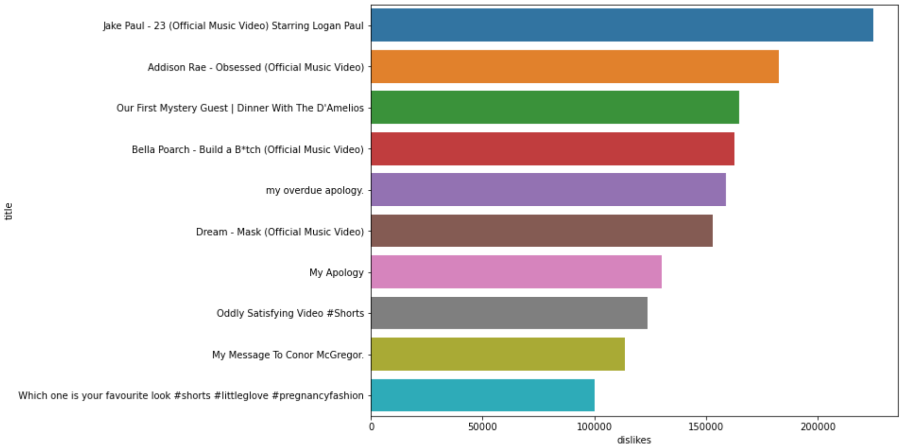

Figure 6. Top 10 Most Disliked Videos in Category *People and Blog.*

Based on Figure 6 above, generally videos with high views tend to have negative feedbacks too. For Bivariate Analysis, bicorrelation between views, likes, dislikes, and channels is observed.

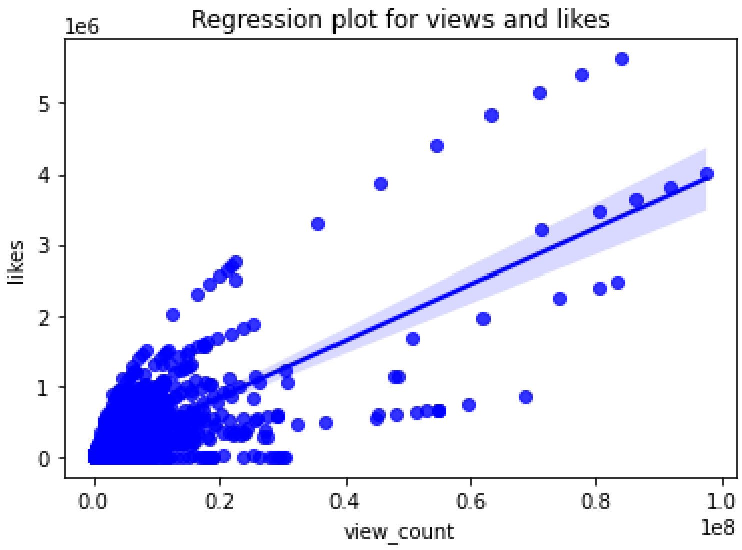

Figure 7. Regression Plot for Views and Likes for Category *People and Blog*.

Figure 7 above shows that the relationship of views and likes is linearly related. From the plot, large percentile of videos is scattered closely to the simple linear regression line while minor percentile of videos is considered outliers. However, residual value of the outliers distributed linearly too. This resulting the videos are having more than one pattern. The simple assumption can be made that likes does not always dependent to the view count. There are some possibilities taking from the scattered plot, where view counts of 2 videos are same, but one highly skewed from line of best fit likes and another one has lowly skewed from line of best fit likes.

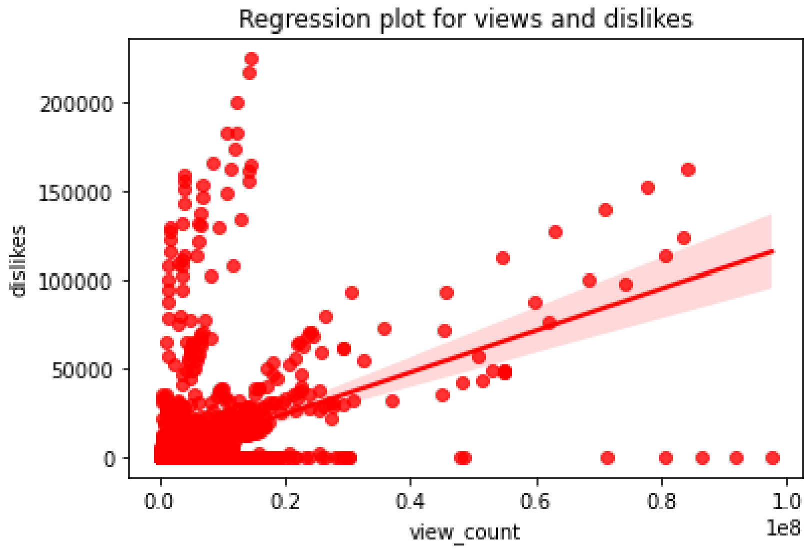

Figure 8. Regression Plot for Views and Dislikes for Category *People and Blog*.

Figure 8 shows that simple linear regression line does not fit views and likes. While large percentile of videos scattered closely to the line, there are outliers that potentially show different behavioral patterns. Usual case is where the greater view counts, the greater amount of dislikes video has. However, the behavioral patterns mentioned earlier is that there are videos that despite having low views, the videos might have great number of dislikes. This can be interpreted as videos is extremely disliked by viewer, but they watch it anyway. Other behavioral pattern is there are videos that despite having high views, the videos might have lesser number of dislikes (compared to best of fit line). This can be interpreted as the videos is extremely liked by the viewers. Notes that dataset are originally extracted from only high-interacted videos where the proposition videos of low views in the statistic is assumed higher than average videos on YouTube. From assumptions made earlier, the deduction is the video does not particularly need to be likable to viewers to reach high interactions.

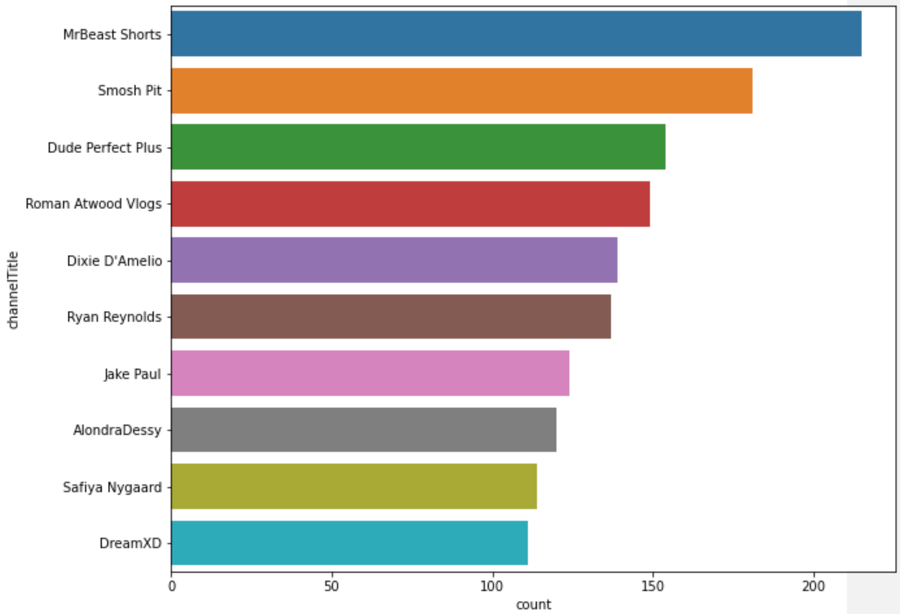

Figure 9. Top 10 Channels with Most Videos on Trending.

Figure 9 above shows that channels that has higher subscribers tends to have more their uploaded videos on trending. For Multivariate Analysis the Pearson Correlation Chart is observed.

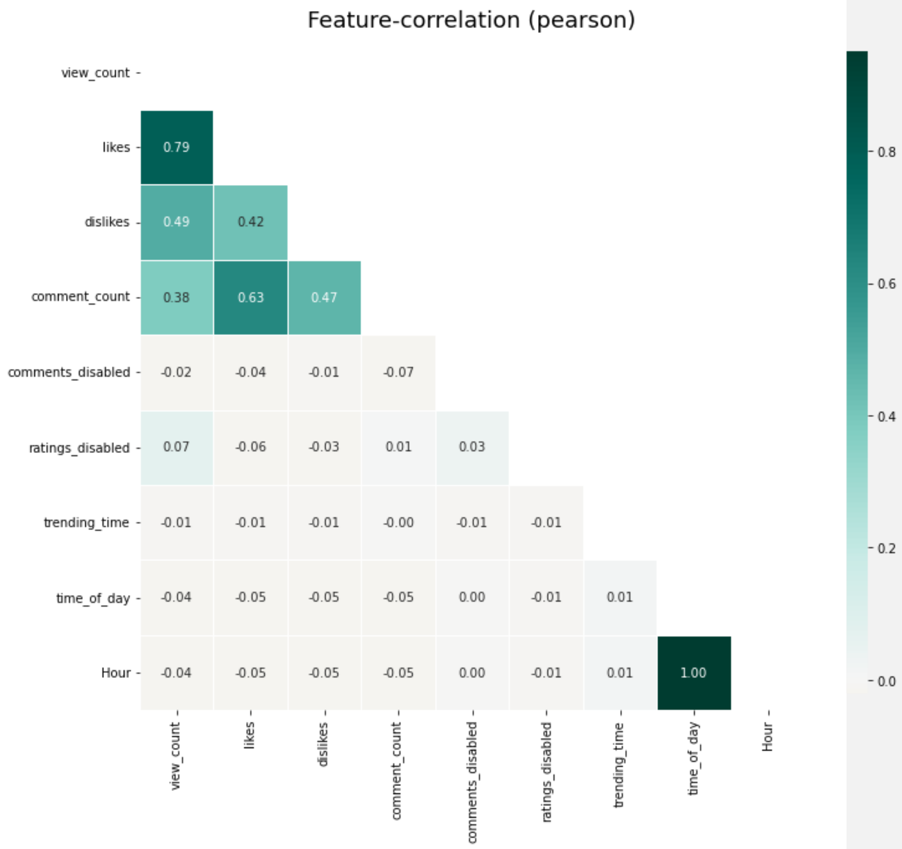

Figure 10. Pearson Feature Correlation Chart for Category *People and Blog*.

Figure 10 shows the Pearson correlation chart for numerical features in the dataset. From the chart, some assumptions can be made. First assumption, likes is highly positive correlated to view counts with 0.79 score meanwhile dislikes is in average positively correlated to view counts. Comments counts is highly positive correlated to likes, however it has average influence on view counts and dislikes.

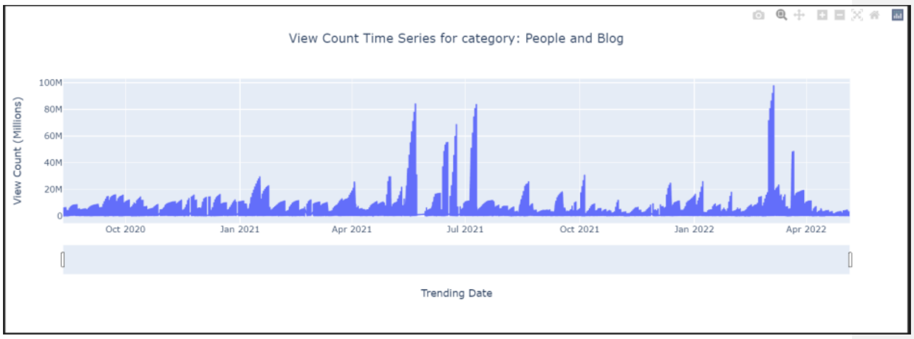

Figure 11. Overall Time Series of View Count for Category *People and Blog*.

Figure 11 above shows that there are some unusual spikes of view counts over 80 million views at several timeframe.

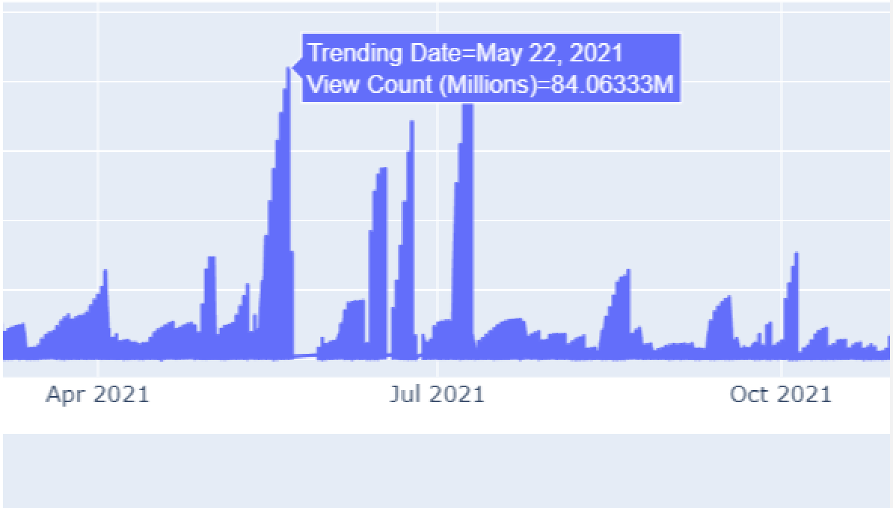

Figure 12. Views Spiked up to 84.0633 million on May 22, 2021.

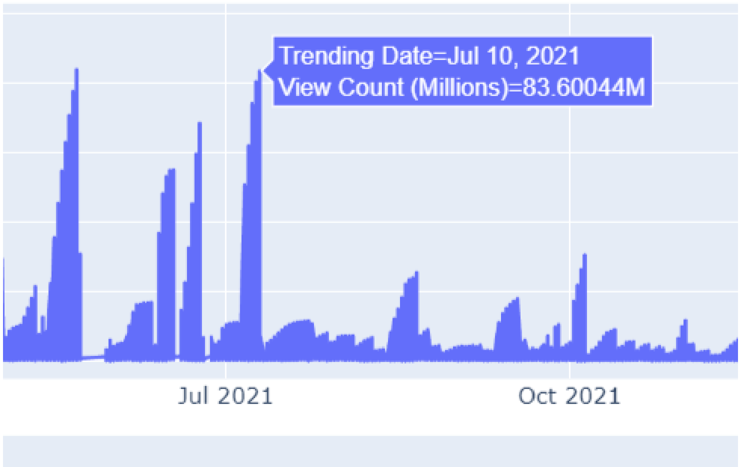

Figure 13. Views Spiked up to 83.6 million on July 10, 2021.

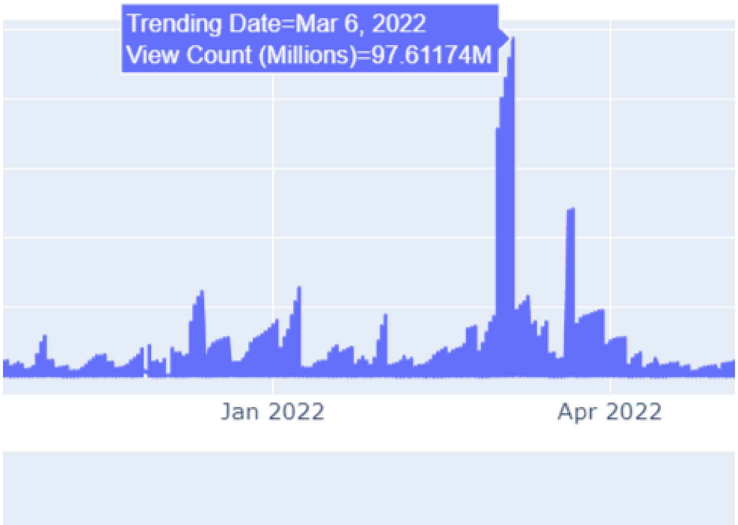

Figure 14. Views Spiked up to 97.611 million on March 6, 2021

Figure 12, Figure 13, and Figure 14 shows the exact date view count spiked on time series plot. The timeframe is:

16th May 2021 to 23rd May 2021

6th July 2021 to 10th July 2021

22nd March 2022 to 26th March 2022

The unusual spikes may indicate that the view counts are influenced by events on real world happened at the timeframes. WordCloud is used to illustrate what events happened on the timeframe.

Figure 15. *WordCloud* for Timeframe between May 16, 2021 to May 23, 2021.

Figure 15 above shows the WordCloud for timeframe that illustrate what videos are posted in timeframe from 16th May 2021 to 23rd May 2021. From the WordCloud, the words “Babish”, “kitchen”, and “Andrew rea” indicates that the video about cooking is trending in this timeframe. First assumption is the views are influenced by first pandemic wave that take places in the timeframe. People around the world were stuck at home and that explain why cooking videos reached high interactions at that time.

Figure 16. *WordCloud* for Timeframe between July, 6 2021 to July, 10 2021.

Figure 16 above shows the WordCloud that illustrate what videos are posted in timeframe from 6th July 2021 to 10th July 2021. Some familiar words from earlier timeframe discussed are spotted as this timeframe is not too far from first timeframe discussed earlier. Word “pasta”, “Babish”, “Al pesto”, “chopped cheese”, “luca” and “trenette” indicates that the video about cooking is trending in this timeframe. Second assumption is the views are influenced by second pandemic wave that take places in the timeframe. People around the world were stuck at home once again like first pandemic wave and that explain why cooking videos reached high interactions at that time.

Figure 17. *WordCloud* for Timeframe between March 22 2022 to March, 26 2021.

Figure 17 above shows the WordCloud that illustrate what videos are posted in timeframe from 22nd March 2022 to 26th March 2022. Words “Kelly Clarkson” and “Stray Kids” indicates singer. There are many assumptions can be made about why sudden spike of views based on words related to singer. For example, the singers produce new songs or reality TV shows about the singers. Words “Quitting” and “quit” may indicate many people decided to change or quit their current jobs. Word “Cardboard” may indicate the videos about manipulating cardboards are trending at the time.

From 3 generated WordCloud, the finding is real world events does impact greatly on view counts. The videos that are related to exact events of real world tends to get more interactions than videos that are not related.

Topic Modelling

For base model, the number of topics is randomly assigned for $k = 10$ (in Python num_of_topic = 10). Result of base LDA model is shown in Figure 18 below.

```
# Compute Perplexity
print('\nPerplexity: ', lda_model.log_perplexity(dt_corpus))  # a measure of how good the model is. lower the better.

# Compute Coherence Score
coherence_model_lda = CoherenceModel(model=lda_model, texts=data_lemmatized, dictionary=id2word, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score: ', coherence_lda)
```

```
Perplexity:  -8.075182183006767

Coherence Score:  0.35944259744198476
```

Figure 18. The Perplexity and Coherence Score for Base LDA Model.

Based on Figure 18 above, the perplexity and coherence score are observed. Observation is made to measure how good the LDA model is. For perplexity score, the lower the score, the higher chance that the base LDA model is good. Coherence score is used to evaluates a single topic's score by gauging the degree of semantic similarity between the topic's top-scoring words. The higher coherence score, the better chance LDA model is good.

Model Tuning

The base LDA model is compared to base LDA Mallet model that uses Gibbs Sampling method. Figure 19 below shows the result of coherence score for base LDA Mallet model.

```
# Compute Coherence Score for mallet
coherence_model_lda = gensim.models.CoherenceModel(model=ldamallet, texts=data_lemmatized, dictionary=id2word, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score: ', coherence_lda)
```

```
Coherence Score:  0.365307724851063
```

Figure 19. The Perplexity and Coherence Score for Base LDA Mallet Model.

Based on Figure 19, coherence score for base LDA Mallet model is slightly higher than base LDA model. Hence, LDA Mallet model is chosen to compute optimal model. Then, the next step is tuning the model by finding the optimal number of topics. This is done by computing multiple LDA Mallet models and calculate their coherence score ‘c_v’ vs number of topic, k for k is in range of $2 \leq k \leq 40$. The results are plotted in graph as shown in Figure 20 below.

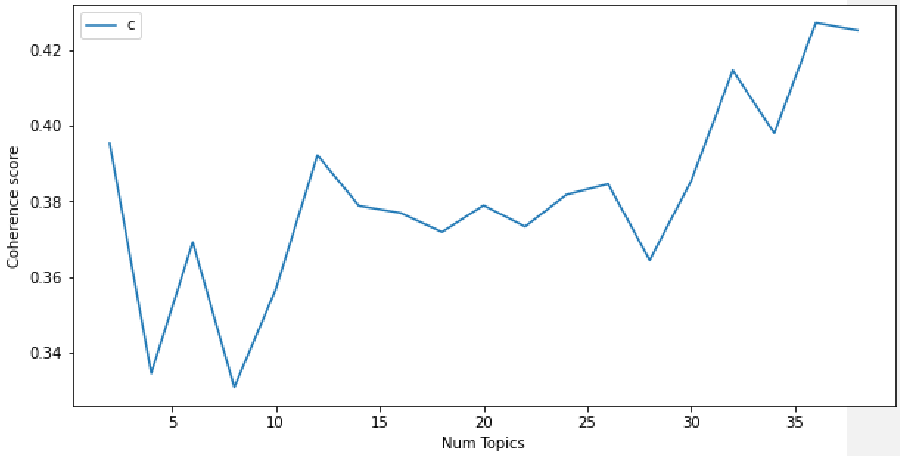

Figure 20. Graph of LDA Mallet Model's Coherence Score vs Number of Topic**.**

Based on graph in Figure 20, the optimal number of topics is chosen subjectively at the point where the coherence score is at the peak before flattening out. Hence the chosen optimal number of topics, $k = 12$. By choosing $k = 12$, the optimal LDA Mallet is then built. Next, each dominant topic is computed with weightage of the keywords as shown in Table 1 below.

Table 1. The Dominant Topic Computed based on Number of Topics, $k = 12$

Keywords	Dominant Topic
baby, vlog, pregnant, code, boy, couple, find, pregnancy, girl, birth	1
tiktok, music, show, official, love, song, happy, hope, fun, tik_tok	2
dream, kid, official, make, support, time, day, long, update, back	3
short, funny, story, part, prank, among_us, viral, challenge, moment, game	4
good, watch, friend, click, big, play, scene, dude_perfect, laugh, comedy	5
family, vlog, day, life, surprise, challenge, royalty, wedding, house, watch	6
eat, make, food, cook, asmr, recipe, cooking, babish, steak, minecraft	7
link, find, free, home, store, business, worth, content, speed, episode	8
move, makeup, year, fashion, check_out, give, car, world, beauty, make	9
season, watch, voice, live, podcast, team, episode, series, clip, highlight	10
getty, make, dog, room, hour, leave, real_life, challenge, turn, man	11
life, home, camera, build, water, house, live, cabin, off_grid, tiny	12

The dominant topics is then used to predict each video in the dataset. To evaluate performance of the optimal model, confusion matrix is calculated by feeding the dominant topic and combined text (video title, video description, and video tags) in each video into two model, Bag-of-Word model and TF-IDF model using supervised Random Forest Classification.

The classification report for Bag-of-model is shown in Figure 21 below.

	precision	recall	f1-score	support
0	0.34	0.88	0.49	40
1	0.66	0.80	0.72	50
2	0.74	0.37	0.49	38
3	0.57	0.77	0.66	53
4	0.92	0.50	0.65	24
5	0.58	0.66	0.61	29
6	1.00	0.65	0.79	37
7	0.96	0.57	0.72	40
8	0.48	0.34	0.40	32
9	0.65	0.39	0.49	28
10	0.85	0.71	0.77	31
11	0.87	0.45	0.59	29
accuracy			0.61	431
macro avg	0.72	0.59	0.61	431
weighted avg	0.70	0.61	0.62	431

Figure 21. Bag-of-Word Model Classification Report.

Based on Figure 21, the performance of BoW model is 0.61. The purpose of using BoW model is to measure how well the document classification based on frequency occurrence of each word when it is used as feature in the model classifier. While having a decent precision of model, the low recall score made F-1 score is lowered to 0.61. For TF-IDF model evaluation, the classification result is shown in Figure 22 below.

	precision	recall	f1-score	support
0	0.30	0.90	0.45	40
1	0.67	0.84	0.74	50
2	0.61	0.37	0.46	38
3	0.59	0.70	0.64	53
4	0.92	0.46	0.61	24
5	0.65	0.59	0.62	29
6	1.00	0.62	0.77	37
7	0.96	0.60	0.74	40
8	0.64	0.44	0.52	32
9	0.64	0.32	0.43	28
10	0.96	0.74	0.84	31
11	0.87	0.45	0.59	29
accuracy			0.61	431
macro avg	0.73	0.59	0.62	431
weighted avg	0.72	0.61	0.62	431

Figure 22. TF-IDF Model Classification Report.

The purpose of using TF-IDF model is to evaluate how well the document classification based on most relevance word in the document. Based on Figure 22, the performance of TF-IDF model is also 0.61. While having a decent precision of model, the low recall score made F-1 score is lowered to 0.61.

Based on two evaluation model discussed, the performance of model is assumed to be low bias but having have high variance. The model is overfit due to the fact when data pre-processing, only “English” words are selected to be fed into the model while the dataset should consist of different other languages text. In testing phase, the recall score has become low due to incapability of model to predict the text with other languages text. Hence the model only fitted for English Languages text.

CONCLUSION

Exploratory Data Analysis done has shown several findings. Firstly, the usage of the new feature is effective to attract viewers. Secondly, likes are not always dependent to the view count and the deduction is the video does not particularly need to be likable to viewers to reach high interactions. There are some possibilities taking from the analysis where exceptional videos are made without having many likes. Next, the finding is real world events does impact greatly on view counts. The videos that are related to exact events of real world tends to get more interactions than videos that are not related. Using unsupervised machine learning, the main topic of trending video is discovered. Content creators might use the main topic produced as a guideline to generate title, description, and tags for their video as it was used by other successful content creators to make their video appearance stands out.

REFERENCES

Snickars, P., & Vonderau, P. (2009). *The youtube reader*. Kungliga biblioteket. Holmbom, M. (2015). The YouTuber: A qualitative study of popular content creators.

Rowley, J. (2004). Online branding. Online Information Review, 28(2), 131-138.

Biel, J. I., & Gatica-Perez, D. (2010, November). Vlogcast yourself: Nonverbal behavior and attention in social media. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction* (pp. 1-4).

Biel, J. I., Aran, O., & Gatica-Perez, D. (2011, July). You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 5, No. 1).

Luers, W. (2007). *Cinema without show business: A poetics of vlogging* (Vol. 5, No. 1). Ann Arbor, MI: MPublishing, University of Michigan Library.

Wauters, Robin. (2010, January 5). State of the vlogosphere. Retrieved from http://techcrunch.com/2010/01/05/mefeedia-state-of-the-vlogosphere-2010/

Mitchell, J. C. (1969). The concept and use of social networks. In J. C. Mitchell (Ed.), Social networks in urban situations. Manchester, England: University of Manchester Press, 1969.

Arthurs, J., Drakopoulou, S., & Gandini, A. (2018). Researching youtube. *Convergence*, *24*(1), 3-15.

Dehghani, M., Niaki, M. K., Ramezani, I., & Sali, R. (2016). Evaluating the influence of YouTube advertising for attraction of young customers. *Computers in human behavior*, *59*, 165-172.

Usmani, Z. (2017). *Kaggle for Beginners: with Kernel Code*. Gufhtugu Publications.

Ghojogh, B., Samad, M. N., Mashhadi, S. A., Kapoor, T., Ali, W., Karray, F., & Crowley, M. (2019). Feature selection and feature extraction in pattern analysis: A literature review. *arXiv preprint arXiv:1905.02845*.

Everitt, B. S., Landau, S., & Leese, M. (2001). Cluster Analysis, ch. 4.

Wu, K. (2016). YouTube marketing: Legality of sponsorship and endorsements in advertising. *JL Bus. & Ethics*, *22*, 59.

Firat, D. (2019). YouTube advertising value and its effects on purchase intention. *Journal of Global Business Insights*, *4*(2), 141-155.

Bendersky, M., Garcia-Pueyo, L., Harmsen, J., Josifovski, V., & Lepikhin, D. (2014, August). Up next: retrieval methods for large scale related video suggestion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1769-1778).

Choudhury, S., & Breslin, J. G. (2010). User sentiment detection: a YouTube use case. Rogers, R. (2013). *Digital methods*. MIT press.