Data Mining

Crafty Clickbait

Team 6
Julio Alves de Oliveira
Doug Greaves
Satoshi Taniguchi

School of Graduate Professional Studies
Data Analytics

SWENG-545 – Data Mining

Fall II, 2021

Abstract
Data mine video titles of the popular YouTube channel, "5-Minute Crafts," to gain insights on the impact of word choice on the number of views and analyze effectiveness of clickbait key words and topics on video popularity.

TABLE OF CONTENTS

Section 1: Introduction

The continuing shift in consumer behavior moving from traditional media to new media for entertainment has resulted in more marketing and advertising dollars spent on digital content platforms such as YouTube. To maximize earning potential, a content creator must be focused on increasing metrics such as number of views and subscribers for their channel. There are many strategies that exist to capture an audience including posting on multiple social platforms, targeted subject matter, producing universal shareable/viral content, user interaction, and selection of genre whether news/opinion, educational, or storytelling.

This study will analyze the titles of the successful YouTube channel, "5-Minute Crafts," to gain insights on the impact of word choice on the number of views and analyze effectiveness of clickbait key words and topics on video popularity. Without including YouTube's managed channels, "5-Minute Crafts" is the 11th most subscribed to channel with almost 75 million subscribers as of December 2021. (1)

Problem Statement:

- Identify the most effective key words and phase patterns that can increase total views on YouTube
- Answer if total views are impacted by other factors such as duration, number of characters, number of words and sentiment

Section 2: Main Characteristics of the Dataset

The dataset used in this analysis was collected via API from YouTube on 2021-10-16 by Shivam Bansal with public domain license (Figure 1). It is published on Kaggle at the site: https://www.kaggle.com/shivamb/5minute-crafts-video-views-dataset (2)
The dataset is a CSV file comprised of 4904 YouTube video title records with 15 characteristics each including attributes such as length and views along with calculated fields describing the count of words and characters in the titles from the 5-Minute Crafts channel. This data presents the opportunity to mine text insights relative to video performance controlled for channel.

title	active_sin	duration_s	total_view	num_chars	num_word	num_punc	num_word	num_word	num_stop	avg_word
Wow! Let's go live! Epic decorations and DIYS	1	558	10825	45	8	3	1	4	1	5.62
EXTREME ROOM TRANSFORMATION || Cool Design Ideas For Your Place	1	1020	184374	63	10	2	3	0	2	6.
LATE SUMMER HACKS TO SAVE YOUR DAY	2	629	478170	34	7	0	7	0	2	4.85714
EVERY SMART PARENT KNOWS THESE USEFUL HACKS #shorts	2	41	197359	51	8	1	7	1	1	6.37
SMART HACKS TO SAVE YOUR WEDDING DAYðŸ¤µâ€ðŸ'°ðŸ•Šï¸	3	784	162025	43	7	0	7	0	2	6.14285
MINI FOOD COOKING || Amazing Mini Crafts & Miniature House Models	3	1080	187127	65	11	3	3	0	0	5.90909
Emergency Hacks For Smart Moms And Dads || Everyday Parenting Tips	4	900	187048	66	11	2	0	0	2	
Best DIY Jewelry Ideas || 3D Pen & Epoxy Resin Crafts	5	960	650103	53	11	3	2	0	0	4.81818
TRENDY FIDGET TOYS YOU NEED TO SEE || DIY Fidget Toys Ideas	6	696	717778	59	12	2	8	0	2	4.91666
Wow! Useful Gadgets, Toys And Hacks For Smart Pet Owners	5	900	1569346	56	10	2	0	0	2	5.

Figure 1: Representative data

Insights from high level exploration of the dataset:
Named 5-Minute Crafts, but 96% of videos longer than 6 minutes and 75% longer than 11 minutes.
75% of the videos have more than 567K views. Most viewed video has 280M views.
The mean number of words in the video title is 8.
The average word length of 5.46 vs. the average English word length 5.1 per WolframAlpha (3).
93% of the videos have a numerical digit in the title.

Section 3: Data Preprocessing Steps

When reviewing and preparing the data for analysis, first the video_id was checked for duplicate videos and it was confirmed that all ids are unique. Then checking for duplicate titles, 8 records were identified as duplicates. The dataset was checked for null/missing data and none were found. Visualizing the data, we saw that different scales are used to represent time series data. This is a YouTube setting where active_since_days are shown to increment by 1 until 6 days, then 10, 15, 20, then from 30 increment by 30 (month) to 330, then by 365 (year) - see Figure 2. The number of records under 365 were significantly lower than 1 to 5 years. The decision was made to omit 289 or 6% of videos of less than 1 year because they are in a dynamic state, represent a small number of the overall videos and may distort the results.

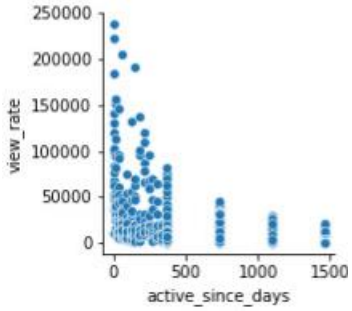

Figure 2: YouTube delineation of days for active_since_days variable less than 1 year.

Because the plan is to cluster the data based on key performance metrics of active_since_days, duration_seconds and total_views, StandardScaller was used to normalized these values to account for varying magnitudes and different units of measure between these continuous variables. Some feature reduction was accomplished by transforming num_words_uppercase and num_words_lowercase into one feature: perc_uppercase. Also, because the focus of the final analysis is on the text in the video title, stop words were removed and using TF-IDF (term frequency weighted by inverse document frequency) calculated the relevance of the words.

Section 4: Description of Data Mining Model

The methodology of this mining project can be followed in the attached Jupyter Notebook.
In the first step, the number of clusters was determined based on the goal of grouping videos based on performance metrics active_since_days, duration_seconds and total_views. This was done by using the Kmeans elbow technique to determine an optimal number of clusters, in this case 9. Looking at the cluster size, three of the clusters had less than 100 videos associated. These clusters were determined to be either outlier data or specific segment not of interest in our analysis. The larger clusters represent more repeatable characteristics that can be analyzed (Figure 3).

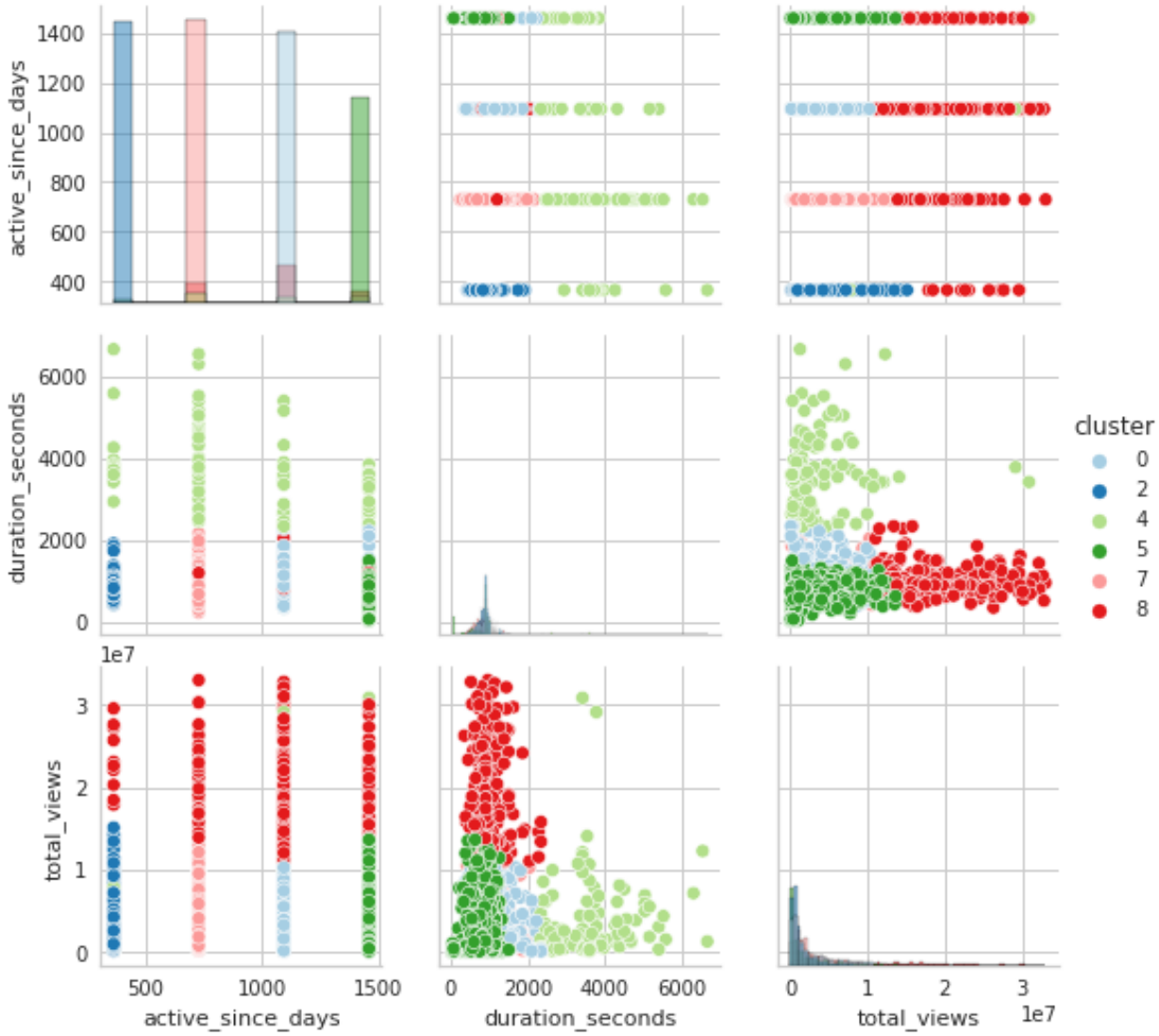

Figure 3: Clusters of interest as related to active_since_days, duration_seconds and total_views

Our intuitive observations were applied to each cluster to define each group by how it presents against our KPIs.

0: 3 years
2: 1 years
4: Long duration
7: 2 years
8: Top performers
5: 4 years

To begin the understanding of text, first observations were made by cluster against characteristics of the title in terms of num_words, num_punctuation, num_stopwords, avg_word_len, contain_digits, and perc_uppercase. Some generalizations were found that might be representative of the channel's naming strategy over time:

- 1 year videos: more words, more punctuation

- 3 year videos: more uppercase words
- 4 year videos: more titles containing digits
- Top performers: less punctuation

Next, focus turned to the title words and the most relevant words of each cluster were analyzed giving additional insights (Figure 4). Words: life, hacks, tricks, know, cool, ideas, and make are common to all clusters and relevant to the brand of 5-Minute Crafts. Some interesting observations in clusters are:

- Long duration has relevant words like: "Live", "Compilation"
- 4 years has "craft" and "minute" used the most
- Top performers relatively flat with no one topic word obvious

8: Top performers		4: Long duration		2: 1 years		7: 2 years		0: 3 years		5: 4 years	
know	0.02	top	0.04	know	0.02	easy	0.02	try	0.02	tricks	0.02
crazy	0.02	time	0.04	crazy	0.03	make	0.02	ideas	0.02	know	0.02
simple	0.02	epic	0.04	cool	0.03	cool	0.02	cool	0.02	ways	0.03
ideas	0.02	crafts	0.04	home	0.03	know	0.03	tips	0.02	tips	0.03
cool	0.02	beauty	0.05	try	0.03	tricks	0.03	make	0.02	make	0.03
tricks	0.03	compilatio	0.08	life	0.03	beauty	0.03	know	0.03	life	0.04
beauty	0.03	life	0.09	tricks	0.03	crazy	0.04	tricks	0.03	hacks	0.05
make	0.03	best	0.1	diy	0.04	ideas	0.04	easy	0.03	I	0.06
life	0.07	live	0.11	ideas	0.06	life	0.05	life	0.05	minute	0.06
hacks	0.1	hacks	0.11	hacks	0.07	hacks	0.1	hacks	0.09	crafts	0.06

Figure 4: Top 10 most relevant words of each cluster

Now we will look for the most positive correlated feature of the dataset including the TF-IDF word features with the total_views to understand if there is a strong indicator of performance (Figure 5). Interesting to note is that the word “fortune” is higher correlated to total_views than active_since_days. Upon closer look, the top video of 280 million views uses the word “fortune”: 42 HOLY GRAIL HACKS THAT WILL SAVE YOU A FORTUNE. The phrase “save you a fortune” does average higher views than the population mean for a given year’s videos.

Feature	R
I	-0.08
num_chars	-0.04
num_punctuation	-0.04
ideas	-0.04
minute	-0.04
	...
hacks	0.1
active_since_days	0.1
cluster	0.13
fortune	0.15
total_views	1

video_id	title	total_views
2820	40 life hacks that will save you a fortune	439203
3086	30 cheap home repair hacks that will save you ...	1496512
3218	30 smart everyday tips that will save you a fo...	432835
3527	27 super easy diy clothing hacks that'll save ...	1075226
3549	42 holy grail hacks that will save you a fortune	283031109
4154	16 life hacks that'll help you save a fortune	1369195
4377	11 cool life tips that will save you a fortune	21118335

Figure 5: Total_views feature correlation

Section 5: Results

After analysis on the 5-Minute Crafts dataset, the winning topic and title pattern for videos on this channel include the words life, hacks, tricks, know, cool, ideas. These words define the concept of the channel, but further success focused on title subjects like beauty, home, diy, crafts appear as strong topics for videos. Using these words in new videos, and following the direct, 6- to 10-word title averaging 5.5 characters per word with numerical digit can be tested as a title template to increasing views and subscriber counts. Also discovered is that top performance cluster has video length between 6 and 35 minutes regardless of how long the video has been published.

Section 6: Comments on Limitations and Perspectives Project

During analysis, many additional questions were brought to light for future work on the subject. One area of study is to determine the outside factors that can cause a video view count to increase beyond the baseline. Some of these reasons may be that a video is promoted/advertised by the channel, or views are redirected from another source or social media platform with more users. We feel visibility of time series data of views could point to anomalies or define a view signature that can forecast a video's lifetime view count. This could also provide insight on seasonality of different topics for example summer activities vs holiday subjects. Data about the viewer type could also help further the understanding of whether the subscribers are watching all videos or select topics, whereas unique viewers might be interested in others, and how likely is a user going to view multiple channel videos.

All of these insights in addition to those discovered in this project can then be applied to performance of other YouTube channels and possibly a sample of all YouTube videos to find if this knowledge is universal across all topics.

Section 7: References

(1) *Top 100 Youtubers Sorted by Subscribers - Socialblade ...* https://socialblade.com/youtube/top/100/mostsubscribed.

(2) Bansal, Shivam. "5-Minute Crafts: Video Clickbait Titles?" *Kaggle*, 29 Nov. 2021, https://www.kaggle.com/shivamb/5minute-crafts-video-views-dataset.

(3) "Average+English+Word+Length - Wolfram: Alpha." *WolframAlpha Computational Knowledge AI*, https://www.wolframalpha.com/input/?i=average%2Benglish%2Bword%2Blength.

(4) Greeneltch, Nathan. *Python Data Mining Quick Start Guide : A Beginner's Guide to Extracting Valuable Insights from Your Data*, Packt Publishing, Limited, 2019. *ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/pensu/detail.action?docID=5761233.

(5) "Learn: Machine Learning in Python - Scikit-Learn 0.16.1 Documentation." *Scikit*, https://scikit-learn.org/.

(6) Agrawal, Rohit. "Analyzing Text Classification Techniques on YouTube Data." *Medium*, Towards Data Science, 15 Apr. 2020, https://towardsdatascience.com/analyzing-text-classification-techniques-on-youtube-data-7af578449f58.