International Journal of Internet, Broadcasting and Communication Vol.13 No.2 224-230 (2021)
http://dx.doi.org/10.7236/IJIBC.2021.13.2.224

IJIBC 21-2-28

A Study on Story propose model based on Machine Learning - Focused on YouTube

[1]Sanghun CHUN, [2] Seung-Jung SHIN

[1]Doctor Candidate, Department of IT Convergence, Hansei University, Korea
odissay@daum.com
[2]Professor, Department of ICT, Hansei University, Korea
expersin@gmail.com (corresponding author)

Abstract

YouTube is an OTT service that leads the home economy, which has emerged from the 2020 Corona Pandemic. With the growth of OTT-based individual media, creators are required to establish attractive storytelling strategies that can be preferred by viewers and elected for YouTube recommendation algorithms. In this study, we conducted a study on modeling that proposes a content storyline for creators. As the ability for Creators to create content that viewers prefer, we have presented the data literacy ability to find patterns in complex and massive data. We also studied the importance of compelling storytelling configurations that viewers prefer and can be selected for YouTube recommendation algorithms. This study is of great significance in that it deviated from the viewer-oriented recommendation system method and proposed a story suggestion model for individual creaters. As a result of incorporating this story proposal model into the production of the YouTube channel Tiger Love video, it showed a certain effectiveness. This story suggestion model is a machine learning text-based story suggestion system, excluding the application of photography or video.

Keywords: *Youtube, Storyline, Data Literacy, Python, Data Visualization, algorithm*

1. Introduction

The most popular streaming service among domestic OTT viewers is Youtube, which is used by 38.4% of the total OTT users as of 2018 [1]. According to the results of a big data research in September 2020, 83% of all citizens of Koreans, or 43.18 million people, spend an average of 30 hours per person watching YouTube for a month. With the growth of individual creators on SNS, the ability of content creators to compose stories is emerging as a key factor in channel activation. Both individuals and companies are jumping into the value of emotional information in social media activities utilizing big data. In addition, the composition of new stories accumulated through these activities is predicted to lead to the birth of new media such as YouTube in

Manuscript Received: April. 19, 2021 / Revised: April. 24, 2021 / Accepted: April 27, 2021
Corresponding Author: expersin@gmail.com
Tel: +82-31-4505274, Fax: +82-31-450-5137
Professor, Department of ICT, Hansei University, Korea

the future IT field [2]. As online media-based content moves from text-based to image and video content markets, the importance of stories is becoming more prominent [3]. Among the elements of video content, the most important element that attracts viewers' attention is moving from the beauty of the video to the storytelling [4]. In summarizing these prior studies, the composition of a highly complete story in single-person media, including YouTube, is emerging as a key element of invigorate channel. However, most individual content creators lack the ability to analyze data patterns on which stories should be produced to draw viewers' preferences. Prior research on YouTube recommendation algorithm system for creators is poor compared to research on recommendation system for viewers.

2. Data Literacy

2.1. Data Literacy

Data literacy is the ability to read, analyze, apply, and think about data in order to properly interpret and apply it to the desired field [5]. The ability to find, analyze, and evaluate the information desired by data users (data literacy) is an important foundation in terms of content production [6]. In order to produce and process data, it is important to have the ability to understand the meaning of data and overall understanding according to the situation through the results of research [7]. Content creators have to produce reliable storytelling content that can empathize with consumers on SNS and have data literacy capabilities that can interpret the needs of these consumers [8]. In this study, data literacy is defined not only as the ability to use data to interpret data, but also to accurately grasp the full context behind data in vast amounts of data and apply it throughout the data usage process. The target of data literacy is YouTube's individual content creator.

Creator provides content that viewers prefer to YouTube by grasping the reaction analysis of viewers' content based on their data literacy capabilities. YouTube provides data about viewers to Creators. In addition, it recommends content that viewers prefer through an AI recommendation algorithm. The most important thing in this correlation diagram is how well creators understand the AI recommendation algorithm exposure system and have ability to create content that viewers prefer. It is creator's data literacy ability that determines the success or failure of this capability as shown in Figure 1.

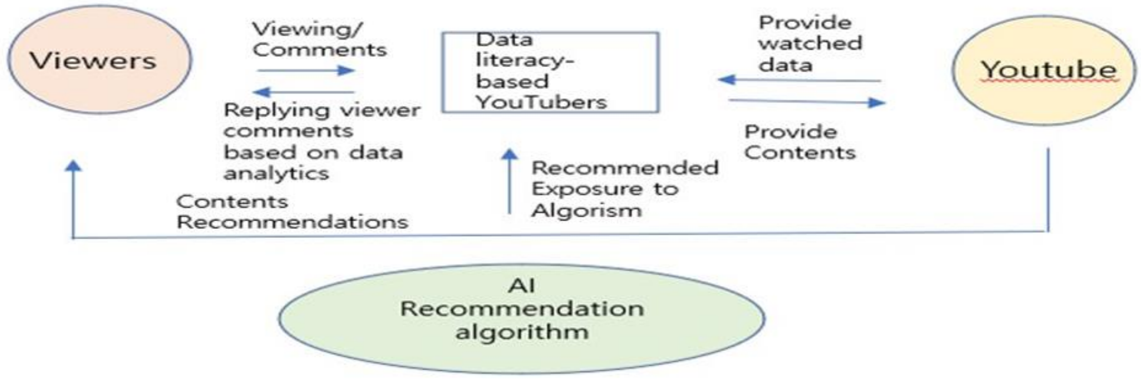

Viewers
Viewing/ Comments
Replying viewer comments based on data analytics
Data literacy-based YouTubers
Provide watched data
Provide Contents
Youtube
Contents Recommendations
Recommended Exposure to Algorism
AI Recommendation algorithm

Figure 1. Diagram of the correlation between creators, viewers, and YouTube algorithms

2.2. Factors required of creators to improve Data Literacy

Platforms including YouTube collect a lot of data and optimize the content recommendation algorithm by individual creator is the core of the business model [9]. Content stories such as videos, photos, texts, and emotions on SNS are also converted into data, playing a key role in applying an AI platform (algorithm) to the content field [10]. The content preferences of viewers, an important factor in the YouTube algorithm, include the types of content that viewers primarily click, the order of viewers who click YouTube, and changes in viewers' viewing disposition due to changes in the external environment.

Digital Set factors refer to Python and Accel abilities as essential abilities to handle data. Cognitive factors include collaboration that can analyze data on viewers' preferences through collaboration with editors, Critical Thinking, Problem Solving, and Ability to see the entire flow of data to analyze the backside of the data. In addition, it is a data visualization ability that can show text response data for viewers' content in a single flow. It includes a recommendation candidate generation model and the ability to analyze ranking networks, which are how YouTube algorithms work with viewers' reactions. As non-Cognitive factors, even if viewers react negatively to content created in comments, it is necessary to have self-control ability as shown figure 1.

Table 1. Data literacy capabilities that YouTubers should have

Division	Data literacy capabilities
Raw Data	Critical Thinking Creative Thinking Problem Solving Collaboration Ability to see the entire flow of data Data Visualization YouTube algorithm analysis
Non-Cognitive factors	Self-control
Digital Set factors	Python/Excel

3. YouTube story suggestion model based on machine learning

3.1. Story Proposal Modeling Method

The method of proposing stories based on data literacy crawls the titles, comments, and likes of videos that have recorded a certain number of views, which are estimated to be recommended contents of the YouTube algorithm, among videos of a YouTube channel. After that, data visualization is performed by referring to the number of views provided by YouTube Studio, and the frequency and graph are created, and the words or keywords that show the most views and viewer preference among them are recommended as shown in Figure 2**.**

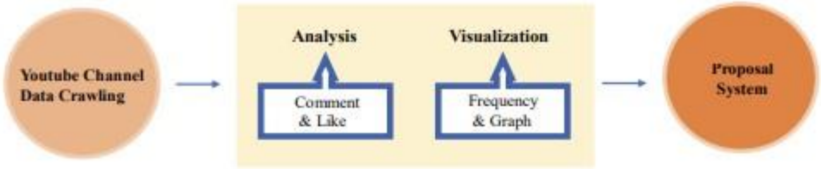


Figure 2. The process of story proposal modeling

3.2. Story Proposal System Application Technology and Operation Process

It is needed to crawl the video's comments and titles to get the data to proceed with the story recommendation. At this time, Google API and Python are used. After organizing the collected data into Excel and Txt files, save them. To analyze the saved files, morpheme analysis is performed using a Python Korean information processing package called KoNLPy. By extracting keywords, checking the frequency of keywords and drawing 'Word Cloud' by keyword ranking to determine which keywords are included as shown in Figure 3.

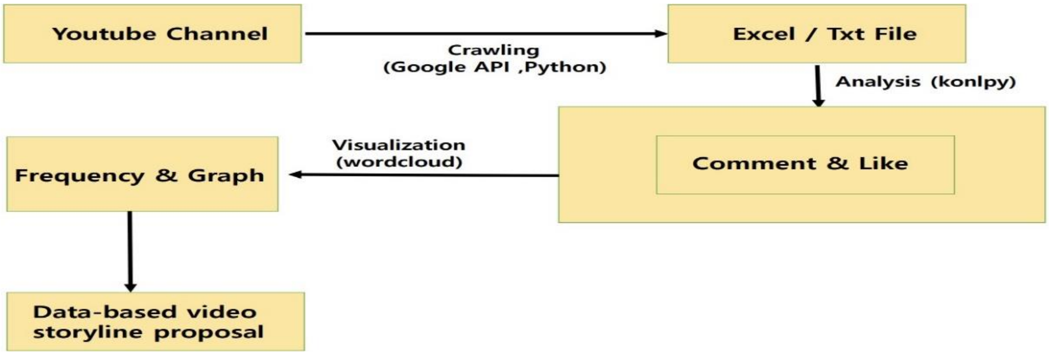


Figure 3. Youtube Storyline Proposal Model Conceptual Diagram

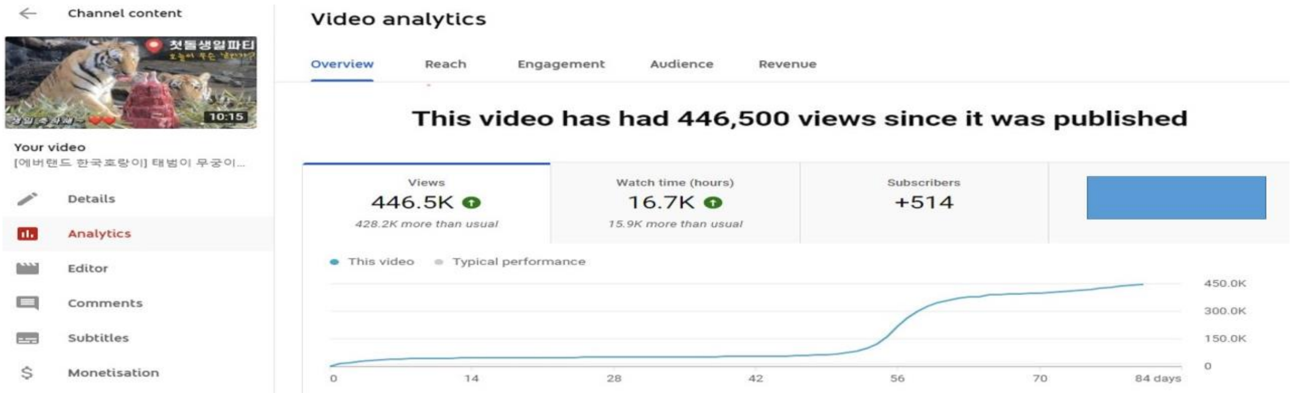
The next step is to analyze the YouTube statistical analysis data. In YouTube Studio, video views, average viewing time, subscriber growth, video revenue, real-time views, impression click-through rate, traffic source type, how impressions and impressions affect watch time, average watch duration, net number of viewers, the average number of viewers, and the number of subscribers are provided. The final decision is made by combining the proposed story and analysis of YouTube studio statistics, and creators' literacy skills.

3.3. Story Proposal System Application Examples and Results Analysis

We applied and analyzed Tiger Love, which is currently operating.

Channel name: Tiger Love,
Number of subscribers: 9,400
Grand Prize:'[Everland Korean Tiger] Taebeom and Mugung's first birthday video!'
Upload date: February 20, 2020

According to statistics from Tiger Love YouTube Studio,'[Everland Korean Tiger] Taebeom and Mugung's first birthday video!' was 43,000 on March 1, 2021 after 9 days of uploading the video. And 50 days by April 11 by day was 69,000. By the way, from April 12th to April 22nd, the total number of views increased more than four times to 283,000 as shown in Figure 4.



Your video
[에버랜드 한국호랑이] 태범이 무궁이...

Video analytics

This video has had 446,500 views since it was published

Views	Watch time (hours)	Subscribers
446.5K	16.7K	+514
428.2K more than usual	15.9K more than usual	

Figure 4. YouTube Studio of '[Everland Korean Tiger] TaeBeom and Mugung's first birthday video statistics'

Considering that the average number of video views of SBS TV 'Animal Farm x AnimalBa' of 3.89 million subscribers, which is the No. 1 YouTube ranking in the domestic animal field, was 29.44 million. Therefore, it can be considered that the total number of views in the last 10 days was exposed to the YouTube recommendation algorithm. After this statistical analysis, as shown in Figure 4, after analyzing the title and comment likes using Python, the data was visualized through the work crowd. After that, I made a story by selecting the keywords and words that were exposed the most to viewers by creating a frequency and graph. The data visualization result of '[Everland Korean Tiger] Taebeom and Mugung's first birthday video!' is shown in figure 5.

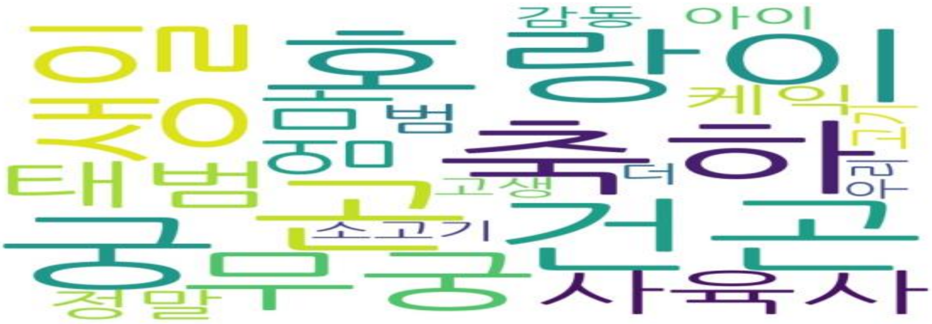
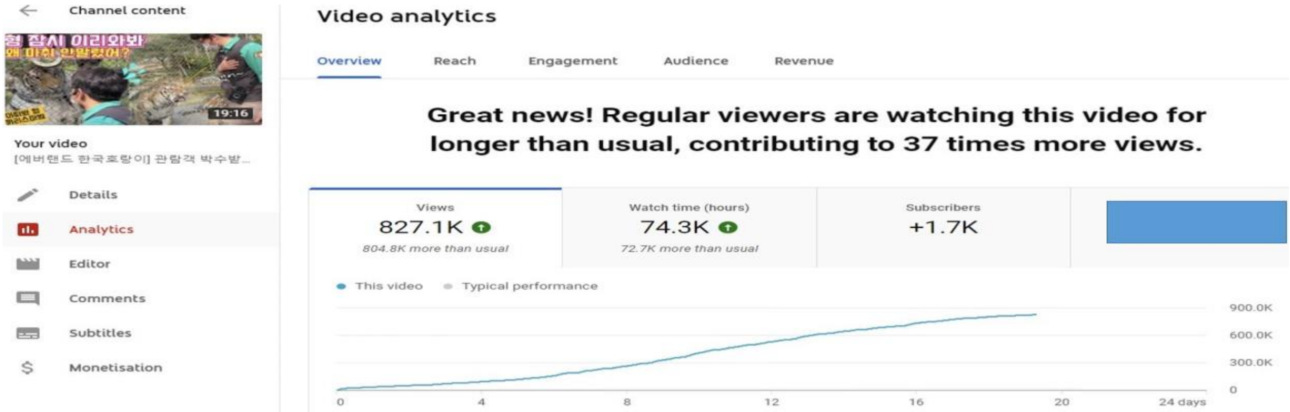


Figure 5. The Visualization result of Taebeom and Mugung's first birthday video data

Based on the results of keywords from the 1st to 10th place derived from this visualization, a story was conceived and uploaded on April 22, 2021 under the title of "Impressive Animal Talk, which was applauded by viewers of [Everland Korean Tiger]". As of May 11, 2021, the total number of views was 820,000 and the average time was 74,000 hours, far higher than the average number of views of Tiger Love video of 20,000 and viewing time of 2,000 hours. According to the statistics analyzed by YouTube, "Fixed viewers have watched this video longer than usual, resulting in a 37times increase in views." Considering these results, the YouTube story suggestion system was found to be partially effective is shown in figure 6**.**



Great news! Regular viewers are watching this video for longer than usual, contributing to 37 times more views.

Views	Watch time (hours)	Subscribers
827.1K 804.8K more than usual	74.3K 72.7K more than usual	+1.7K

Figure 6. The YouTube Studio of '[Everland Korean Tiger] Impressive Animal Talk video statistics'

4. Conclusion

Data literacy is a skill that finds a certain pattern of big data using digital sets including artificial intelligence. It is an essential skill that must be equipped in order for various videos produced by individual YouTube creators to be exposed to the YouTube recommendation algorithm. In this study, under the premise of having such data literacy, creators used the library provided by python. And extracted keywords based on the frequency of titles, comments, and likes for popular video contents they produced. After that, we dealt with a system that extracts the frequency and graph and suggests keywords that can be grafted to the content storyline. As the limitations of this study, first, modeling was conducted using only Everland Korean tiger images among various media contents. This has a limitation in the representativeness of the content production area. Second, since it is based on machine learning, the question of whether it is possible to apply the photo or video section as a text-based story suggestion system can be pointed out. In order to overcome these limitations, a deep learning system rather than machine learning will be needed in the future to apply a story proposal system model through analysis of thumbnails and image data to various content fields.

References

[1] Hee-yoon Noh, “Analysis of domestic OTT service usage status -Focused on Youtube”, Information and Communication Broadcasting Policy, KISDI, 2019.

[2] Seung-Jung Shin, “SNS using Big Data Utilization Research.”, The Journal of The Institute of Internet, Broadcasting and Communication (JIBC), Vol. 12. No.6. pp 267-272. Dec 2012.
DOI: http://dx.doi.org/10.7236/JIWIT.2012.12.6.267

[3] YooJung Kim, “The interaction effect of product type and contents genre on consumer responses: Focused on YouTube branded contents.”, Innovation studies, Vol.14. No.4. pp. 97 – 117. 2019.
DOI: https://doi.org/10.46251/INNOS.2019.11.14.4.97

[4] Hye-yung Kim, “ Type and strategy of storytelling in mobile video contents about science and technology : focused on activity-centered video on YouTube”, Kocon, pp. 177–178, 2019, ISSN 2234-2001, May 2019

[5] D’Ignazio, “Approaches to building big data literacy.”, Proceedings of the Bloomberg data for good exchange conference, 2015

[6] Eun-mee Kim, “The Effects of Online Networks and Internet Literacy on Adolescents’ Online Participation.”, Korean Journal of Journalism & Communication Studies (KJJCS), Vol. 61. No.3. pp 121-154. 2017.
DOI: http://doi.org/10.20879/kjjcs.2017.61.3.004

[7] Sang Woo Han, “A Study about the Concept of Data Literacy based on Digital Humanities.”, Journal of the Korean Society for Information Management (JKOSIM), Vol. 35. No.4. pp 223-236. 2018.
DOI: https://doi.org/10.3743/KOSIM.2018.35.4.223

[8] Eun-Ji Bae, “The Effects of SNS Storytelling Composition Factors on Para-social Interaction, Attitude and WOM Intention: A Case Study of Beauty YouTube.”, The Journal of the Korea Contents Association, Vol. 20. No.1. pp 16-24. 2020.
DOI: https://doi.org/10.5392/JKCA.2020.20.01.016

[9] Mi-Kyung Kim, “The data gap in the platform data ecosystem: beyond digital inequality”, Communication Theory.”, Vol. 16. No.4, pp.5-45. 2020.
DOI: https://doi.org/10.20879/ct.2020.16.4.005

[10] Seoeyon CHOI, “A Study on Expansion Proposal of Data Dividend Qualification Based on the Contributio n of Platform Workers.”, The Journal of The Institute of Internet, Broadcasting and Communication (JIBC), Vol. 12. No. 6. pp 267-272. Apr 2021.