University of Tartu

Faculty of Science and Technology

Institute of Computer Science

Ismayil Tahmazov

I hate you like I love you: A case study of likes and dislikes on YouTube

Master's Thesis (30 ECTS)

Software Engineering Curriculum

Supervisor:

Rajesh Sharma PhD

Tartu 2019

Abstract:

I hate you like I love you: A case study of likes and dislikes on YouTube

YouTube is one of the most popular and largest video sharing websites with Social Network features of the World . Therefore YouTube is also the biggest video promotional and advertisement platform on the internet. While some content is popular, some other content is not successful at all. Popularity in YouTube interesting research area. The aim of this paper is investigate the popularity of YouTube videos and channels to find a simple formula for popularity on YouTube. Our data set contains over 100,000 videos. We'll analyze this data to gain insight into popular Videos on YouTube, to see what's common among those videos. We try to analyze the key features of YouTube popularity using metrics like views, likes, dislikes, comments and video duration time. We use text mining and statement to analyze comments on videos.

Keywords: Comments, View Count, Likes, Dislikes

Contents

1 Introduction

YouTube was created in 2005, and since then it gained billion of users and user content generators including bloggers, musicians, comedians, educators and scholars. After YouTube was taken by Google the number of professional channels significantly increased. Now being popular on YouTube is essential for people who want to promote own products or just want to advertise what they do. Unfortunately, YouTube does not give any global statistics about videos and channel views, likes, comments so that , even simple questions like "How many videos exist on YouTube?" cannot find answer. Our goal is to understand the fundamental properties that make a YouTube video popular.

We collected data using the YouTube API (Application Programming Interface), by using analytical sampling methods. We used view Count (the number of views of video) as the most important parameter for popularity and associations with comments, likes, and dislikes.

1.1 Research objectives

The objectives to achieve through this study are following:

General objective: Investigating video popularity on YouTube.

Specific objectives:

1. Comment frequency analysis.
2. Like and dislike analysis.
3. Video duration analysis.
4. Comment sentiment analysis.
5. Network analysis.

The research questions we want to answer are :

1. When can we say that a video or a channel is popular?
2. What are the reasons of popularity in YouTube?

2 Related work

Recent studies have been focused on understanding YouTube's popularity from various perspectives. Researchers analyzed YouTube in perspective of spam content [2], view count analysis [8], content analysis [12], and geography related analysis[5].In [3] the authors examined YouTube videos evolution during the last decade. We can see this evaluation of channels uploads and views from table 1:

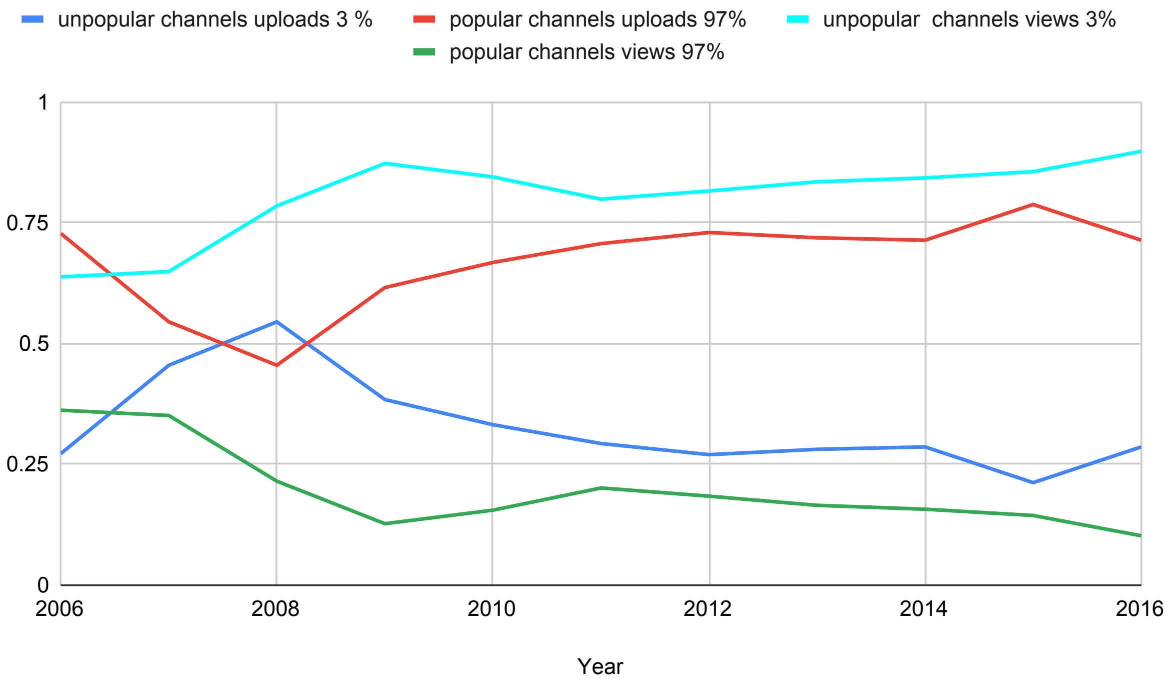


Figure 1: YouTube progress in last decade

As shown in figure 1 and table 1, uploads and views progress are different. During the last 10 years uploads increased several times, but video viewers did not increase in the same way. YouTube is promoting older channels in the search algorithms and does not give chance to new channels and videos to be shown . From the results, we can say that views the channels related to People and Blog, Gaming, Music, Entertainment categories increased of 6 times during 2006-2016.

YouTube has several customer categories, one of the largest category is about contents for kids. In [1] the authors investigated the collective behavior of user for this category and what kind of videos they like. They used techniques like sentiment analysis and classification. Additional to

Year	unpopular channels uploads 3%	popular channels uploads 97%	less viewed channel views 3%	popular channels view 97%
2006	0.272	0.272	0.638	0.362
2007	0.455	0.455	0.649	0.351
2008	0.545	0.545	0.785	0.215
2009	0.384	0.384	0.873	0.127
2010	0.332	0.332	0.845	0.155
2011	0.293	0.293	0.799	0.201
2012	0.27	0.27	0.816	0.184
2013	0.281	0.281	0.835	0.165
2014	0.286	0.286	0.843	0.157
2015	0.212	0.212	0.856	0.144
2016	0.286	0.286	0.898	0.102

Table 1: Evolution of YouTube.

that also did some advertisement and audience analysis. The result of research follows: Different age groups, geographic locations like different contents and popularity of video depends on thumbnails, faces in the video. A data set of this analysis is below:

Country	#channels	#videos	#views	#comments	#commenters	#faces
Brazil	24	7,664	4 M	10M	2 M	129K
US+UK	17	5,184	37M	3M	3 M	1K

Table 2: Analysed channels data set in YouTube Kids

Another perspective of YouTube video analysis is to do it by Geographic location [5]. As we can understand in different countries, people like different things. In [5] authors analyzed the relationship between geography and video categories. In " [9] author analyzed the relationship between view Count and comments, upload frequency of videos and other aspects. From this should we can understand that viewCount is an important metric for measuring popularity, but other metrics are also important.

In [12] the author created a simple way for analyzing channel metrics. He analyzed the "MIN Official " channel. First the author gathered data from the YouTube API with the "tuber" library. This is a powerful library in R to work with the YouTube API. The format of the data set is shown table in 3:

Channel	Subscriptions	Views	Videos	Likes	Dislikes	Comments
MIN OFFICIAL	655795	266544606	54	1409734	58752	109800

Table 3: Example channel data set.

Then he created the plot for comparing like Count, dislike Count, comment Count vs View Count.

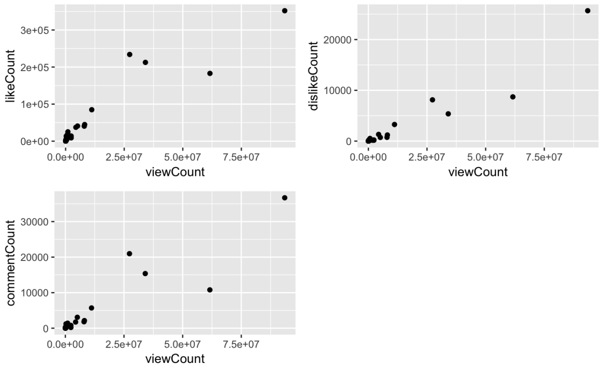


Figure 2: Relationships between YouTube main metrics

The last method which the author of [12] used is the plot of comments frequency by date. With this plot, we can find the most commented period of a channel. In [16] author used API search method. He also used the "tuber " library for gathering data and manipulations. The author did analyze by plotting upload videos during the months. He also talked a little bit about video titles. The video title is one of the important aspects to attracts user. Spam detection problem was analyzed in [2]. Spams is one of the most important problems of YouTube. Spams appears when watching videos in the form unrelated content. Some researchers also study YouTube as a Social Network [10][4][18][17] [21]. Another interesting research line is YouTube video recommendation system analysis [11] [19] [6] .The gaming industry is gained more and more popularity. Every day we can find new games in the online game catalogs. With the popularity of this new trend game videos also started to be popular[15].

From Table 4 we can see that Kjellberg's channel (game category) is most popular one because this field is new in YouTube and user content generators have big chance to show themselves.

YouTube has a big impact on consumer behavior during shops[14]. In the article "[7] authors examined illegal uploads. On the platforms like Netflix, Hulu, Amazon prime we don't have problem, since have checking system for preventing it.

When we are surfing on YouTube we're trying to watch related content that we see before.

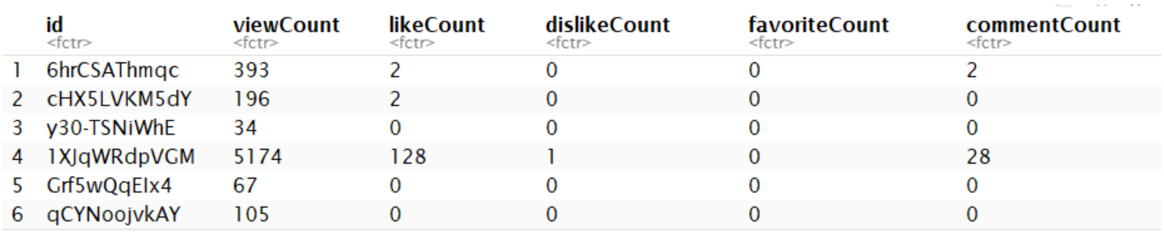
YouTuber	Subscribers	Channel views	views per month	Estimated yearly income
Kjellberg	40,315,481	10,341,94,335	29,6 M	1M-16 M dollar
Sugg	9,455,481	586, 711,156	22,95 million	64,6K-1M dollar
Helbig	2,781,292	155, 687,601	7,51 million	22,6K-361,1K dollar

Table 4: Different category channel's data set

For example, you are looking interesting videos about the "Marvel" films, what you will watch next depends on comments, tags, ratings, and likes.In [20] author did a perfect analysis of all of this. From network of video comments we can examine users' behaviors, which content they are like and how work this kind of recommendation systems. If we can Understand all of this in the recommendations our video can again appear. Yonghyun Ro, Han Lee, Dennis Won also investigated YouTube recommendation systems and network.

3 Data description

We gathered data directly from the YouTube API using the "tuber" library and combined YouTube video pages crawling with Selenium."tuber" is a special library in written R created to use the YouTube API. The extracted data set from YouTube with the "tuber" library is shown Figure 3 :



	id <fctr>	viewCount <fctr>	likeCount <fctr>	dislikeCount <fctr>	favoriteCount <fctr>	commentCount <fctr>
1	6hrCSAThmqc	393	2	0	0	2
2	cHX5LVKM5dY	196	2	0	0	0
3	y30-TSNiWhE	34	0	0	0	0
4	1XJqWRdpVGM	5174	128	1	0	28
5	Grf5wQqElx4	67	0	0	0	0
6	qCYNoojvkAY	105	0	0	0	0

Figure 3: Simple data set from Tuber library

Full data set description shown in Table 5:

Video Count	Size	Comment Count	Time Frame
20000	200MB	800000	2011-2019

Table 5: Full dataset

We now give a brief introduction of YouTube data and discuss few important features. After watching a video, users can give feedback about it . Users can give like or dislike and some comments to the video. These feedbacks can be measured using popularity metrics like commnentCount, viewCount, likeCount, dislikeCount. In Table 6 we can see an example data set from tuber library which we will use in our experiments:

Metrics	Minimum	Average	Maximum	Short Description
ViewCount	2M	3.5M	1B	number of views
likeCount	0	26308	10683228	number of likes
dislikeCount	0	840	511921	number of dislikes
commentCount	0	1088	309570	number of comments

Table 6: Used metrics

In this study, we focus an few metrics, provided by YouTube, likeCount, dislikeCount and the number of comments. All these metrics can reflect video popularity. For the study we are selected 10 most popular video categories from YouTube . You can see our dataset overview in the figure 4:

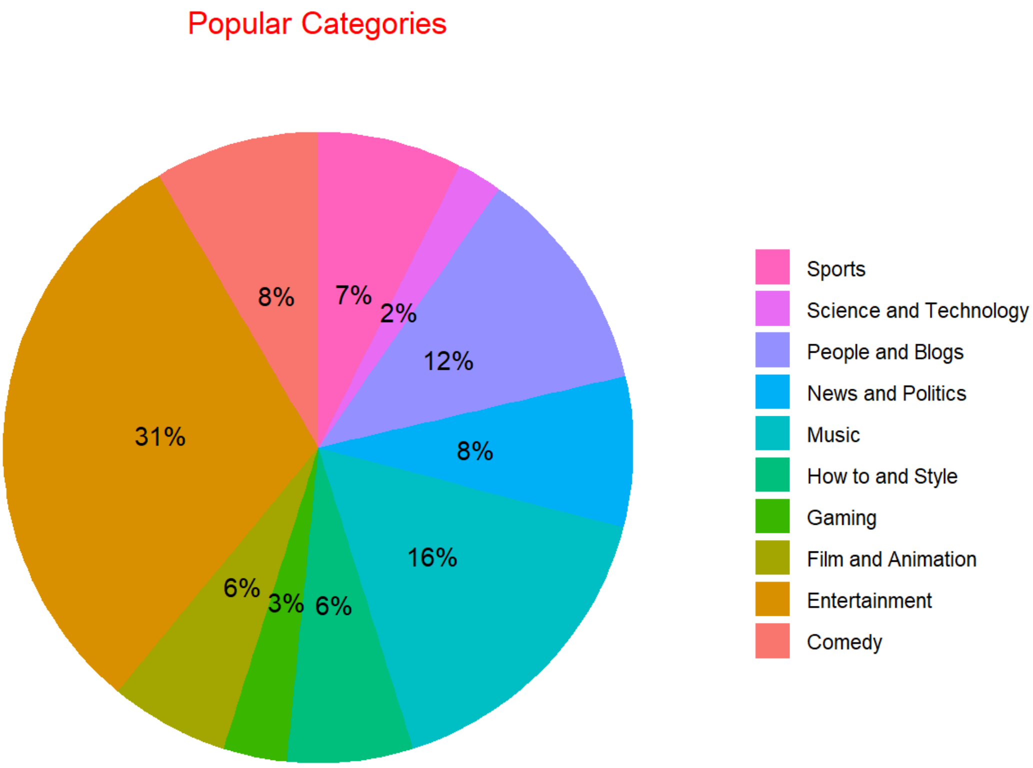


Figure 4: Popular Video categories

4 Methods

One of the important aspects when working with datasets is to have a plan. Our plan is shown in Figure 5. In the next sections we explain different stages in details.

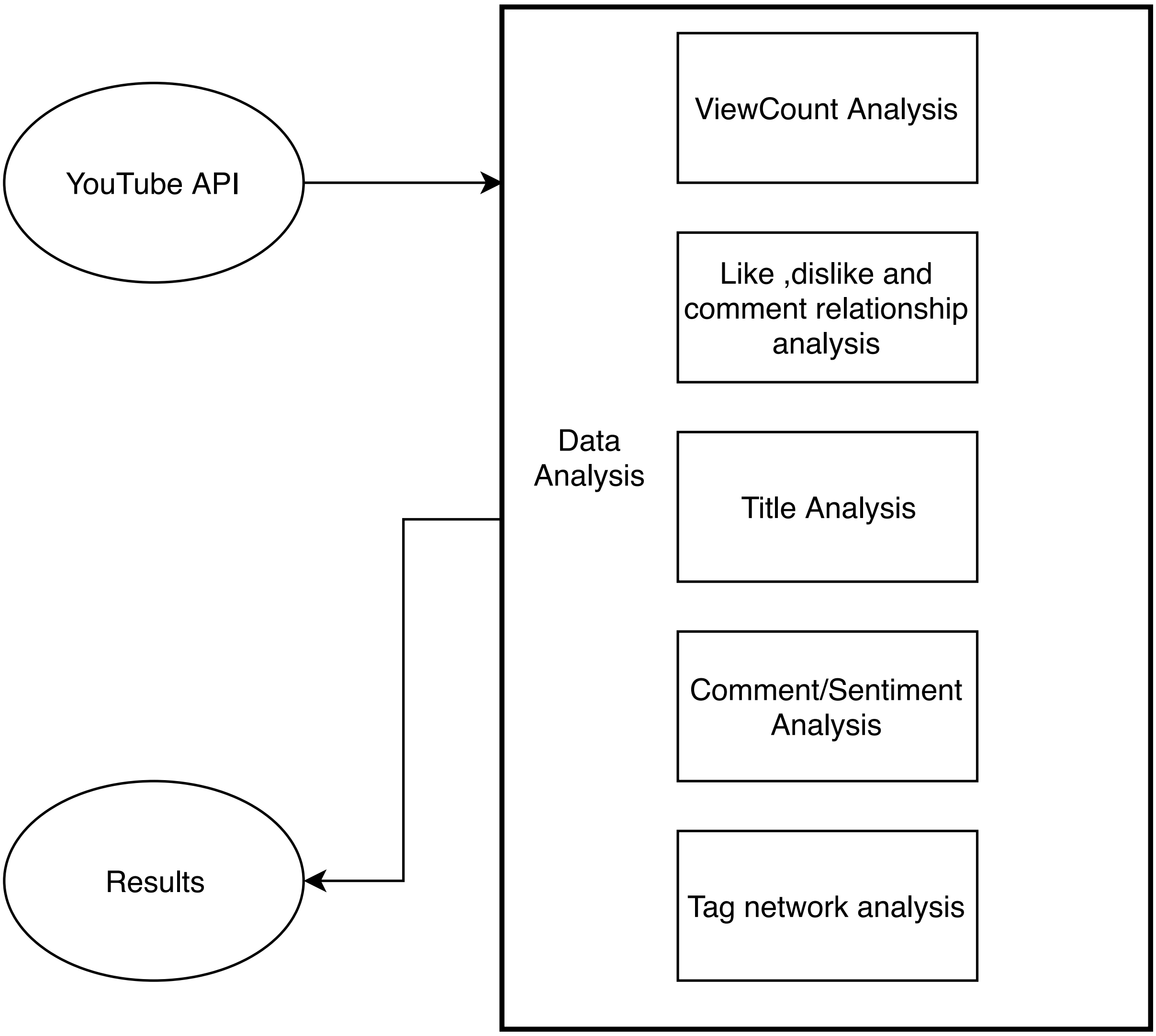


Figure 5: Thesis data Evolution flowchart

4.1 YouTube API

YouTube has an API for developers and researchers. We are used this for the direct access datasets of the last 3 years.

4.2 Selenium Crawling

Selenium is a web browser automation tool initially designed to simplify web applications for testing purposes. This is also used for many other uses, such as automating web-based admin tasks, communicating with non-Api systems, and also Web Crawling.

4.3 Video Description Analysis

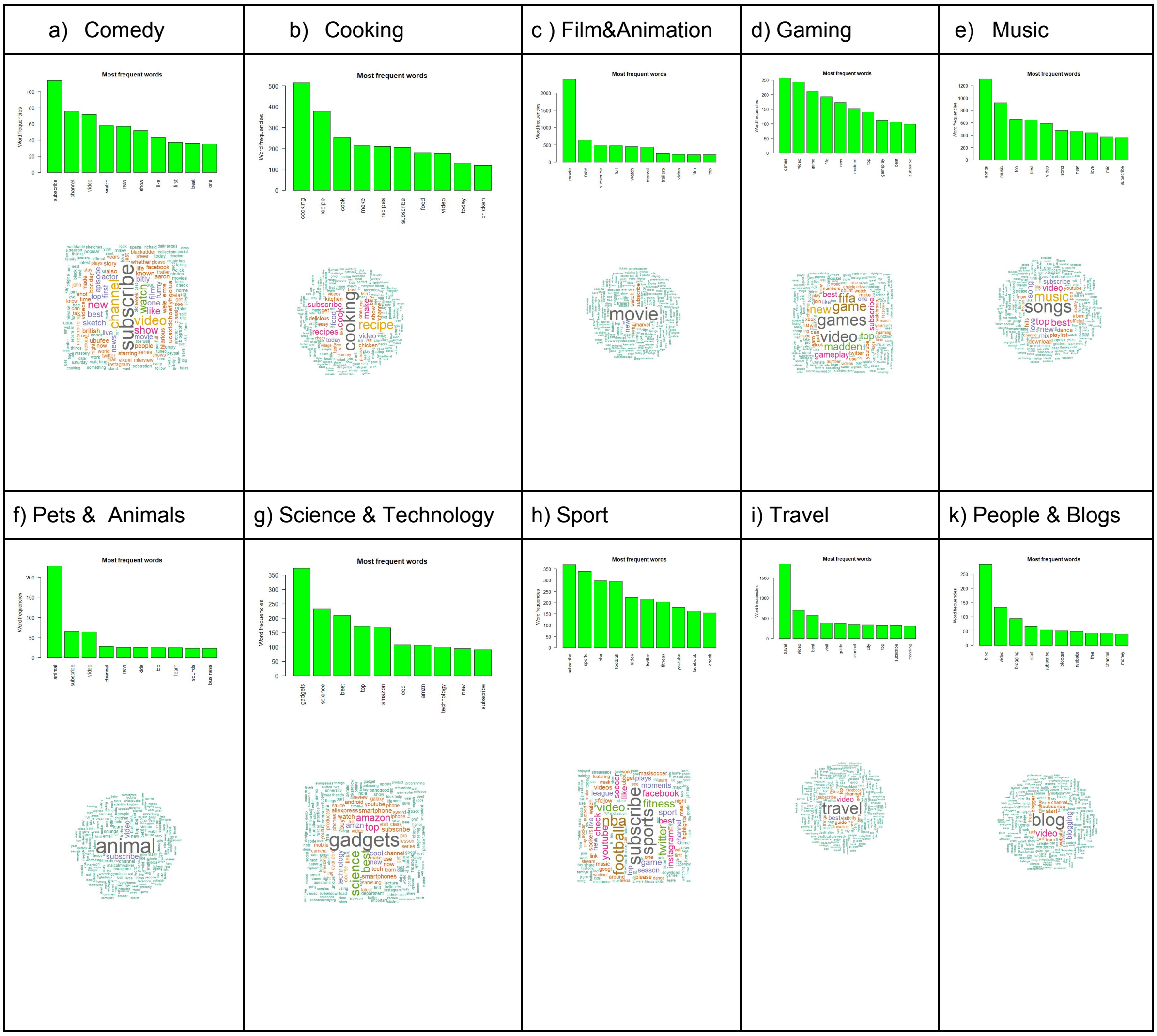
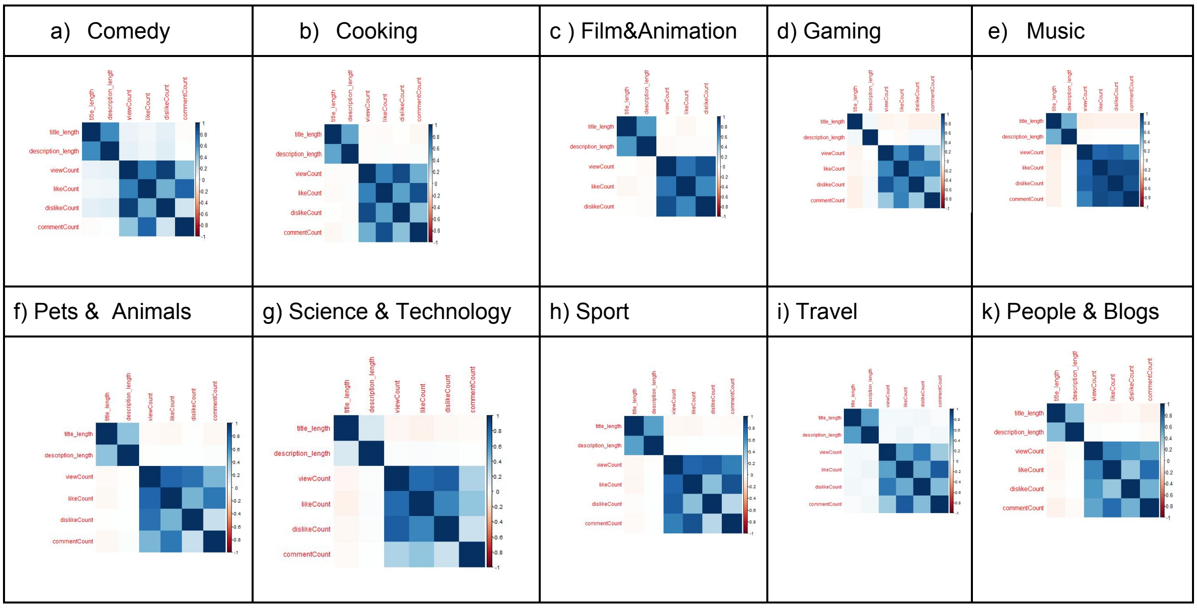

a) Comedy
b) Cooking
c) Film&Animation
d) Gaming
e) Music
f) Pets & Animals
g) Science & Technology
h) Sport
i) Travel
k) People & Blogs

Figure 6: Video description Analysis analysis

4.4 Cor relationship between YouTube Metrics



a) Comedy

b) Cooking

c) Film&Animation

d) Gaming

e) Music

f) Pets & Animals

g) Science & Technology

h) Sport

i) Travel

k) People & Blogs

Figure 7: Correlation between YouTube metrics and title/description length

4.5 Title Analysis

To most content creators that post videos to YouTube, one of the main goals is to be higher on the suggested YouTube ranking system. But, the fact is that 70 percent of YouTube videos are discovered through searching. That's why the keywords used for YouTube videos are highly important to achieve better user friendliness. The first aspect people look for when they're searching for a video is a title that will show them the answer they're looking for. But it's not the only purpose that the title of your video serves. In figure 8 we can see the categories videos and most used words in the title. It also lets YouTube realize what the video is about and why it will be seen as a search result. That's why the recognition and use of keywords in description is one of the easiest ways to automate videos. Another factor of video popularity is publishing time. For example, Christmas related videos popular during November, December and January and also holidays time is collapse of video viewing. In the figure 8 we can see most used words in the titles by categories. As we say before title is important for attract people in your content. YouTube keywords are intended for three essential components and used as metadata during the search. Every category in YouTube have special most used keywords for example in Comedy category most used words : " stand up" ," funny " , " top" ,"movie" ," sketch" and so on . When content creators upload videos to YouTube in the title at least should be have one popular keyword, as YouTube searching algorithm work with this keywords in the title.

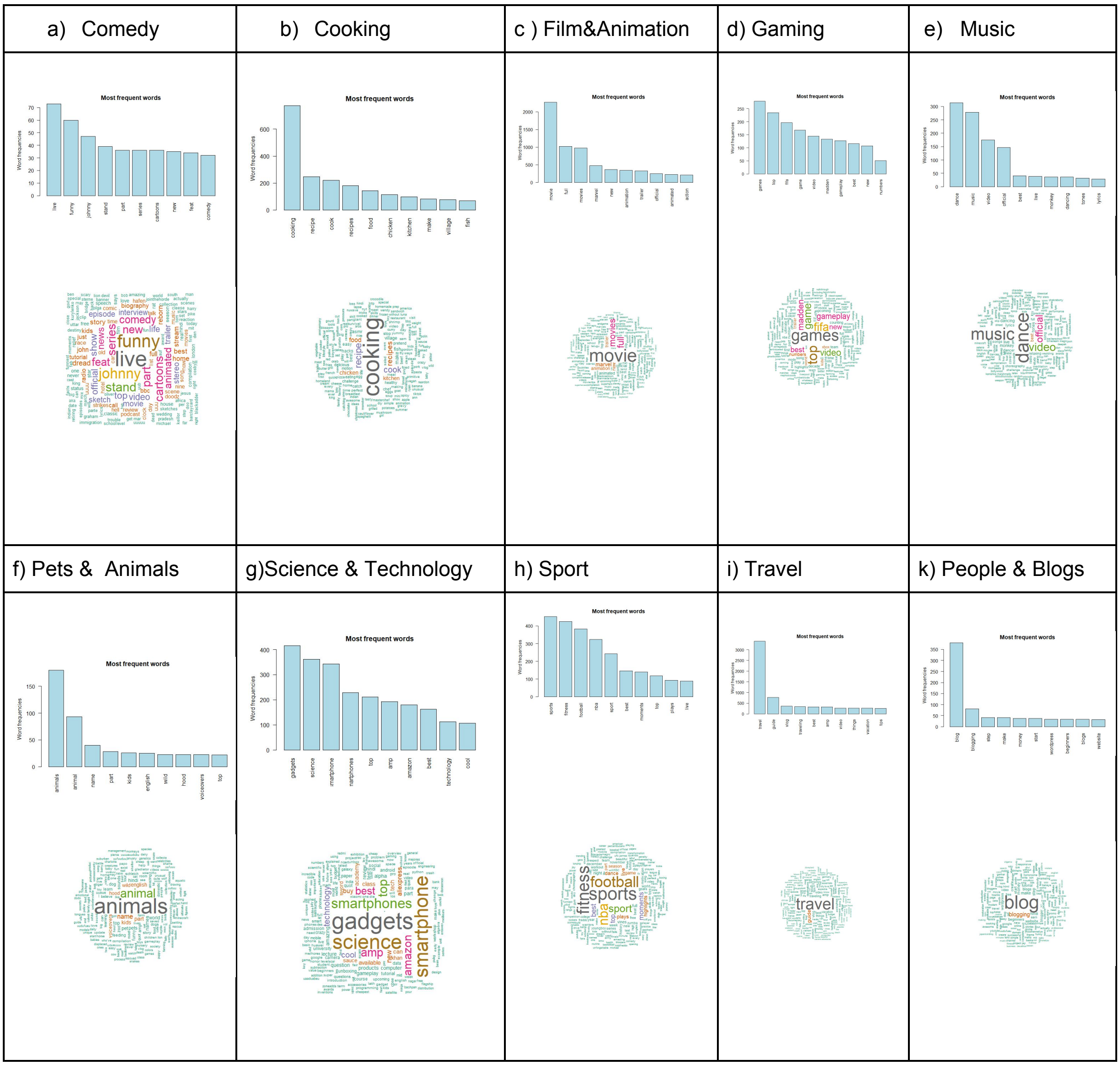


Figure 8: Most used words in titles

4.6 Video duration analysis

A big difference with traditional media web providers is the length of YouTube posts. Although most typical sites contain a small to medium number of long videos, YouTube videos usually have a length of between 5 minutes and 2 hours and are mainly made up of short videos as shown in figure 9.

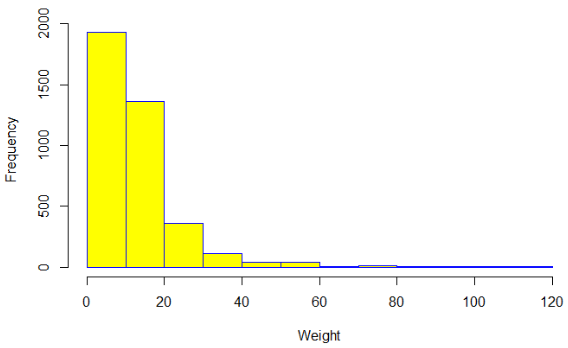


Figure 9: YouTube videos length from dataset

YouTube statistics show that the best duration for a video max is 15 minutes. However this depends on the video category. For example for music content the best duration is 3-4 minutes, for a comedy is 5 minutes. Figure 10 displays the distribution of video length for the top ten most popular categories.

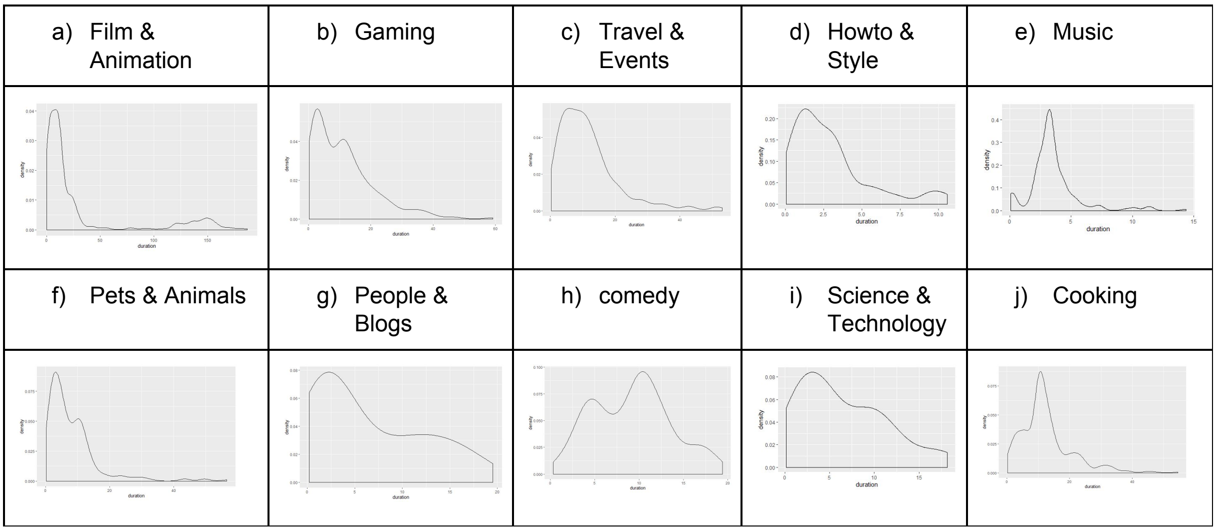


Figure 10: YouTube video duration Analysis plot

4.7 Comment Analysis

A useful tool for popularity analysis is sentiment analysis[13].In sentiment analysis, we gather comments from a channel and try investigate popular channels. What people are think about the videos, authors. Sentiments describe the audience rate about the video. YouTube is also using this tool for video verification. If total sentiment analysis positive the video is promoted to the Main page. Frequency of comments is one of the important popularity metric. If video comments are frequent this means that the video is popular (figure **??**).

The next investigation step is world cloud analysis. Most used words in video comments is key feature for the investigation video popularity.Because during search queries in YouTube also taking account this comments like tags.

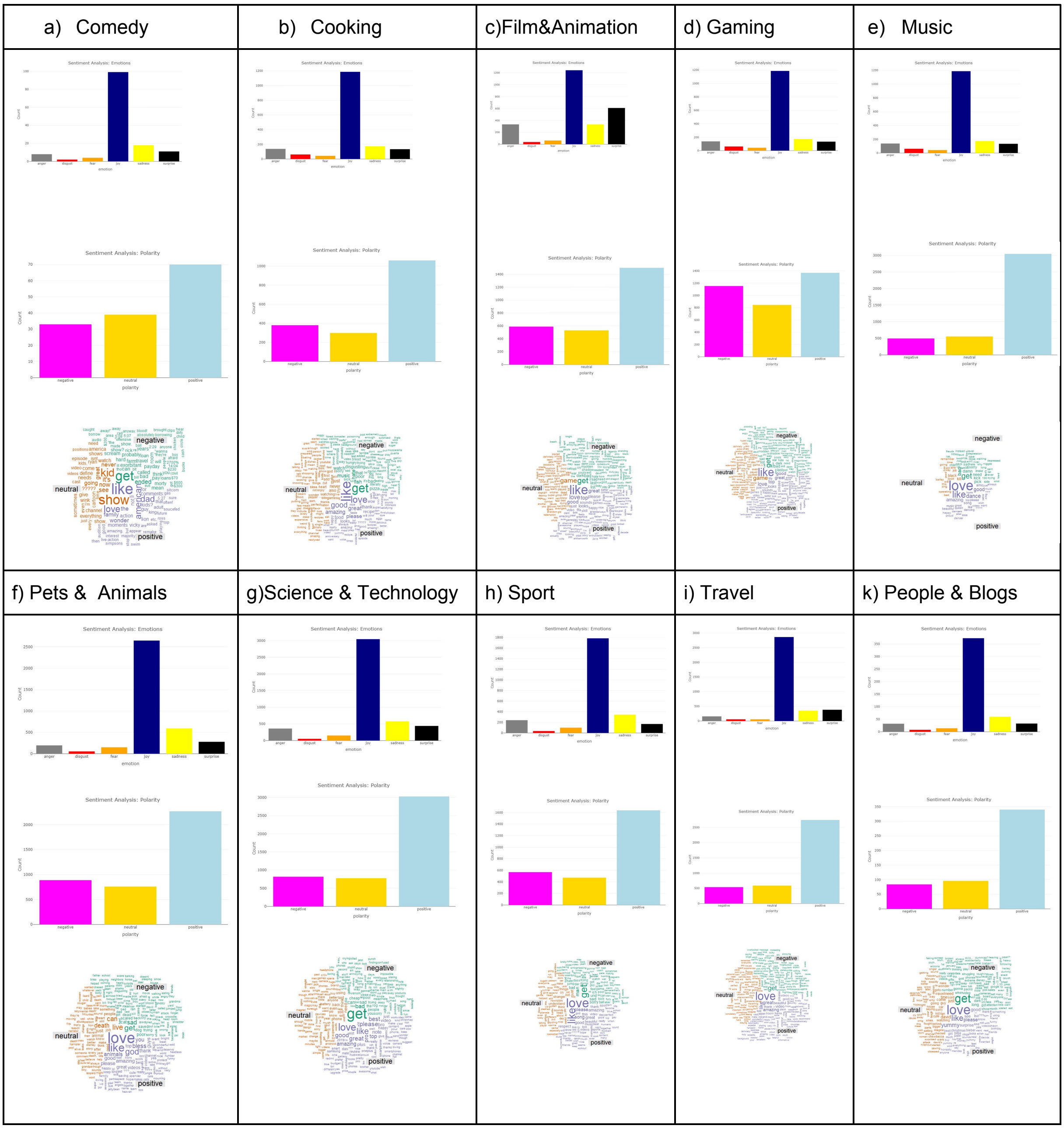


Figure 11: Comment Table

4.8 Popular videos analysis in the different countries

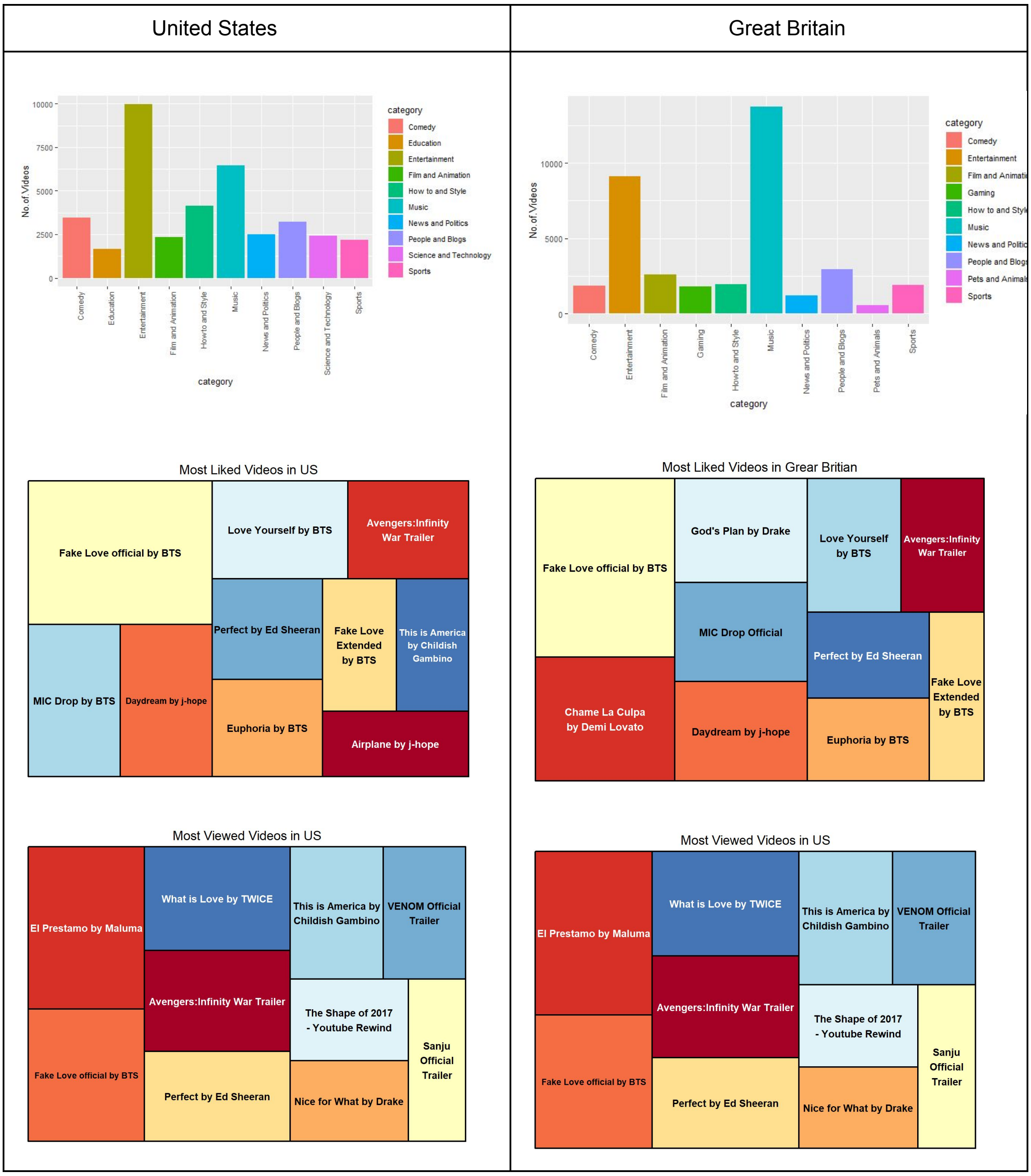


Figure 12: US vs GB

5 Results

When one channel has more than a million subscribers, we can say that the channel is popular. Sentiment analysis of the comments very important for the investigation video popular or not. After analyses of more than ten articles about video duration, we can say that the best duration is 15 minutes for the average. But this can change based on different video categories. For example, for music, this is 3-4 minutes, comedy 5 minutes, movies 1 hour. We are currently only doing preliminary analysis of datasets for that we can not say too much about the video popularity reasons.

6 Conclusion

This research is a first step for understanding YouTube popularity, which provides the initial foundation for the future explorations. We studied principles of YouTube popularity from simple data sets. We analyzed the relationship between key popularity metrics (viewcount, likecount, comments). Based this work we identified future research directions.

References

[1] C. S. Araújo, G. Magno, W. Meira, V. Almeida, P. Hartung, and D. Doneda. Characterizing videos, audience and advertising in youtube channels for kids. In *International Conference on Social Informatics*, pages 341–359. Springer, 2017.

[2] N. Ashar, H. Bhatt, and S. Mehta. A framework for detection of video spam on youtube. 2015.

[3] M. Bärtl. Youtube channels, uploads and views: A statistical analysis of the past 10 years. *Convergence*, 24(1):16–32, 2018.

[4] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Video interactions in online video social networks. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 5(4):30, 2009.

[5] A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web*, pages 241–250. ACM, 2012.

[6] F. Cena, E. Chiabrando, A. Crevola, M. Deplano, C. Gena, and F. N. Osborne. A proposal for an open local movie recommender. In *PATCH 2013: Personal Access to Cultural Heritage*, pages 1–6. CEUR Workshop Porceedings, 2013.

[7] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14. ACM, 2007.

[8] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking*, 17(5):1357–1370, Oct 2009. doi: 10.1109/TNET.2008.2011358.

[9] G. Chatzopoulou, C. Sheng, and M. Faloutsos. A first step towards understanding popularity in youtube. In *2010 INFOCOM IEEE Conference on Computer Communications Workshops*, pages 1–6. IEEE, 2010.

[10] X. Cheng, C. Dale, and J. Liu. Statistics and social network of youtube videos. In *2008 16th Interntional Workshop on Quality of Service*, pages 229–238. IEEE, 2008.

[11] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296. ACM, 2010.

[12] Q. Do. Youtube channel analysis. *RPubs*, 2018.

[13] R. Feldman. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4): 82–89, 2013.

[14] L. Hamzaoui Essoussi and D. Merunka. Consumers' product evaluations in emerging markets: does country of design, country of manufacture, or brand image matter? *International Marketing Review*, 24(4):409–426, 2007.

[15] M. Holland. How youtube developed into a successful platform for user-generated content. *Elon journal of undergraduate research in communications*, 7(1), 2016.

[16] A. Hossain. Youtube analysis on 'bangladesh'. *RPubs*, 2018.

[17] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[18] J. C. Paolillo. Structure and network in the youtube core. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, pages 156–156. IEEE, 2008.

[19] A. Sobecki. Service recomendation on wiki-ws platform. 2015.

[20] A. Susarla, J.-H. Oh, and Y. Tan. Social networks and the diffusion of user-generated content: Evidence from youtube. *Information Systems Research*, 23(1):23–41, 2012.

[21] M. Wattenhofer, R. Wattenhofer, and Z. Zhu. The youtube social network. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Ismayil Tahmazov**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   I hate you like I love you: A case study of likes and dislikes on YouTube,

   Supervised by Rajesh Sharma

2. I grant the University of Tartu a permit to make the work specified on p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons license CC BY NC, ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive license does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Ismayil Tahmazov
07/09/2019