Discovering the Special Characteristics of YouTube Users via Data Mining

Richard Huang, Caitlyn Tran, Esther Yi, Mayra Carrera, Polia Donova

Nomenclature

ID3	Iterative Dichotomiser 3
SQL	Structured Query Language

Abstract

This paper attempts to find the characteristics that allow YouTube channels that did not originally have a fanbase prior to the creation of the channel to eventually become popular. Once these characteristics are found, they can be applied to several other fields such as advertising or informational videos, including announcements and tutorials. The first tool we used was StatSheep, which collects statistics about YouTube users and compares different channels. Social Blade, which also collects information about the progress of YouTube channels, was another tool that was used. After using data mining and ID3 to sift through the vast amount of information about YouTube, we concluded that interactivity between users had the most impact on the popularity of certain channels.

Introduction

As globalization accelerates, connecting millions around the world, several media outlets of all genres have arisen, ranging from news and entertainment to music. Popular channels like Fox, CBS, and Disney consistently reach over one million viewers every day [1]. Yet, new forms of broadcasting have been created. Multiple content-sharing websites host an increasing number of content creators.

Of these websites, YouTube has risen to be one of the most popular for video sharing on the Internet. YouTube offers all of their users, who refer to themselves as YouTubers, a chance to make and post their own videos on the site, which other users can then see, comment on, and rate. Users are able to subscribe to other users, which lets the people who subscribed, known as subscribers, receive notifications whenever the content creator releases a new video.

Subscriber count is an accurate measure of the popularity and quality of a channel. For example, channels like *The Tonight Show With Jimmy Fallon* have the money and resources to ensure quality entertainment. However, this makes the fact that more than half of the top 100 youtube channels in terms of subscriber count are Although this may seem like a small portion, after discounting the 68 channels created by celebrities or large companies, who already would have had a significantly large fan-base, the 17 gaming channels make up over half of the remaining channels [2]. It is important to keep in mind that a majority of these channels simply feature people

playing games or acting out a short skit in front of a camera. For instance, the most subscribed channel on YouTube, user *PewDiePie*, may release a 12-minute video of playing a game, and receive upwards of 2-3 million views.

The primary goal of this paper is to determine the reasons as to why channel that began with unknown hosts hold such popularity among large corporations and celebrities that have much more resources and a much stronger foundation when beginning their channel. Companies have a need to understand the power, influence, and methods of these YouTubers, as they are the drivers of a new age of media, away from the days of regular television. Following them are millennials and Generation Z youth. As such, understanding what makes YouTubers popular allows us to know what younger generations are interested in, and what they look for in those they follow. These patterns have potential applications within fields of education, political science, and advertising, to garner attention and interest, also allowing organizations to gain popularity using similar means and techniques. Ultimately, this results in increased public support for what the organization stands for. Even for companies largely unrelated to the field of media and entertainment, these YouTubers have large audiences, and because they tend to be more relatable than large corporations or celebrities, sponsorships and advertisements attract a higher percentage of viewers than in regular cable TV [3]. As mentioned before, these methods can be applied to other organizations to increase public support for their cause.

This paper is organized as follows. Related literature works are reviewed in the next section, followed by a project description, along with the proposed methods used in this paper and the experiment description. Afterwards, data from the experiment is discussed and, in the last section, the conclusion is given.

Literature Review

On the YouTube Help Site [4], created by the hosts of the website itself, several points on how channels are promoted on YouTube were explained. YouTube itself commonly advertises the content users post. This is based on a retention-based algorithm, in which the more Watch Time, which is how long users watch videos in one session, channels bring, the more they will be promoted by YouTube. These promotions occur through YouTube's front-page system, in which the user is brought to the front page where the most successful videos can be seen when they first access the website. The other way is personalized towards each user, based on watching habits, in which YouTube will present a user with several recommended videos. If a channel wishes to earn money, they are able to participate in YouTube's partnership program, which involves giving YouTube permission to broadcast ads on a channel's videos. These pay on a per-click basis.

One of the most important topics the site covers is the distribution of recommended videos. Channels in one region of the world are typically recommended to viewers within the same region. However, there is still a greater priority on Watch Time.

Figueirdo et al. [5] have conducted research regarding the growth of videos and the patterns that arise based on how YouTube's algorithms suggest videos to viewers. They came to the conclusion that listing related videos are the most effective at bringing in viewers within similar genres except for videos from content creators without a fan base, in which search referrals will be the main source of views.

Additionally, one of the most cited papers on this topic provides further information about what makes YouTube the one of the most successful video distribution platforms. Since the website was set up in 2005, YouTube has added several features aiming to appeal to users, such as the links of related videos mentioned on the YouTube Help Site. Cheng, Dale, and Liu [6] investigated further to find the secret behind YouTube's success by spending four months looking through the video data of over 3 million users. After their research, they concluded that the reason behind YouTube's popularity is their social network of videos, based on the preferences of the user. Using this information, the individual popularity of videos can be predicted.

A recent study by Hou et al. [7] attempts to predict the likes and dislikes of movie trailers. Their observations can be applied to YouTube videos. A total of 725 trailers were downloaded from YouTube to conduct the experiment. According to the study, there are three main components of movie trailers that decide their success or failure, including color, motion and emotion, and shot features, known as low level feature extraction. A conclusion made after the experiment states the shot length variance affects the likes and dislikes the most.

Another relevant article written by Cheng, Fatourechi, Ma, Zhang, and Liu [8] attempted to understand what YouTube partnership businesses sought in YouTube channels. Though their purpose differed from our own, their research relates to the features that allowed YouTube to become as popular as it is today. They concluded that premium partners are responsible for the increased number of YouTube viewers and predicted that the amount of views could be increased further if we expanded our knowledge about the distribution of decay in video views, characteristics of user visiting behavior, impact of user engagement, and video sharing services in written works.

Using the results of these related works, we were able to compare them with our own findings of the attributes that allow certain YouTubers to gain popularity. We then utilized these characteristics and can now apply them to other domains of research, such as advertising and education.

Project Description

This research was aimed to discover the hidden correlations between the most popular YouTube channels that acquired their start on YouTube. This was done by gathering aspects of the most popular YouTube channel's videos as well as features that make a channel unpopular. To carry this out, we observed the most and least popular channels, along with instances in time where subscribers were lost. Then, using the Iterative Dichotomiser 3 (ID3) Decision Tree to analyze and

compare the data, correlations about what made certain channels more popular than others, which were not visible with general analysis, were found.

Method

One tool we used was StatSheep [9], a statistics collecting site that tracks the contents, subscribers and view changes, related networks, and overall subscriber and view count rate of YouTubers. In addition to that, this tool allows us to compare the graphs of two channels. The graphs display the channel's video views and subscriber difference in numbers, showing the contrast between their uploads, subscriptions, views, earnings, and contacts. StatSheep then decides a *winner*.

Another tool we used was Social Blade [10], a tool that collects statistics for eight million accounts on Youtube. According to their *About* section, they are a news blog focused on updating their readers about social media, online videos, and events. This tool tracks progress and growth of Youtube channels by calculating the number of subscribers and percent change of subscriber count over time with graphs and charts, the number of total channel views, and the total number of video uploads.

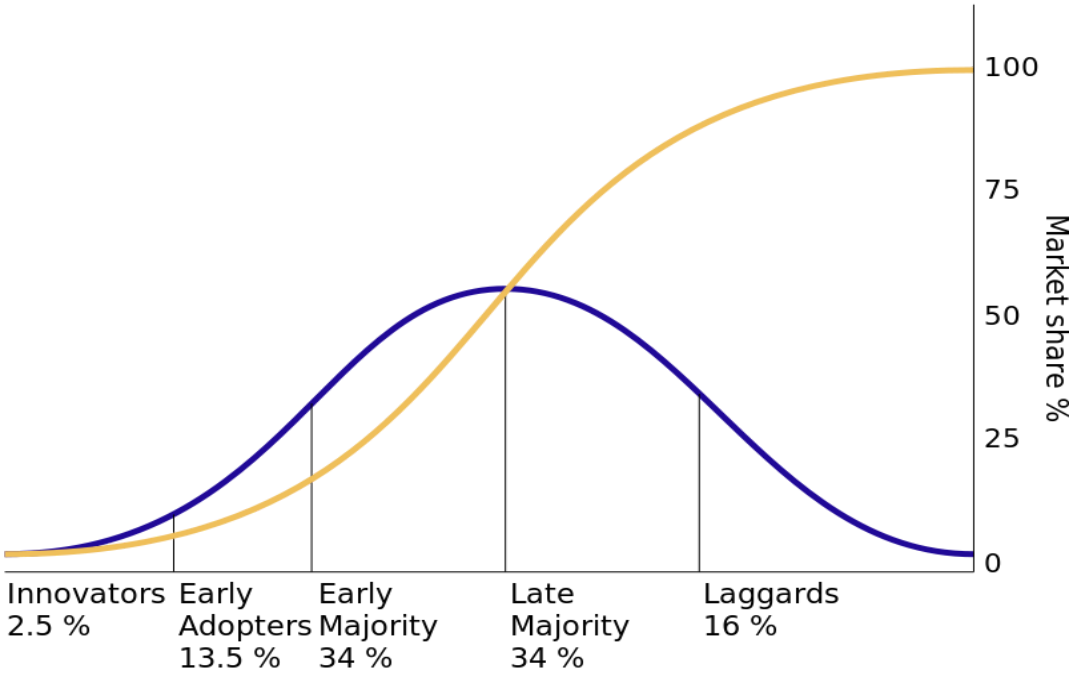

Fig. 1. Everett Roger's Diffusion of Innovation Model. In this graph, a blue line, sections divided based on Roger's adopter classifications, and a yellow market share line display the state of diffusion in the market as more people adopt an innovation. [11]

In Figure 1, a form of classification used within the process was Everett Roger's Diffusion of Innovations model, also known as the technology cycle or technology curve, which is based on a normal distribution, in which innovators, early adopters, early majority, late majority, and laggards

are classified, as shown in Figure 1 above. Based on the technology cycle, in order for some form of technology to truly become popular, it must pass the *Innovator* and *Early Adoption* stages of growth, into *Early Majority*. Those with sufficient financial resources are willing to become *Innovators*, the earliest adopters. Videos of this nature usually attract the least viewers, as the common viewer is unwilling to spend time watching a video by a user they have not heard of before. However, the video provides a push through the technology adoption curve into *Early Adoption*, where several other smaller channels, seeing less risk, decide to make a similar video. Based on the reception of these videos, the YouTuber either falls back, failing to become popular, or pushes onwards to *Early Majority*. From here, the YouTuber's videos become a trend and will carry itself to the end of the curve. When this happens, initial *Innovators* also receive a burst of popularity, as their videos are the first and most viewed when the trend first begins, and in this way, a channel can quickly become popular.

At first, we considered using Structured Query Language (SQL). SQL is a programming language designed specifically for organizing data in databases. This language has the ability to combine different tables and select certain attribute values that follow specified conditions. This way, the data can be searched for with key attributes. This could have been applied to the topic at hand by organizing data about each YouTuber into a database and finding keys that may have a correlation with their popularity. We discovered that a YouTuber's primary method of retaining and gaining subscribers is to associate the channel with a positive item or idea, such as giveaways, charity, good gameplay, a good personality, or any combination of these, to instill a desire within viewers to engage and support the channel more. However, this method proved to be inefficient because many of the personal characteristics about each YouTuber are subjective. For example, among the attributes we came up with, there was the personality of the YouTuber. However, while one person might find the YouTuber to be funny, someone else may think otherwise. Thus, we decided to use ID3 [12], which is an algorithm used to create a decision tree from a dataset, aiming to create pure subsets, which have aligning classes. Beginning with a single set of data, the algorithm analyzes the distinct attributes of the set and, for each one, calculates the entropy, or information gain, of that attribute, using the equation in Figure 2. The closer the resulting values of the equation is to 0, the greater entropy the data will have. The algorithm determines the attribute with the greatest entropy and divides the set into subsets with distinct instances of that attribute. It then repeats this process with each non-pure subset, or each set that does not lead to a single decision, until every subset is a pure set. Using this algorithm, one can look at specific instances of attributes in a group of data and predict the data's class.

These tools that we chose to use in the end aided in gathering and relating data during the *Process*, which helped in classifying and establishing patterns within the observed sets.

$$H = -\sum_i p_i(\log_2 p_i)$$

Fig. 2. This is the algorithm used to calculate entropy from the Computer Science Source [13].

Process

In general, larger YouTube channels will constantly be gaining new subscribers and views, rarely seeing a net loss of subscribers in a day. As such, the first thing we did was look at instances among channel history in which subscribers were lost, rather than gained. Looking at why subscribers dislike a channel's content enough to unsubscribe is pivotal to understanding why a general user would enjoy the content.

After this, we watched the most and least popular videos on each channel and compared their lengths, titles, comments, and dates, and linked them to each other to find the preference of viewers. However, these videos will be restricted to those released within the current year of this paper, 2016, as the growth of a channel can signify differences of several million views, and older videos have more opportunity to be seen than newer videos.

Between these instances of data gathering, we found and connected related information, such as trends at the time of video release, comments of general users, and reaction videos that express a popular opinion.

Data Results and Discussion

Table I . Change in subscribers over a period of 11 days using data from Socialblade [10].

Channel	PewDiePie	HolaSoyGerman	elrubiusOMG	VanossGaming	markiplierGAME
June 23, 2016	40,046	15,049	22,579	18,796	17,570
June 24, 2016	35,120	19,141	22,916	18,602	18,633
June 25, 2016	37,869	18,194	26,216	17,617	17,190
June 26, 2016	44,926	17,058	26,861	19,745	17,128
June 27, 2016	51,469	14,358	23,176	19,108	17,559
June 28, 2016	-558,666	-12,001	-169,753	-243,895	-295,361
June 29, 2016	55,100	13,064	17,508	15,579	16,506
June 30, 2016	74,316	11,558	16,632	13,195	14,362
July 1, 2016	58,495	13,235	15,585	12,513	13,350
July 2, 2016	47,327	14,565	19,497	12,816	11,753
July 3, 2016	38,562	13,467	21,661	12,398	12,386

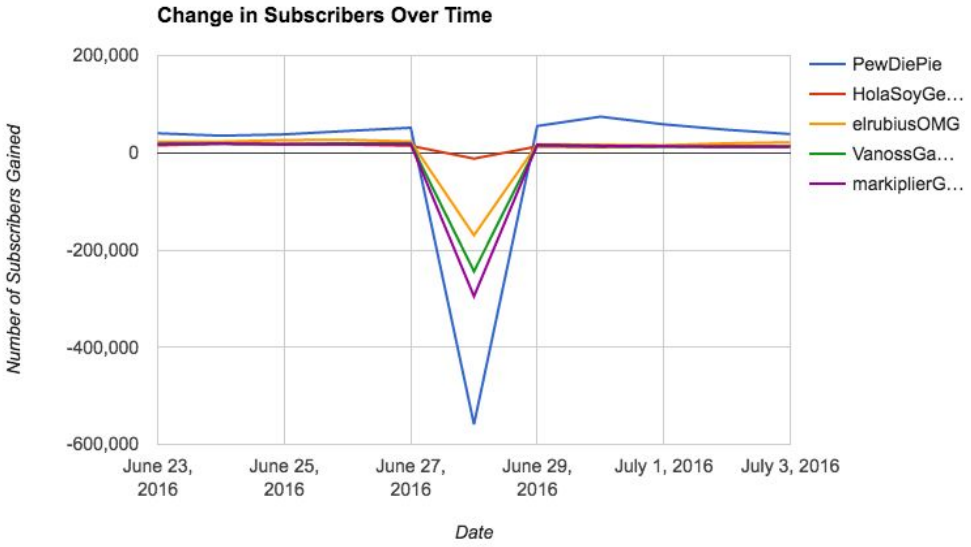

Fig. 3. The change in subscribers over an 11 day period based on Table I. There is a sudden drop in subscriber gain among all of these top channels on June 28, 2016.

Table II. Attributes that make YouTubers popular using data from YouTube [14].

Channel	Fanbase/ Support of a Company	Genre	Channel Format	Gender	MPV Views	MPV Topic	Upload Schedule	Top 100
PewDiePie	no	games	individual	M	73628098	compilation	daily	yes
HolaSoyGerman.	no	entertainment	individual	M	62910497	song	monthly	yes
YouTube Spotlight	yes	news	group	B	109927460	song	monthly	yes
JustinBieberVEVO	yes	music	individual	M	1602366553	song	monthly	yes
Smosh	no	comedy	group	M	101909858	skit	weekly	yes
Sebastián Villalobos	no	people	individual	M	4630183	howto	weekly	no
netd müzik	yes	music	group	B	190282943	song	weekly	no
Mister V	yes	comedy	individual	M	14374151	story	yearly	no
El Reino Infantil	yes	music	group	B	236196068	animation	weekly	no
YRF	yes	music	group	B	97427759	song	weekly	no

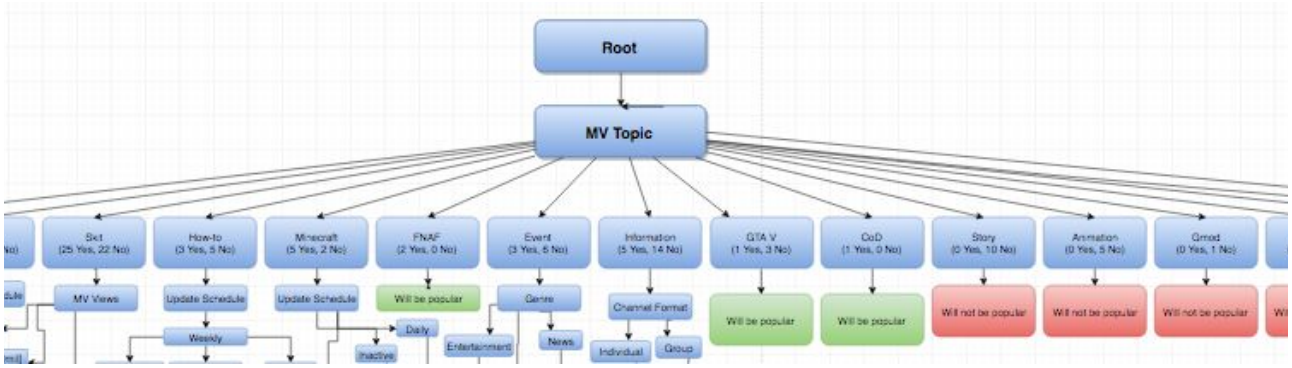

Fig. 4. A sample of the ID3 decision tree we made.

Among several of the top YouTube channels, such as *PewDiePie* and *markiplierGAME*, a recent notable instance of the loss of subscribers was on June 28, 2016. These instances stand out, because they had not occurred in weeks and never to such a degree. As shown in table 1, the channel *PewDiePie* lost about 558,666 subscribers while *markiplierGAME* lost about 295,361 subscribers. These were the channels that suffered the greatest losses in terms of both numbers and percentages, along with other gaming channels. Other channels, such as *HolaSoyGerman*, who have a similar number of subscribers, only lost 12,001 subscribers, which is significantly fewer in comparison.

Corresponding with this date was the release of Youtuber *FaZe*'s video on the topic of *clickbait*, a strategy for views in which a channel titles their video misleadingly or has a false thumbnail (video titular picture), in order to attract viewers, who typically expect videos to be something related to the title and thumbnail.

In a personal investigation, we found that it is indeed true that, among the several genres of channels, gaming channels have a higher frequency of using misleading titles, thumbnails unrelated to the title, and exaggerating the magnitude of what occurs on the video. This correlation between the date of subscriber loss and the release of *FaZe'*s video indicates that common users believe this to be true as well, causing many viewers to leave these channels.

However, within comments of videos in response to these subscriber losses, one notable argument for fans was that the use of clickbait titles do not matter, as they subscribed for the personality of the YouTuber. For example, YouTuber *PewDiePie* posted a video about clickbait as an ironic joke, and one of the comments displayed how his subscribers did not care about the use of clickbait titles, as shown in Figure 2 below. These comments have a lower occurrence in gaming channels than in video blogs.

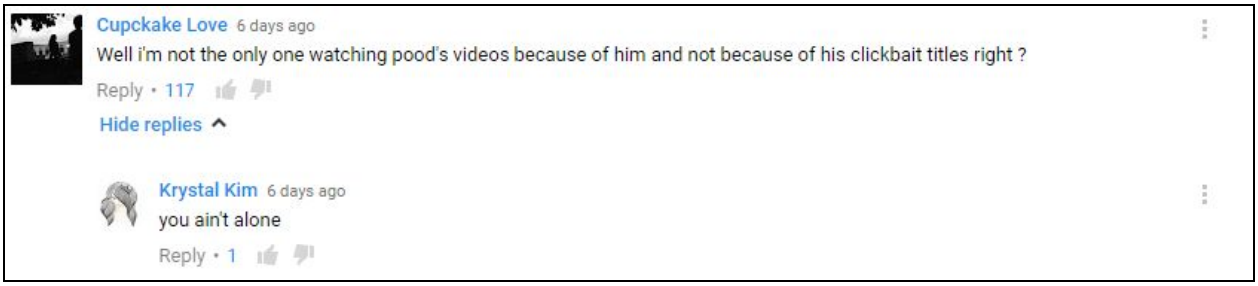

Fig. 4. This is an example of how clickbait did not affect subscribers on Pewdiepie's channel on YouTube [14].

As displayed by all top 50 YouTube channels, a channel will generally remain with the genre it started with in terms of subscriber count. For example, although the most popular channels are typically a mix between gaming and video blog channels, there is a distinction between the subscribers of these two types of channels. Channels that initially became popular through video blogs may create a separate channel for gaming, but it will always remain within their initial genre. This is vital to understanding why subscribers are lost. In observing the comments of old gaming videos and old blog videos of popular channels, we have noted that, within gaming channels, viewers comment about the reaction of the player to the game, while in video blog channels, viewers comment about their own reaction to the channel's host. A gaming channel relies on a relationship between the viewer, a game, and the host to gain new subscribers while a video blog channel relies solely on the relationship between the viewer and its host's personality. Though gaming channel viewers may come to appreciate a gaming channel host's personality as well, there is still a dependence on the video game itself to facilitate this appreciation. Thus, in general, when a YouTuber switches to another genre, they lack the relationships viewers are used to, which is what causes a loss in subscribers.

There are advantages and disadvantages to this reliance, as a channel may receive a burst in popularity if they have taken a successful risk when covering an unknown topic that eventually becomes popular, as outlined Roger's Diffusion of Innovation. If the topic receives a positive reception by users, being associated with that topic may improve a host's own reception.

When observing the patterns of the most popular videos and least popular videos with this knowledge in mind, we noticed that both the most and least popular videos cover topics when they are in the stage of *Innovators*, in which the idea has not become popular yet. This correlation between videos of opposite view counts highlights a notable strategy that can be used for popularity.

Conclusion

In this paper, we have searched for patterns regarding the reason YouTube channels that began without a fan-base managed to become as popular as they are now. We concluded that there are many attributes that cause a YouTube channel to become popular, which come with certain rules and patterns. For example, video topics, such as stories, animations, and interviews, proved to

be quite unpopular, as none of the top channels had those topics as their most viewed videos. Using observations such as these, we made a decision tree to determine whether a YouTuber's channel will become popular.

We faced many limitation. First of all, there was a limited reliability of several websites in researching the topic, due to its influence largely remaining solely on the Internet. Also, we used a fairly small sample size, in comparison to the hundreds of other popular channels with several million subscribers that we could have also observed, which also somewhat reduces the reliability of our patterns and conclusions. In addition, another limitation faced was the fact that the data is unstable, as it is constantly changing. For instance, the number of subscribers and views a video gets is not a static variable. As time goes on, the number of subscribers will fluctuate, and the number of video views will increase. Not to mention, it was very difficult to complete the ID3 decision tree in addition to collect and analyze all the data we managed to gain during the given 6-week period. Thus, we only included a sample of the ID3 tree we were working on.

In the future, we hope to complete the tree we started and encourage other researchers to use our data to make their own ID3 decision tree. Once we have accomplished this, we will investigate a larger sample size, with the hopes of both making our conclusions more credible and finding more patterns amongst channels that are not in the list of top 100 most viewed YouTube channels. We anticipate this line of research will bring about a change in how companies and organizations advertise or broadcast themselves to the younger generations on this growing media platform as well as the strategies and forms educators and teachers will bring forth to convey the most knowledge.

Acknowledgements

This research team would like to thank the U.S. Department of Defense for funding the STEM summer program, Dr. Justin Zhan for directing this program, and Sweta Gurung for managing it. Due to their efforts, we were able to meet Dr. Ming-Tai Wu, Dr. Jain-Shing Wu, and Mr. Payam Ezatpoor, who taught us all about data mining and introduced us to its applications to real world concepts, such as YouTube. Without the program and our mentors, this paper would not be possible.

References

[1] M. Schneider. (2015, December 28). *Most Watched Television Networks: Ranking 2015's Winners and Losers (1st Edition)* [Online]. Available: http://www.tvinsider.com/article/62572/most-watched-tv-networks-2015/.

[2] "YouTube Top 100 Most Subscribed Channels List - Top by Subscribers". (n.d). *Vidstatsx.com*. [Online]. Available: http://vidstatsx.com/youtube-top-100-most-subscribed-channels. Accessed Jul. 10, 2016.

[3] J. Klima. (2014, June 14). *Click it or Skip it: Does YouTube Advertising Actually Work? (2nd Edition)*. [Online]. Available: http://newmediarockstars.com/2014/06/click-it-or-skip-it-does-youtube-advertising-actually-work/

[4] "YouTube Help". (n.d). *Support.google.com*. [Online]. Available: https://support.google.com/youtube/?hl=en#topic=4355266. Accessed Jun. 28, 2016.

[5] F. Figueiredo, F. Benevenuto and J. Almeida, "The tube over time", *Proceedings of the fourth ACM international conference on Web search and data mining* - WSDM '11, 2011.

[6] X. Cheng, C. Dale and J. Liu, "Statistics and Social Network of YouTube Videos," *Quality of Service*, 2008. IWQoS 2008. 16th International Workshop on, Enschede, 2008, pp. 229-238.

[7] Y. Hou et al., "Predicting Movie Trailer Viewer's "Like/Dislike" via Learned Shot Editing Patterns," *in IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 29-44.

[8] X. Cheng, F. Mehrdad, X. Ma, C. Zhang and J. Liu, "Understanding the YouTube partners and their data: Measurement and analysis," *in China Communications*, vol. 11, no. 12, pp. 26-34, Dec. 2014.

[9] "Welcome to StatSheep - YouTube Channel Statistics". (n.d). *Statsheep.com*. [Online]. Available: http://www.statsheep.com/. Accessed Jun. 29, 2016.

[10] "All About Social Blade". (n.d). *Socialblade.com*. [Online]. Available: https://socialblade.com/info. Accessed Jul. 3, 2016.

[11] E. Rogers, *Diffusion of innovations*. New York, NY, USA: Free Press, 2003, pp. 16-30

[12] J. Quinlan, "Induction of decision trees", *Mach Learn*, vol. 1, no. 1, pp. 81-106, 1986.

[13] BADGERATI, "Machine Learning – Entropy". *Computersciencesource.files.wordpress.com*. N.p., 2016. Web. 11 July 2016. [Online]. Available: https://computersciencesource.wordpress.com/2010/01/10/year-2-machine-learning-entropy/

[14] "YouTube". (n.d). *Youtube.com*. [Online]. Available: http://youtube.com Accessed Jul. 6, 2016