YouTube Video Popularity:
Predicting Video View Count From User-Controlled Features

Jordy Snijders
ANR 386791

Master’s Thesis
Communication and Information Sciences
Specialization Business Communication and Digital Media

Faculty of Humanities
Tilburg University, Tilburg

Supervisor: Dr. G.A. Chrupala
Second Reader: Dr. C.S. Shahid
July 2015

Table of contents

Acknowledgements

This Master's thesis concludes my educational career at Tilburg University. It all started for me with the pre-master program in *Communication and Information Sciences* of one year at Tilburg University. At the start of that academic year, I hardly knew what a research paper was, let alone how important statistics are. I successfully concluded that program and ventured my way through the *Business Communication and Digital Media* track of the Master program *Communication and Information Sciences* where I got to apply statistics, language processing and machine learning in a practical setting. So, I can honestly say that I have learned a great deal while studying at Tilburg University.

I would like to express my appreciation for Grzegorz Chrupała as a person and for his guidance during the time in which I had to write this Master's thesis. As my supervisor, he reminded me time and time again of the things that I absolutely had to write besides the things that I so eagerly wanted to write. He also flawlessly spotted even the tiniest of mistakes that I made, so, for his eye for detail, I am grateful. Machine learning is as much a precise science as it is a form of art and Grzegorz showed this to me, even though I am fairly sure that I have only scratched the surface of what machine learning is about.

Next to this, I would like to thank my second reader, Suleman Shahid, as he too had to read my wall of text and assess it. Finally, a 'thank you' goes out to my friends and family for being supportive of my choices and endeavors, and to my fellow students who made attending Tilburg University as enjoyable as it was.

Abstract

An increasing number of studies are investigating popularity prediction on Social Networking Sites such as Facebook, Twitter and YouTube. Also, many people and businesses are getting more interested in becoming popular in the digital world and for various reasons ranging from personal and social achievements to financial benefits among other justifications. This research contributes to the topic of online popularity prediction by predicting the view count of YouTube videos based only on user-controlled features. A data set of 1.5 million YouTube videos is used to construct a Stochastic Gradient Descent-based linear regression model. With an explained variance of 53.3% on test data, the predictive model, containing features that were selected through experimentation, shows a good performance. This finding presents a challenge for a predictive machine learning system to be created that analyzes videos and returns advice on how to improve the video before it is uploaded on YouTube.

1 Introduction

1.1 Motivation

Social Networking Sites (SNS) allow people to create a personal profile, make a list of friends and family, and look at lists of friends, family and others (Boyd, & Ellison, 2007). Next to this, people can typically share videos, photos and texts, and show their appreciation for what others are doing amongst other things (Lin, & Lu, 2011).

Many SNSs exist with Facebook, Twitter, LinkedIn, YouTube and Instagram among the most used SNSs at present day (Boyd, & Ellison, 2007). This first major SNS, Six Degrees, was founded in 1997 with others to follow in 1999 (e.g., LiveJournal and AsianAvenue). So far, more than 30 other SNSs have been added to that list but they all adhere to the premise of people having a profile and making lists of friends (Boyd, & Ellison, 2007). Differences between SNSs exist such as with some SNSs focusing on certain types of media (e.g., videos for YouTube, photos for Flickr, résumé documents for LinkedIn). Also, some SNSs allow a user to mark their profile as private whereas other SNSs have no such option. Another difference is that, even though most SNSs allow for sending messages via profiles, some also feature direct private messaging. What most SNSs have in common though is that they are accessible through the World Wide Web as only a few SNSs are solely available through a specific type of means (e.g., via a smartphone app) (Boyd, & Ellison, 2007). Finally, a limited number of SNSs only accept people from specific geographical regions.

People are motivated for all kinds of reasons in their use of these different SNSs. The most important reasons, however, are that people mostly use SNSs for their enjoyment, the number of friends and family that already make use of the SNS, and the usefulness of the medium itself (Lin, & Lu, 2011). With regards to the social reasons that motivate people, contacting new people, staying in touch with friends and family, and socializing in general are the most important reasons (Brandtzæg, & Heim, 2009). Apparently, the fact that other people make use of an SNS attracts more people towards that SNS.

Once people have a preference for a SNS and are motivated to use it two kinds of people can be identified, those who actively participate and those who lurk (i.e., people who merely observe with the absence of any two-way interaction) (Burke, Kraut, & Marlow, 2011). Some of the people who actively participate crave for themselves to be more popular on that particular SNS and this is referred to as the Need for Popularity (NfP) (Utz, Tanis, & Vermeulen, 2012).

However, this concept is not tied to SNS usage nor is it new. Creating and maintaining large social networks has always been important to people. Many consumers dream of being popular among their friends and family, and businesses are continuously attempting to enlarge their customer reach. If done properly, even nonprofit organizations grow as a result of the increased awareness of the organization among people (Waters, Burnett, Lamm, & Lucas, 2009). Some of the reasons for wanting large networks are the social benefits, such as high self-esteem and satisfaction with life (Steinfield, Ellison, & Lampe, 2008) or reduced levels of depression and anxiety (Zimet, Dahlem, Zimet, & Farley, 1988). Next to this, one's increased popularity can result in more access to information, having more power, and receiving more solidarity from others (Adler, & Kwon, 2000). Other financial benefits that are associated with having many connections are the social skills and competences that are learned which can result in more effective interaction with others (Baron, & Markman, 2000). For example, one may get better at reading others, make better first impressions, or become adaptive, persuasive, or influential.

The benefits of being popular extend beyond the physical world and into the digital world. For SNSs, popularity is typically measured by the number of friends someone has (e.g., on Facebook), the number of time an image has been liked (e.g., on Instagram), or the number of times a video has been watched (e.g., on YouTube). Many studies have investigated the numbers behind existing SNSs such as Flickr for photos (Cha, Mislove, & Gummadi, 2009; Lipczak, Trevisiol, & Jaimes, 2013), Twitter for messaging (Hong, Dan, & Davison, 2011; Bandari, Asur, & Huberman, 2012; Kupavskii, Umnov, Gusev, & Serdyukov, 2013), YouTube for videos (Cheng, Dale, & Liu, 2007; Cha, Kwak, Rodriguez, Ahn, & Moon, 2009; Figueiredo, Benevenuto, & Almeida, 2011), and Slashdot for news (Kaltenbrunner, Gomez, & Lopez, 2007). In this regard, all kinds of angles have been taken in previous research, from metrics related to popularity to the characteristics and the potential of such SNSs as well as their technical feasibility.

This research focuses specifically on popularity on YouTube which is a widely used SNS where people share videos and have full control over video content, titles and descriptions among other video properties. Viewers can watch, like, dislike, or comment on these videos or add them to playlists. To stay in the loop of new videos, viewers may also subscribe to those who upload videos. However, YouTube has been altered considerably over the last five years (Evans, 2013) and some of the metrics that were used for prediction analyses in earlier work have been modified or removed. Next to this, YouTube has added new metrics that show promises for new predictors for video popularity. By extracting user-controlled features from present day video metrics of 1.5 million YouTube videos, this study takes some of the new metrics and constructs a truly predictive regression model. Such a supervised machine learning setup analyzes videos after they have been uploaded on YouTube but it presents a challenge as well. A system could be created that also gives advice on how people can improve their videos before they have been uploaded on YouTube. This would be a welcome tool for those who want to become popular on YouTube.

1.2 Contribution

From a theoretical perspective, any findings related to online popularity prediction give additional insight into social interactions. The introduction of SNSs have made it easy for people to interact socially and to form both strong and weak ties (Ellison, Steinfield, & Lampe, 2007). The statistics for many of these successful SNSs reinforce this as Facebook has more than 1.39 billion monthly active users (Facebook, 2015), more than 3 billion check-ins have been registered by Foursquare (The Next Web, 2015), the popular video *Gangnam Style* by *Psy* has received more than 5.3 million reactions from people (YouTube, 2015a), more than 500 million new Tweets are sent per day (Twitter, 2015), and Instagram processes more than 2.5 billion likes of photos daily (Instagram, 2015). These numbers clearly indicate the reach of these platforms. The metrics involved are interesting to consumers, businesses and organizations alike as they can use these to methodically improve their search for fame or for attracting more customers.

Predicting popularity on SNSs is also important from a practical point of view. More effective communication tools can be developed as a result of the findings of such research. This means that the design of products and services are influenced by the use and performance of existing SNSs (Figueiredo, 2013; Pinto, Almeida, & Gonçalves, 2013). By looking at usage patterns of SNSs (Gill, Arlitt, Li, & Mahanti, 2007), how information spreads on such platforms (Cha et al., 2009), and by deciphering what draws people’s attention (Yang, & Leskovec, 2011; Wu, & Huberman, 2007), many commercial benefits arise. Users of such services are increasingly put in a position that allows them to be more creative and to use this position for business related endeavors (Cha et al., 2009). For instance, businesses can use information on what is popular (e.g., products) on SNSs to improve their own websites’ search capabilities. This could lead to better product findability, or more personalized recommendation systems for content (Kong, Ye, & Feng, 2014). It also enables the companies that deal with a wealth of digital content to better filter their content and make such content better accessible (Lerman, & Hogg, 2010).

1.3 Outline

The *Introduction* chapter has already defined the subject of this research, provided the motivation, and has shown its relevance. In the text to follow, the underlying theory is first presented in the *Theoretical background* chapter along with this study's research questions. After this, the setups for several regression analyses that were conducted are elaborated on in the *Method* chapter. This is followed by the *Data* chapter in which a descriptive presentation is shown of the data that were gathered for the regression analyses. Then, several linear regression models with different setups are experimented with in the *Experiments* chapter and their evaluation scores influenced the construction of the final regression models. The results of those final regression models are presented in the *Results* chapter and their meanings are explained in the *Discussion* chapter. This includes an evaluation of a linear regression model from previous research as well. Finally, the main findings of this research and their implications are iterated on in the *Conclusion* chapter along with limitations and suggestions for future research.

2 Theoretical background

2.1 Video view count as YouTube popularity metric

Studies that deal with popularity on YouTube have focused on video properties (e.g., category, duration, size) (Abhari, & Soraya, 2010); referrals (Figueiredo, Benevenuto, & Almeida, 2011); usage patterns, file properties and social network maintenance (Gill, Arlitt, Li, & Mahanti, 2007); social interactions (Halvey, & Keane, 2007); related videos (Cheng, Dale, & Liu, 2007); video duplicates and popularity classes (Pinto, Almeida, & Gonçalves, 2013); and many other topics. Most of these works agree that video view count is the most important metric related to video popularity, followed by the average user rating and the number of ratings, because the number of views a video receives is indicative of access patterns (Abhari, & Soraya, 2010). Since 25% of videos on YouTube are viewed more than once (Zink, Suh, Gu, & Kurose, 2008) and cannot be downloaded from YouTube (Gill et al., 2007; Abhari, & Soraya, 2010), these access patterns are highly informative for investigating video popularity.

View counts of videos on YouTube can be characterized by a Zipfian distribution (Cha, Kwak, Rodriguez, Ahn, & Moon, 2007; Gill et al., 2007) and the same holds true for various other online media (Chesire, Wolman, Voelker, & Levy, 2001; Almeida, Krueger, Eager, & Vernon, 2001; Cherkasova, & Gupta, 2004; Sripanidkulchai, Maggs, & Zhang, 2004; Yu, Zheng, Zhao, & Zheng, 2006). Zipf's law applied to YouTube videos means that the number of views (F) a video receives is related to its rank (R) according to,

$$F \sim R^{-\beta}$$

where β is a constant number of approximately one (Gill et al., 2007). Plotting the number of views of YouTube videos and their ranks on a log-log scale would produce a straight line, which is indicative of such a Zipfian distribution.

2.2 YouTube video popularity over time

Unlike certain other media types (e.g., user-contributed news articles on Digg), videos on YouTube keep getting views throughout their lifetime. A strong positive correlation even exists between videos' early (log-transformed) view counts and their future view counts (Szabo, & Huberman, 2010). However, this only holds true for videos have been online for half a year because videos' ranks on YouTube change significantly their early in their lifetime (Borghol, Ardon, Carlsson, Eager, & Mahanti, 2012). Instead, for videos younger than half a year, the uploader’s characteristics, the number of keywords associated with the video, and the video’s quality are indicative of its future view count (Borghol et al., 2011).

One can look at a video’s view count before, at, or after a popularity peak to assess its popularity over time. This reveals that about 75% of videos on YouTube peak in popularity within the first six weeks after uploading (Borghol et al., 2011) and that videos with the highest view counts typically remain popular in between 1 hour and 41 hours, and 12.2 hours on average (Abhari, & Soraya, 2010). Long-term popular videos (i.e., videos listed in the weekly or monthly top 100 most viewed videos) tend to have durations in the 2.5 and 3.5 minutes range and are always shorter than 10 minutes as well as shorter than unpopular videos (Gill et al., 2007). Finally, YouTube videos have a daily peak of video view count at 13:00h (24 hour format) as well as a weekly peak on Sunday (Chatzopoulou, Sheng, & Faloutsos, 2010).

2.3 Content-related factors affecting YouTube video popularity

People on YouTube are facilitated with options for interacting with others such as the ability to share User Generated Content (UGC). UGC can be defined as content that is developed by normal people as opposed to people who get paid to do so (professionals) (Daugherty, Eastin, & Bright, 2008). When YouTube is tasked with presenting UGC, it typically favors items ranked from most popular to least popular (Cho, & Roy, 2004; Mossa, Barthelemy, Stanley, & Amaral, 2002). On YouTube, people share so much data that an easy way to display relevant items is to show what other people have already seen and appreciated (Borghol et al., 2012; Crane, & Sornette, 2008). A side effect of this is that content with a lot of appeal usually gets an enormous amount of attention whereas items from the long tail (i.e., fresh and innovative content of a smaller scale) tend to get pushed back (Cha et al, 2007; Lerman, & Hogg, 2012; Brodersen, Scellato, & Wattenhofer, 2012).

Several methods, related to video content, exist for people who want to get as many video views on YouTube as possible. Drawing people’s attention is a good initial step which can be achieved by publishing new content that is novel (Carmel, Roitman, & Yom-Tov, 2010); however, people’s attention towards a UGC item on YouTube typically fades after a while. This fade of attention follows a natural time scale of the stretched-exponential type (Wu, & Huberman, 2007) which means that attention never truly fades out entirely but eventually relaxes where low levels of attention remain. So, drawing attention is a challenge and quite necessary for getting popular whereas keeping attention is an almost impossible achievement. Besides the finding that novel content draws people's attention (Carmel et al., 2010), putting up novel content also increases one's chances of having their content featured on YouTube's front page (Cha, Kwak, Rodriguez, Ahn, & Moon, 2009). Next to this, it helps for people to link to their videos from external websites (Cha et al., 2009) and uploading new videos when interested people are online helps in getting an optimal number of video views (Ruhela et al., 2011). Importantly, such strategies only work in the early phase after uploading a video on YouTube, as a video that is unpopular on its first day is very likely to remain unpopular (Cha et al., 2009).

Other strategies for increasing one's video views on YouTube require a more methodical approach. Typically, people do not stumble upon a YouTube video but, instead, are referred to it. Out of all types of sources leading to a video (internal, external, search, mobile, featured, social, and viral), YouTube's internal mechanisms (e.g., lists of related videos) and its search functionality are the most important sources of traffic to videos (Figueiredo, Almeida, Gonçalves, & Benevenuto, 2014; Cheng, Dale, & Liu, 2008). Since most search engines rank items by usefulness (Chen, Shih, Chen, & Chen, 2011), a Search Engine Optimization (SEO) strategy such as only uploading useful videos on YouTube should be applied. Next to SEO, one should attempt to get their videos listed in the related videos list for other videos, however, this the most difficult popularity aspect to exert influence over because such a list is personalized for every visitor of YouTube (Zhou, Khemmarat, & Gao, 2010). To be precise, the related videos list is created through associating rule mining (i.e., finding relationships where there appear to be none) (Gupta, & Lehal, 2009) as well as by looking at the personal activity of the individual who is presented with the list of related videos (Davidson, Liebald, Liu, Nandy, & Van Vleet, 2010).

More methods exist that allow people to influence the popularity of their own videos on YouTube. However, popularity on YouTube has never been studied with only user-controlled variables so that is what this research contributes to the body of literature on popularity prediction. To investigate the accuracy of such a prediction model, the following research questions is proposed:

RQ 1 *How accurately can a supervised machine learning setup predict video view count from user-controlled variables?*

2.4 Content-agnostic factors affecting YouTube video popularity

Next to content-related factors, content-agnostic factors (e.g., the size of the video uploader's social network, the video's previous total view count, the number of keywords associated with the video, or the video's age) influence whether people will view videos on YouTube (Borghol et al., 2012). Unlike the former type, this type is difficult to study because people with large social networks, for instance, may not be popular because of their large social networks but because they happen to produce interesting video content. Of the content-agnostic factors, the video's total previous view count is the most important factor, followed by a combination of the video's age and its total previous view count. However, video age as a factor on its own is simply not highly related to video view count (Borghol et al., 2012).

When controlling for content-related factors, content-agnostic factors reveal that popularity of YouTube videos follow a rich-get-richer model that is scale-free and has a power-law exponential of about one (Borghol et al., 2012). The concept of rich-get-richer means that a video's view rate is proportional to its current view count. This is due to the first-mover-advantage which holds that the first video with certain content has typically already gotten so much attention that video clones or other videos covering the same content only strengthen the view rate of the original video (Borghol et al., 2012). The power-law property means that few people are able to accumulate many subscribers on YouTube and that many people accumulate few subscribers (Clauset, Shalizi, & Newman, 2009; Faloutsos, Faloutsos, & Faloutsos, 1999). The exponent of one is the decay rate which limits the power-law behavior for people with a very high number of subscribers. Finally, the concept of scale-free states that most people tend to be subscribed to only a few people that have many subscribers (Pastor-Satorras, & Vespignani, 2001). Also, people who just start out on YouTube and want to subscribe to other people tend to subscribe to people who already have many subscriptions (Barabási, 2009).

2.5 Predicting YouTube video popularity

In-depth analyses have been conducted to assess the core metrics related to popularity on YouTube (Chatzopoulou et al., 2010; Szabo, & Huberman, 2010). In Chatzopoulou et al.'s study, a database of more than 37 million YouTube videos was created, which consisted of many different properties, patterns and measurements. Four of those metrics (view count, comment count, favorite count, and rating count), all highly positively correlated with each other, were considered the popularity metrics on YouTube. Two linear regression models were constructed, a complete model and a simplified version of the complete model. The mathematical representation of the simplified model follows (Chatzopoulou et al., 2010),

$$viewcount = 3201.588F^2 - 0.014R^2 + 14.059F + 391.011R + 5220.755$$

where F is the number of times a video has been favorited and R is the number of ratings the video has received. Both estimation models performed equally well for predicting view count of YouTube videos (both R^2 = .768) but there was no mention of the approach taken (e.g., cross validation, development data set, etcetera).

View counts of videos on YouTube have also been predicted directly from previous view counts (Szabo, & Huberman, 2010). The S-H regression model, named after its creators, predicts a video's view count at a target date t_t based on a linear function of its view count at reference date t_r, where $t_r < t_t$. On a data set of 7,146 YouTube videos, this models produces a Relative Squared Error of about .6 for videos that have just been uploaded but approximates zero for videos that are one month old. This model was improved by a Multivariate Linear (ML) regression model which assigns different weights to different days (Pinto et al., 2013). The ML model scored 13% to 15% better than the S-H model with a Mean Relative Squared Error of 0.202 and 0.184 on two data sets of 16,123 and 5,813 YouTube videos. However, these models were improved even further by a model based on Radial Basis Functions (RBF). The RBF model, which adds measures of video similarity, performed 19% to 21% better than the S-H model and produced a Mean Relative Squared Error of 0.189 and 0.172 on the same two data sets (Pinto et al., 2013). So, the accuracy of predicting a video's future view count was improved by assigning different weights to different samples of videos as well as by including features for similarities between videos. An additional important finding is that a general model is better than models that are created for specific video categories; this is because videos in virtually all categories follow the same view count patterns (Pinto et al., 2013).

All these studies included metrics that were available at the time that they were conducted; however, as more people use Social Networking Sites (SNS), more challenges arise for improving such an SNS (O'Reilly, 2007; Mislove, Marcon, Gummadi, Druschel, & Bhattacharjee, 2007). Many SNSs followed up on this and have been modified over the years (e.g., Evans, 2013). Such modifications include visual layout changes as is the case with Twitter's update to the presentation of user's profiles (Twitter, 2014) or a change in the inner workings of the SNS itself such as with YouTube replacing their video rating system for a system where people can either like or dislike videos (Rajaraman, 2009). This could imply a decrease in the reliability of existing popularity prediction models since the variables leading to these findings may have changed or may have been removed. Whether previous literature is still accurate can be investigated by including metrics from earlier regression models in new regression models in this research. Therefore, the following research question is proposed:

RQ 2 To what extent have modifications to YouTube had an impact on predictive models from earlier studies?

To answer both research questions, new linear regression models can be created where video view count is predicted from several features; however, there will always be differences between the real view count and the predicted view count. Loss functions (more on this later) try to minimize these differences (loss) (Bottou, 2010). The performance of these functions can be measured in two ways, either with empirical risk (i.e., performance on training data) or with expected risk (i.e., estimated performance on unseen data) (Bottou, 2010).

Gradient Descent (GD) minimizes the empirical risk by measuring in steps (the gradient) and by adjusting the weights of features while moving towards the lowest loss value (Zhang, 2004),

$$w_{t+1} = w_t - \gamma_t \frac{1}{n} \sum_{i=1}^{n} \nabla_w Q(z_i, w_t)$$

where w_t is the current weight, γ_t is the learning rate, z_i are all the items from the sample, and $Q(z_i, w_t)$ is the loss as calculated by the function $\ell(f_w(x), y)$ (Bottou, 2010).

Stochastic Gradient Descent (SGD) makes this process less complex, and lighter in terms of computation, by estimating the gradient instead of computing it like GD does. In doing this, it randomly takes one item without taking into account whether that item was used in previous estimations (Bottou, 2010),

$$w_{t+1} = w_t - \gamma_t \nabla_w Q(z_i, w_t)$$

where z_i is a randomly chosen item from the sample with its corresponding weight w_t and where the other variables are as the same as earlier described for GD.

This is how SGD improves the performance of the regression model when handling large data sets. The performance can be improved even further by optimizing the number of regression passes (epochs). According to the official Scikit-Learn documentation (2010), one can achieve this by setting the number of passes to $\left\lceil\frac{10^6}{n}\right\rceil$ where n is the size of the training data set.

All this comes with a few caveats, however. The SGD algorithm requires lots of experimenting with the regularization parameter, alpha, for optimal prediction quality (Scikit-Learn Developers, 2010), and can be affected by the scaling of features. This means that a feature with a large value range (e.g., 0-1,000,000) will dominate a feature with a small value range (0-10). Thus, properly scaling features before fitting the regression model is necessary.

SGD regression uses a penalty to regularize complex models which helps to prevent overfitting of the regression model (i.e., sticking close to the training data as opposed to generalizing to the test data) (Breheny, 2011). The alpha value is the regularization parameter of SGD-based linear regression. It functions as a multiplier that either strengthens or weakens the penalty, so, different alpha values yield different evaluation scores of the regression model. In finding the optimal alpha value, evenly spaced values on the $\log_{10}$ scale in the 10^1-10^{-5} range are typically considered as per Scikit-Learn's official documentation (Scikit-Learn Developers, 2010). This represents the following values: 10, 1, 0.1, 0.01, 0.001, 0.0001, and 0.00001. An alpha value close to zero means a weak regularization of the model whereas a high value indicates a strong regularization.

Next to optimizing alpha, a loss function and penalty have to be decided on. *Ordinary Least Squares* (*OLS*) tries to linearly minimize the squared differences between the predictions and the true values (*RSS* or *Residual Sum of Squares*) (Schmidt, 2005; Owen, 2007),

$$OLS = \frac{1}{2}\sum_{i=1}^{n}(\bar{y} - y_i)^2$$

where $\bar{y}$ is a predicted value and y_i is an observed true value.

Huber uses the *Sum or Squared Error* (*SSE*) but changes to linear loss past a set threshold (e.g., a value of one),

$$Huber = \begin{cases} z^2 & |z| \leq 1 \\ 2|z| - 1 & |z| \geq 1 \end{cases}$$

where *z* is the loss (Owen, 2007). Thus, it focuses less on correctly predicting outliers.

Epsilon insensitive, ignores error lower than a set threshold (e.g., a value of one). Above that threshold, it is linear. Epsilon insensitive loss uses L1 (or Lasso) loss which reduces as many weights (*w*) as possible to values close to zero while still being able to approximate the true values (Owen, 2007). In doing this, it tries to construct a sparse model, which means a simpler, more interpretable model. Formally, the Lasso is known as (Scikit-Learn Developers, 2010),

$$L1 = \sum_{i=1}^{n} |w_i|$$

where w_i is the feature's coefficient (weight).

Squared epsilon insensitive is similar to epsilon insensitive. However, it uses squared loss instead of linear loss above a set threshold (e.g., a value of one). Squared epsilon insensitive uses L2 (or Ridge) loss which increases the weighting for coefficients as parameter values get lower. Ridge is defined as (Scikit-Learn Developers, 2010),

$$L2 = \frac{1}{2} \sum_{i=1}^{n} w_i^2$$

where w_i is the feature's coefficient (weight).

Finally, *ElasticNet* combines L1 and L2 in that it creates a model with a minimal number of coefficients like the Lasso. Next to this, it also reduces the strength of the coefficients for correlated features (Scikit-Learn Developers, 2010),

$$ElasticNet = \frac{\rho}{2} \sum_{i=1}^{n} w_i^2 + (1 - \rho) \sum_{i=1}^{n} |w_i|$$

where w_i is a feature's coefficient (weight) and ρ is a mixing parameter that is typically set to 0.15. Next to the options of L1, L2, or ElasticNet, it is possible to have no penalizing at all.

3 Method

3.1 Collecting the data

The programming language Python and the official libraries for the YouTube Data API (hereafter referred to as API) were used. This API let one search for, add, modify, or delete any videos for which one had the appropriate access permissions (Google Developers, 2014).

Two different programs were developed to collect the data, and both programs are available at http://msc.jordysnijders.com. All the data were stored in SQLite database files. The first program used the search list, and videos list methods from the API to retrieve video data for up to 500 videos for each hour of every day in between February 14, 2005 and May 4, 2015. The number 500 is the maximum number of results for any query's response, without a possibility to retrieve more. This resulted in data on 1.5 million unique YouTube videos. The second program used the search list and videos list methods from the API to collect repeated measures data. It only retrieved view count, comment count, dislike count, and like count statistics. Upon starting this program, these four statistics were gathered for up to 500 videos that were 1 hour old, then for up to 500 videos that were 2 hours old, then for up to 500 videos that were 3 hours old, etcetera. It did this for every hour of every day for videos that were no older than 31 days.

3.2 Retrieved metrics

Even though many more metrics could be retrieved, the following metrics for videos were gathered by the first program as these were the metrics to be included in the descriptive analysis and/or regression models. Many of these metrics were under the direct influence of the video owner. Such metrics are interesting because they allowed for a predictive regression model to be constructed. The descriptions of the following metrics are as per the YouTube Data API documentation (Google, 2015):

- *View count*: the number of times the video had been viewed.
- *Comment count*: the number of comments the video had received.
- *Dislike count*: the number of users who had indicated that they disliked the video.
- *Like count*: the number of users who had indicated that they liked the video.
- *Definition*: whether the video was available in high definition or in standard definition.
- *Caption availability*: whether the video had captions or not.

- *Duration*: the length of the video in seconds.
- *Licensed content*: whether the video contained licensed content.
- *Dimension*: whether the video was available in 2D or in 3D.
- *Description*: the video's description.
- *Publishing date*: the date and time that the video was uploaded.
- *Title*: the video's title.
- *Category*: the video category associated with the video.
- *Embeddability*: whether the video could be embedded on a third-party website.
- *Availability of statistics*: whether the video's statistics could be viewed by anyone.
- *Latitude*: the latitude coordinate of the video's origin in degrees.
- *Longitude*: the longitude coordinate of the video's origin in degrees.

3.3 Post-processing the data

The numerical values of video view count, video like count, video dislike count, and video comment count showed extreme values, ranging from zero to several billions. A data transformation had to be applied in those cases before inserting them as features in any of the regression models. This was necessary because features with a high numerical range would have overwhelmed features with a lower numerical range in a regression model.

The square root, $\log_2$, natural log, $\log_{10}$ and inverse transformation are the most common data transformations for these purposes but the $\log_{10}$ transformation is the appropriate transformation for this situation (Osborne, 2005). This is because of how well $\log_{10}$ transformations deal with extreme values. To clarify this, having a value range such as [0, 1] or [-1, 1] in a regression models makes all features in the model equally important. The $\log_{10}$ of 1,000 is 3, for instance, whereas the $\log_2$ of 1,000 is 9.97. So, because $\log_{10}$ values tend to produce a narrow range which is closer to zero and one, this logarithm was the preferred choice. However, no $\log_{10}$ transformation was applied to values of zero.

3.4 New metrics from existing ones

New metrics were created from metrics retrieved via the YouTube Data API such as by counting, adding or subtracting from existing metrics. Next to this, it was also be possible to extract textual features. These were derived from video titles, channel names, and video descriptions. As any video owner could manipulate these texts, the following new metrics were under the control of the video owner as well.

3.4.1 Count metrics

The following count metrics (i.e., mappings from words or characters to numerical values) were derived from video titles, channel names, and video descriptions.

- *Part of Speech (POS) tags*: part of speech tagging is a method of processing natural language where two types can be distinguished: rule-based and stochastic (based on statistics) (Brill, 1992). Such a tagger identifies words based on the context in which they are used, so, by counting these occurrences the predictability of these words can be calculated (Schmid, 1994). It was expected for part of speech tag counts from video titles, channel names, and descriptions of 1.5 million YouTube videos to have an impact upon insertion in one of the regression models.
- *Word ngrams*: ngrams are sequences of items (characters or words) that allow for machines to learn similarities in topics from such sequences (Damashek, 1995). These sequences (i.e., mapping the sequences to numerical values) are then counted or otherwise weighted. As with POS tagging, an ngram system improves as the number of sequences it can compare grows (Lesher, Moulton, & Higginbotham, 1999). With 1.5 million video titles, channel names, and video descriptions in this data set on YouTube videos; it was expected for word ngrams to yield good results in a regression model. *Term Frequency* (TF) ngrams and *Term Frequency-Inverse Document Frequency* (TF-IDF) ngrams can be distinguished. TF simply counts terms whereas TF-IDF counts terms in a text but adjusts their weights by the terms' counts in all texts. This means that the weights of frequent and unimportant terms are decreased while the weights of rare and insightful terms are increased (Ramos, 2003). Also, TF-IDF automatically filters out many of a language's stop words (i.e., words that occur in many sentences but are not indicative of anything). No list of stop words for all the languages in the world is available, so, since many texts in many different languages are written on YouTube, it was expected for TF-IDF to perform better than TF.
- *Character ngrams*: besides word ngrams, character ngrams exist. This type of ngram spots chunks of words such as letters and syllables. Character ngrams can also extend to

finding words. As with word ngrams, it was expected for character ngrams to have an influence in regressing on video view count. For this reason, character ngrams were derived from video titles and channel names.

3.4.2 Dichotomous metrics

The following dichotomous metrics (i.e., discrete/binary values such as a value of ‘0’ representing false or a value of ‘1’ representing true) were derived from video descriptions.

- *Websites presence*: whether the video description contained one or more links to external websites. Any *Uniform Resource Locator* (URL) qualifies for this. The thought behind this metric was that it could have been meaningful in one of the regression models since URLs in video descriptions on YouTube are colored blue which makes them stand out next to the standard black text. This might have drawn people’s attention and triggered them to view a video.
- *Social media presence*: whether the video description contained one or more links to profiles and/or media on Social Networking Sites (SNS) such as Facebook, Twitter, and Instagram. This includes shortened URLs for these SNSs (e.g., http://fb.me, http://t.co, and http://instagr.am). The reason for this metric was that people who link to their other SNS profiles from their videos’ description might have been popular on those SNSs. Such high connectedness on SNSs could have indicated that these people were also popular on YouTube in terms of video views.

3.4.3 Numerical metrics

The following numerical metrics (i.e., continuous values such as 0.25 or 39) were derived from video titles, channel names, and video descriptions.

- *Video age*: the publishing date’s timestamp (the number of seconds as of January 1, 1970; also known as the *Unix epoch*) was subtracted by the timestamp at which the video’s statistics were retrieved. The resulting value was the video’s age in seconds.
- *Character counts*: the length of the video title as counted by the number of characters in it, whitespaces excluded.
- *Word counts*: the length of the video title as counted by the number of words in it.

- *Uppercase/lowercase ratio*: a value in between zero and one representing the number of capital letters in the video title relative to the number of lowercase letters.
- *Alphanumerical ratio*: a value in between zero and one representing the number of alphabetical and numerical characters in the video title relative to the number of other types of characters (e.g., symbols).
- *Alphabetical ratio*: a value in between zero and one representing the number of alphabetical characters to other character types in the video title
- *Numerical ratio*: a value in between zero and one representing the number of digits to other character types in the video title.

3.5 Linear regression with Stochastic Gradient Descent

A machine learning algorithm such as multiple linear regression can be used to predict how well future video uploads on YouTube will perform based on features such a regression model has seen from samples in a training data set (i.e., the data from which the regression model learns). So, in this research, linear regression was used to predict video view counts from features as determined by the results of several regression experiments.

The downside of using regression with large data sets, where N > 100,000 (Scikit-Learn Developers, 2010), is the increased complexity for computationally processing this high amount of data. As explained in section 2.6, Stochastic Gradient Descent (SGD) is an optimization algorithm that improves performance in that case. Therefore, in this research, SGD based linear regression was used to predict video view count with the features that were listed. In doing this, the best parameters for the alpha, the loss function, and the penalty had to be found. This meant that every combination of these three parameters were tested in the regression models. However, before presenting the results of these regression models, the data are discussed.

4 Data

In this chapter, an exploratory analysis is presented. This is done to give the reader an impression of the most important statistics for YouTube videos. An overview of the descriptive statistics is shown for this purpose. Importantly, by inspecting the data early on, one can investigate what impact certain metrics might have in the regression models as well as which metrics are not that useful (i.e., metrics where the values are mostly similar) and should not be included in such regression models. This is due to the law of parsimony which states that simpler models should be favored over more complex ones in the case that they both predict similarly (Domingos, 1999).

The last section addresses the changes of key metrics for video popularity (number of views, comments, likes, and dislikes) for YouTube videos over a period of time. This was a suggestion by Chatzopoulou, Sheng, and Faloutsos (2010).

4.1 Descriptive statistics

As discussed in the section 3.1, two data set were created. The primary data set contained data on 1,500,094 YouTube videos, and is used for the descriptive statistics and regression models that follow. The secondary data set contains repeated measures on 3,014,962 videos and is only used for the repeated measurements section of this chapter. Both data sets are available at http://msc.jordysnijders.com. None of the videos that were retrieved were live broadcasts. This means that only static content that was uploaded to YouTube was analyzed. Also, every video in the dataset had its privacy setting set to ‘public’. This is important, not just because public videos are the focus of this research but also because ‘private’ content artificially restricts the reach of such videos in terms of views, comments, likes, and dislikes.

4.1.1 Categories

YouTube features 32 video categories. These are *Film & Animation, Autos & Vehicles, Music, Pets & Animals, Sports, Short Movies, Travel & Events, Gaming, Videoblogging, People & Blogs, Comedy, Entertainment, News & Politics, Howto & Style, Education, Science & Technology, Nonprofits & Activism, Movies, Anime/Animation, Action/Adventure, Classics, Comedy, Documentary, Drama, Family, Foreign, Horror, Sci-Fi/Fantasy, Thriller, Shorts, Shows,* and *Trailers*.

Some of the categories are more popular than others (see Figure 1). The categories with the most videos are Music (14.0%), Education (13.6%), Entertainment (12.1%), and People & Blogs (11.5%). The least popular video categories are Shows (0.6%), Movies (0.1%), and Trailers (0.02%). However, some of the categories are not represented in the data set. These are Short Movies, Videoblogging, Anime/Animation, Action/Adventure, Classics, Comedy, Documentary, Drama, Family, Foreign, Horror, Sci-Fi/Fantasy, Thriller, and Shorts. The majority of these absent categories have only existed as of 2011 and are for full-length movies that one can rent, purchase or stream (YouTube Official Blog, 2011).

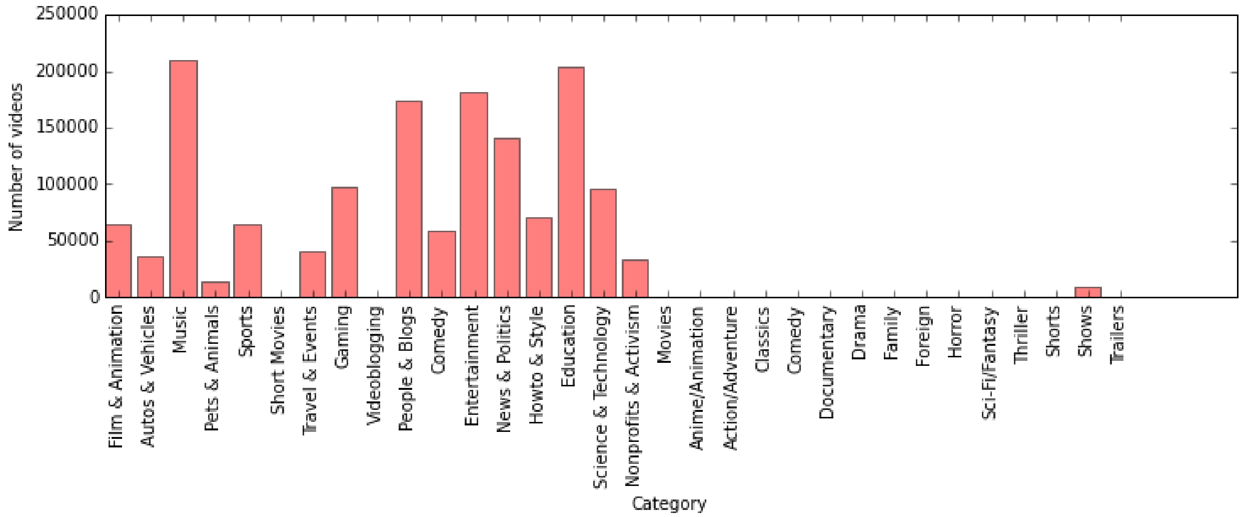

Figure 1. Distribution of videos over all categories on YouTube

4.1.2 Geographical location

Figure 2 shows a world map with the locations of the top 1,000 videos in terms of view count. Not all videos have geographical coordinates (i.e., longitude and latitude) associated with them so the points do not represent the 1,000 most viewed videos but rather the most viewed 1,000 videos with geographical coordinates on where the videos were shot. On this plot, North America and Europe are responsible for the majority of the most popular videos but every continent is represented with at least a few videos. Regardless, it shows that video popularity is not bound to certain locations. Therefore, other factors must play a role in determining video view count.

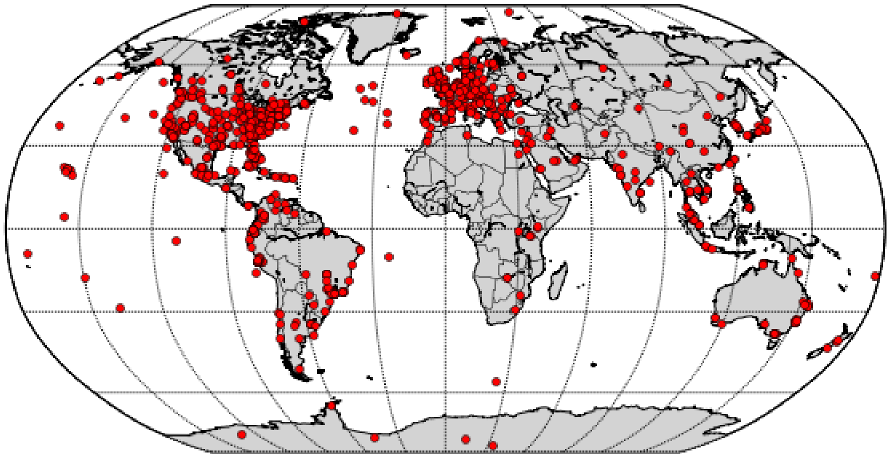

Figure 2. Origins of the 1,000 most popular videos with a location on YouTube

4.1.3 Definition

Videos can either be uploaded on YouTube in Standard Definition (SD) or High Definition (HD). A video resolution of 144p, 240p, 360p, or 480p means it is SD whereas videos with a resolution of 720p, 1080p, 1440p, 2160p, or 4320p are considered HD quality. Figure 3 shows that most videos in the data set are of SD quality (54.5%).

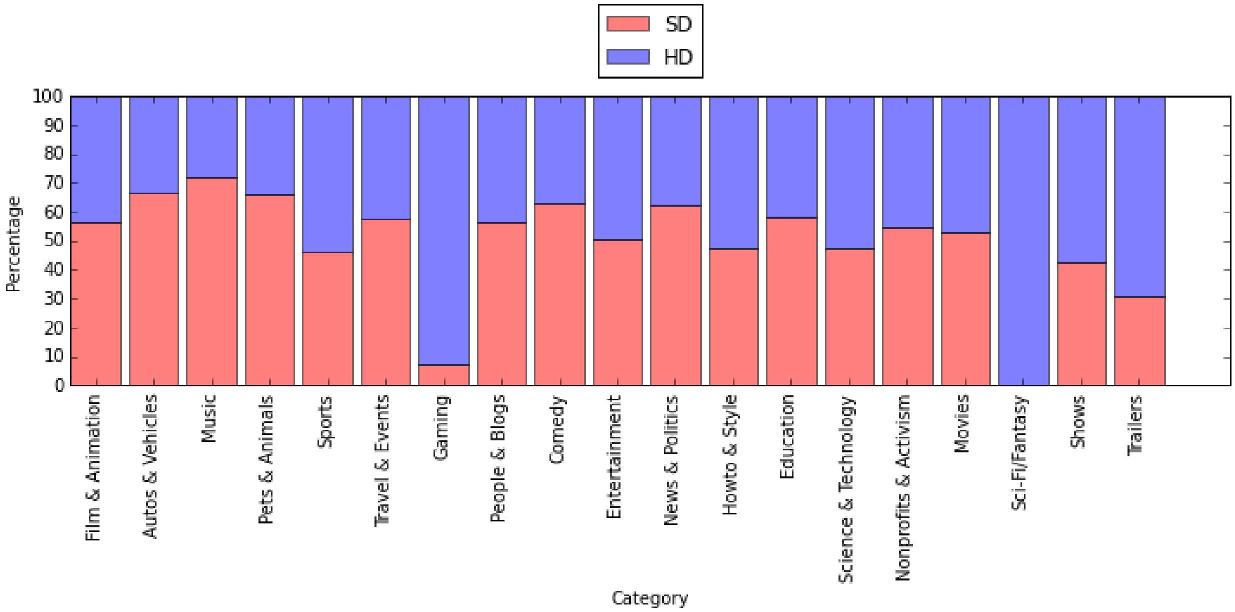

Figure 3. Video definition per category on YouTube

4.1.4 Dimension

There are 2D and 3D videos on YouTube but most of them are 2D (99.5%). Two categories in particular show signs of being more popular for 3D content (see Figure 4). These are Movies (3.8%) and Film & Animation (1.1%). Less than one per cent of the videos in other categories are 3D videos.

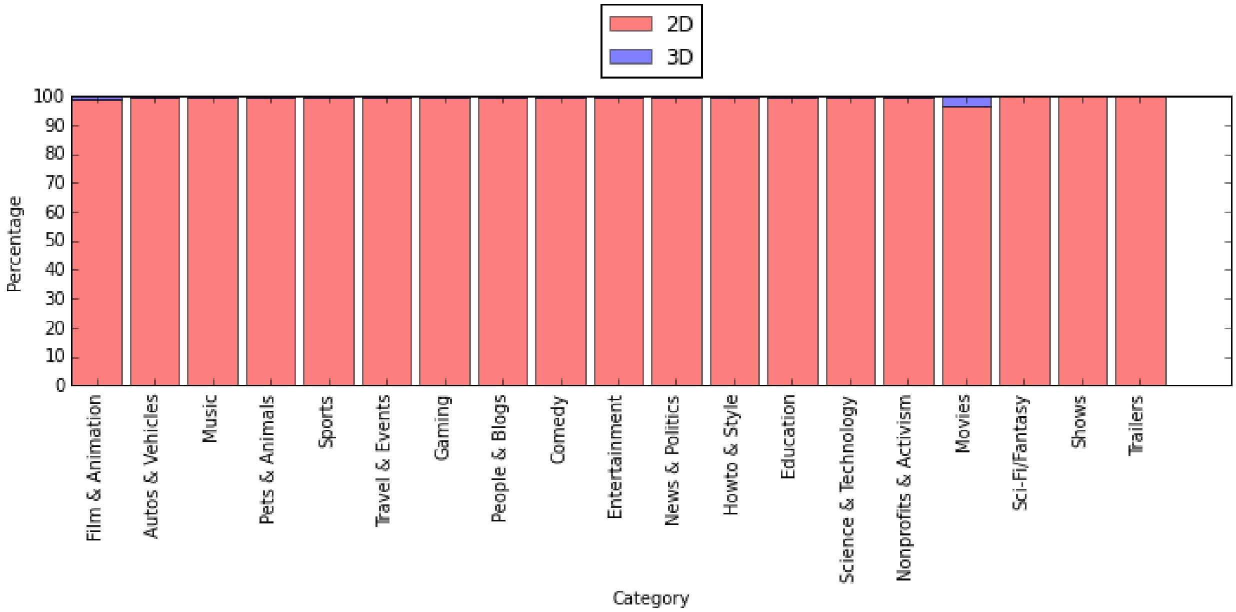

Figure 4. Video dimension per category on YouTube

4.1.5 (Closed) captions and subtitles

Both captions and subtitles can be added to YouTube videos. Subtitles are transcriptions of dialogue. Captions are descriptions of all hearable content. Figure 5 shows that neither are very popular on YouTube with most video uploader opting not to use them (3.8%).

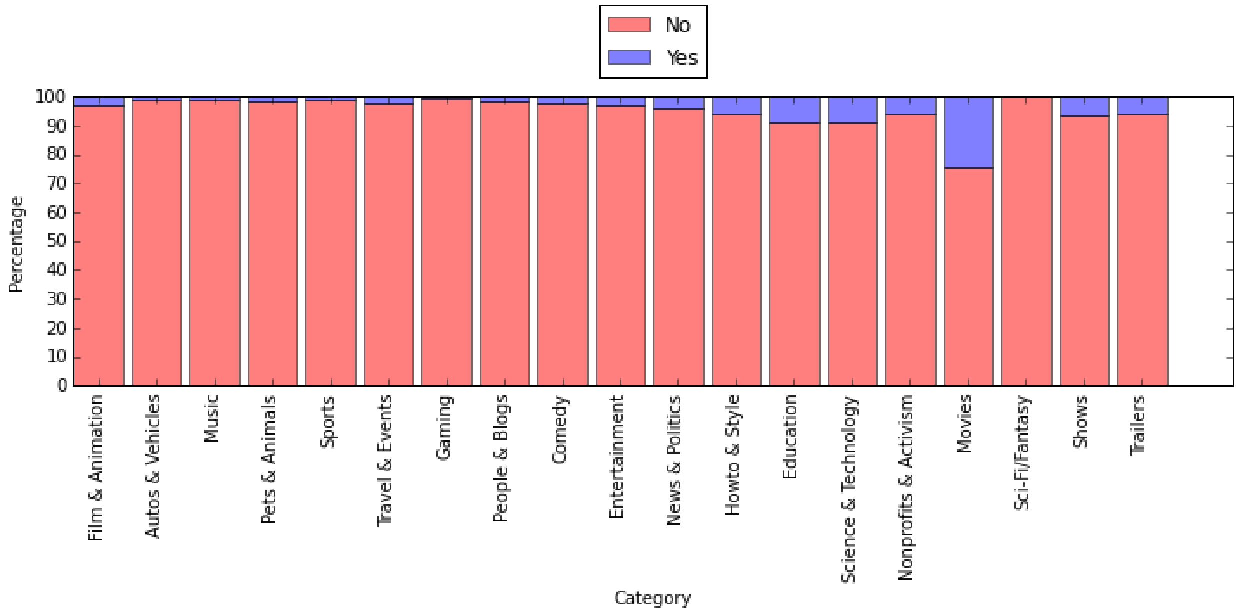

Figure 5. Availability of subtitles per category on YouTube

4.1.6 Duration

Short, medium, and long videos exist on YouTube. Videos no longer than 4 minutes in length are considered short, videos in between 4 and 20 minutes are medium in length, and those longer than 20 minutes are long. Most videos on YouTube are medium long (44.3%), then short (40.9%), and long videos are underrepresented (14.8%).

Six video categories have the longest videos on average in seconds (see Figure 6). These are Movies (*M* = 5,700.12, *SD* = 2,197.99), Education (*M* = 1,246.61, *SD* = 1,921.39), Shows (*M* = 1,146.71, *SD* = 1,304.21), Gaming (*M* = 1,117.98, *SD* = 2,151.82), News & Politics (*M* = 1,114.34, *SD* = 95,989.74), and Film & Animation (*M* = 1,027.92, *SD* = 1,979.59).

The longest video in the data set is titled *If you do business in Russia (Full version)* in the News & Politics category and is reported to be 36,057,368 seconds long (417.33 days). However, upon viewing the video on YouTube, the duration initially shown is 417 days but then switches to 1 minute and 43 seconds, so this is a bug on YouTube.

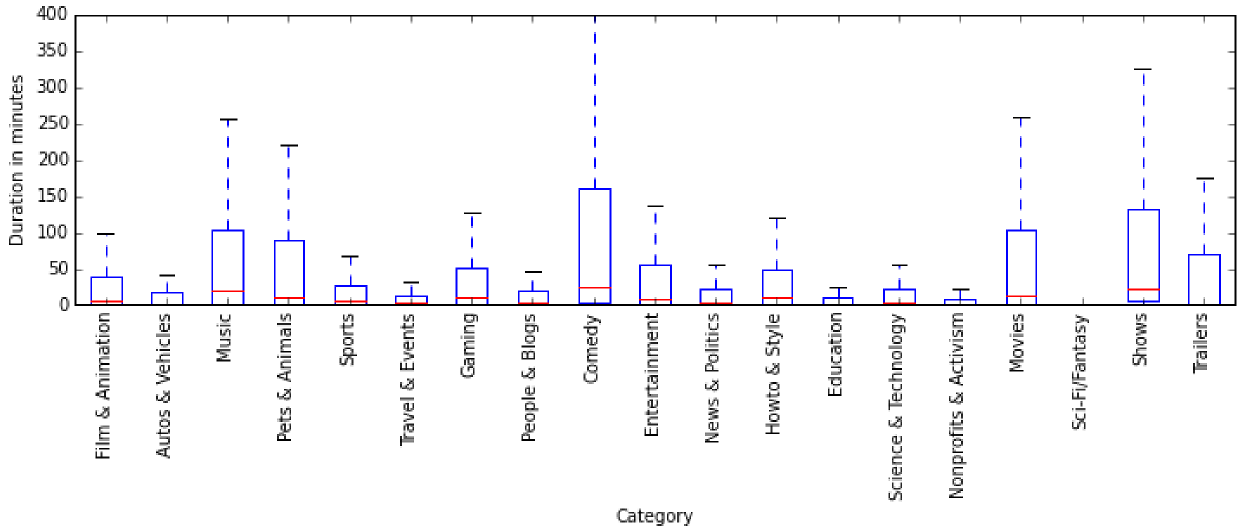

Figure 6. Video duration per category on YouTube

4.1.7 Licensed content

Figure 7 shows that most videos on YouTube do not contain licensed material (59.3%). Also, people who upload videos to YouTube can specify under which license type they would like their videos to be published. A choice has to be made between YouTube and Creative Commons. Figure 8 shows that the majority of videos is published under the YouTube license (98.7%).

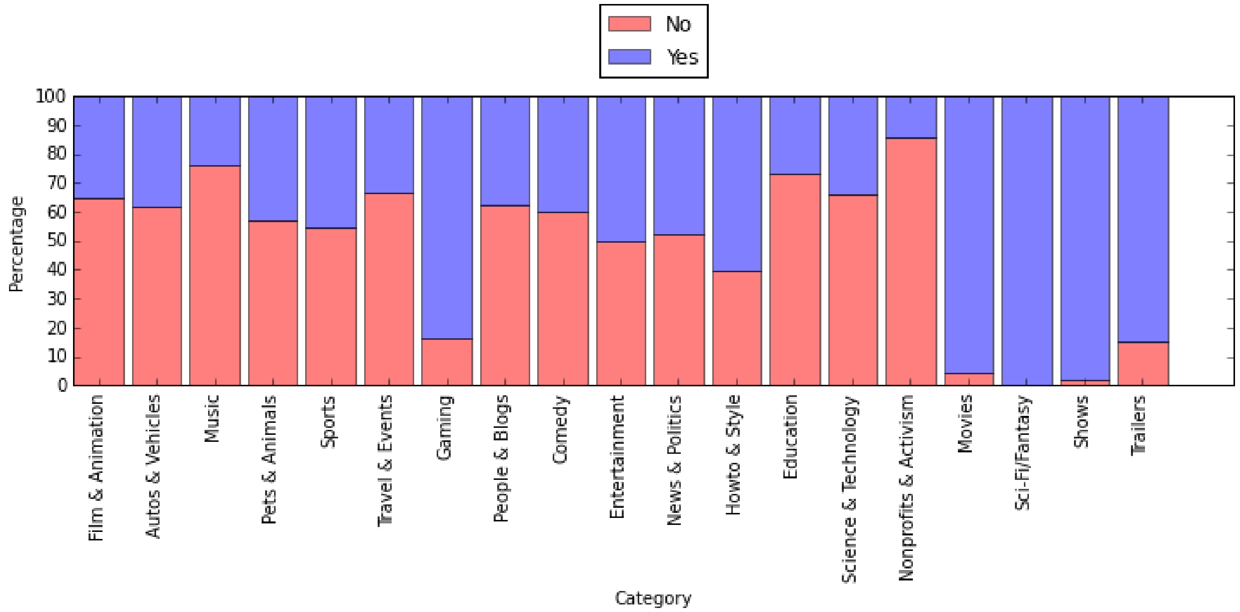

Figure 7. Relative amount of licensed video material per category on YouTube

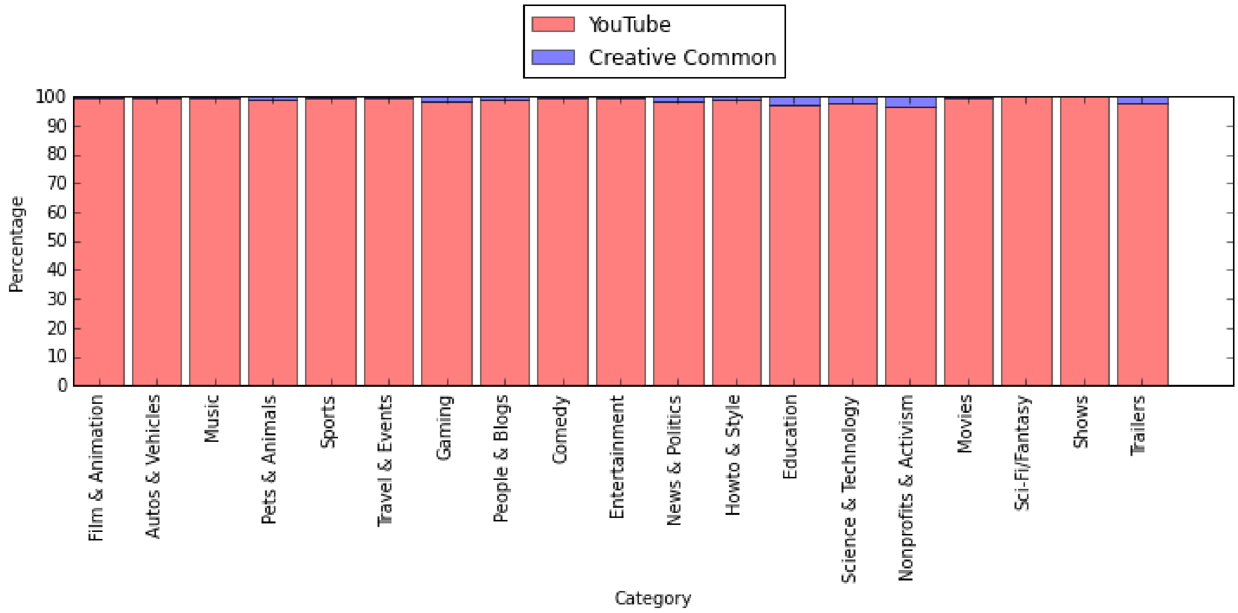

Figure 8. Video publishing license per category on YouTube

4.1.8 Availability of statistics

Video uploaders can specify whether the statistics next to their videos are visible to the public. Such statistics include graphs of number of views, number of comments, and number of favorites over time, as well as inbound links (i.e., websites referring to the video) and associated views. Most people (89.2%) allow for their videos' statistics to be seen (see Figure 9).

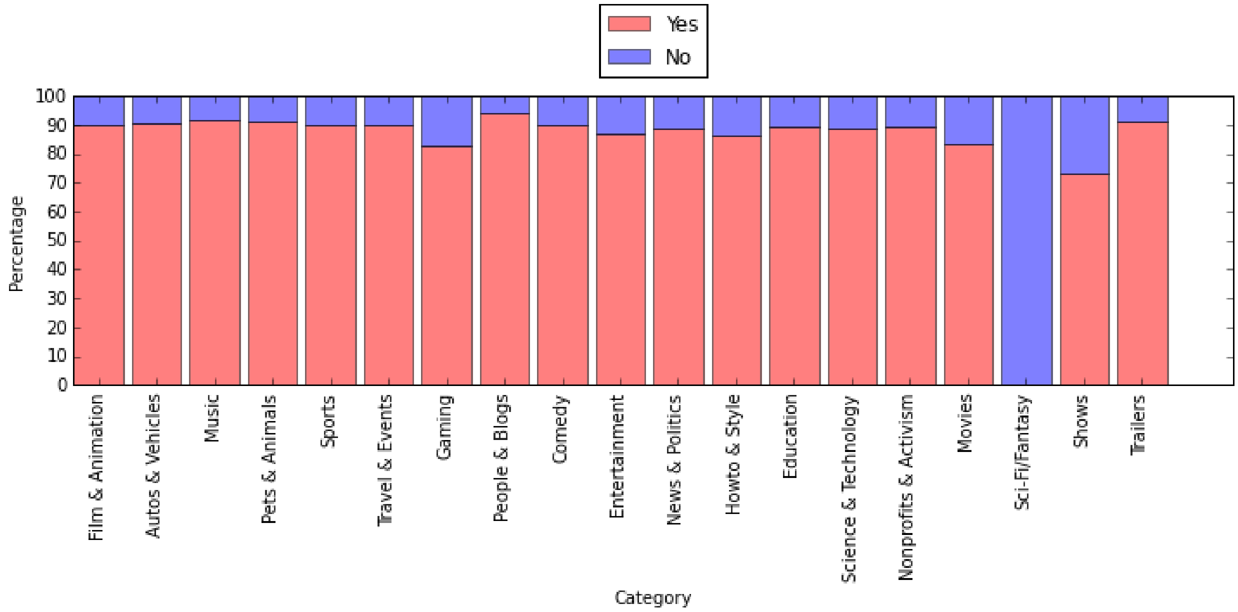

Figure 9. Availability of publically viewable statistics per category on YouTube

4.1.9 Embeddability

Those who publish videos on YouTube can also enable or disable the ability for third parties to include (embed) a video on websites other than YouTube. The majority of people (99.0%) allow for this (see Figure 10).

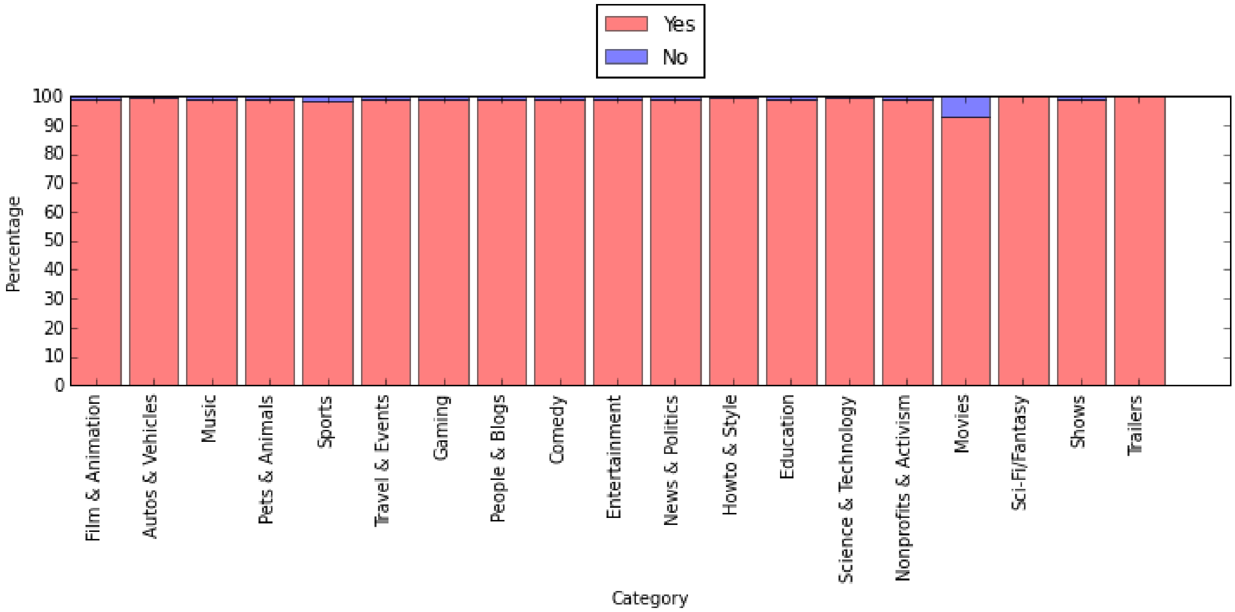

Figure 10. Embeddability of videos per category on YouTube

4.1.10 Video age

The average video age in the data set is 1,207.12 days (*SD* = 938.17). Almost all categories have an equal distribution of video age with the exception of the Gaming category (*M* = 245.65, *SD* = 459.97). This is because videos related to gaming are more popular nowadays than they were years ago. The opposite can be said for videos in the Music (*M* = 1,831.57, *SD* = 1,138.19) and Pets & Animals (*M* = 1,715.48, *SD* = 1,149.02) categories. Also, almost all categories have videos as far back as 9.68 years ago (see Figure 11).

As the data set is ordered by video age, having a normally distributed population is of importance for the regression model. If one would sample the beginning part of the data set for the training set, the middle part for the validation set, and the last part for the test set, then all data sets would clearly be different and the change of metrics over time would be lost in the machine learning process. So, shuffling the data set before splitting into training, validation and test sets was required for the regression models.

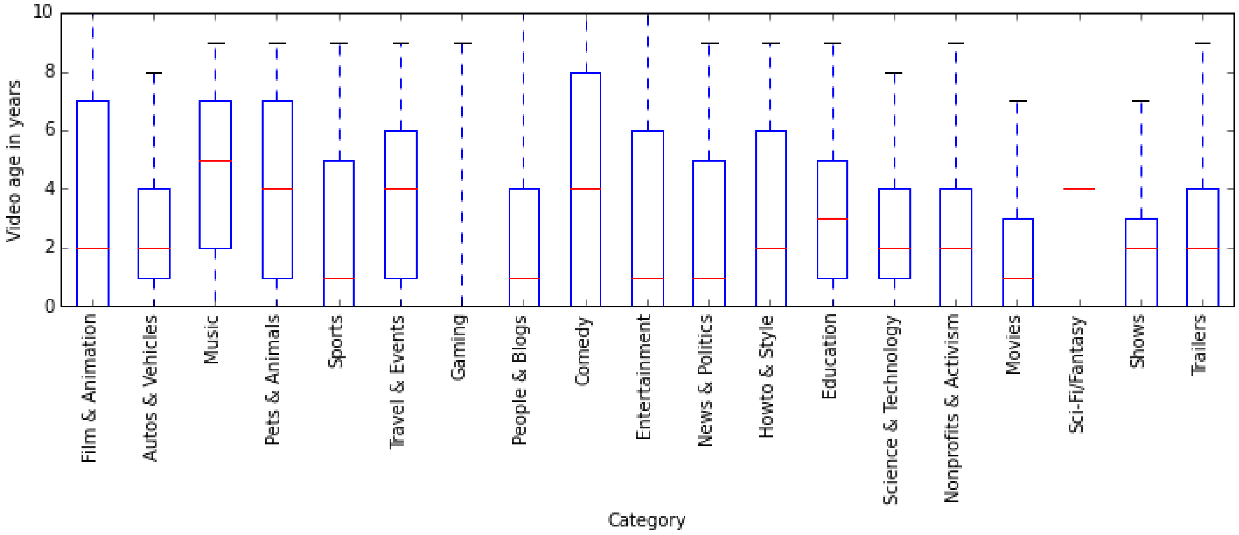

Figure 11. Video age per category on YouTube

4.1.11 Views

The videos in the data set get a few as zero views to as many as two billion view (*M* = 709,348.79, *SD* = 3,787,854.91). The most viewed video is *PSY - GANGNAM STYLE (강남스타일) M/V* in the Music category with 2,319,515,787 views, 5,400,599 comments, 9,368,896 likes, and 1,256,307 dislikes. The second most viewed video is *Justin Bieber - Baby ft. Ludacris*, also in the Music category with 1,161,556,092 views, 6,774,237 comments, 2,906,339 likes, and 4,600,610 dislikes. The third most viewed video is *Katy Perry - Dark Horse (Official) ft. Juicy J*, another video in the Music category with 931,129,321 views, 388,137 comments, 2,841,777 likes, and 458,998 dislikes.

Most views (see Figure 12) by far belong to videos in the Music (*M* = 2,647,923.71, *SD* = 16,893,287.03), Pets & Animals (*M* = 1,224,914.98, *SD* = 4,766,564.99), Comedy (*M* = 1,261,300.93, *SD* = 6,073,158.15), Shows (*M* = 2,146,790.09, *SD* = 11,633,804.01), and Trailers (*M* = 1,269,501.49, *SD* = 3,564,546.87) categories. The category with the least number of views is Sci-Fi/Fantasy (*M* = 1,728.00, *SD* = 458.00).

When the videos in the data set are sorted by view count in descending order (see Appendix A), one can clearly see that videos from the Music category dominate the top 50. Only five videos in that list do not belong to the Music category. This reinforces the finding that music videos get most views.

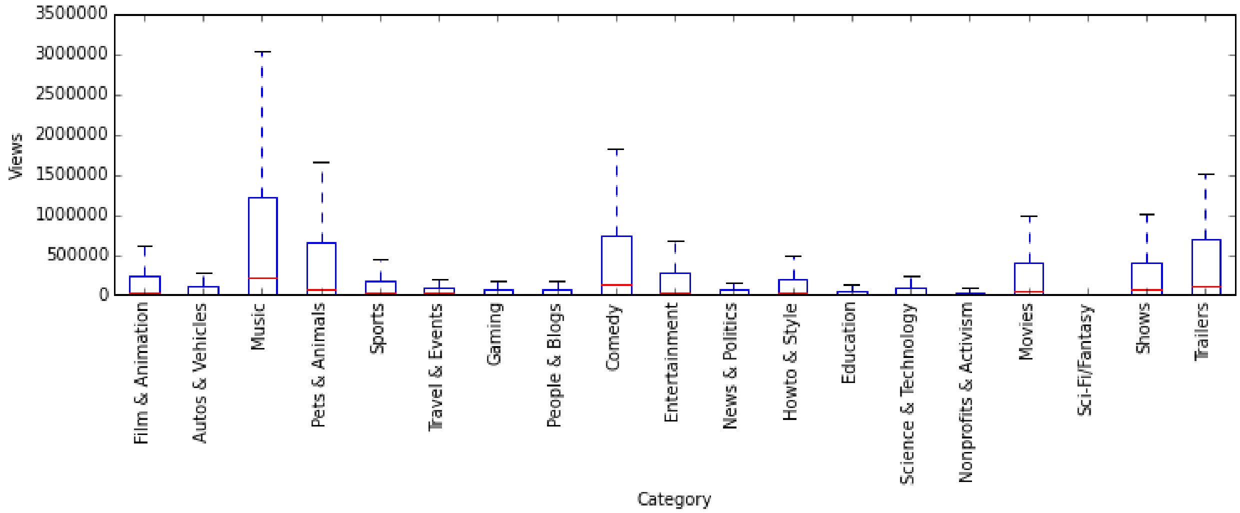

Figure 12. Video view count per category on YouTube

4.1.12 Comments

On average, videos in the data set get 683.93 comments (*SD* = 5,152.55). The two videos that received the most views (as presented before) also received the most comments. The video with the third most comments is *Super Junior 슈퍼주니어 Mr.Simple MUSICVIDEO*, a video from the Music category with 1,291,836 comments, 90,887,184 views, 458,334 likes, and 24,546 dislikes.

The categories that typically have the most comments (see Figure 13) for their videos are Music (*M* = 1,716.00, *SD* = 23,119.61), Comedy (*M* = 1,646.17, *SD* = 8,448.74), Shows (*M* = 2207.54, *SD* = 11995.56), and Trailers (*M* = 1,015.50, *SD* = 4,325.91). The least comments can be found for videos in the Sci-Fi/Fantasy (*M* = 21.00, *SD* = 9.00), Education (*M* = 184.40, *SD* = 1,592.87), and Travel & Events (*M* = 169.19, *SD* = 1,274.53) categories.

As was found for views, when the videos in the data set are ordered by comment count (see Appendix B), the majority of videos are music videos. Only 17 out of the 50 videos with most comments do not belong to the Music category.

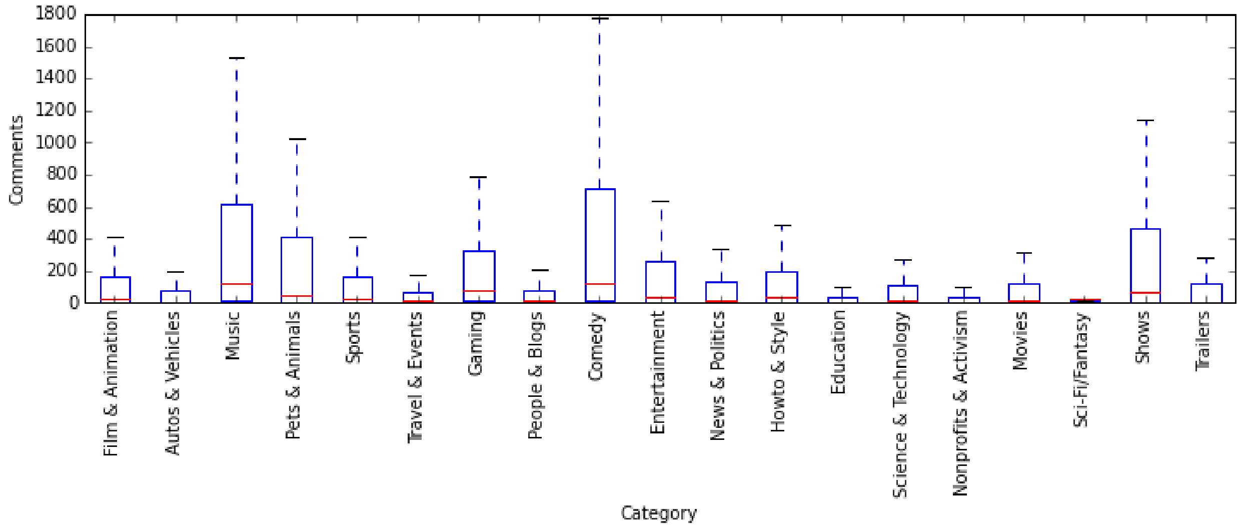

Figure 13. Video comment count per category on YouTube

4.1.13 Likes

The mean number of likes that videos get on YouTube is 2,785.20 (*SD* = 14,454.30) (see Figure 14). As with number of views, *PSY - GANGNAM STYLE (강남스타일) M/V* is also the video to have to highest number of likes (9,368,896). This video is followed by *Ylvis - The Fox (What Does The Fox Say?) [Official music video HD]* from the Entertainment category with 3,668,785 likes, 422,446 dislikes, 510,573,787 views, and 873,591 comments. The third most liked video is *Taylor Swift - Blank Space*, a music video with 3,498,860 likes, 240,502 dislikes, 810,375,030 views, and 352,753 comments.

The dominance of the Music category was noticeable in the top 50 YouTube videos by view count and comment count. The top 50 as sorted by like count (see Appendix C) only strengthens this trend with only one entertainment video, and 49 videos belonging to the Music category.

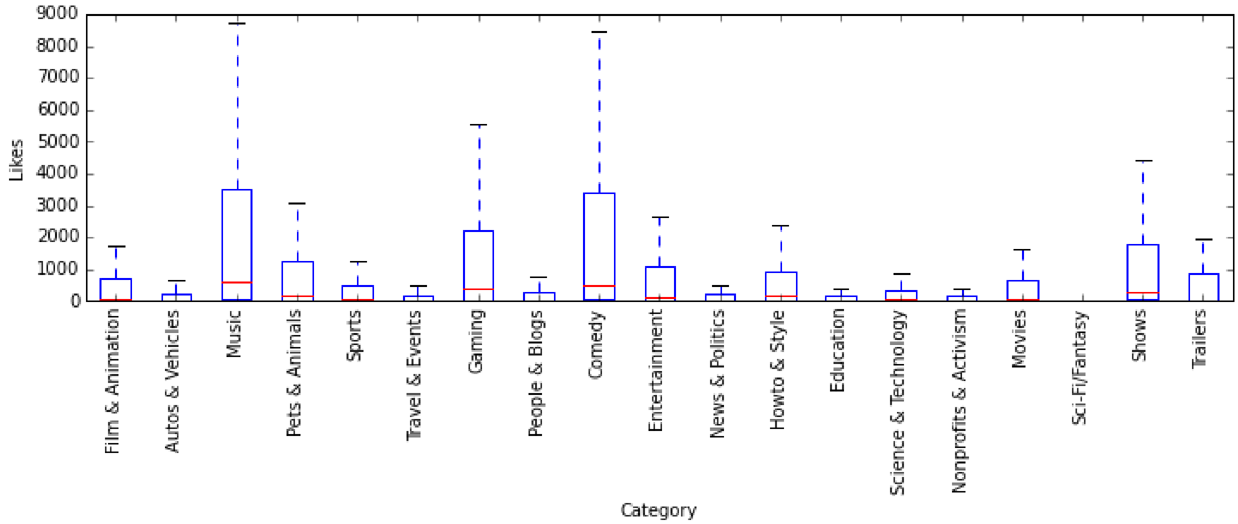

Figure 14. Video like count per category on YouTube

4.1.14 Dislikes

Out of view count, comment count, like count, and dislike count, the number of dislikes is the lowest of all (*M* = 180.72, *SD* = 1,890.15) (see Figure 15). The second most viewed video (as described earlier), *Justin Bieber - Baby ft. Ludacris*, is also the video with the highest number of dislikes (4,600,610). In second place is *Friday - Rebecca Black - Official Music Video*, a music video with 1,571,316 dislikes, 438,485 likes, 78,566,908 views, and 712,093 comments. Third place belongs to *PSY - GANGNAM STYLE (강남스타일) M/V* (also mentioned earlier) with 1,256,307 dislikes.

The top 50 YouTube videos, when sorted by dislike count (see Appendix D) continues the trend of videos from the Music category outnumbering videos from other categories. In the list of 50 videos to receive most dislikes, 38 videos are music videos. The remaining videos are mostly comedy and entertainment videos.

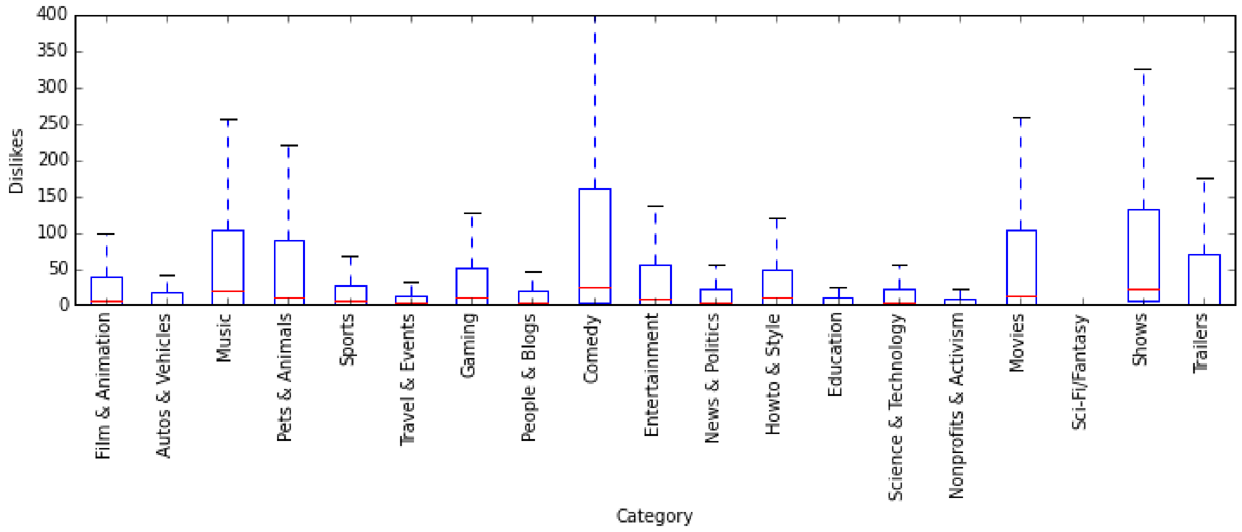

Figure 15. Video dislike count per category on YouTube

4.2 Evolution of the four key metrics over time

The four key metrics for video popularity are view count, comment count, like count, and dislike count (Chatzopoulou et al., 2010). Among these, view count is typically considered the primary metric (Szabo, & Huberman, 2010). Chatzopoulou et al. suggested for research to investigate the changes in the view, comment, like, and dislike count metrics over time. To address this, the secondary data set (*N* = 3,014,962) with the repeated measures data was used, as explained in section 3.1. Figure 16 shows the change in average view, comment, like, and dislike counts over a time span of 31 days for over three million videos. A $\log_{10}$ transformation was applied to the original values unless the original value was zero.

The tendency for view count to fluctuate more erratically over time than the other metrics can be seen on this figure as well as how view count is always the metric with the highest value. More importantly, one can see that view, comment, like, and dislike counts share a relationship. In their research, Chatzopoulou et al. wrote about these four metrics being highly correlated. Table 1 shows the exact correlations between these metrics from the primary data set and, indeed, the key metrics are highly positively correlated (all $r > .8$, all $p < .001$).

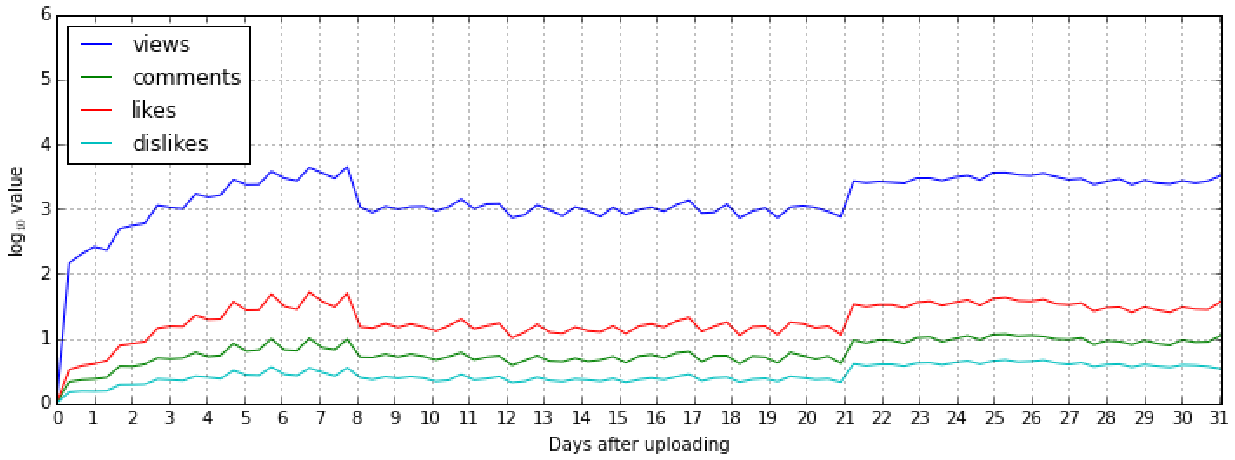

Figure 16. View, comment, like, and dislike counts of YouTube videos over 31 days

This insight also affects one's expectations for the linear regression models that were performed. One of the regression models consisted of view count as the metric to be predicted with comment and rating counts as the model's features. The rating count metric had to be created by adding the like and dislike counts together. By doing this, the new metric (rating count) correlates even more positively with view count (r = .870, p < .001) than the correlations for view count with like count and dislike count separately. So, with the understanding that view count correlates highly positively with comment count, rating count, one can already expect for that regression model to perform very well when it predicts video view count.

Table 1

Pearson's Correlation Coefficients for View Count, Comment Count, Like Count, and Dislike Count (N = 1,500,096).

	View count	Comment count	Like count	Dislike count
View count				
Comment count	.818			
Like count	.857	.896		
Dislike count	.821	.858	.866	

Notes: all p < .001.

5 Experiments

This chapter focuses on the experiments with Stochastic Gradient Descent-based (SGD) linear regression models in predicting the number of views of YouTube videos. For this purpose, the machine learning library for Python called Sci-kit Learn was used (http://scikit-learn.org). As discussed in section 2.1, the view count metric (the total number of views) is chosen to represent popularity. This is done because out of all the key metrics related to popularity (view count, comment count, like count, dislike count) view count is seen by most other research as the most important one (Chatzopoulou, Sheng, & Faloutsos, 2010). The justifications for the features used in the experiments were elaborated on in sections 3.2 to 3.4.3. The fact that the dependent variable (video view count) is known in advance in this study makes all the following experiments of the supervised learning type.

This chapter's first part describes how the data were divided over the training, validation, and test sets. The second part explains how the scores from the linear regression experiments were evaluated. The third part reports on the scores of those experiments that were conducted. This chapter concludes with an overview of the compositions of the final regression models, which are a direct result from these experiments.

5.1 Different data sets

The data were collected as described in the method section, resulting in a data set of 1.5 million YouTube videos to be used in the linear regression models. As is typical in machine learning, the data set had to be divided for training and testing purposes. A training set, a validation set, and a test set were preferred over the use of cross-validation with a training and test set, in which case one reuses data by splitting the test data in multiple equally sized parts and testing the performance of one part on the remaining parts (Karlsson, Andersson, & Norman, 2015). The reason for this was that the data set was sufficiently large (i.e., one typically opts for cross-validation in situations in which data is sparse) (Karlsson et al.., 2015).

Two aspects were important in splitting the data over the different sets. First, the data set had to be randomly shuffled before dividing it into training, validation, and test sets because the entire data set was ordered by video age (as explained in section 4.1.10). This shuffling process was done with a pre-set seed value to ensure that for each experiment, as well as for the final regression models, the same (randomly shuffled) training data were used to training with and the same (randomly shuffled) validation and test data were predicted on.

Secondly, the entire data set was not exceptionally big but was big enough to warrant the use of learning optimizations such as SGD (as discussed in section 3.5). Also, there was a need for resource heavy computations (e.g., ngrams and Part of Speech tags) preceding the regression analysis. So, a medium to large training set was appropriate. As such, the training set consists of 60% of the entire data set. The remaining 40% was split equally over the validation set and the test set (see Table 2).

Table 2

Allocation of Data (N = 1,500,096) to the Training, Validation, and Test Data Sets.

	n	Percentage
Training data set	900,056	60%
Validation data set	300,019	20%
Test data set	300,019	20%

5.2 Types of evaluation scores

After a regression model has learned to predict a target value from training data, its effectiveness in predicting can be measured. Several criteria can be used to compare different regression models. First, the coefficient of determination (*R²*) is the proportion of variance explained by the model, which means that the *R²* value is an indication of how well the model fits the data it has seen (goodness of fit) (Field, 2013). Formally, the *R²* metric is constructed by dividing the *Sum of Squared Errors* (*SSE*) by the *Total Sum of Squares* (*SST*), and subtracting that number from 1 (Field, 2013; Johnson, 2014),

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

Where N is the number of observations, y_i is an observed true value, $\bar{y}$ is the mean of all observed true values, $\hat{y}_i$ is a predicted value, and $(y_i - \hat{y}_i)$ is the difference between a true value and a predicted value (the loss) (Johnson, 2014). As such, an *R²* value lies within zero and one with one representing a perfect score and zero being an imperfect score. It is possible, however, to find negative *R²* values or *R²* values higher than one in certain cases such as when no intercept is included in the regression model (i.e., the model is worse than simply fitting a horizontal line) (Coster, 2009).

Next to this is the *Mean Squared Error* (*MSE*), the average of all the squared loss values for all predictions, as well as the Mean Absolute Error (*MAE*), the mean value of all the loss can be used to evaluate prediction scores. Formally, the *MSE* and *MAE* metrics are given by,

$$MSE = \frac{SSE}{N}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Where all other variables are as defined earlier (Field, 2013). The values for *MSE* and *MAE* can be zero (a perfect score) or higher, however, unlike the *R²* metric, the *MSE* and *MAE* metrics can yield extremely high values as both *MSE* and *MAE* have no predetermined range.

How well a regression model predicts is not solely determined by its parameters but also by the data that it predicts from. Typically, predicting on training data after a regression model has been trained on that same data will yield high results. Therefore, it is a common practice to use a validation data set to optimize the model's parameters before presenting it with unseen (test) data and evaluating its accuracy. Note that predicting on training data does not always result in the highest scores. It is possible for predictions on validation data or even on test data to result in higher evaluation scores than predictions on training data.

The experiments that follow have been trained using the aforementioned training data set (60%). The parameters and scores that follow have been obtained by having the regression model predict from the validation data set (20%). If a feature predicts poorly (e.g., *R²* = .001) from training data on validation data it is deemed an uninformative feature. Consequently, such an ill-performing feature will not be included in any of the final regression models. The final regression models have been trained on training data, optimized using validation data, and predict on the test data set (20%).

5.3 Word ngrams as features

To test whether Term Frequency Inverse Document Frequency (TF-IDF) ngrams—word grams as well as character ngrams—are better quality predictors than TF ngrams both types were tested separately in an SGD regression model. Next to this, as explained in the section 3.4.1, word ngrams can either be words or word combinations derived from texts. This is determined by the word count range that is set (i.e., a min and max value). Therefore, different ngram ranges were also explored. Additionally, the *Document Frequency* (DF) of an ngram could play a role in ngram effectiveness so different DF values were tested as well.

In regressing on video view count, word ngrams were inserted as the only features in a regression model with different ngram ranges and DF values for both TF and TF-IDF ngrams. The ngram ranges tested were (1, 1) to (1, 7), (2, 2) to (2, 7), (3, 3) to (3, 7), (4, 4) to (4, 7), (5, 5) to (5, 7), (6, 6) to (6, 7), and (7, 7) in the (min, max) format. For example, an ngram range of (1, 1) means that one word in total constitutes an ngram, whereas an ngram range of (3, 5) means that 3 words, 4 words, or 5 words are seen as an ngram. The DF values tested were (1, 100%), (3, 100%), (5, 100%) in the (min, max) format. A max value of 100 per cent indicates that there is no upper boundary for ngram occurrences.

5.3.1 TF-IDF word ngrams

In testing the effectiveness of TF-IDF versus TF features, first, TF-IDF word ngrams of video titles in the above mentioned ranges were inserted in a regression model with the use of the TfidfVectorizer from Sci-kit Learn. This resulted in a best score of R^2 = .175, *MAE* = 0.948, *MSE* = 1.412 on the validation set (see Figure 17). The parameters that produced this result were an ngram range of (1, 1), the squared epsilon insensitive loss function with an L2 penalty, an alpha of 10^{-5}, and a DF of (5, 100%).

The same was done with TF-IDF word ngrams for channel names. Again, the TfidfVectorizer was used. Here, the best obtained results with data from the validation set were R^2 = .136, *MAE* = 0.972, *MSE* = 1.479. An ngram range of (1, 1), a DF of (1, 100%), a squared epsilon insensitive loss function, an L2 penalty and an optimized alpha of 10^{-5} were responsible for this result.

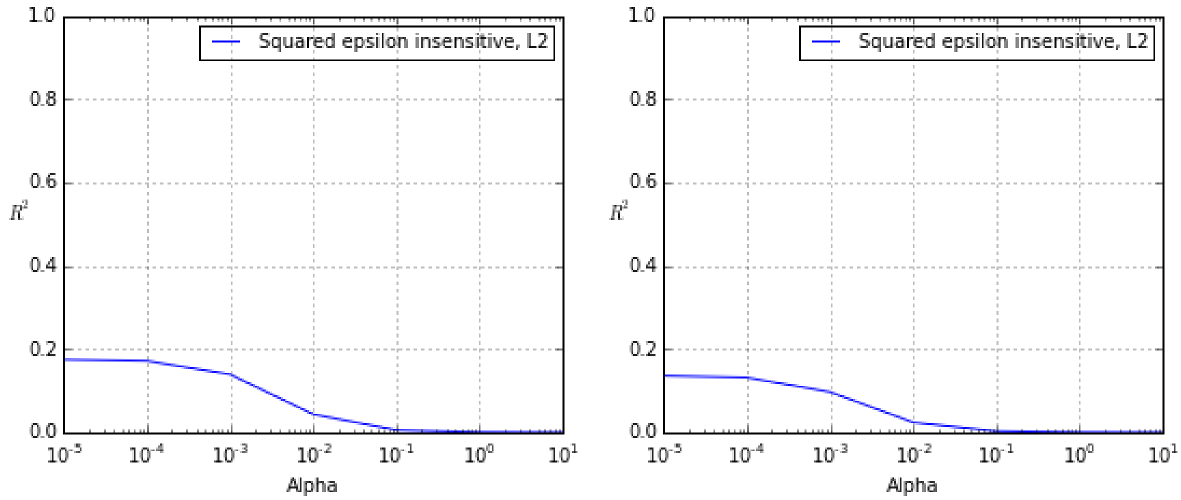

Figure 17. Explained variance scores for different alpha values with the best regression parameters for TF-IDF video title word ngrams with ngram range (1, 1) (left) and for TF-IDF channel name word ngrams with ngram range (1, 1) (right) in the validation set

5.3.2 TF word ngrams

For video titles, the CountVectorizer from Sci-kit Learn was used for TF word ngrams. The best score was associated with an ngram range of (1, 7), and a DF of (1, 100%). This resulted in R^2 = .277, *MAE* = 0.876, *MSE* = 1.238 on validation data (see Figure 18). These scores are considerably higher than those obtained with TF-IDF counterparts. The regression parameters used were the squared epsilon insensitive loss function with an L2 penalty. The optimal alpha for this result was 10^{-5}. DF ranges of (1, 100%), (3, 100%), and (5, 100%) were tested as well. However, regardless of loss function, penalizer and alpha, a DF range of (1, 100%) consistently scored higher than a DF of (3, 100%) and a DF of (5, 100%).

By using the CountVectorizer for channel names, the same word ngram range of (1, 7) as with video titles returned the highest scores. These were R^2 = .176, *MAE* = 0.944, *MSE* = 1.409 on the validation set when the squared epsilon insensitive loss function was used with an L2 penalty and an alpha of 10^{-5}. Again, the scores for TF word ngrams were higher than the scores for TF-IDF word ngrams. The word ngram ranges of (1, 4), (1, 5), and (1, 7) all resulted in the same R^2 of .176. Different DF ranges such as (3, 100%) and (5, 100%) did not improve the scores similarly to the word ngrams for video titles. So, a DF range of (1, 100%) was used for these scores.

For descriptions, the HashingVectorizer from Sci-kit Learn was used instead of the CountVectorizer. Video descriptions produced high numbers of ngrams and to compensate for this, the computationally lighter variant of the ngram vectorizer was used. The downside of this vectorizer, however, is that the coefficients cannot be mapped to the actual ngrams afterwards. Regardless, with this ngram vectorizer, word ngram ranges of (1, 1), (1, 2), and (1, 3) scored higher than (1, 4), (1, 5), (1, 6), and (1, 7). More precisely, word ngrams with range (1, 2) gave the highest scores on the validation data set, R^2 = .220, *MAE* = 0.920, *MSE* = 1.335, followed by ngram ranges of (1, 3) with an R^2 of .216 and an ngram range of (1, 1) scoring an R^2 of .215. The squared epsilon insensitive loss function was used with an L2 penalty and an alpha of 10^{-5}.

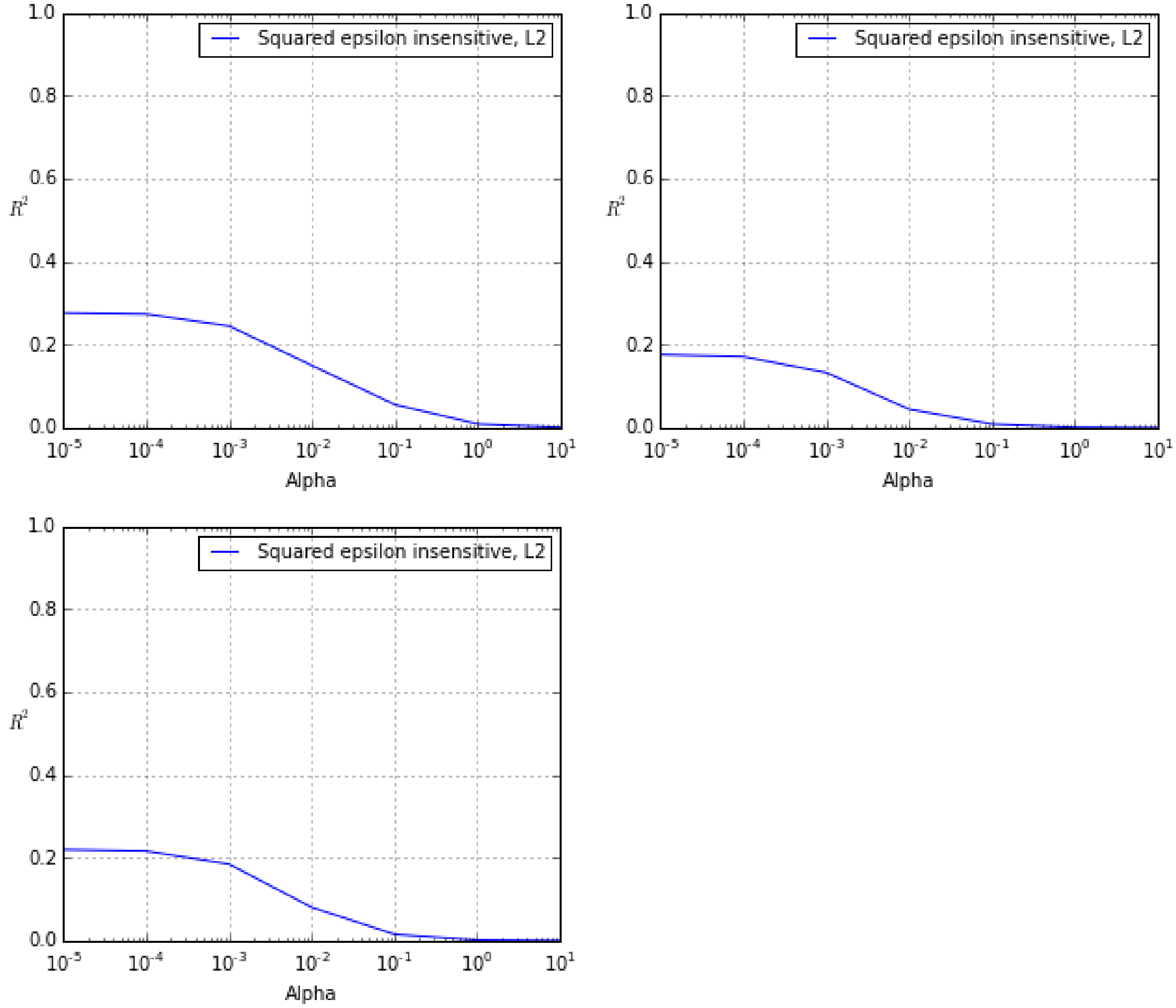

Figure 18. Explained variance scores for different alphas values with the best regression parameters for TF video title word ngrams with ngram range (1, 7) and DF (1, 100%) (top left),

TF channel name word ngrams with ngram range (1, 7) and DF (1, 100%) (top right, and video description word ngrams with ngram range (1, 2) (bottom left) in the validation set

5.4 Character ngrams as features

Similar to word ngrams, character ngram ranges and DFs could both play a role in their efficiency in predicting video view count; so, regression models with different values for both these parameters were tested. For video titles, the highest scores produced by the CountVectorizer were found by using the squared loss function with an L2 penalty. By setting the alpha to the optimal value of 10^{-4} the scores of R^2 = .320, *MAE* = 0.843, *MSE* = 1.163 were found for the validation set (see Figure 19). This was with a character ngram range of (5, 7) and a DF of (5, 100%).

For channel names, the evaluation scores were higher at R^2 = .364, *MAE* = 0.809, *MSE* = 1.089 by using the CountVectorizer on validation data. These results were obtained with the squared epsilon insensitive loss function and an L2 penalty. The optimal alpha value was 10^{-5}, the optimal character ngram range was (3, 7), and the right DF was (3, 100%).

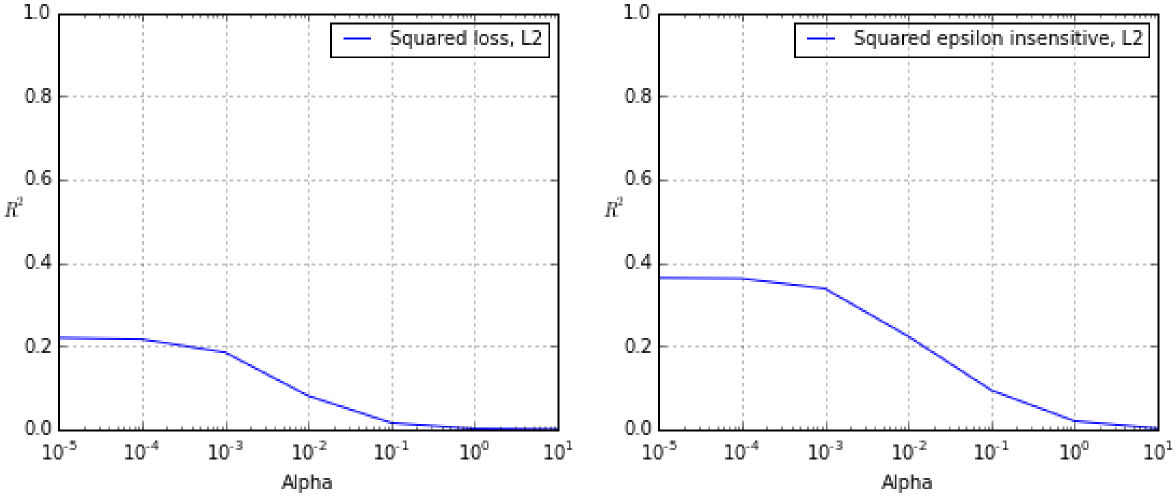

Figure 19. Explained variance scores for different alphas values with the best regression parameters for TF video titles character ngram with ngram range (5, 7) and DF (5, 100%) (left) and TF channel name character ngrams with ngram range (3, 7) and DF (3, 100%) (right) on the validation set

5.5 Part of Speech frequencies as features

As explained in the method section, parts of speech (POS) are types of words such as nouns, verbs, adjectives, adverbs, and pronouns. It was expected for POS frequencies to have an impact on predicting video view count. For example, nouns could be more important than verbs when they are used in video titles. To test this, POS frequencies were derived from video titles and channel names on YouTube. This was done by using the Natural Language Processing (NLTK) library (http://www.nltk.org) for Python. Subsequently, these frequencies were inserted as the only features in the regression model.

POS frequencies for video titles performed very poorly on validation data. The highest obtained scores were *R²* = .004, *MAE* = 1,164,718.03, *MSE* = 46,444,401,478,400 with the squared epsilon insensitive loss function, an L2 penalty, and an optimal alpha of 10^{-3}.

To exclude the possibility that just POS frequencies for video titles result in low evaluation scores, a regression model with POS frequencies for channel names was constructed. Again, NLTK was used for this. Here, the highest scores were *R²* = .001, *MAE* = 1,147,839.73, *MSE* = 46,552,040,216,900 with validation data. This was with the squared epsilon insensitive loss function, an L1 penalty, and an optimal alpha of 1.

5.6 Custom textual features

The final experiment consisted of the insertion of the features websites presence,
social media presence, character counts, word counts, uppercase/lowercase ratio, alphanumerical ratio, alphabetical ratio, numerical ratio for video titles, channel names and video descriptions. The thought behind these metrics was that people exert a certain influence on the number of views their videos will get by, for instance, using short and shouty texts (i.e., many uppercase characters) in their video titles.

This was tested by using these metrics as the only features included in a regression model in predicting view count. The highest score on the validation data set was *R²* = .085, *MAE* = 0.992, *MSE* = 1.566 (see Figure 20). This was with the squared loss function and an L1 penalty. The optimal alpha here was 10^{-5}. Similar scores were obtained by predicting on training data: *R²* = .083, *MAE* = 0.991, *MSE* = 1.563. The exact same loss function, penalizer, and alpha values were used for these results.

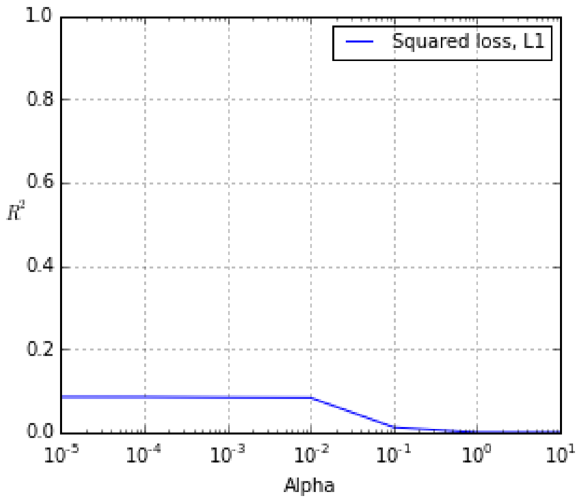

Figure 20. Explained variance scores for different alphas values for textual features of video titles, channel names, and video descriptions in the validation set

5.7 Composition of the final regression models

The experiments indicated which feature worked well and which did not. With this knowledge, four final regression models were constructed (see Table 3). Model 0 consisted of the intercept. Thus, the mean view count from the training data set was used for this model's predictions.

Model 1 was an approximation of the minimal model by Chatzopoulou et al. to investigate whether the regression models used in earlier literature on YouTube popularity prediction are still valid. Therefore, the features that were used by Chatzopoulou et al. in their research were used for this model. In their research, Chatzopoulou et al. used the following features in their model: number of favorites, number of comments, and number of ratings. However, it was found that these metrics have changed. For instance, the number of favorites metric no longer exists and could not be included in this model. Next to this, the number of ratings metric no longer exist on YouTube. Up until April 2010, YouTube allowed users with an account to rate videos on a scale from 1 star to 5 stars. This system was replaced because five star ratings dominated (Rajaraman, 2009). Nowadays, users can either 'like' or 'dislike' a video. The solution to this was that the number of ratings could be constructed by combining the like count and dislike count. This is because whereas users were restricted to casting one rating in the past, they are now restricted to casting either one like or one dislike. To conclude, the features included in this model were comment count and rating count (numerical features).

Model 2 contained all metrics that were obtained via the YouTube Data API, so, the earlier constructed rating count was replaced by the original like and dislike counts. Also, whereas model 1 consisted only of features that could not be influenced by those who upload videos to YouTube, this model had a combination of user-controllable features (e.g., video duration, video category) and features that could not be influenced directly (e.g., comment count, video age). Additionally, this was the first model to include categorical features next to numerical features.

Finally, model 3 consisted only of features that were under the direct influence of those who uploaded the videos in the data set to YouTube (see sections 3.2 to 3.4.3). As such, word counts were included next to categorical and numerical features.

Table 3

Overview of the Regression Models

Model	Description and features
0	**Intercept.** The mean view count from the training data set was used for the predicted values.
1	**A minimal model as proposed by Chatzopoulou, Sheng, and Faloutsos (2010).** The features: comment count, and rating count (composed of like and dislike counts).
2	**All the metrics of YouTube videos that were retrieved.** The features: video age, duration, comment count, like count, dislike count, video category, video definition, video dimension, availability of captions, whether licensed material was included, the license type, whether the video is embeddable on third-party websites, and whether the video's statistics are available.
3	**Only user-controlled features.** The features: video duration; video category, video definition, video dimension, availability of captions, whether the video contains licensed content, license the video is published under, whether the video is embeddable on third-party websites, whether the video's statistics are available, whether the description contains links to websites, whether the description contains links to social media, word ngrams for video titles, channel names, and video descriptions; character ngrams for the video titles and character ngrams; character counts of video titles, channel names, and video descriptions; word count of video titles, channel names, and video descriptions; ratios of uppercase to lowercase characters in video titles, channel names, and video descriptions; ratios of alphanumerical characters to other character types in video titles, channel names, and video descriptions; ratios of alphabetical characters to other character types in video titles, channel names, and video descriptions; and ratios of digits to other character types in video titles, channel names, and video descriptions.

6 Results

In this chapter, the evaluation scores for each of the final linear regression models are presented and compared. Also, an error analysis for model 3 is performed.

Table 4

Coefficient of Determination, Mean Absolute Error, and Mean Squared Error Scores for Predictions for Each Data Sets of all Regression Models

Data set	*R²*	*MAE*	*MSE*
Model 0 training set	0	1.040	1.706
Model 0 validation set	0	1.042	1.711
Model 0 test set	0	1.041	1.707
Model 1 training set	.753	0.487	0.420
Model 1 validation set	.755	0.487	0.420
Model 1 test set	.753	0.488	0.421
Model 2 training set	.818	0.382	0.311
Model 2 validation set	.819	0.382	0.310
Model 2 test set	.818	0.383	0.310
Model 3 training set	.639	0.595	0.616
Model 3 validation set	.536	0.680	0.795
Model 3 test set	.533	0.681	0.797

6.1 Comparison of the final regression models

Table 4 shows the evaluation scores for all regression models. Out of all models, model 2 is clearly the model where the proportion of variance explained is the highest for estimating video view count. Figure 21 shows that model 2 predicts better for higher view counts than for low view counts, and the worst for view counts below the mean view count. A straight line from the bottom left to the top right would be indicative of perfect prediction scores.

The model contains all 13 metrics that are directly associated with videos which were retrieved via the YouTube Data API and it produced *R²* = .818, *MAE* = 0.373, and *MSE* = 0.310 scores on the test data set. This was achieved with a squared loss, L1 and L2 penalizers (ElasticNet), and a strong regularization parameter of 10^{-5}. With those same parameters, it scored a similar *R²* = .818 on the training set and slightly better from validation data (*R²* = .819).

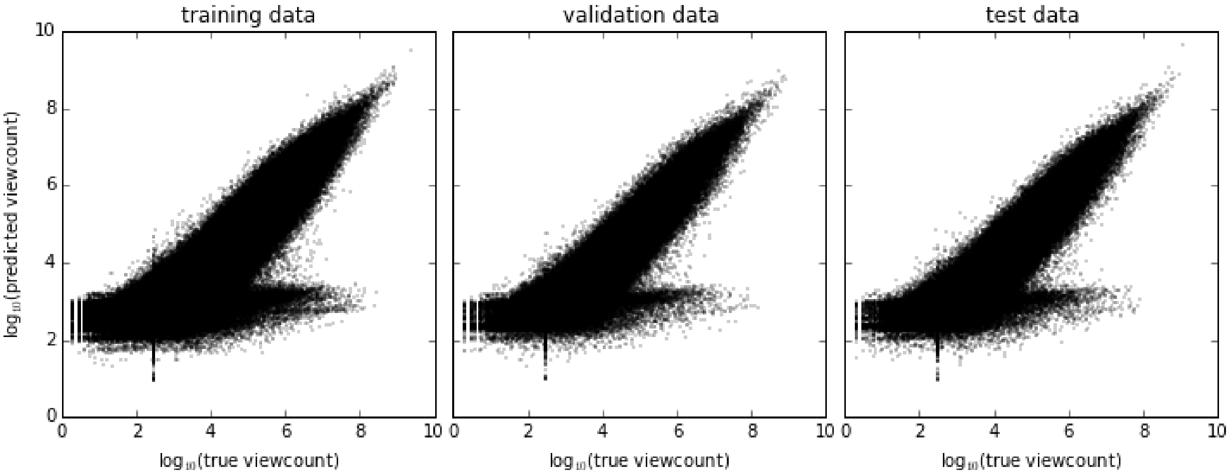

Figure 21. Predicted values versus true values for model 2 for all 3 data sets

Model 1 (see Figure 22), which consists of 2 features (comment count and rating count) is the second to highest scoring regression model (in terms of explained variance, mean absolute error, and mean squared error) for predicting YouTube video popularity as measured by view count. Figure 22 shows that model 1 performed equally well on the test data set (both *R²* = .753) and training data and only slightly better on validation data (*R²* = .755). In all these scores, the squared loss function, along with an L1 penalizer, and an optimal alpha of 10^{-3} were used. The difference in *R²* between model 1 and 2 is as low as .063 or 6.3%.

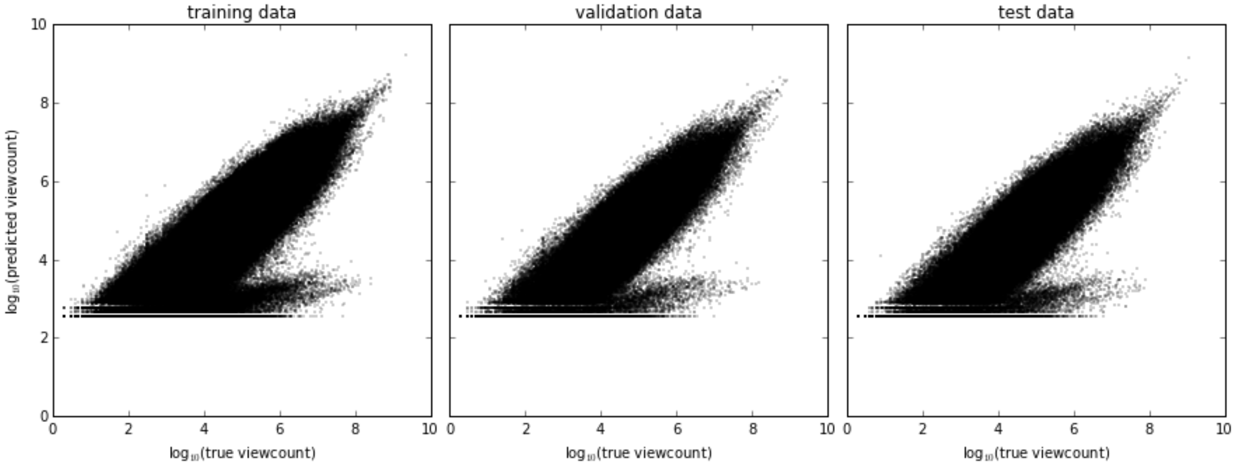

Figure 22. Predicted values versus true values for model 1 for all 3 data sets

Model 3 (see Figure 23), which includes only features that are under the direct control of the person who uploads video to YouTube, is also the model with the lowest *R²* scores. Also, the difference (Δ.103) between predicting data on the training data (*R²* = .639) and the test data (*R²* = .533) was smaller than the difference (Δ.106) between the training data and the validation data set (*R²* = .536). The parameters associated with these scores are the squared loss function, an L2 penalty and an optimal alpha of 10^{-5}. To compare this with the other models, no other model had a drop in performance this big from training to other data sets. Figure 23 shows this difference in performance, but also that the model predicts the best for high (e.g., 10^{8}) and low view counts (e.g., 10^{2}) and the worst for view counts in between (e.g., 10^{5}).

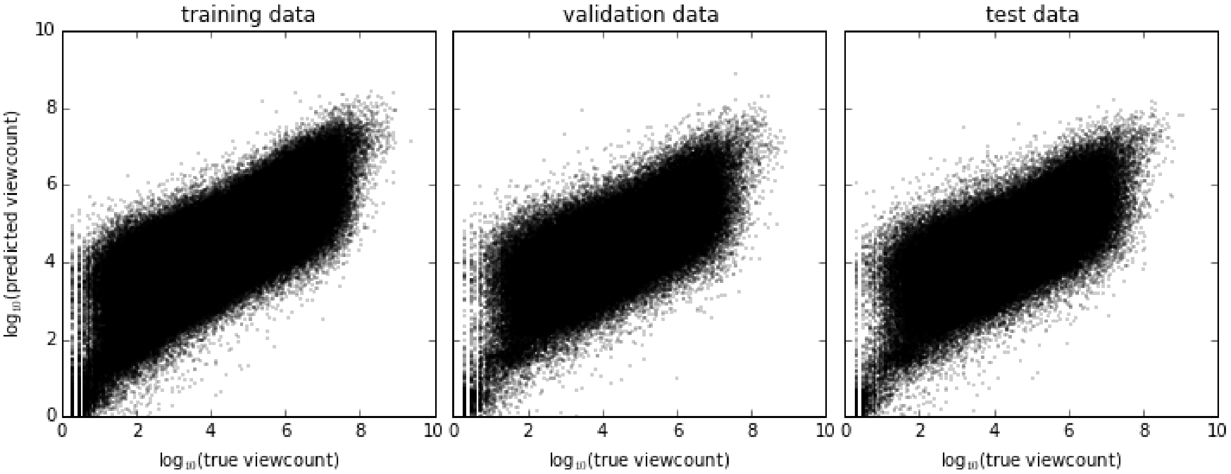

Figure 23. Predicted values versus true values for model 3 for all 3 data sets

Model 0 takes the mean view count from the training data set as its predicted values. Compared to this baseline model, model 3 still performs a great deal better with a change in R^2 of .456, an improved *MAE* (Δ0.315) as well as an improved *MSE* (Δ0.779) when the accuracy of the regression models on the test sets are compared.

6.2 Strongest features of model 3

Model 3 produced a total of 20,752,106 coefficients which are shown on Figure 24. These coefficients consist of 19 numerical features (e.g., video duration, uppercase to lowercase ratio of the video's title), 63 categorical features (e.g., video category, availability of subtitles); 15,287,587 word ngrams for video titles; 658,367 word ngrams for channel names; 1,048,576 word ngrams for video descriptions; 2,406,139 character ngrams for video titles; and 1,351,355 character ngrams for channel names. No coefficients for numerical and categorical features can be seen on Figure 24 because these features (19 and 63 features respectively) represent a tiny portion of the coefficients on the far left side of the figure. Instead, the coefficients for numerical and categorical features are displayed on Figure 25.

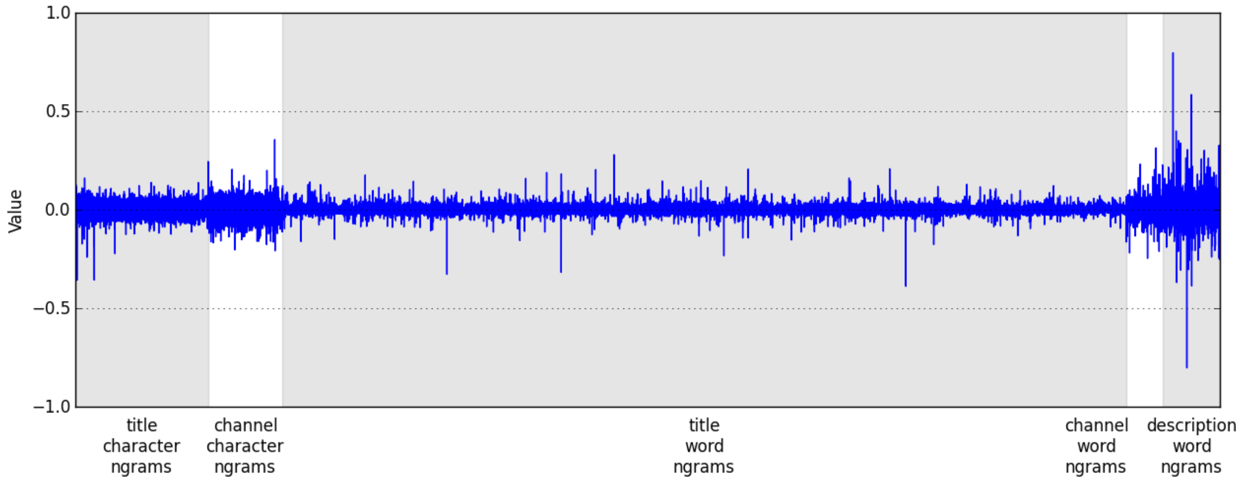

Figure 24. Regression coefficients for character and word ngram features of model 3

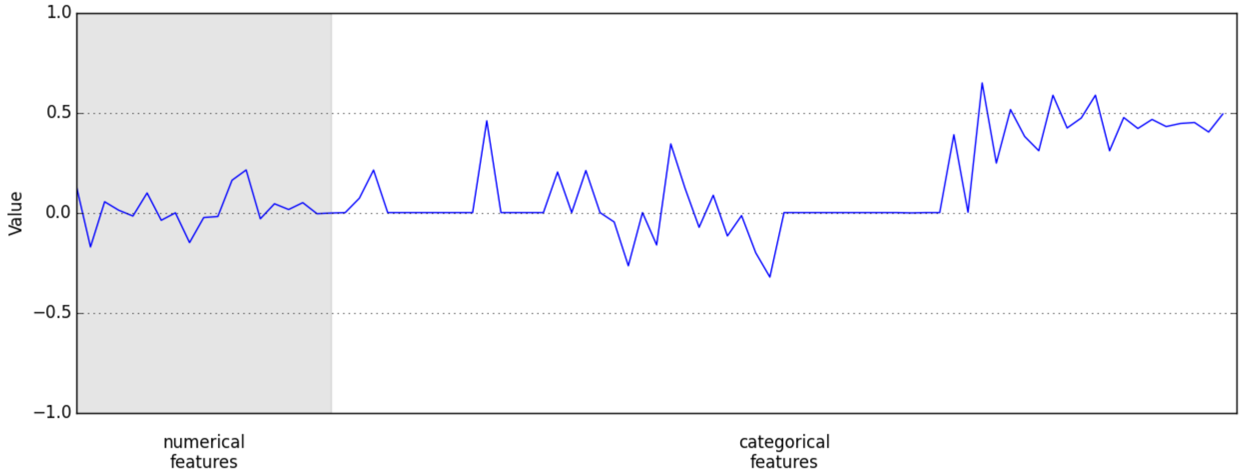

Figure 25. Regression coefficients for numerical and categorical features of model 3

Tables 5 and 6 show the top 10 strongest positive and negative coefficients of model 3 (see also Appendices E to H). However, as explained in section 5.3.2, a ‘hashed’ variant of ngram extraction was used, but only for video descriptions. The ngrams for video descriptions improved the regression model but the coefficients it produced could not be mapped to the actual ngrams. As a result, the exact content of such ngrams were not listed in Tables 5 and 6.

Table 5

Strongest Negative Coefficients

Feature	Coefficient value
Description word ngram	-0.802
Video title word ngram: ‘sold’	-0.388
Description word ngram	-0.386
Description word ngram	-0.369
Video title character ngram: ‘ - sold’	-0.357
Video title character ngram: ‘- sold’	-0.356
Video title word ngram: ‘cover’	-0.327
Video in the category Anime/Animation	-0.322
Video title word ngram: ‘full movie’	-0.317
Description word ngram	-0.306

Table 6

Strongest Positive Coefficients

Feature	Value
Description word ngram	0.798
HD quality	0.648
Published under the Creative Commons license	0.587
Contains no licensed material	0.587
Description word ngram	0.586
3D video	0.515
Links to SNSs in the video’s description	0.493
Embedding on third-party websites allowed	0.475
Published under the YouTube license	0.473
Availability of video statistics	0.466

6.3 Error analysis for model 3

For the purpose of discussing the prediction accuracy of model 3, the strongest outliers, as measured by the ratio (or proportion) of the predicted score to the true score, are analyzed. Such a ratio lies between zero and infinity. The videos closest to a ratio of zero and the videos with the highest ratios are used for this.

A visual inspection of these outliers for model 3 showed a trend for their inaccurate predictions of the total number of times a video has been viewed. Most of the worst predictions with ratios higher than one belonged to a single YouTube channel called *Hall Hyundai Newport News*. This channel has uploaded hundreds of car advertisement videos such as *Cadillac SRX #G5797A in Norman Oklahoma City OK 73069 - SOLD* (ratio = 141.57), *2012 Honda CR-Z #M2005620 in Westminster MD Baltimore MD - SOLD* (ratio = 123.23), and *2012 Honda Accord #2120633 in Virginia Beach VA Norfolk VA - SOLD* (ratio = 99.03). Most of videos from this channel have a view count of zero. Also, this channel is responsible for virtually all bad predictions with a true to predicted score ratio higher than five.

On the other end of the spectrum for true to predicted ratios, the same YouTube channel is also responsible for most of the bad predictions. Here, the channel takes up almost all videos with a ratio of zero since other videos, unrelated to *Hall Hyundai Newport News* have a ratio of at least 0.05. Car advertisements titled *2012 Jeep Wrangler #1821480 in Virginia Beach VA Norfolk – SOLD* can be found among the bad predictions for this particular YouTube channel for ratios close to zero. Most of this channel's video titles share a similar pattern with a year, a car brand and type, an identifier, a location and an indication of whether the car has been sold.

In general, the metric values associated with videos that are responsible for bad predictions produce a clear picture. Most of the videos with bad predictions have a title with a word count around 11 and a character count of about 56 where about 33.3% of these characters is in uppercase, and close to one-sixth of the characters are digits. Also, a word count ranging from 160 to 230 words with a character count in between 1,000 and 1,400 characters for video descriptions are associated with poorly predicted videos.

Other reasons for poor predictions can be found in bugs related to YouTube. Some videos, such as *Baú EQUI: Igreja orgânica em Curitiba no Paraná* (ratio = 0), have a view count that is frozen at zero. This means that the view count of that video never increases regardless of the actual number of times that people have watched the video. Meanwhile, other metrics for that video, such as the number of times people have commented on the video, do function as intended. Manually looking up the video on YouTube confirmed this unexpected behavior. As such, the view count for such a video is completely unrelated to any of the metrics that the model tries to predict from, which results in poor predictions for videos of this kind.

7 Discussion

In this chapter, first, the results of the regression experiments are discussed. This is followed by an interpretation of the results of the final regression models.

It has to be noted that all regression models predict from metrics that were measured at a certain point in time, a snapshot. So, view counts for the same videos over time are not modeled. However, a different model could be constructed where view count is predicted by linear regression with repeated measures (i.e., a linear mixed model). In such a case, multiple measurements are taken from every video in the data set at different times. Thus, a regression model that handles repeated measures data could potentially find trends in video view count for specific videos by assessing the change in certain metrics over time. Such a regression model was not the focus of this research, however.

7.1 Word and character ngrams

The initial regression experiment revolved around comparing the effectiveness of Term Frequency-Inverse Document Frequency (TF-IDF) and plain Term Frequency (TF) in a regression model. TF-IDF decreases the weights of frequently occurring yet unimportant ngrams (e.g., stop words) and increases the weight of rarely occurring ngrams (Manning, Raghavan, & Schütze, 2008). TF simply counts ngram occurrences. So, it was expected for TF-IDF ngrams to outperform TF ngrams upon insertion in the regression model. However, the results showed the opposite with TF-IDF scoring almost twice as low as TF in terms of explained variance regardless of the parameters (e.g., ngram ranges and document frequency ranges) that were used.

The finding that TF outperforms TF-IDF could be due to a high amount of noise in the bag of words that TF-IDF produces (Metzler, 2007). To illustrate this, it is highly likely that many people list as many terms as possible in their videos' descriptions or channel name in an attempt to get their videos to rank higher because their videos will match more queries from those who use YouTube's search functionality. Such terms will produce high TF-IDF scores because they are wrongfully considered as rare and informative. This could explain why the regression models with TF ngrams as its features scored better than the regression models with TF-IDF ngrams.

With the understanding that TF ngrams were to be used, different Document Frequency (DF) ranges were tested. For titles of YouTube videos as well as for names of YouTube channels, TF ngrams with a DF range of (1, 100%) scored better than DF ranges of (3, 100%) and (5, 100%).

The finding that a low minimum DF value yielded the best results can be explained by the number of features that are produced based on this parameter. To clarify, a high minimum DF value results in less features than when a low minimum DF value is used because this parameter functions as a cut-off point. Thus, by using the lowest possible value of one for this parameter, the most features are returned. This implies that either the total number of features that were found by using a high minimum DF value was too low or that these features were simply less informative than those that were found with a low minimum DF value.

7.2 Part of Speech frequencies

Part of Speech (POS) tag frequencies for video titles and channel names were inserted as features in separate regression models as experiments. The evaluation scores, in terms of explained variance, for these models (R^2 = .004 and R^2 = .001 respectively) can be considered as extremely low. No follow-up experiment with POS tag frequencies for video descriptions was conducted. This is because the low evaluation scores for POS tag frequencies for video titles and channel names made it very unlikely that such an experiment would produce better results.

Apparently, POS tag frequencies are not very predictive of video popularity. It was expected that, for example, nouns are significantly more important than verbs when they are used in video titles or channel names. The results of these experiments indicate that this is not the case however. This is unfortunate since the computation of POS tag frequencies are relatively resource inexpensive. Regardless, all these findings combined meant that POS tag frequencies were not included in any of the final regression models.

7.3 Custom textual features

Several user-controlled numerical and categorical features, such as social media presence, character counts, word counts and uppercase/lowercase ratios, were derived from video titles, channel names and video descriptions. These features were inserted in an experimental regression model which resulted in an R^2 of .085 or 8.5% of the explained variance. Even though this score is not particularly high, one can still consider it decent considering the fact that the features responsible for this score require little computing resources. This score can possibly be improved even further by adding more user-controlled features such as the channel category, the length and word count of the channel's description or the total number of videos one has uploaded on YouTube. Such features are just as easy to compute as the aforementioned features.

7.4 Final regression models

The optimal parameters for most of the regression experiments and final regression models show that a low alpha value and an L2 penalty produce the best results. The low alpha value indicates that little to no regularization was needed to prevent overfitting and generalization of the model. The L2 penalty means that the model favored features with similar to equal coefficients over a sparse model (i.e., a model with many coefficients of zero). Figures 24 and 25, which show the coefficients of model 3, support this by displaying many non-zero coefficients that are spread out over the entire model.

Going purely by the evaluation scores that these parameters produced, regression model 2 (with all retrieved metrics included as its features) is the best performing model and the clear winner out of all four models. In that same line of thought, model 1 (with only 2 metrics as its features) takes second place, and model 3 performs a bit worse and deserves only a third place. Finally, model 0 (with the mean view count from the training data) performs the worst. However, low R^2 scores are not necessarily an indication of bad models, especially in the social sciences (Moksony, 1999) so the R^2 score is not the Holy Grail here. As model 3 consists only of features that are under the control of the persons who uploaded the videos, this makes this particular model interesting from a social perspective. This very fact also makes it difficult for the model to predict because, simply put, people can behave in unpredictable ways. So, instead of positing that model 3 is a bad model, this research suggests that model 3 is a good model because it is predictive as opposed to descriptive.

The fluctuations in the results for different data sets that are only present for model 3 are also worth commenting on. Model 3's results indicate that it performed best on training data but a little worse on validation and test data. This was not the case for models 1 and 2 where the evaluation scores for predicting on different data sets were almost the same. This means that models 1 and 2 will likely yield similar scores when new data is gathered and presented to these models but this is not the case for model 3. The results for model 3 indicate that the model is difficult to train in such a way that its evaluation scores are similar to its scores on training data.

Model 3's strongest positive and negative coefficients, which are shown in Tables 5 and 6. These tables indicate that the model's strongest positive coefficients consist mostly of categorical features. It is not surprising that YouTube videos with high resolutions (HD) which are original (unlicensed material) and can be distributed even further by anyone (embedding on third-party websites) are associated with predicted high view counts because these properties are indicative of high quality videos. On the other hand, ngrams are primarly found among the model's strongest negative coefficients. Many of the features that were responsible for these negative coefficients (e.g., the word ngram 'sold') were discussed in section 6.3. It was shown that most of the videos connected with these features had a view count of zero and were responsible for many of the poorly predicted view counts. Fortunately, the model can be updated to deal with videos that share such characteristics. For instance, besides the possibility of removing the YouTube channel from the data set entirely, all the car advertisement videos, that were responsible for many of the bad predictions, had a similar video title pattern (e.g., the year, a car brand, a car type, a location) in a specific order. The videos in the data set could be filtered for titles that are constructed in this particular way. Also, the data set contained videos where certain metrics that are supposed to change over time did not change. Such an oddity could also be avoided by improving the features of the regression model.

Model 3 might be the most important model of this research but model 1 was the start-off point for this study. In its essence, model 1 makes an attempt at validating one of the regression models that were found by Chatzopoulou, Sheng, and Faloutsos (2010). Their results showed two regression models that predicted video view count, a model with many features (R^2 = .786), and a simplified version of that model (R^2 = .786). This research attempted to reconstruct the simplified version (i.e., with the number of times a video has been favorited by people, and the number of ratings people have given a video as features) but this was hindered. In between Chatzopoulou et al.'s research and this research, YouTube had removed the favorite and rating count metrics. In approximating Chatzopoulou et al.'s model, model 2 had of rating count (i.e., by summing like count and dislike count) and comment count as its features. Interestingly, the new model (R^2 = .753) scores relatively similarly to Chatzopoulou et al.'s model (R^2 = .786). The difference was a mere .033, or 3.3%, of the explained variance so these findings are in line with the findings from previous literature.

8 Conclusion

In this final chapter, the most important findings of this study are first given, followed by the possible implications of these findings. Then, limitations of this research that arose during this study are presented and elaborated on. Finally, suggestions for future work are provided.

8.1 Main findings

The following research questions were formulated:

1. How accurately can a supervised machine learning setup predict video view count from user-controlled variables?
2. To what extent have modifications to YouTube had an impact on predictive models from earlier studies?

The answer to the first research question was provided by the main regression model of this research. Regression model 3 has shown that the popularity of a YouTube video, in terms of view count, can be fairly well predicted from features that are under the full control of the person responsible for uploading the video. In this regard, model 3's prediction were substantially better than simply taking the mean view count as a video's view count prediction. Also, the error analysis of model 3 has clearly presented options for improving the model that easy to implement as well. This includes pattern recognition for video titles and descriptions, as well as checking for metrics that are supposed to change when other metrics have changed as well, and both these suggested optimizations are very resource inexpensive.

Before answering the second research question, one must realize that all major Social Networking Sites (SNS) are constantly undergoing changes, YouTube included (e.g., O'Reilly, 2007). A study from five years ago (Chatzopoulou, Sheng, & Faloutsos, 2010) found a good performing descriptive model in a supervised learning setup by regressing on video view count with the features favorite count (i.e., the number of times a video had been favorited by user of YouTube) and rating count (i.e., the number of times people had given a video a rating on a scale from 1 to 5 stars). However, in this research, it was found that neither favorite count nor rating count still exist. As an approximation of rating count, like count and dislike count were summed to form rating count but no substitute for favorite count exists. So, the findings of this research indicate that machine learning projects that focus on SNSs need to frequently update the metrics from which they predict.

8.2 Limitations and future research

Several limitations arose while conducting this research. First, this research aimed to include features for video popularity prediction that are under the control of the person who uploaded the video. The most prominent user-controlled aspects of videos on YouTube are the video title, the name of the channel to which the video belongs, the video's description, and the image thumbnail that the video owner can set for the video. However, the features used for the machine learning models in this research only relate to the first three aspects. So, future research should look into deriving features from image thumbnails. This requires retrieving, storing and extracting features from large numbers of images. Extraction of values for red, green, blue, hue, saturation, brightness could be considered but such values might be too generic. More information-rich features (e.g., object detection and identification) are more computationally heavy but may improve the accuracy of a regression model (Viola, & Jones, 2001). Perhaps, a presence of people's faces in video thumbnails results in more video views than when this is not the case.

Secondly, in line with the previous limitation, no features were gathered from the videos themselves for inclusion in the linear regression models. It is very likely that certain relationships exist between video content and popularity metrics. Perhaps, the number of comments or dislikes a video gets is more related to particular video content than view count. This could shift the focus from a one-way video aspect (i.e., viewing a video) to a two-way aspect that is interactive and social (i.e., commenting on a video and replying to other people’s messages).

Finally, the number of observations in the data set of this research, although not a limitation as proven with this research's findings, should be increased in future research. As explained in the section 4.1.1, some of the video categories were absent or underrepresented. Also, YouTube hosted an estimated video count in between 141 and 144 million in 2012 (Russakovskii, 2012) so the exact number of videos that YouTube hosts ought to be higher than that nowadays. Thus, this data set of 1.5 million videos represents slightly less than 1.04 per cent of all YouTube’s videos. Even though this is representative of YouTube is general, one cannot find short-term trends in such a data set which potentially could be spotted with a data set where all videos from every day are captured. With the understanding that metrics from SNSs, such as YouTube, are occasionally removed, altered or added, a challenge arises. A new study could construct a supervised machine learning project which keeps itself updated on changing metrics, is constantly gathering video data, with historical data as well as with time series for all videos, and which is continuously improving its evaluation scores in predicting the number of views a newly uploaded video will receive. The practical implication could be that those interested in uploading videos on YouTube could analyze their videos before they are uploaded. They could then be provided with scientifically based information on which aspects of their videos are successful along with advice on how to improve their videos. People's increase in popularity might be one reason for wanting such a system but there are financial reasons as well. YouTube has paid out over one billion dollars in total to more than one million people who participate in its partner program (YouTube, 2015b). This partner program allows people to monetize their videos by attaching advertisements to them which means that more money is made when their videos are viewed more often (Cheng et al., 2014). So, those involved in YouTube's partner program will very likely be interested in a system that tells them how to improve their videos because the advice they are given might affect their wallet in a positive way.

References

Abhari, A., & Soraya, M. (2010). Workload generation for YouTube. *Multimedia Tools and Applications, 46*(1), 91-118.

Adler, P. S., & Kwon, S. W. (2000). *Social capital: The good, the bad, and the ugly*.

Almeida, J. M., Krueger, J., Eager, D. L., & Vernon, M. K. (2001). Analysis of educational media server workloads. In *Proceedings of the 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, 21-30. ACM.

Bandari, R., Asur, S., & Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity. In *ICWSM*, 26-33.

Barabási, A. L. (2009). Scale-free networks: A decade and beyond. *Science*, *325*(5939), 412.

Baron, R. A., & Markman, G. D. (2000). Beyond social capital: How social skills can enhance entrepreneurs' success. *The Academy of Management Executive, 14*(1), 106-116.

Borghol, Y., Ardon, S., Carlsson, N., Eager, D., & Mahanti, A. (2012). The untold story of the clones: Content-agnostic factors that impact Youtube video popularity. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1186-1194. ACM.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, 177-186. Physica-Verlag HD.

Boyd, D. M. & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer‐Mediated Communication, 13*(1), 210-230.

Brandtzæg, P. B., & Heim, J. (2009). Why people use social networking sites. In *Online Communities and Social Computing*, 143-152. Springer Berlin Heidelberg.

Breheny, O. (2011). *Penalized regression: Introduction* [Powerpoint slides]. Retrieved from http://web.as.uky.edu/statistics/users/pbreheny/764-F11/notes/8-30.pdf

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Workshop on Speech and Natural Language*, 112-116. Association for Computational Linguistics.

Brodersen, A., Scellato, S., & Wattenhofer, M. (2012). YouTube around the world: Geographic popularity of videos. In *Proceedings of the 21st International Conference on World Wide Web*, 241-250. ACM.

Burke, M., Kraut, R., & Marlow, C. (2011). Social capital on Facebook: Differentiating uses and users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 571-580. ACM.

Carmel, D., Roitman, H., & Yom-Tov, E. (2010). On the relationship between novelty and popularity of user-generated content. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 1509-1512. ACM.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., & Moon, S. (2007). 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*, 1-14. ACM.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., & Moon, S. (2009). Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking (TON), 17*(5), 1357-1370.

Cha, M., Mislove, A., & Gummadi, K. P. (2009). A measurement-driven analysis of information propagation in the Flickr social network. *In Proceedings of the 18th International Conference on World Wide Web*, 721-730. ACM.

Chatzopoulou, G., Sheng, C., & Faloutsos, M. (2010). A first step towards understanding popularity in YouTube. In *INFOCOM IEEE Conference on Computer Communications Workshops, 2010*, 1-6. IEEE.

Chen, C. Y., Shih, B. Y., Chen, Z. S., & Chen, T. H. (2011). The exploration of internet marketing strategy by search engine optimization: A critical review and comparison. *African Journal of Business Management*, *5*(12), 4644-4649.

Cheng, X., Dale, C., & Liu, J. (2007). Understanding the characteristics of internet short video sharing: YouTube as a case study. *ArXiv preprint arXiv:0707.3670*.

Cheng, X., Dale, C., & Liu, J. (2008). Statistics and social network of Youtube videos. In *16th International Workshop on Quality of Service, 2008*, 229-238. IEEE.

Cheng, X., Fatourechi, M., Ma, X., Zhang, C., Zhang, L., & Liu, J. (2014). Insight data of YouTube from a partner's view. In *Proceedings of Network and Operating System Support on Digital Audio and Video Workshop*, 73. ACM.

Cherkasova, L., & Gupta, M. (2004). Analysis of enterprise media server workloads: Access patterns, locality, content evolution, and rates of change. *IEEE/ACM Transactions on Networking*, *12*(5), 781-794.

Chesire, M., Wolman, A., Voelker, G. M., & Levy, H. M. (2001). Measurement and analysis of a streaming media workload. In *USITS*, *1*, 1-1.

Cho, J., & Roy, S. (2004). Impact of search engines on page popularity. In *Proceedings of the 13th International Conference on World Wide Web*, 20-29. ACM.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661-703.

Coster, A. (2009). *Goodness-of-fit statistics*. Retrieved from web.maths.unsw.edu.au/~adelle/Garvan/Assays/GoodnessOfFit.html

Crane, R., & Sornette, D. (2008). Viral, quality, and junk videos on YouTube: Separating content from noise in an information-rich environment. In *AAAI Spring Symposium: Social Information Processing*, 18-20.

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science*, *267*(5199), 843-848.

Daugherty, T., Eastin, M. S., & Bright, L. (2008). Exploring consumer motivations for creating user-generated content. *Journal of Interactive Advertising*, *8*(2), 16-25.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet (2010). The YouTube video recommendation system. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, 293-296. ACM.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data mining and Knowledge Discovery*, *3*(4), 409-425.

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, *12*(4), 1143-1168.

Evans, T. (2013). *How YouTube's updated platform has successfully embraced emerging consumer behavior*. Retrieved from www.economistgroup.com/leanback/channels/how-youtubes-updated-platform-has-successfully-embraced-emerging-consumer-behavior/

Facebook (2015). *Company Info | Facebook Newsroom*. Retrieved from http://newsroom.fb.com/company-info/

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the internet topology. *ACM SIGCOMM Computer Communication Review*, *29*(4), 251-262.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Figueiredo, F. (2013). On the prediction of popularity of trends and hits for user generated videos. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, 741-746. ACM.

Figueiredo, F., Almeida, J. M., Gonçalves, M. A., & Benevenuto, F. (2014). On the dynamics of social media popularity: A YouTube case study. *ACM Transactions on Internet Technology (TOIT), 14*(4), 24.

Figueiredo, F., Benevenuto, F., & Almeida, J. M. (2011). The tube over time: Characterizing popularity growth of Youtube videos. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 745-754. ACM.

Gill, P., Arlitt, M., Li, Z., & Mahanti, A. (2007). YouTube traffic characterization: A view from the edge. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, 15-28. ACM.

Google (2015). *YouTube Data API*. Retrieved from https://google-api-client-libraries.appspot.com/documentation/youtube/v3/python/latest/

Google Developers (2014). *YouTube Data API (v3)*. Retrieved from https://developers.google.com/youtube/v3/

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in web Intelligence*, *1*(1), 60-76.

Halvey, M. J., & Keane, M. T. (2007). Exploring social dynamics in online media sharing. In *Proceedings of the 16th International Conference on World Wide Web*, 1273-1274. ACM.

Hong, L., Dan, O., & Davison, B. D. (2011). Predicting popular messages in Twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web*, 57-58. ACM.

Instagram (2015). *Press Page • Instagram*. Retrieved from https://instagram.com/press/

Johnson, C. (2014). *Linear regression with Python*. Retrieved from http://connor-johnson.com/2014/02/18/linear-regression-with-python/

Kaltenbrunner, A., Gomez, V., & Lopez, V. (2007). Description and prediction of Slashdot activity. In *Web Conference, 2007. LA-WEB 2007. Latin American*, 57-66. IEEE.

Karlsson, C., Andersson, M., & Norman, T. (Eds.). (2015). *Handbook of Research Methods and Applications in Economic Geography*. Edward Elgar Publishing.

Kong, S., Ye, F., & Feng, L. (2014). Predicting future retweet counts in a microblog. *Journal of Computational Information Systems, 10*(4), 1393-1404.

Kupavskii, A., Umnov, A., Gusev, G., & Serdyukov, P. (2013). Predicting the audience size of a tweet. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 693-696.

Lerman, K., & Hogg, T. (2010). Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th International Conference on World Wide Web*, 621-630. ACM.

Lerman, K., & Hogg, T. (2012). Using stochastic models to describe and predict social dynamics of web users. *ACM Transactions on Intelligent Systems and Technology (TIST), 3*(4), 62.

Lesher, G. W., Moulton, B. J., & Higginbotham, D. J. (1999). Effects of ngram order and training text size on word prediction. In *Proceedings of the RESNA'99 Annual Conference*, 52-54.

Lin, K. Y., & Lu, H. P. (2011). Why people use social networking sites: An empirical study integrating network externalities and motivation theory. *Computers in Human Behavior*, *27*(3), 1152-1161.

Lipczak, M., Trevisiol, M., & Jaimes, A. (2013). Analyzing favorite behavior in Flickr. In *Advances in Multimedia Modeling*, 535-545. Springer Berlin Heidelberg.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*, 496. Cambridge: Cambridge University Press.

Metzler Jr, D. A. (2007). *Beyond bags of words: Effectively modeling dependence and features in information retrieval*. ProQuest.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, 29-42. ACM.

Moksony, F. (1999). Small is beautiful. The use and interpretation of R^2 in social research. *Review of Sociology*, 130-138.

Mossa, S., Barthelemy, M., Stanley, H. E., & Amaral, L. A. N. (2002). Truncation of power law behavior in “scale-free” network models due to information filtering. *Physical Review Letters*, *88*(13), 138701.

O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies, 1*, 17.

Osborne, J. (2005). Notes on the use of data transformations. *Practical Assessment, Research and Evaluation*, *9*, 42-50.

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics, 443*, 59-72.

Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters*, *86*(14), 3200.

Pinto, H., Almeida, J. M., & Gonçalves, M. A. (2013). Using early view patterns to predict the popularity of YouTube videos. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, 365-374. ACM.

Rajaraman, S. (2009). *Five Stars Dominate Ratings*. Retrieved from http://youtube-global.blogspot.nl/2009/09/five-stars-dominate-ratings.html

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*.

Ruhela, A., Tripathy, R. M., Triukose, S., Ardon, S., Bagchi, A., & Seth, A. (2011). Towards the use of online social networks for efficient internet content distribution. In *2011 IEEE 5th International Conference on Advanced Networks and Telecommunication Systems (ANTS)*, 1-6. IEEE.

Russakovskii, A. (2012). *How to find out the number of videos on Youtube*. Retrieved from beerpla.net/2008/08/14/how-to-find-out-the-number-of-videos-on-youtube/

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on new Methods in Language Processing*, 12, 44-49.

Schmidt, M. (2005). Least squares optimization with L1-norm regularization. *CS542B Project Report*.

Scikit-Learn Developers (2010). *Stochastic Gradient Descent*. Retrieved from http://scikit-learn.org/stable/modules/sgd.html

Sripanidkulchai, K., Maggs, B., & Zhang, H. (2004). An analysis of live streaming workloads on the internet. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, 41-54. ACM.

Yu, H., Zheng, D., Zhao, B. Y., & Zheng, W. (2006). Understanding user behavior in large-scale video-on-demand systems. In *ACM SIGOPS Operating Systems Review*, 40(4), 333-344. ACM.

Steinfield, C., Ellison, N. B., & Lampe, C. (2008). Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, *29*(6), 434-445.

Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. *Communications of the ACM*, *53*(8), 80-88.

The Next Web (2015). *Foursquare counts its 3 billionth check-in*. Retrieved from http://thenextweb.com/location/2012/11/21/foursquare-has-its-3-billionth-check-in-seeing-growth-of-x/

Twitter (2014). *Your new web profile is here*. Retrieved from https://blog.twitter.com/2014/your-new-web-profile-is-here

Twitter (2015). *About Twitter, Inc. | About*. Retrieved from https://about.twitter.com/company

Utz, S., Tanis, M., & Vermeulen, I. (2012). It is all about being popular: The effects of need for popularity on social network site use. *Cyberpsychology, Behavior, and Social Networking*, *15*(1), 37-42.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1, 511-518. IEEE.

Waters, R. D., Burnett, E., Lamm, A., & Lucas, J. (2009). Engaging stakeholders through social networking: How nonprofit organizations are using Facebook. *Public Relations Review*, *35*(2), 102-106.

Wu, F., & Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences*, *104*(45), 17599-17601.

Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 177-186. ACM.

YouTube (2015a). *PSY - GANGNAM STYLE (강남스타일) M/V*. Retrieved from http://www.youtube.com/watch?v=9bZkp7q19f0

YouTube (2015b). *Statistics – YouTube*. Retrieved from https://www.youtube.com/yt/press/statistics.html

YouTube Official Blog (2011). *Get more into movies on YouTube*. Retrieved from http://youtube-global.blogspot.nl/2011/05/get-more-into-movies-on-youtube.html

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-first International Conference on Machine Learning*, 116. ACM.

Zhou, R., Khemmarat, S., & Gao, L. (2010). The impact of YouTube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, 404-410). ACM.

Zimet, G. D., Dahlem, N. W., Zimet, S. G., & Farley, G. K. (1988). The multidimensional scale of perceived social support. *Journal of Personality Assessment, 52*(1), 30-41.

Zink, M., Suh, K., Gu, Y., & Kurose, J. (2008). Watch global, cache local: YouTube network traffic at a campus network: Measurements and implications. In *Electronic Imaging 2008*, 681805-681805. International Society for Optics and Photonics.

Appendix A

Top 50 Videos on YouTube, Sorted by Number of Views

#	Title	Category	**Views**	Comments	Likes	Dislikes
1	PSY - GANGNAM ...	Music	**2,319,515,787**	5,400,599	9,368,896	1,256,307
2	Justin Bieber ...	Music	**1,161,556,092**	6,774,237	2,906,339	4,600,610
3	Katy Perry - D ...	Music	**931,129,321**	388,137	2,841,777	458,998
4	Katy Perry - R ...	Music	**902,332,616**	383,383	3,180,803	255,107
5	LMFAO - Party ...	Music	**861,694,735**	565,397	2,688,535	133,673
6	Eminem - Love ...	Music	**854,664,643**	614,649	2,358,439	85,075
7	Enrique Iglesi ...	Music	**847,003,236**	101,620	2,212,815	147,290
8	Shakira - Waka ...	Music	**844,000,601**	777,939	1,451,466	105,065
9	PSY - GENTLEMA ...	Music	**830,031,169**	1,134,567	3,278,938	636,394
10	Jennifer Lopez ...	Music	**827,935,713**	432,427	1,573,929	127,931
11	Charlie bit my ...	Comedy	**818,368,377**	626,957	1,387,191	219,016
12	Taylor Swift - ...	Music	**810,375,030**	352,753	3,498,860	240,502
13	Meghan Trainor ...	Music	**790,501,987**	348,628	3,170,558	310,631
14	OneRepublic - ...	Music	**771,094,470**	135,938	2,642,349	82,921
15	Taylor Swift - ...	Music	**770,094,429**	333,435	3,050,395	352,500
16	Miley Cyrus - ...	Music	**764,902,467**	1,142,983	2,730,753	1,178,257
17	Sia - Chandeli ...	Music	**698,967,208**	248,396	2,599,138	171,211
18	MACKLEMORE & R ...	Music	**698,627,001**	433,156	3,054,337	126,006
19	Carly Rae Jeps ...	Music	**674,796,029**	612,942	2,191,388	167,098
20	Romeo Santos - ...	Music	**668,434,864**	69,189	1,235,465	92,773
21	Bruno Mars - T ...	Music	**659,005,870**	459,022	2,143,534	95,973
22	Adele - Rollin ...	Music	**643,233,943**	380,165	2,249,312	66,771
23	Pharrell Willi ...	Music	**642,972,019**	138,166	2,832,358	134,182
24	Gotye - Somebo ...	Music	**634,150,407**	454,857	2,474,754	116,119
25	One Direction ...	Music	**632,693,879**	1,055,857	2,595,385	280,957
26	Michel Teló - ...	Music	**628,432,058**	321,639	1,195,394	121,344
27	Lady Gaga - Ba ...	Music	**621,885,423**	1,143,346	1,051,401	200,447
28	Avicii - Wake ...	Music	**616,329,802**	166,819	1,988,523	79,630
29	Eminem - Not A ...	Music	**615,871,729**	1,077,030	2,545,875	71,389
30	Passenger - Le ...	Music	**611,564,803**	136,302	2,381,654	63,840
31	Mark Ronson - ...	Music	**583,044,402**	167,806	3,076,163	126,917
32	Pitbull - Timb ...	Music	**582,909,430**	89,127	1,457,928	99,194
33	Pitbull - Rain ...	Music	**580,430,471**	185,954	1,172,018	55,617
34	Маша и Медведь ...	Shows	**565,631,590**	7,964	355,251	233,117
35	Rihanna - Diamonds	Music	**564,835,892**	211,339	1,828,110	111,850
36	Katy Perry - F ...	Music	**564,834,439**	598,409	1,607,853	89,296
37	PSY (ft. HYUNA ...	Music	**564,387,286**	405,199	1,397,990	343,334
38	Wheels On The ...	Education	**554,665,206**	3,536	369,264	288,494
39	The Gummy Bear ...	Film & Animation	**554,184,908**	168,284	661,117	193,341
40	PULCINO PIO - ...	Music	**553,282,193**	93,614	784,433	238,778
41	Bruno Mars - J ...	Music	**546,172,832**	211,120	1,470,579	38,763
42	Martin Garrix ...	Music	**545,883,369**	119,534	2,121,965	83,764
43	Jason Derulo - ...	Music	**539,573,306**	82,507	1,869,821	133,831
44	Miley Cyrus - ...	Music	**537,378,893**	1,016,324	1,812,910	1,134,866
45	Iggy Azalea - ...	Music	**532,898,109**	187,042	1,791,749	182,921
46	Prince Royce - ...	Music	**529,504,739**	76,975	1,154,178	74,634
47	Shakira - Can' ...	Music	**520,417,387**	177,955	1,690,102	181,841
48	Naughty Boy - ...	Music	**516,771,790**	192,063	1,512,714	103,940
49	Ellie Goulding ...	Music	**511,308,316**	115,312	1,765,798	59,858
50	Ylvis - The Fo ...	Entertainment	**510,573,787**	873,591	3,668,785	422,446

Appendix B

Top 50 Videos on YouTube, Sorted by Number of Comments

#	Title	Category	Views	**Comments**	Likes	Dislikes
1	Justin Bieber ...	Music	1,161,556,092	**6,774,237**	2,906,339	4,600,610
2	PSY - GANGNAM ...	Music	2,319,515,787	**5,400,599**	9,368,896	1,256,307
3	Super Junior 슈 ...	Music	90,887,184	**1,291,836**	458,334	24,546
4	Huge Fall Give ...	Howto & Style	1,320,158	**1,281,392**	33,294	539
5	DIY Backpack + ...	Howto & Style	5,131,373	**1,251,754**	423,133	2,537
6	Lady Gaga - Ba ...	Music	621,885,423	**1,143,346**	1,051,401	200,447
7	Miley Cyrus - ...	Music	764,902,467	**1,142,983**	2,730,753	1,178,257
8	PSY - GENTLEMA ...	Music	830,031,169	**1,134,567**	3,278,938	636,394
9	10 questions t ...	Howto & Style	6,737,750	**1,121,115**	41,749	21,827
10	Eminem - Not A ...	Music	615,871,729	**1,077,030**	2,545,875	71,389
11	One Direction ...	Music	632,693,879	**1,055,857**	2,595,385	280,957
12	Miley Cyrus - ...	Music	537,378,893	**1,016,324**	1,812,910	1,134,866
13	One Direction ...	Music	341,497,561	**891,870**	2,400,358	226,592
14	Ylvis - The Fo ...	Entertainment	510,573,787	**873,591**	3,668,785	422,446
15	Shakira - Waka ...	Music	844,000,601	**777,939**	1,451,466	105,065
16	Goku VS Superm ...	Shows	28,574,416	**722,221**	178,442	111,328
17	Friday - Rebec ...	Music	78,566,908	**712,093**	438,485	1,571,316
18	Miley Cyrus - ...	Music	452,226,563	**681,260**	1,111,244	147,843
19	Justin Bieber ...	Music	402,032,683	**652,185**	1,684,266	663,760
20	Charlie bit my ...	Comedy	818,368,377	**626,957**	1,387,191	219,016
21	Lady Gaga - Ap ...	Music	206,448,888	**620,897**	1,063,795	125,788
22	Super Junior_S ...	Music	42,978,638	**619,923**	356,488	16,220
23	Eminem - Love ...	Music	854,664,643	**614,649**	2,358,439	85,075
24	Carly Rae Jeps ...	Music	674,796,029	**612,942**	2,191,388	167,098
25	Nyan Cat [orig ...	Comedy	120,166,678	**604,871**	983,211	68,362
26	5 Seconds of S ...	Music	1,571,167	**598,733**	122,058	923
27	Katy Perry - F ...	Music	564,834,439	**598,409**	1,607,853	89,296
28	Justin Bieber ...	Music	475,860,799	**595,095**	1,974,463	525,221
29	GINGERS DO HAV ...	People & Blogs	37,875,859	**591,081**	232,584	94,791
30	Barack Obama v ...	Entertainment	106,384,302	**572,329**	755,752	20,699
31	LMFAO - Party ...	Music	861,694,735	**565,397**	2,688,535	133,673
32	Nicki Minaj - ...	Music	89,809,066	**564,946**	432,945	745,101
33	Steve Jobs vs ...	Film & Animation	94,705,123	**551,980**	648,697	12,526
34	One Direction ...	Music	276,312,550	**546,701**	1,620,029	98,564
35	Justin Bieber ...	Music	400,574,599	**545,261**	922,840	513,299
36	One Direction ...	Music	354,147,098	**540,831**	2,540,375	119,970
37	LEAVE BRITNEY ...	Entertainment	49,768,876	**535,962**	187,851	278,242
38	Rebecca Black ...	Music	39,857,438	**534,162**	437,773	701,157
39	LMFAO - Sexy a ...	Music	322,878,266	**531,535**	1,390,817	140,258
40	One Direction ...	Music	376,909,506	**530,712**	1,852,878	122,133
41	Minecraft Live ...	Gaming	616,451	**513,420**	83,529	1,593
42	Justin Bieber ...	Music	470,118,317	**513,238**	1,172,889	375,987
43	Back to School ...	Howto & Style	2,513,981	**511,836**	230,314	1,695
44	Girls' Generat ...	Music	127,653,645	**509,806**	653,908	89,278
45	KONY 2012	Nonprofits & Activism	100,406,983	**506,881**	1,395,152	200,567
46	Итоги Самого М ...	Gaming	1,043,600	**505,365**	72,987	4,814
47	Lady Gaga - Judas	Music	206,276,873	**501,070**	700,309	180,819
48	One Direction ...	Music	385,395,002	**500,327**	1,562,051	102,481
49	"Chocolate Rai ...	News & Politics	102,040,064	**484,329**	653,229	97,102
50	Huge Summer Gi ...	Howto & Style	1,903,513	**482,162**	265,757	1,094

Appendix C

Top 50 Videos on YouTube, Sorted by Number of Likes

#	Title	Category	Views	Comments	**Likes**	Dislikes
1	PSY - GANGNAM ...	Music	2,319,515,787	5,400,599	**9,368,896**	1,256,307
2	Ylvis - The Fo ...	Entertainment	510,573,787	873,591	**3,668,785**	422,446
3	Taylor Swift - ...	Music	810,375,030	352,753	**3,498,860**	240,502
4	PSY - GENTLEMA ...	Music	830,031,169	1,134,567	**3,278,938**	636,394
5	Katy Perry - R ...	Music	902,332,616	383,383	**3,180,803**	255,107
6	Meghan Trainor ...	Music	790,501,987	348,628	**3,170,558**	310,631
7	Mark Ronson - ...	Music	583,044,402	167,806	**3,076,163**	126,917
8	MACKLEMORE & R ...	Music	698,627,001	433,156	**3,054,337**	126,006
9	Taylor Swift - ...	Music	770,094,429	333,435	**3,050,395**	352,500
10	Justin Bieber ...	Music	1,161,556,092	6,774,237	**2,906,339**	4,600,610
11	Katy Perry - D ...	Music	931,129,321	388,137	**2,841,777**	458,998
12	Pharrell Willi ...	Music	642,972,019	138,166	**2,832,358**	134,182
13	Miley Cyrus - ...	Music	764,902,467	1,142,983	**2,730,753**	1,178,257
14	LMFAO - Party ...	Music	861,694,735	565,397	**2,688,535**	133,673
15	OneRepublic - ...	Music	771,094,470	135,938	**2,642,349**	82,921
16	Sia - Chandeli ...	Music	698,967,208	248,396	**2,599,138**	171,211
17	One Direction ...	Music	632,693,879	1,055,857	**2,595,385**	280,957
18	Eminem - Not A ...	Music	615,871,729	1,077,030	**2,545,875**	71,389
19	One Direction ...	Music	354,147,098	540,831	**2,540,375**	119,970
20	Gotye - Somebo ...	Music	634,150,407	454,857	**2,474,754**	116,119
21	One Direction ...	Music	341,497,561	891,870	**2,400,358**	226,592
22	Ed Sheeran - T ...	Music	451,867,734	116,914	**2,389,226**	63,744
23	Passenger - Le ...	Music	611,564,803	136,302	**2,381,654**	63,840
24	Eminem - Love ...	Music	854,664,643	614,649	**2,358,439**	85,075
25	Maroon 5 - Sugar	Music	379,959,097	95,298	**2,314,462**	67,869
26	John Legend - ...	Music	478,471,439	83,203	**2,267,147**	67,760
27	Adele - Rollin ...	Music	643,233,943	380,165	**2,249,312**	66,771
28	Enrique Iglesi ...	Music	847,003,236	101,620	**2,212,815**	147,290
29	Wiz Khalifa - ...	Music	171,175,119	112,411	**2,210,781**	25,792
30	Carly Rae Jeps ...	Music	674,796,029	612,942	**2,191,388**	167,098
31	Bruno Mars - T ...	Music	659,005,870	459,022	**2,143,534**	95,973
32	Martin Garrix ...	Music	545,883,369	119,534	**2,121,965**	83,764
33	Ariana Grande ...	Music	503,862,483	179,159	**2,107,995**	151,944
34	MAGIC! - Rude	Music	506,068,627	99,363	**2,040,146**	90,069
35	Avicii - Wake ...	Music	616,329,802	166,819	**1,988,523**	79,630
36	Justin Bieber ...	Music	475,860,799	595,095	**1,974,463**	525,221
37	Nicki Minaj - ...	Music	442,079,822	436,589	**1,874,187**	933,219
38	Jason Derulo - ...	Music	539,573,306	82,507	**1,869,821**	133,831
39	One Direction ...	Music	376,909,506	530,712	**1,852,878**	122,133
40	Adele - Someon ...	Music	475,454,673	236,755	**1,842,198**	55,562
41	Ellie Goulding ...	Music	331,724,483	68,483	**1,838,773**	60,593
42	Rihanna - Diamonds	Music	564,835,892	211,339	**1,828,110**	111,850
43	Miley Cyrus - ...	Music	537,378,893	1,016,324	**1,812,910**	1,134,866
44	Imagine Dragon ...	Music	333,060,859	203,112	**1,812,722**	54,874
45	Iggy Azalea - ...	Music	532,898,109	187,042	**1,791,749**	182,921
46	Demi Lovato - ...	Music	358,964,850	130,577	**1,782,209**	124,195
47	Ellie Goulding ...	Music	511,308,316	115,312	**1,765,798**	59,858
48	Shakira - Can' ...	Music	520,417,387	177,955	**1,690,102**	181,841
49	Shakira - La L ...	Music	502,530,160	155,023	**1,689,444**	144,450
50	One Direction ...	Music	153,759,309	203,771	**1,687,872**	72,574

Appendix D

Top 50 Videos on YouTube, Sorted by Number of Dislikes

#	Title	Category	Views	Comments	Likes	**Dislikes**
1	Justin Bieber ...	Music	1,161,556,092	6,774,237	2,906,339	**4,600,610**
2	Friday - Rebec ...	Music	78,566,908	712,093	438,485	**1,571,316**
3	PSY - GANGNAM ...	Music	2,319,515,787	5,400,599	9,368,896	**1,256,307**
4	Miley Cyrus - ...	Music	764,902,467	1,142,983	2,730,753	**1,178,257**
5	Miley Cyrus - ...	Music	537,378,893	1,016,324	1,812,910	**1,134,866**
6	Nicki Minaj - ...	Music	442,079,822	436,589	1,874,187	**933,219**
7	Strong	News & Politics	9,095,768	70	28,397	**810,563**
8	Nicki Minaj - ...	Music	89,809,066	564,946	432,945	**745,101**
9	Rebecca Black ...	Music	39,857,438	534,162	437,773	**701,157**
10	Justin Bieber ...	Music	402,032,683	652,185	1,684,266	**663,760**
11	PSY - GENTLEMA ...	Music	830,031,169	1,134,567	3,278,938	**636,394**
12	Justin Bieber ...	Music	475,860,799	595,095	1,974,463	**525,221**
13	Justin Bieber ...	Music	400,574,599	545,261	922,840	**513,299**
14	Katy Perry - D ...	Music	931,129,321	388,137	2,841,777	**458,998**
15	Ylvis - The Fo ...	Entertainment	510,573,787	873,591	3,668,785	**422,446**
16	PSY - HANGOVER ...	Music	197,145,550	162,918	1,098,234	**389,156**
17	Justin Bieber ...	Music	470,118,317	513,238	1,172,889	**375,987**
18	Kanye West - B ...	Music	48,157,702	102,508	192,769	**374,738**
19	Taylor Swift - ...	Music	770,094,429	333,435	3,050,395	**352,500**
20	PSY (ft. HYUNA ...	Music	564,387,286	405,199	1,397,990	**343,334**
21	Meghan Trainor ...	Music	790,501,987	348,628	3,170,558	**310,631**
22	Bieber After t ...	Comedy	43,970,603	197,245	134,684	**293,495**
23	Wheels On The ...	Education	554,665,206	3,536	369,264	**288,494**
24	Mike WiLL Made ...	Music	415,122,357	255,697	1,509,221	**283,077**
25	Miley Cyrus - ...	Music	99,374,784	150,618	660,643	**281,644**
26	One Direction ...	Music	632,693,879	1,055,857	2,595,385	**280,957**
27	LEAVE BRITNEY ...	Entertainment	49,768,876	535,962	187,851	**278,242**
28	Jennifer Lopez ...	Music	127,764,248	119,800	527,223	**275,274**
29	Avril Lavigne ...	Music	79,582,841	145,294	628,049	**272,896**
30	Katy Perry - R ...	Music	902,332,616	383,383	3,180,803	**255,107**
31	Justin Bieber ...	Music	301,452,516	348,171	784,840	**248,957**
32	Taylor Swift - ...	Music	810,375,030	352,753	3,498,860	**240,502**
33	PULCINO PIO - ...	Music	553,282,193	93,614	784,433	**238,778**
34	Money Boy - Dr ...	Entertainment	21,425,084	159,117	97,823	**235,836**
35	Маша и Медведь ...	Shows	565,631,590	7,964	355,251	**233,117**
36	Justin Bieber ...	Music	201,542,907	392,834	1,063,650	**231,685**
37	Best Sex Ever!!!	Entertainment	161,201,483	3,128	74,012	**231,040**
38	Justin Bieber ...	Music	138,174,322	183,009	1,059,070	**228,896**
39	Rihanna - Pour ...	Music	158,841,475	158,773	572,512	**227,496**
40	One Direction ...	Music	341,497,561	891,870	2,400,358	**226,592**
41	Justin Bieber ...	Music	246,171,879	406,469	1,377,599	**225,643**
42	We Are One (Ol ...	Music	370,585,960	147,235	1,219,612	**223,405**
43	Nicki Minaj - ...	Music	484,177,791	449,417	1,279,331	**219,171**
44	Charlie bit my ...	Comedy	818,368,377	626,957	1,387,191	**219,016**
45	30 Surprise Eg ...	Entertainment	472,982,464	21,645	247,680	**209,991**
46	Neutral Response	Comedy	3,896,276	24,191	207,516	**207,516**
47	Nicki Minaj - ...	Music	129,933,855	187,037	471,305	**207,233**
48	Cher Lloyd - S ...	Music	79,883,653	235,294	527,295	**206,294**
49	~YouTube Worst ...	Comedy	97,000,506	41,344	25,651	**203,473**
50	#SELFIE (Offic ...	Music	324,426,647	136,376	1,485,050	**201,507**

Appendix E

Top 100 Most Positive and Negative Coefficient Values for Video Title Word Ngrams

Negative Coefficients		Positive Coefficients	
Value	Ngram	Value	Ngram
-0.388	sold	0.280	hq
-0.327	cover	0.209	sex
-0.317	full movie	0.207	mv
-0.233	mep	0.205	hd
-0.175	tag	0.190	flv
-0.160	2015	0.183	full movies
-0.155	equi	0.177	baby
-0.153	open	0.161	rc
-0.149	amv	0.160	ever
-0.143	movie	0.148	lyrics
-0.142	mw3	0.147	kid
-0.133	lps	0.147	real
-0.131	vlog	0.145	cat
-0.125	hangout	0.142	3d
-0.113	dr	0.130	ti
-0.112	panel	0.129	ad
-0.112	fl	0.129	2pac
-0.111	phd	0.124	say
-0.110	st	0.123	you
-0.109	eu	0.122	000
-0.109	season episode	0.119	te
-0.106	closed	0.117	top
-0.105	map	0.117	mi
-0.102	cod	0.117	lego
-0.100	haul	0.115	best
-0.097	ii	0.115	los
-0.097	ep1	0.113	man
-0.095	05	0.112	gun
-0.095	web	0.109	scene
-0.094	ep	0.107	ufo
-0.094	領袖成長共生家園	0.106	dvd
-0.091	oc	0.106	nba
-0.090	pvp	0.105	jet
-0.090	chat	0.105	cop
-0.089	box	0.105	lion

-0.088	part	0.104	tu
-0.088	cc	0.103	car
-0.086	HD	0.102	boy
-0.086	ca	0.101	je
-0.086	pt	0.101	si
-0.086	15	0.101	911
-0.085	tx	0.101	ktv
-0.085	audio	0.099	gay
-0.085	md	0.099	vs
-0.085	pr	0.098	bbc
-0.085	dj	0.096	die
-0.084	blog	0.095	girl
-0.084	roblox	0.094	tsunami
-0.084	bag	0.094	500
-0.084	prod	0.093	u2
-0.083	testimonial	0.092	720p
-0.083	ict	0.091	dad
-0.082	vid	0.091	100
-0.082	uni	0.091	funny
-0.078	ieee	0.091	guy
-0.078	talk	0.089	el
-0.078	ny	0.089	ali
-0.078	jam	0.088	prank
-0.077	radio	0.088	www
-0.077	mba	0.088	robot
-0.076	og	0.088	in the world
-0.076	uc	0.086	90
-0.076	ace	0.085	80
-0.076	ph	0.085	300
-0.075	may 2015	0.084	2001
-0.075	wmv	0.083	obama
-0.074	iii	0.083	min
-0.074	by	0.082	ipad
-0.073	tax	0.082	ellen
-0.073	law	0.082	navy
-0.073	ask	0.082	les
-0.072	piano	0.082	bmw
-0.071	remix	0.081	original
-0.070	art	0.081	giant
-0.070	lab	0.081	101

-0.069	sxsw	0.080	se
-0.069	lp	0.080	very
-0.069	hack	0.080	haka
-0.068	original mix	0.080	engine
-0.068	mn	0.080	adele
-0.067	eco	0.079	iphone
-0.067	ceo	0.079	1980
-0.067	tuto	0.079	10
-0.066	小邑	0.079	bow
-0.065	play	0.079	birth
-0.065	live	0.078	how
-0.064	instrumental	0.077	747
-0.064	sale	0.077	eminem
-0.064	original song	0.077	akon
-0.063	prof	0.076	niño
-0.063	arts	0.076	cómo
-0.063	edit	0.076	2015 full movie
-0.063	img	0.076	trailer hd
-0.062	wcdrr	0.076	sexy
-0.062	sub	0.076	asi
-0.062	let	0.075	las
-0.062	intro	0.075	banned
-0.062	zone	0.075	full movie english
-0.062	hope	0.075	scary
-0.061	rb3	0.075	ne

Appendix F

Top 100 Most Positive and Negative Coefficient Values for Channel Name Word Ngrams

Negative Coefficients		Positive Coefficients	
Value	Ngram	Value	Ngram
-0.246	music	0.315	ted
-0.219	bers	0.232	ign
-0.210	vmax	0.205	smosh
-0.163	227pop	0.182	vice
-0.162	sites	0.180	wwe
-0.159	top sites	0.168	rt
-0.137	vevo	0.164	google
-0.136	akel	0.131	nba
-0.136	akel edin	0.127	tmz
-0.133	edin	0.125	ksi
-0.126	qello	0.119	hulu
-0.116	latin45	0.107	我愛康熙
-0.115	qello movies	0.105	dnews
-0.115	khaotic1	0.103	vsauce
-0.114	yin	0.101	bfvsgf
-0.112	mlb	0.098	make
-0.112	julian correa	0.096	letsplay
-0.111	三万発	0.092	enchufetv
-0.109	吳老師教學部落格	0.091	shane
-0.102	lineup	0.085	scishow
-0.102	сми	0.082	w2s
-0.098	k45mm	0.082	kipkay
-0.094	謝光忠	0.082	pewdiepie
-0.092	mc	0.081	smtown
-0.091	allure	0.081	石涛tv
-0.091	david fry	0.080	gopro
-0.089	yowa19	0.080	espn
-0.088	spirithyper	0.078	van
-0.088	tedley	0.077	apple
-0.087	fry	0.077	and
-0.086	datinglogic	0.076	nameless
-0.086	entertainment	0.075	nigahiga
-0.084	uwtv	0.075	kids
-0.083	kcscrag	0.074	lol
-0.080	ndtv	0.074	faze

-0.080	david carter	0.072	zeetv
-0.080	tehmonitor	0.072	asapscience
-0.079	talk	0.071	ea
-0.079	bu	0.071	oscars
-0.078	24	0.070	collegehumor
-0.078	menchfer	0.070	kids channel
-0.077	montagedrift	0.070	ultra music
-0.076	kaidamma	0.069	vat19
-0.076	aj	0.069	abc news
-0.074	freedom club	0.069	markiplier
-0.073	veronica senegal	0.068	canal
-0.073	denjp	0.068	yapyap
-0.072	yell	0.067	react
-0.071	ak	0.067	channel
-0.071	osu	0.067	pocoyo
-0.070	mfm	0.067	hot
-0.070	misenda	0.066	sam
-0.070	mich	0.065	disney
-0.070	cool	0.065	grey
-0.069	konstantin	0.065	kidstv123
-0.069	masterd	0.064	simon
-0.068	袁腾飞历史课	0.064	gaming channel
-0.068	rymi64	0.064	谷阿莫
-0.068	mike96881	0.064	emimusic
-0.068	glory patriot	0.064	yuya
-0.068	g00dfriends	0.064	jan
-0.067	mu osu	0.063	car
-0.066	yin yin	0.062	ali
-0.066	the new yorker	0.062	tested
-0.066	plays	0.062	team coco
-0.066	1nbz	0.062	phan
-0.066	case	0.061	michelle phan
-0.066	alpha news	0.061	top trending
-0.066	cristian	0.061	veritasium
-0.065	will	0.061	popularmmos
-0.065	senegal	0.061	cgp
-0.065	hamikka	0.061	cgp grey
-0.065	chun0523	0.060	space
-0.064	ace	0.060	digitalrev
-0.064	unisdr	0.060	digitalrev tv

-0.064	liveleak	0.060	awe
-0.064	new yorker	0.060	joe
-0.064	yorker	0.060	rob
-0.064	venad	0.059	pbs
-0.064	venad news	0.059	talking
-0.064	cris	0.059	gameplayrj
-0.064	douglitas	0.059	swoozie
-0.063	digi	0.058	bethany mota
-0.063	patriot	0.058	tomska
-0.063	details	0.058	en
-0.062	ak lectures	0.058	cbs news
-0.062	appleburner	0.057	vanossgaming
-0.062	style com	0.057	norman
-0.061	josh galt	0.057	christian
-0.061	573luis	0.057	elrubiusomg
-0.061	mage	0.057	mota
-0.061	galt	0.057	cnn
-0.060	karaokeonvevo	0.056	ed
-0.060	zee tv	0.056	lin
-0.060	jr	0.056	mbc
-0.060	serve entertainment	0.055	danger dolan
-0.060	konstantin bass	0.055	izuniy
-0.060	gaming mage	0.055	vitalyzdtv
-0.060	musical gaming	0.055	tim
-0.060	musical gaming mage	0.054	dc

Appendix G

Top 100 Most Positive and Negative Coefficient Values for Video Title Character Ngrams

Negative Coefficients		Positive Coefficients	
Value	Ngram	Value	Ngram
-0.357	- sold	0.162	movie
-0.356	- sold	0.139	lyrics
-0.240	sold	0.139	yrics
-0.222	cover	0.130	(198
-0.189	- sol	0.125	ovie
-0.186	- sol	0.123	iest
-0.173	full m	0.121	movie
-0.149	mep	0.117	prank
-0.125	equi:	0.117	9/11
-0.124	equi:	0.117	(hd)
-0.123	qui:	0.111	ft.
-0.122	hangou	0.110	- 19
-0.121	angout	0.109	kiss
-0.121	hangout	0.109	beyonc
-0.119	hangou	0.109	eyonc
-0.118	ngout	0.107	real
-0.116	angou	0.105	anima
-0.112	ull mov	0.104	(197
-0.109	ll movi	0.103	путин
-0.106	movie	0.102	an -
-0.105	hango	0.101	www.
-0.100	ll mov	0.100	movie)
-0.099	hang	0.100	ovie)
-0.097		0.100	pocoy
-0.097	audio	0.099	top 5
-0.096	full mo	0.098	super
-0.095	losed	0.098	song
-0.094	mod-0	0.098	best
-0.094	共生家園	0.096	كامل
-0.094	成長共生家	0.095	interv
-0.094	成長共生家園	0.094	ele -
-0.094	成長共生家園	0.094	ele -
-0.094	袖成長共生	0.094	soldi
-0.094	袖成長共生家	0.094	movie)
-0.094	袖成長共生家園	0.093	ever

-0.094	長共生家園	0.092	beyon
-0.094	長共生家園	0.091	ni -
-0.094	領袖成長共	0.090	soldie
-0.094	領袖成長共生	0.090	cial)
-0.094	領袖成長共生家	0.089	(196
-0.094	ull mo	0.088	llor
-0.094	promo	0.087	teen
-0.093	closed	0.087	gest
-0.093	\\\\\	0.087	baby
-0.092	hango	0.086	amor
-0.091	roblo	0.086	adele
-0.090	lec-	0.086	(hq)
-0.089	mix)	0.085	prank
-0.089	blog	0.085	(full m
-0.087	oblox	0.085	traile
-0.087	roblox	0.085	1080p
-0.087	on 2015	0.084	ga -
-0.086	aking o	0.083	adele
-0.086	(audio)	0.083	crash
-0.082	pvp	0.083	llora
-0.082	top	0.083	lion
-0.081	(cove	0.083	niño
-0.081	(cover	0.083	sunam
-0.080	vlog	0.083	sunami
-0.079	andom	0.083	unami
-0.079	sale -	0.082	9/11
-0.078	, - s	0.082	y guy
-0.078	binar	0.082	bebe
-0.078	equi:	0.082	lyric
-0.078	hack	0.081	haka
-0.078	- 11	0.081	twin
-0.078	lyrics	0.081	tedx
-0.078	yrics	0.081	funny
-0.078	sale	0.081	prost
-0.078	fest	0.081	2pac
-0.077	equi:	0.080	x fact
-0.077	sale -	0.080	giant
-0.077	sale -	0.080	motor
-0.077	, - so	0.080	- go
-0.077	, - sol	0.080	navy

-0.077	hines	0.079	obama
-0.076	.v.i.	0.079	(video)
-0.076	ar ac	0.079	plane
-0.075	haul	0.079	x facto
-0.075	ú equi:	0.079	15042
-0.074	mark	0.079	tsunam
-0.074	recto	0.079	tsunami
-0.074	.o.v.i	0.078	el -
-0.074	.o.v.i.	0.078	2015/
-0.074	.v.i.e	0.078	ya -
-0.074	m.o.v.i	0.078	x fac
-0.074	o.v.i	0.077	basic
-0.074	o.v.i.	0.077	domino
-0.074	o.v.i.e	0.077	niña
-0.074	v.i.e	0.077	emine
-0.074	va -	0.077	ly guy
-0.073	m.o.v	0.076	e (19
-0.073	m.o.v.	0.076	y guy -
-0.073	forum	0.075	li -
-0.073	sale	0.075	scary
-0.073	humb	0.075	tsuna
-0.073	intro	0.075	in 1
-0.072	rics	0.075	ii -
-0.072	haos	0.075	zoella
-0.072	(cover)	0.075	danc

Appendix H

Top 100 Most Positive and Negative Coefficient Values for Channel Name Character Ngrams

Negative Coefficients		Positive Coefficients	
Value	Ngram	Value	Ngram
-0.208	vma	0.357	vevo
-0.206	vmax	0.316	vev
-0.203	gov	0.246	-
-0.176	sites	0.206	evo
-0.166	27p	0.201	smosh
-0.164	uni	0.175	ksi
-0.163	227p	0.159	wwe
-0.163	227po	0.152	&
-0.163	227pop	0.150	mosh
-0.163	27po	0.147	
-0.163	27pop	0.144	ign
-0.163	7pop	0.143	smos
-0.162	sites	0.136	vsau
-0.161	7po	0.136	vsauc
-0.159	op sit	0.136	vsauce
-0.159	op site	0.135	/
-0.159	p sites	0.133	smo
-0.159	top sit	0.130	kip
-0.158	op si	0.129	vsa
-0.158	top si	0.127	kid
-0.158	bers	0.124	tmz
-0.157	p si	0.122	flo
-0.157	p sit	0.117	zie
-0.157	p site	0.117	007
-0.154	227	0.116	music
-0.154	p s	0.113	yap
-0.151	ites	0.111	hik
-0.145	site	0.110	urr
-0.144	pop	0.109	ump
-0.144	top s	0.107	愛康熙
-0.137	el edi	0.107	我愛康
-0.136	akel e	0.107	我愛康熙
-0.136	akel ed	0.105	nha
-0.136	el edin	0.104	huf

-0.136	kel ed	0.103	fvs
-0.136	kel edi	0.103	vsg
-0.136	l edin	0.102	mim
-0.134	qel	0.102	auce
-0.132	op s	0.102	sauce
-0.132	kel e	0.102	hulu
-0.129	el ed	0.101	jer
-0.126	l edi	0.101	eki
-0.126	qell	0.101	bfvs
-0.126	qello	0.101	bfvsg
-0.126	qello	0.101	bfvsgf
-0.126	akel	0.101	fvsg
-0.122	tail	0.101	fvsgf
-0.121	hao	0.101	vsgf
-0.120	yos	0.101	sgf
-0.119	kel	0.101	stf
-0.119	vod	0.100	bfv
-0.119	edin	0.100	kak
-0.118	akel	0.100	yta
-0.118	sit	0.100	ip
-0.116	ic1	0.098	sauc
-0.116	tic1	0.098	nba
-0.116	atin4	0.098	nfl
-0.116	atin45	0.097	tit
-0.116	latin4	0.097	iepie
-0.116	latin45	0.096	a
-0.116	tin45	0.096	isn
-0.115	ello mo	0.095	music
-0.115	llo mov	0.095	cau
-0.115	qello m	0.095	nee
-0.115	aotic1	0.095	dnew
-0.115	haotic1	0.095	usic
-0.115	otic1	0.094	ius
-0.114	khaot	0.093	198
-0.114	khaoti	0.093	chufe
-0.114	khaotic	0.093	enchuf
-0.113	eup	0.093	enchufe
-0.113	in45	0.093	nchuf
-0.112	em	0.093	nchufe
-0.112	tin4	0.093	pka

-0.112	edi	0.093	auc
-0.112	llo mo	0.093	iepi
-0.111	三万発	0.093	hufe
-0.111	lps	0.092	mot
-0.111	ulian c	0.092	nchu
-0.111	lian co	0.092	dnews
-0.110	khao	0.092	chufet
-0.110	ello m	0.092	chufetv
-0.110	100	0.092	hufet
-0.109	吳老師	0.092	hufetv
-0.109	吳老師教	0.092	nchufet
-0.109	吳老師教學	0.092	ufetv
-0.109	吳老師教學部	0.092	sic
-0.109	吳老師教學部落	0.091	ztv
-0.109	學部落	0.091	asap
-0.109	學部落格	0.091	ufet
-0.109	師教學	0.091	by
-0.109	師教學部	0.090	pi
-0.109	師教學部落	0.090	ted
-0.109	師教學部落格	0.090	fail
-0.109	教學部	0.089	ake:
-0.109	教學部落	0.089	ke:
-0.109	教學部落格	0.089	make:
-0.109	老師教	0.089	ndy
-0.109	老師教學	0.089	iha
-0.109	老師教學部	0.089	antu