Analysis of Buzzfeed YouTube Video Views

Amelia Chu (ac4119), Cora Hyun Jung (hj1399), and Dee Hahm (drh382)

Team Ramen

New York University

Business Understanding

BuzzFeed, a leader in digital content production, has been heralded as the "most influential news organization in America today" (La France & Meyer, 2015). The company is an expert at viral content, popularizing online quizzes, "listicles", and short-form video. Its content ranges from silly to serious, from "What Famous Internet Cat Are You?" to "Poison in the System", a Pulitzer Prize finalist news investigation (Burton, 2012; Wilson, Banerjee, Hernandez, Abramson, & Bronner, 2018). Therefore it was quite unexpected when, earlier this year, BuzzFeed appeared in headlines for not innovation and expansion, but a 15% layoff of their workforce (Lee, 2019).

In an ever-competitive market, digital content producers, such as BuzzFeed, need to continuously increase their profit margins to maintain relevance in the digital media space. They need to not only cut costs, but also find ways of increasing revenue. For BuzzFeed, like many other media companies, revenue means advertising, whether it's native (on BuzzFeed.com, etc.), creating sponsored content, or using ad services on BuzzFeed's myriad of social platforms (Robischon, 2016).

One of these social platforms is YouTube, where BuzzFeed hosts the majority of their video content. On YouTube, revenue can come in the form of pre-/mid-/post-roll advertisements, channel memberships, or sponsored content (YouTube, 2019). This means generated revenue is largely tied to video consumption or views obtained. Therefore, identifying levers that can increase viewership will, in turn, increase revenue.

A data mining solution can assist content creators in quickly identifying levers to adjust to maximize viewership for their content. It can help creators answer: Will there be an increase in views if there are specific people, places, or things in the video?

Data Understanding

A portion of BuzzFeed viewer behavior can be obtained with the public YouTube Data API. To control for varying subscriber counts, only the BuzzFeed main YouTube channel (BuzzFeedVideo) will be used for this project. This channel was chosen because the main channel hosts a variety of content, as opposed to specific

topics, e.g. BuzzFeed Unsolved (Supernatural & True Crime Mysteries), BuzzFeed Celeb (Celebrity), BuzzFeed Multiplayer (Gaming). Additionally, new series or properties are typically on the main channel first, then with a proven track record are spun out (e.g. LadyLike, Tasty). These niche channels cater to specific audiences so findings may not be as translatable across properties (opposed to the main channel).

The video resource from the YouTube Data API, provides information on a specific video at the time of the API call. For each video, the resource includes viewership and engagement metrics (e.g. number of video views) as well as video metadata (e.g. video title, publish time, duration, tags). The data was retrieved in six batches due to API call limits on October 27, 2019. At the time, there were a total of 6,148 videos on the BuzzFeedVideo channel (Table 1). From these videos, 42,129 unique tags were extracted. The most frequent tags associated the video as BuzzFeed property, e.g. 'buzzfeed' (n = 5,129), 'buzzfeedvideo' (n = 4,902). The top non-BuzzFeed property tag was 'funny' (n = 1,726).

Table 1. *Summary of Key Values in Retrieved Data*

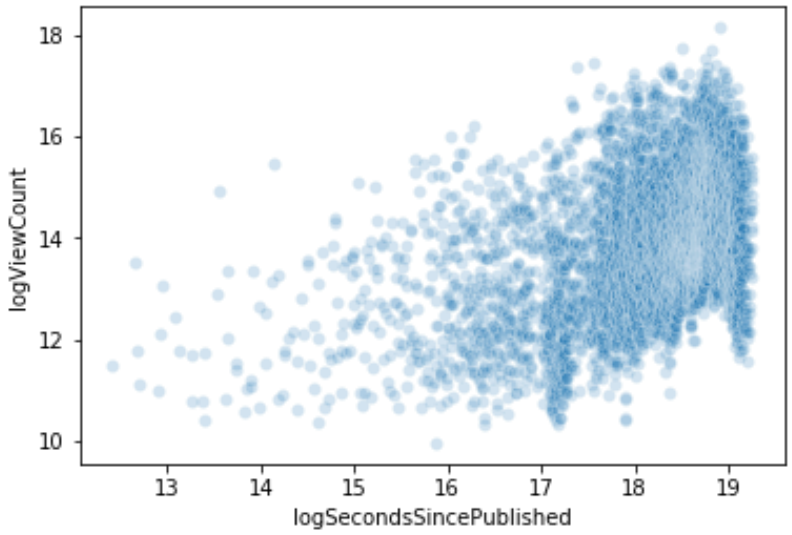
Total Number of Videos	6,418 (18 null rows omitted)
Publish Date Range	2012-07-03 to 2019-10-26
Most Common Publish Day of Week	Saturday (n = 1,110)
Average Video Duration	250 seconds (4.16 minutes)
Total Number of Unique Tags	42,129
Most Frequently Used Tag	'buzzfeed' (n = 5,129), 'funny' (n = 1,726)

Table 2. *Distribution of View Count, Title Length, and Video Duration*

	N	*M*	*SD*	*Min*	*25%*	*50%*	*75%*	*Max*
viewCount	6,418	2,497,455	3,987,847	21,124	497,564.25	1,069,100	2,828,219	75,660,174
titleLength	6,418	40.48348	13.43938	8	32	39	47	100
durationInSeconds	6,418	249.9092	353.4934	24	111	173	296	11990

Note that because data was retrieved from a single point in time, the total number of viewers is correlated to the age of the video ($r = 0.178$, $p < 0.0001$). The log was taken for total number of viewers (viewCount) and age of the video (secondsSincePublished) as both variables were skewed (Figure 3).

Figure 3. *Relationship Between log(viewCount) and log(secondsSincePublished)*



Data Preparation

In order to identify the videos that belong to the BuzzFeedVideo channel, the playlist resource on the YouTube Data API was accessed. A list of all videosIds (i.e., the unique identifier for videos) was retrieved from the 'Uploads' playlist, a default playlist containing all videos for a channel.

To perform data analysis, the data was converted from the nested JSON format provided by the YouTube Data API to a tabular format with each row represented as a single video. Exploratory variables such as length of video title, and day of week published were extracted from existing variables via the API (i.e., videoTitle, publishDate). Length of video title was extracted as previous research suggests that this factor contributes to the success of a video (Briggs, 2018). Initial analysis appeared to reinforce this research (Figure 4a); videos with title lengths above 65-70 characters seem to fare less well compared its shorter titled counterparts. However, this trend disappears when broken out by the target variable (Figure 4b).

Figure 4a. *Total Number of Views vs. Title Length*

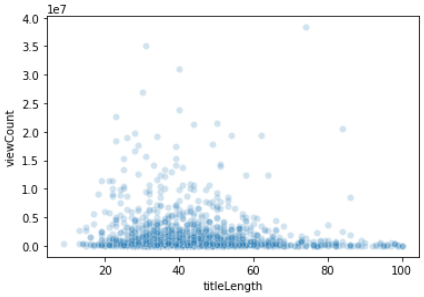
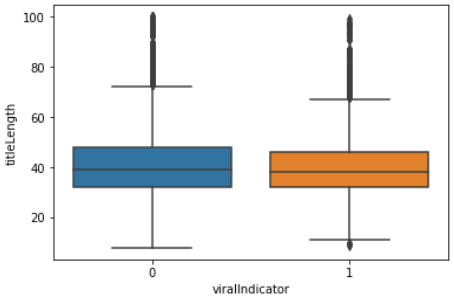


Figure 4b. *Distribution of Title Length by Viral Indicator*



Target Variable: Video View "Viral Indicator"

A straightforward way to determine whether a video is successful or has hit the critical minimum number of views is if the video has gone 'viral'. Today, this is qualified by whether the video recieves "more than 5 million views in a 3-7 day period" (Nalty, 2011). However, we are unable to retrieve historical log-level data. Therefore, we used the historical 'viral video' definition: hitting more than 1 million views (Nalty, 2011).

The specific data point used is statistics.viewCount from the YouTube Data API. This video count has been algorithmically validated by YouTube to remove fraudulent views such as views by bots (YouTube, 2018). The data point is then transformed into the binary variable "ViralIndicator", i.e., 1 for 1 million views or above (n = 3,347) , 0 for less than 1 million views (n = 3,071).

Tag Preparation and Topic Extraction

The tags retrieved from the YouTube Data API required additional processing due to the format and volume (N = 42,129) of the data. There were tags that (1) added no information (e.g. 'video', 'bfvideo', 'buzzfeed'), (2) tags which were essentially duplicates (e.g. 'buzz feed', 'buzzfeed'), (3) tags that were conceptually similar (e.g. 'friends', 'friendships'), and (4) single tags that represent significantly distinct or separate concepts (e.g., 'apple', which was used to represent both the fruit and the brand). To address issues 1 and 4, tags fell into those categories were removed (See Appendix A for list). For issue (2), for each row, all tags were lower-cased, spaces were removed, then the *set*() function was applied to remove duplicates. Finally, to address issue 3, domain knowledge feature extraction was applied to group conceptually similar together.

Due to the high volume of tags, and presumed low return on investment for tags with low feature importance, a decision tree was applied across all tags so only tags with *Feature Importance* != 0 (n = 1,358) had domain knowledge feature extraction applied. The final features were limited to the top 100 tags/topics (See Appendix B for full list).

Modeling & Evaluation

As the target variable is binary, and the features are categorical, classification models will be used for this project. A logistic regression (LR) model was used to set the baseline, as this is a simple, but robust model that could be quickly iterated upon (Chauhan, 2019). The primary evaluation metric will be the AUC score metric as the final model should have a high discrimination capacity to distinguish between videos that will be successful ('viral') or not.

Baseline Model: Logistic Regression

Since the retrieved BuzzFeed video dataset is not very large (n = 6,418) and, after feature engineering, the tags and topics are not expected to be correlated, the LBFGS Algorithm seemed the most appropriate solver for the baseline model (scikit-learn developers, 2019). Setting the solver = 'lbfgs' and using defaults for all other parameters, yielded a test accuracy of 0.646 (AUC = 0.696). With a training accuracy of 0.657, the model appears to only slightly overfit the training data (Table 5).

Table 5. *LR Model Evaluation Metrics by Different Solvers and C*

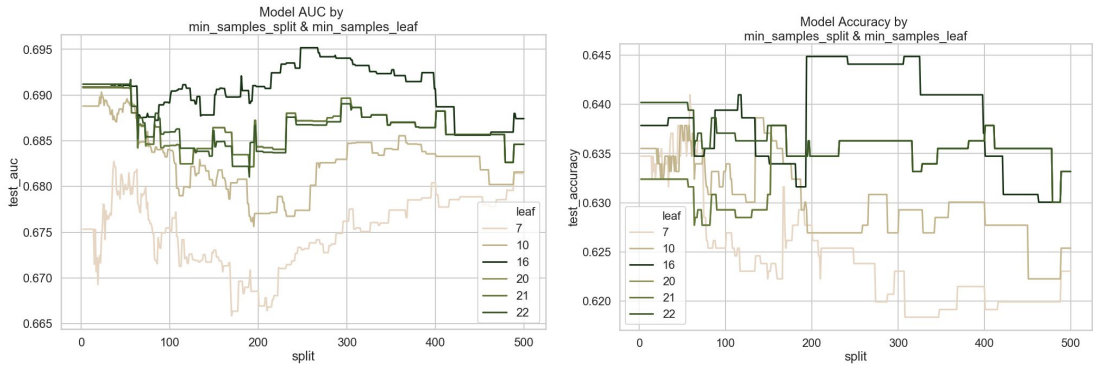
	C = 1				C = 1e30			
Solver	*Accuracy (Training)*	*Accuracy (Test)*	*AUC*	*r2*	*Accuracy (Training)*	*Accuracy (Test)*	*AUC*	*r2*
lbfgs	0.656993	0.646417	0.695641	0.124736	0.657382	0.645639	0.694214	0.123908
newton-cg	0.656993	0.646417	0.695636	0.124735	0.657382	0.645639	0.694209	0.123904
liblinear	0.656993	0.646417	0.695626	0.124708	0.657382	0.645639	0.694190	0.123906
sag	0.656993	0.646417	0.695648	0.124734	0.657382	0.645639	0.694212	0.123903
saga	0.656993	0.646417	0.695638	0.124735	0.657382	0.645639	0.694207	0.123905

All other available solvers for the logistic regression model were attempted and a large C value was included. This is because larger C values ought to lessen regularization (Garcia, 2016). However, when the models were generated, the larger C value appear to have little impact on the results. In fact, $C = 1e30$ overfit the data slightly more that $C = 1$. This could be due to the fact that create complex models (Garcia, 2016). Based on the AUC, specifying parameters $C = 1$, *solver* = 'sag', would be the best logistic regression model.

Decision Tree

The decision tree (DT) classifier model was also considered as it is a useful tool for testing feature interactions as well as ranking features. When the different models were fitted to the data, *min_samples_leaf* = 16 performed best against the evaluation metrics (Figure 6). Multiple minimums splits performed equally well, so the largest split value was taken (*min_samples_split* = 267), amongst those with the highest AUC (*AUC* = 0.69518), to reduce the amount of overfitting (Sief, 2018).

Figure 6. *DT Model AUC (a) and Test Set Accuracy (b) by Different Minimum Split and Leaves*



The amount of features retained in the model was also considered. It is important for the end users of this particular project have access to a wide variety of features (i.e. tags and topics) and their respective importance. The goal is to inform content creators on potential topics, people, places, or things that will result in a 'viral' video, so if there is an extremely performant model, that only look at one feature (e.g. if a video was funny or not) it would be unhelpful.

Initially, as *min_samples_split* and *min_samples_leaf* grew larger, the number of features with importance also greatly decrease (Table 7). Inherently, however, selecting a model with a large amount of

features, makes the model susceptible to overfitting (Elite Data Science, 2019). For example, even though setting *min_samples_split* = 2 and *min_samples_leaf* = 1 considers all features, the train and test accuracy demonstrate that it's greatly overfitted. Thus, the parameters ultimately select would need to balance retaining the most features, and having the least amount of overfitting.

Table 7. *Number of Remaining Features with Importance by Model Parameters*

min_samples_split	*min_samples_leaf*	*Accuracy (Train)*	*Accuracy (Test)*	*Features with Importance*
2	1	0.863264	0.609813	100
10	10	0.67413	0.635514	70
15	50	0.627970	0.630062	27
25	30	0.638098	0.622274	36
267	16	0.648811	0.64408	50

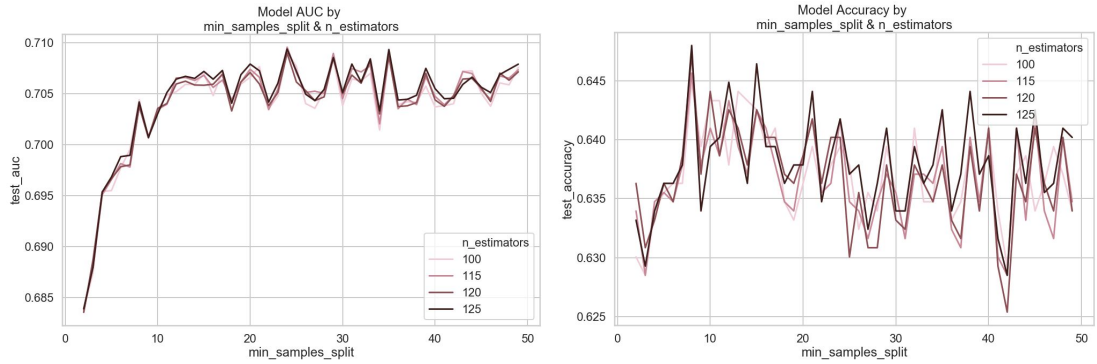
As another check for overfitting, different values for *max_depth* were considered. With parameters *min_samples_split* = 267 and *min_samples_leaf* = 16, the initial value for *max_depth* was 38. So, all values between 1 and 38 were tested to see if it reduced overfitting. The test and training accuracy remained the same through all values of *max_depth (train accuracy = 0.649; test accuracy = 0.644)*. Therefore, a *max_depth* of 20 was selected rather than the minimum or maximum *max_depth* 'possible', as a value too low may reduce accuracy and a value too high may introduce overfitting for future videos (Seif, 2018). Based on the AUC and other considerations, specifying parameters *min_samples_split* = 267 and *min_samples_leaf* = 16, *max_depth* = 20, would be the best decision tree model with *criterion* = 'entropy'.

Random Forest Classifier

Random Forest classification (RFC) was then attempted as it was more robust to overfitting than decision trees (Liberman, 2017). Initially, different values were specified for parameters *n_estimators* and *min_samples_split* . The random forest with the highest AUC and test accuracy had parameters *n_estimators* =

115 and *min_samples_split* = 24 (Figure 8). However, this model appeared to have an overfitting issue (*train accuracy* = 0.797; *test accuracy* = 0.6417).

Figure 8. Initial *RFC Model AUC (a) and Test Set Accuracy (b) by Different Minimum Split and Estimators*



To address the overfitting issue, different values for both *max_depth* and *min_samples_leaf* were tested. When both parameters were added to the model together, the highest possible *AUC* decreased significantly from the initial model. Thus, the parameters were added back in separately. The initial value for *max_depth* was 96, so all values between 1 to 96 were tested. Specifying, *max_depth* increase both *AUC* and *test accuracy.* However, the model was still overfitted to the training set (Table 9). With *min_samples_leaf* = 20, the overfitting issue was greatly reduced and retained, generally, the *AUC* and *test accuracy* from the initial model. Thus, specifying parameters *n_estimators* = 115, *min_samples_split* = 24, and min_samples_leaf = 20, would be the best random forest classifier.

Table 9. *Comparing Initial Random Forest Classifier to New Variations with Highest AUC*

Model	*Accuracy (Training)*	*Accuracy (Test)*	*AUC*	*r2*
Initial Model *n_estimators* = 115 *min_samples_split* = 24	0.797428	0.641744	0.709561	0.132551
Include Depth & Leaves *n_estimators* = 115 *min_samples_split* = 49 max_depth = 25 min_samples_leaf = 25	0.645111	0.641745	0.699146	0.099417
Include Max Depth Only *n_estimators* = 115 *min_samples_split* = 24	0.718543	0.652647	0.714082	0.136114

max_depth = 20				
Include Leaves Only *n_estimators* = 115 *min_samples_split* = 24 min_samples_leaf = 20	0.648811	0.6417445	0.701952	0.105117

Support Vector Machine

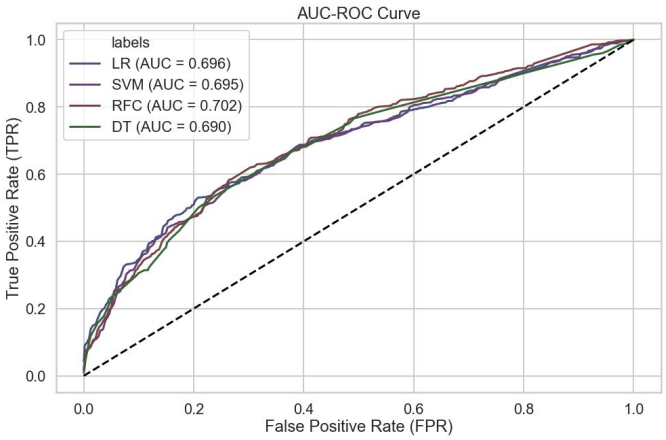
For Support Vector Machine (SVM), three different types of kernels (Linear, Gaussian, Polynomial) and 100 different parameters c were tested, and the one that yielded the highest AUC score was applied for the final modeling (C = 0.9, kernel= "linear"). The parameter C is a trade-off between bias and variance. Consider the equation: $\min_w \text{loss}(w; X, y) + \lambda \, \text{regularization}(w)$ where C is the inverse of λ. It is intuitive that as λ converges to ∞, the loss function will be almost entirely dependent on λ, which leads to high bias and low variance, and vice versa (Cawley, 2007). With the selected parameters applied, the model yielded a test accuracy of 0.647 with AUC of 0.695. The selected kernel is a default kernel("linear") of other kernels tested, which indicates our data points are linearly separable.

It may be worthwhile to compare the results of SVM and LR as both models yielded very similar results (Table 10). Since the optimization problem we have uses linear SVM and regularized LR, the loss function of those two models should be very similar, with SVM minimizing hinge loss function while LR minimizing log loss function (Drakos, 2018). This is why AUC scores of SVM(0.695) and LR(0.694) of our model show very similar results. Since log loss function diverges faster than hinge loss function, it is more sensitive to outliers. Also, log loss doesn't exactly converge to zero even if the data point is classified with enough confidence, which leads to minor degradation in accuracy (Hastie, 2017). This explains the small difference in accuracy between two models (Table 11). Practically, LR works better when classes are closely separated, while SVM works well when classes are well separated (Hastie, 2017).

Table 11. *Evaluation Metrics for Models*

	AUC	*Test Accuracy*
Logistic Regression	0.694	0.646
Decision Tree	0.695	0.644
Random Forest Classifier	0.709	0.642
Support Vector Machine (SVM)	0.695	0.647

Figure 12. *ROC Curve for Models*



Overall, based on AUC scores from four different models tested, Random Forest Classifier performs the best, while Support Vector Machine has the highest accuracy of all the models. However, models perform very similarly with AUC scores and accuracy scores of all models being very close within 0.01 difference.

Deployment

This model can be deployed below a user interface where BuzzFeed content creators can input (existing) potential topics for future content and see if producing that content would be likely to be 'viral'. Conversely, if the content producers need video content ideas, they could access the interface to see a list of topics most likely to be 'viral' using feature importance.

Based on the nature of digital content, the model would need to be updated periodically, perhaps on a weekly basis, to ensure relevance of topics/tags uses. The model can be evaluated with the same metrics used in this project (*AUC*, *Test Accuracy*) when it's updated.

Awareness and buy-into this project must occur for the model to be successful. It is important to note that the tagging in the data is not always consistent. There are some videos that contain topics, but do not have the appropriate tags. One such instance is the video "People Who Don’t Cook Make Brunch For Their Significant Others", which is about 'food' and 'breakfast', but only has the 'buzzfeedvideo', 'buzzfeed' tags. The

creators who are generating tags must be aware of the importance of having consistent tags and how generating an accurate model to help inform content creation decision can be beneficial to creators.

This is a particularly salient obstacle as many content creators fear that working for large corporations like BuzzFeed stifles creativity (Safi, 2019). In fact, many successful ex-BuzzFeed creators cite this as a primary factor of leaving the company (Nygaard, 2017; Dunn & Raskin, 2016; Habersberger, Fulmer, Kornfeld, & Yang, 2018). Feedback should be obtained from creators to determine the best way of creating rules around generating tags and topics. Once a consensus is reached, creators can help ensure new videos are tagged accordingly. The company can also commission crowdworkers to apply consistent tagging across historical videos and provide ways for it to be easier to apply consistent tags across future videos (e.g. have a dropdown of previous tags used when the creator is inputting tags for a new video).

When implementing this model, it is important to ensure that creators feel comfortable using the tool. Creator names are tags as well in this dataset, so a decision could be easily made to tie viewership of a video to the compensation of a particular creator. This along with the sentiment that algorithms encourage the production of generic content can increase creator attrition (Dominique, 2019). Without talented and innovative creators, who feel they have creative freedom and ownership of their work, there will not be engaging content to create long-term success for the company.

References

Briggs, J. (2018, March 26). YouTube SEO ranking factor study. *Briggsby*, Retrieved from https://www.briggsby.com/

Burton, S. A. (2012, October 10). What famous internet cat are you? *BuzzFeed*, Retrieved from https://www.buzzfeed.com/

Cawley, G.C. & Talbot, N. L. C. (2007, April). Preventing over-fitting in model selection via Bayesian regularisation of the hyper-parameters, *Journal of Machine Learning Research, 8*, 841-861.

Chauhan, N. S. (2019). Real world implementation of logistic regression. *Towards Data Science*, Retrieved from https://towardsdatascience.com/

Drakos, G. (2018, August 12). Support Vector Machine vs Logistic Regression. Towards Data Science, Retrieved from https://towardsdatascience.com/

Dominique, D. (2019, November 15). Let's talk about morality, tinder, youtube fame | with best dressed Ashley [Online Video]. *Red Wine Talks*, Retrieved from https://www.youtube.com/

Dunn, G., & Raskin, A. (2016, November 16). Why we left Buzzfeed (some thoughts on taking risks) [Online Video]. *VlogBrothers*, Retrieved from https://www.youtube.com/

Elite Data Science. (2019). Overfitting in machine learning: What it is and how to prevent it. *Elite Data Science*. Retrieved from https://elitedatascience.com/

Garcia, J. P. (2016). *Tuning parameters for logistic regression*. Retrieved from https://www.kaggle.com/

Habersberger, K., Fulmer, N., Kornfeld, Z. & Yang, E. L. (2018, July 18). Why we started our own company [Online Video]. *The Try Guys*, Retrieved from https://www.youtube.com/

Hastie, T., James, G., Witten, D. & Tibshirani, R. (Eds.) (2017). An Introduction to Statistical Learning. New York, NY: Springer

Lee, E. (2019, January 23). BuzzFeed plans layoffs as it aims to turn profit. *The New York Times*, Retrieved from https://www.nytimes.com

Liberman, N. (2017, January). Decision trees and Random Forests. *Towards Data Science*, Retrieved from https://towardsdatascience.com/

Natly, K. (2011, May 6). How many views do you need to be viral? Retrieved from http://willvideoforfood.com

Nygaard, S. (2017, March 19). *Why I left BuzzFeed* [Online Video]. Retrieved from https://www.youtube.com/

Robischon, N. (2016, March). How BuzzFeed's Jonah Peretti is building a 100-year media company. *Fast Company*, Retrieved from https://www.fastcompany.com

Safi, D. (2019, October 17). Eugene Lee Yang talks putting it all online [Online Video]. *The Feed*, Retrieved from https://www.youtube.com/

Seif, G. (2018, November). A guide to decision trees for machine learning and data science. *Towards Data Science*, Retrieved from https://towardsdatascience.com/

scikit-learn developers. (2019). Logistic regression. In *scikit-learn user guide: Release 0.22* (p. 179). Retrieved from https://scikit-learn.org/

Wilson, T., Banerjee, N., Hernandez, E., Jr., Abramson, J., & Bronner, E. (2018). Finalist: Staff of BuzzFeed News. *The Pulitzer Prizes*, Retrieved from https://www.pulitzer.org/

YouTube. (2018, January 30). How video views are counted. *YouTube Help*. Retrieved from https://support.google.com/

YouTube. (2019, February 25). Lesson: Make money on YouTube. *YouTube Creator Academy*, Retrieved from https://creatoracademy.youtube.com/

Appendix A. *Tags Removed from Analysis*

'Motivationalspeechyoutuberedoriginalseries',

'pl5vtqduum1dmy4t5c_7dycem1zxdkozqq',

'pl5vtqduum1dk_eerei21fzkdndz2ljysh',

'bfvideo',

'bfv',

'apple'

Appendix B. *Tags and Topics Used as Features*

Topics	'sports_topic', 'animal_topic', 'beauty_topic', 'bzfdemployee_topic', 'bzfdseries_topic', 'celebrity_topic', 'challenge_topic', 'drink_topic', 'family_topic', 'firsttimetry_topic', 'food_topic', 'friends_topic', 'funny_topic', 'game_topic', 'identity_topic', 'lgbtq_topic', 'money_topic', 'music_topic', 'relationship_topic', 'scary_topic', 'sex_topic', 'storytime_topic', 'travel_topic', 'work_topic', 'youtuber_topic'
Tags	'tryguys', 'stevenlim', 'fun', 'cheese', 'film', 'pretty', 'halloween', 'truth', 'facts', 'quiz', 'silly', 'health', 'hilarious', 'holiday', 'scary', 'world', 'adorable', 'people', 'crazy', 'best', 'science', 'awkward', 'college', 'americansreact', 'sleep', 'test', 'interview', 'art', 'party', 'diy', 'weird', 'dance', 'sketch', 'women', 'text', 'teens', 'life', 'amazing', 'drugs', 'photo', 'beatit', 'television', 'youtubered', 'school', 'facebook', 'cancer', 'new', 'internet', 'eugene', 'death', 'phone', 'secrets', 'socialmedia', 'healthy', 'news', 'christmas', 'tips', 'cool', 'princess', 'creep', 'face', 'tricks', 'religion', 'versus', 'videoclip', 'traffic', 'environment', 'review', 'diet', 'hunting', 'disney', 'netflix', 'yummy', 'fans', 'red'

Appendix C. Team Ramen Division of Responsibilities

Amelia Chu (ac4119)	Data Collection, One-Hot-Encoding of categorical variables, Random Forest Classifier and Model Analysis Write-up
Cora Hyun Jung (hj1399)	SVM Model, Parameter Selection, and Analysis Write-up
Dee Hahm (drh382)	Logistic Regression and Decision Tree Models Write-up