Analysis of Trending YouTube Videos Across Different Countries

Olivia Long
Princeton University
olong@princeton.edu

Catherine Nguyen
Princeton University
cn5@princeton.edu

Anne Zou
Princeton University
azou@princeton.edu

Abstract

As our world becomes more connected online, data on trending YouTube videos can yield interesting information about our increasingly global society. To explore the similarities and differences across countries, we analyze the Kaggle dataset ”Trending YouTube Video Statistics” from November 2017 to June 2018 for the United States, Canada, and the United Kingdom. In this project, we are interested in the words used in the video titles, tags, and descriptions. We employ Latent Dirichlet Allocation (LDA) to find hidden structure in the words used among the different countries. Using these words as predictive features, we classify the video category using Multinomial Naive Bayes, Linear Support Vector Classification, and Logistic Regression. Through LDA, we found that the words used in the Canada dataset was the most homogeneous across videos and those used in the US dataset were the most diverse. From classification, we found that Logistic Regression performed the worst in terms of precision and recall, and yielded the most variable performance between the different categories. Lastly, we used a Linear Regression model to predict the number of views a video gets using the words appearing in the video title, tags, and description. We get an R^2 score of 0.6561, indicating that this number is moderately predictable.

1 Introduction and Problem Statement

YouTube is one of the most popular social media platforms today. It allows an opportunity to be heard or seen. Videos that have a large number of views can heavily influence others and shape society. In this final project, we will explore datasets that contain information on trending YouTube videos in the United States (US), Canada (CA), and the United Kingdom (UK) to compare the topics that are trending in different countries. To do this, we will first examine the latent structure of the words used in the titles, tags, and descriptions of the videos for each country dataset separately. Then, using only the US dataset, we will attempt to both classify the video category and predict the number of videos views based on the words used in the titles, tags, and descriptions of the videos.

2 Related Works

The increased popularity of YouTube allows many people to use information on trending videos to better understand what people like and dislike. There has been research conducted on how people feel towards products through the reviews they leave and this can also be done with how people feel about certain videos. Companies are interested in having a better understanding of consumers’ likes, dislikes, and behaviors so that they can have more influence in society. In the paper “How Useful are Your Comments? - Analyzing and Predicting YouTube Comments and Comment Ratings”, Siersdorfer analyzed about 6 million comments collected by 67,000 YouTube videos. One of the main questions he was exploring was whether comment rating behaviors depend on topics and categories

[2]. Thus, the relationship between the video category and the words associated with the video is an area of interest.

Due to the interest in using clickbait, others have also investigated the relationship between the words used in YouTube videos and resulting video view counts. Attempted models include neural networks, which did not yield good results [3]. Thus, we are motivated to search for alternative ways to describe this relationship.

3 Methods

3.1 Data Description and Processing

The raw data set was given as a 'csv' file for each country (United States, United Kingdom, and Canada) [1]. The data set included the country's trending videos from Nov 2017 to June 2018. To process the data, we converted the 'title', 'tags', and 'description' features for all videos into a bag-of-words representation and used the 1,000 most frequently appearing words from each of these sources as the vocabulary. The original vocabularies were quite large, being greater than 40,000, which motivated us to perform feature selection. We selected 1,000 because after running Principal Component Analysis (PCA) on the US dataset and examining the explained cumulative variance, we saw that the 1,000 most frequently appearing words in each of the 'title,' 'tags,' and 'description' features sufficiently explained essentially all the variance in the dataset, as shown in Figure 5. Then, the most frequently appearing words for each feature were appended to form a final matrix with 3,000 columns and, for each country, the same number of rows/videos as the original dataset.

For the US dataset, there were 40,949 videos, so the bag-of-words matrix has dimensions 40,949 x 3,000. For the United Kingdom, there were 38,916 videos and the bag-of-words matrix has dimensions 38,916 x 3,000. Lastly, for Canada, there were 40,881 videos and the bag-of-words matrix has dimensions 40,881 x 3,000.

In the original dataset, each video was categorized into one of 16 total categories: Film and Animation, Autos and Vehicles, Music, Pets and Animals, Sports, Travel and Events, Gaming, People and Blogs, Comedy, Entertainment, News and Politics, How-to and Style, Education, Science and Technology, Nonprofits and Activism, and Shows. Each category was labeled using a unique number. To process the data, we extracted the 'category' column as a vector to use as classes in our multi-class classification tasks. The column of video view counts was also extracted as a vector for linear regression.

Thus, for both our classification and linear regression tasks, we use the 40,949 x 3,000 bag-of-words matrix from the US dataset to predict the 40,949 x 1 vectors of video category labels and video view counts. When fitting the models, an 80-20 training-test split was used.

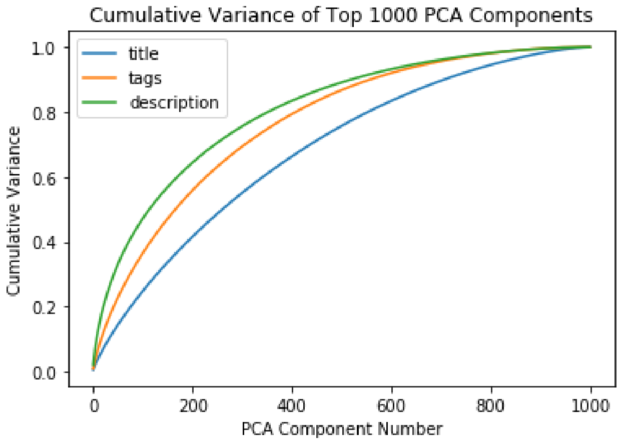


Figure 1: **Cumulative variance explained as a result of Principal Component Analysis (PCA) on the bag-of-words representations for trending video titles, tags, and descriptions in the US.**

3.2 Data Exploration

To have a better understanding of our data, we first wanted to see if the distribution of video categories differed among different countries. From Figure 4, we can see that Entertainment is widely popular among the three countries, especially in Canada and US. It is interesting to see that Music is the most popular category in the UK, as opposed to Entertainment, which is the most popular in the other two countries. Based on the relative heights of the bars, it is also interesting to note that the category distribution in the US and Canada are more similar to each other than to the UK. Figures 2 and 3 display comparisons of the most used words in the US's video descriptions with how often these words are used in the UK and CA. From the plots, it seems that these words have similar frequencies in the CA and UK relative to in the US.

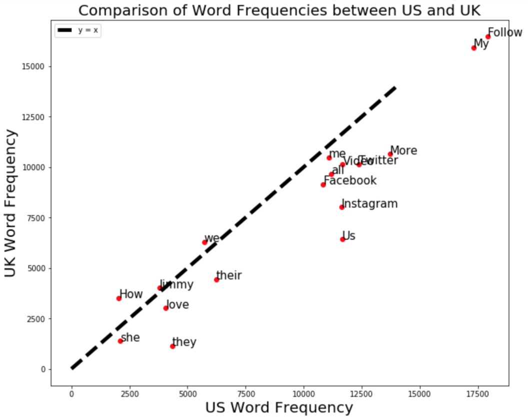


Figure 2: **Comparison of Word Frequency of Popular Words used in the United States (US) and United Kingdom (UK)**. The y=x line is shown as a dotted line.

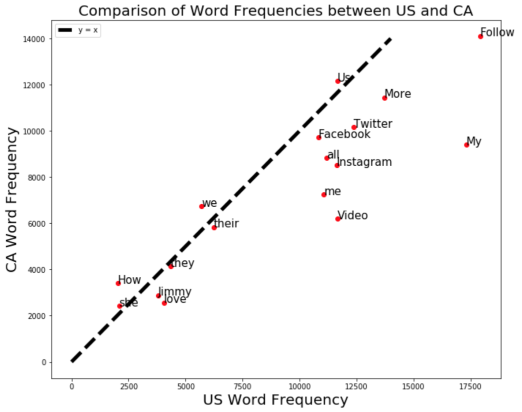


Figure 3: **Comparison of Word Frequency of Popular Words used in the United States (US) and Canada (CA).** The y=x line is shown as a dotted line.

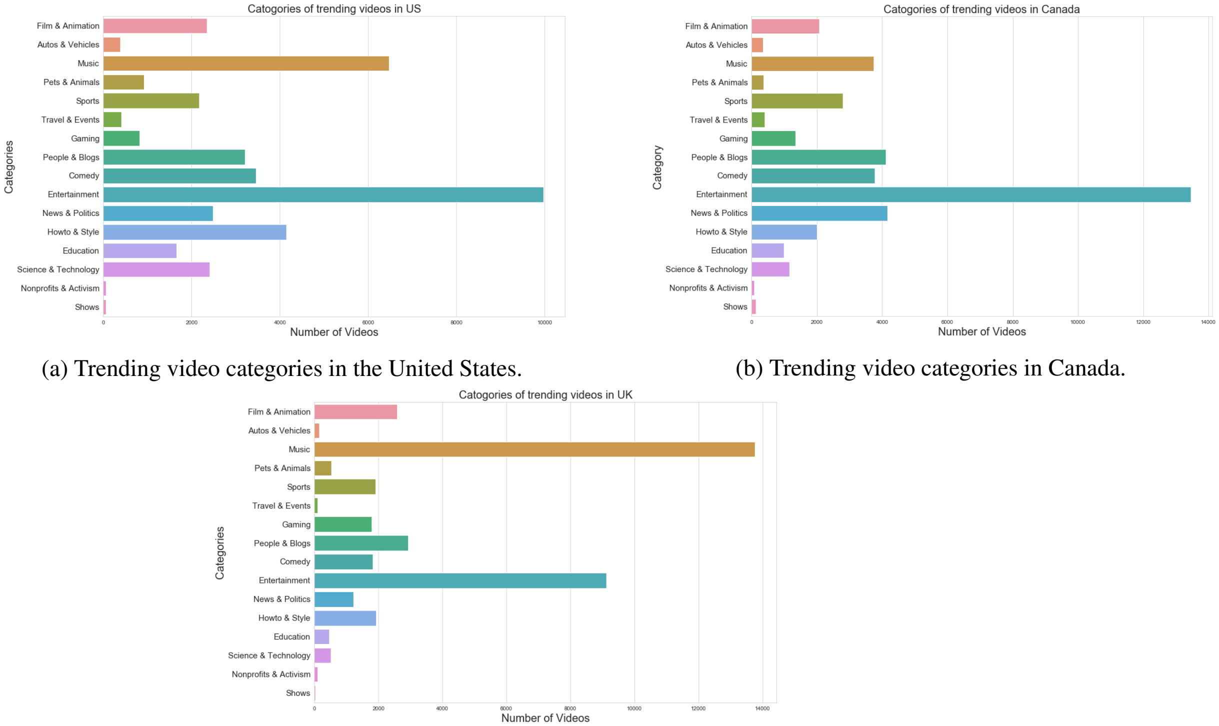
3.3 Classification

In this second part of our project, we wanted to classify videos by category based on the words used in the title, tags, and description (represented by bag-of-words). To explore the performance of different models, we used the Multinomial Naive-Bayes, Linear SVM, and Logistic Regression classifiers. The classifiers were used in a one-vs-rest strategy, which fits one classifier per class. This class is fitted against all of the other classes. In our case, the 'class' corresponds to the video category. Since the Multinomial Naive-Bayes assumes independence of the different categories and uses a multinomial distribution for each category, we used this as a baseline model to compare the other classifiers to. When fitting, we used set the smoothing parameter α to 0 (no smoothing), since that was the optimal value given by performing GridSearchCV.

The Linear SVM uses a different algorithm that attempts to make a line (decision boundary) separating the data points based on whether or not it is in the given category. In contrast to Linear SVM, Logistic Regression uses the logistic sigmoid function to assign probability values of being in two or more discrete classes. When fitting the Linear SVM and Logistic Regression models, 'L2' regularization was used to prevent overfitting. We were motivated to compare these two models, since both use a linear combination of features.

3.4 Regression

In the next part of our project, we wanted to predict the number of views on a video based on the words used in the title, tags, and description (represented by bag-of-words) using a Linear Regression model. In linear regression, we fit a linear model with weights for each feature to minimize the error, measured as a residual sum of squares between the actual and predicted target features. We use this to examine how well the descriptive features predict the target feature.



(a) Trending video categories in the United States.

(b) Trending video categories in Canada.

(c) Trending video categories in the United Kingdom.

Figure 4: **Trending videos by category for the United States, Canada, and the United Kingdom.**

3.5 Evaluation Metrics

To evaluate the performance of the different classifiers, we compared the precision-recall curves, since the sizes of the different categories were imbalanced, as shown in Figure 4. The precision P and recall R are given by the formulas: $P = T_p/(T_p + F_p)$ and $R = T_p/(T_p + F_n)$ where T_p is the number of true positives, F_p is the number of false positive, and F_n is the number of false negatives. Precision describes how well a model positively predicts the given class, and recall describes the sensitivity of the model [4]. Precision-recall pairs were plotted for different probability thresholds. F1-scores (the harmonic means of the precision and recall) were also plotted on the precision-recall curves and area under curve (AUC) was calculated for comparison.

To evaluate the performance of the linear regression model, we look qualitatively at the predictions made as well as at the coefficient of determination R^2 of the prediction. The coefficient is defined as $R^2 = 1 - \frac{u}{v}$, where u is the residual sum of squares and v is the total sum of squares. For reference, the best possible score is 1, and a model that predicts the expected target value has a score of 0.

4 Results

4.1 Latent Dirichlet Allocation (LDA)

LDA is a powerful model that groups text by hidden themes. We observed patterns using LDA with the bag-of-words(BOW) representations of video tags, title and description for three western countries: United States (US), United Kingdom (UK), and Canada (CA). Utilizing GridSearch, we determined that it is best to have 10 components with 0.5 decay. Figure 5 displays how the topics with the top weights based on pseudocount were distributed in each country. A higher the pseudocount for a word means that there are more documents that have the same pattern of the word. From the figure, we can see that in the US, Facebook has the highest pseudocount, followed by Twitter and then Instagram. In the UK, Instagram had the highest pseudocount, followed by Facebook and then Twitter. In Canada, Twitter had the highest pseudocount, followed by Facebook and then Instagram. The US and Canada also had heavier weights on topics related to food while the US and UK had heavier weights on topics related to make-up. For histograms of the latent topic proportions for each country, see Figure 13 in the Appendix.

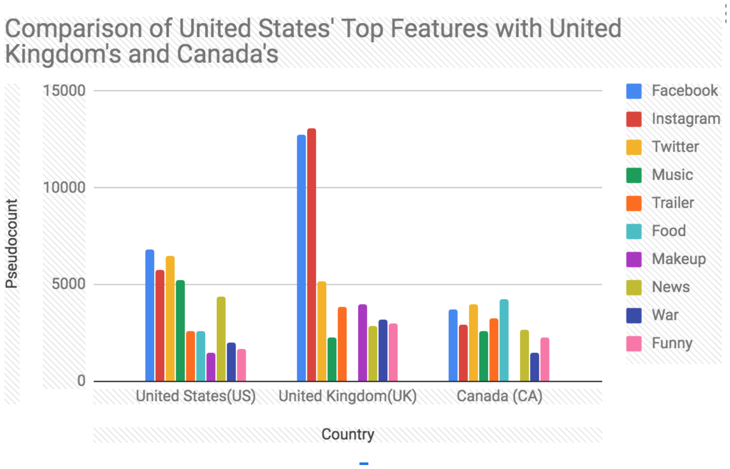


Figure 5: **Distribution of 10-Component LDA pseudocounts across the United States (US), Canada (CA), and the United Kingdom (UK) for the top weighted words across all the topics from US LDA.**

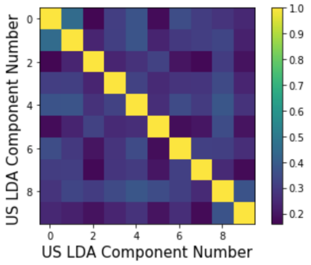


Figure 6: **Correlation between 10 Latent Topics in United States**

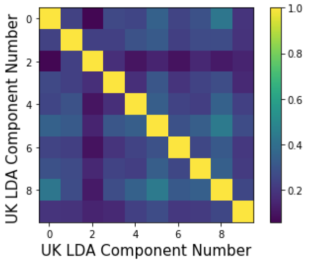


Figure 7: **Correlation between 10 Latent Topics in United Kingdom**

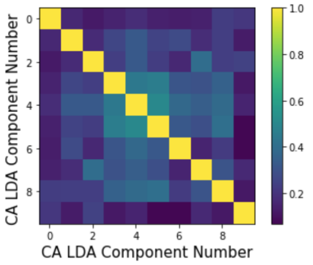


Figure 8: **Correlation between 10 Latent Topics in Canada**

Figure 6 demonstrates how each countries' 10 Latent Topics correlates with each other. The brighter the pixel indicates more similarity between latent topics. From this, we can see that Canada has the most homogeneous set of latent topics. When we shuffled the pixels for the Canada Plot, it is still apparent that Canada has more brighter pixels, compared to US and UK. The US has the least homogenous set of latent topics.

4.2 Classification

To attempt classification of video categories based on the bag-of-words representations of the title, tags, and descriptions of the videos, we compared the performance of three different classifiers: Multinomial Naive Bayes, Linear SVM, and Logistic Regression. Across all the classifiers, we found that the categories 'Music', 'Pets and Animals', and 'Sports' were consistently classified well, as shown in Figures 9, 10, and 11. Categories that were consistently classified badly include 'How-to and Style', 'Education', 'Science and Technology', and 'Shows'. This is most likely because the frequency of these categories was low in the test and training data set, making it more difficult to accurately predict them. It was interesting to find that 'News and Politics' is not classified well across the classifiers (areas of 0.20, 0.20, and 0.22 for NB, LSVM, and LR respectively), despite its high frequency in the training and test data sets, comprising 24% of the videos in both data sets. The

‘People and Blogs’ category yielded an NaN area value for each classifier since no videos in the test set belonged to this category.

Notable differences include the classification on ‘Film and Animation,’ which performed well using the Naive Bayes (area = 0.89) and Linear SVM (area = 1.00) classifiers, but did not perform as well using Logistic Regression (area = 0.52). This was also the case for ‘Autos and Vehicles’, which gave area values of 0.98 and 0.97 for Naive Bayes and Linear SVM, respectively, and a value of 0.21 for Logistic Regression. This is most likely because of the small percentage of videos that fell under these categories, since only 5% of videos were ‘Film and Animation’ and 1% of videos were ‘Autos and Vehicles’ in the training and test sets. From Figure 11, we also see that in the Logistic Regression classifier, there is much more variation among the precision-recall pairs for different probability thresholds. This is most likely because Logistic Regression is known to perform better on low-dimensional data.

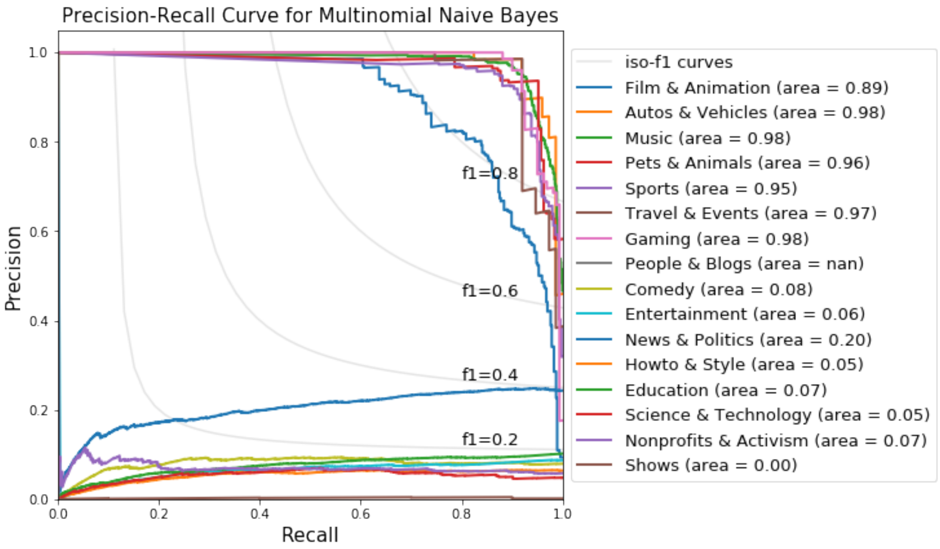


Figure 9: **Precision-recall curve for Multinomial Naive-Bayes used as One-vs-Rest classifier on 16 video categories.** An 80/20 training-test split was used on the US dataset, which contained 40,949 videos in total. Iso-F1 curves are plotted as gray lines.

4.3 Regression

We used a linear regression model on an 80-20 train-test split and got a coefficient of determination R^2 on the test prediction of 0.6561, which is relatively good and which indicates that the number of views is moderately predictable. When standardizing the target feature, we got an R^2 value of 0.6578, which is not much better.

To visualize the predictions of our trained model, we plotted the actual and predicted numbers of views for all videos in the US dataset versus the total word count of certain keywords. We selected keywords that we found to be top weighted in running LDA as well as frequently found in our bag-of-words representations for video titles, tags, and descriptions; namely, we used ’facebook,’ ’instagram,’ and ’twitter.’ These plots (Figure 12) show that our predictions generally capture the trends in the numbers of video views, but they tend to incorrectly predict actual values where the distribution is more sparse, as shown by the thinly dispersed red points in the tops of the scatterplots. This is understandable, as there are few outliers that reach extremely high view counts, so our model may not have had many training instances of that sort to work with.

5 Discussion and Conclusion

In this project, we explored and compared the latent structure of the words used in the titles, tags, and descriptions of trending YouTube videos in each three different countries: the United States,

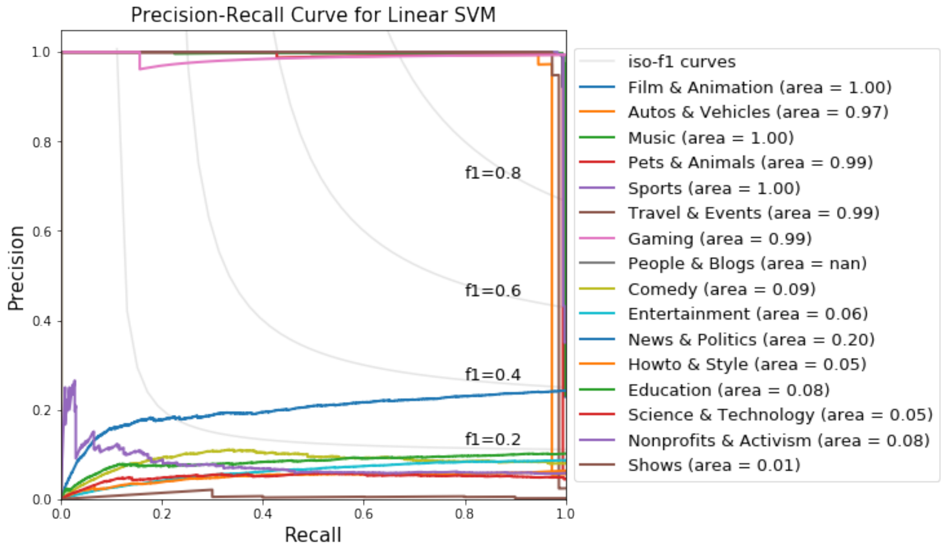


Figure 10: **Precision-recall curve for Linear Support Vector Classification used as One-vs-Rest classifier on 16 video categories.** An 80/20 training-test split was used on the US dataset, which contained 40,949 videos in total. Iso-F1 curves are plotted as gray lines.

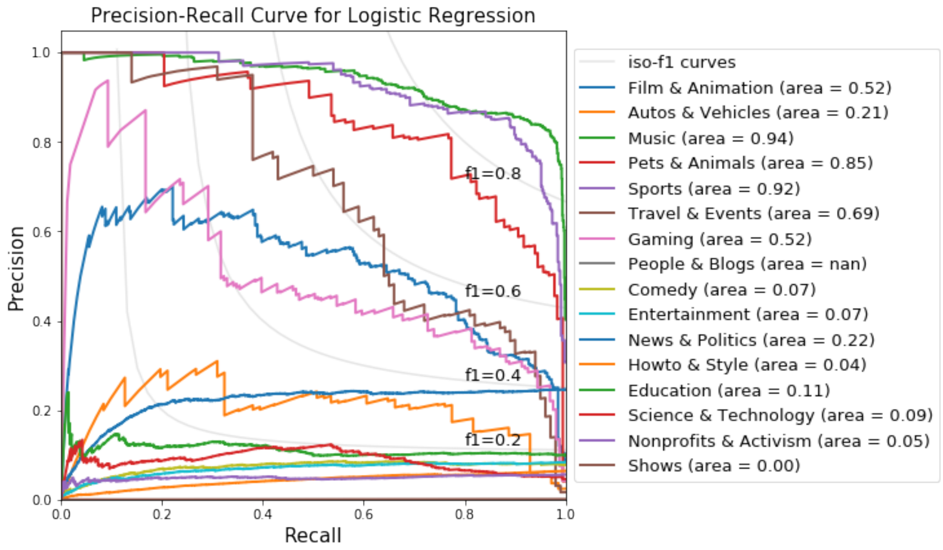
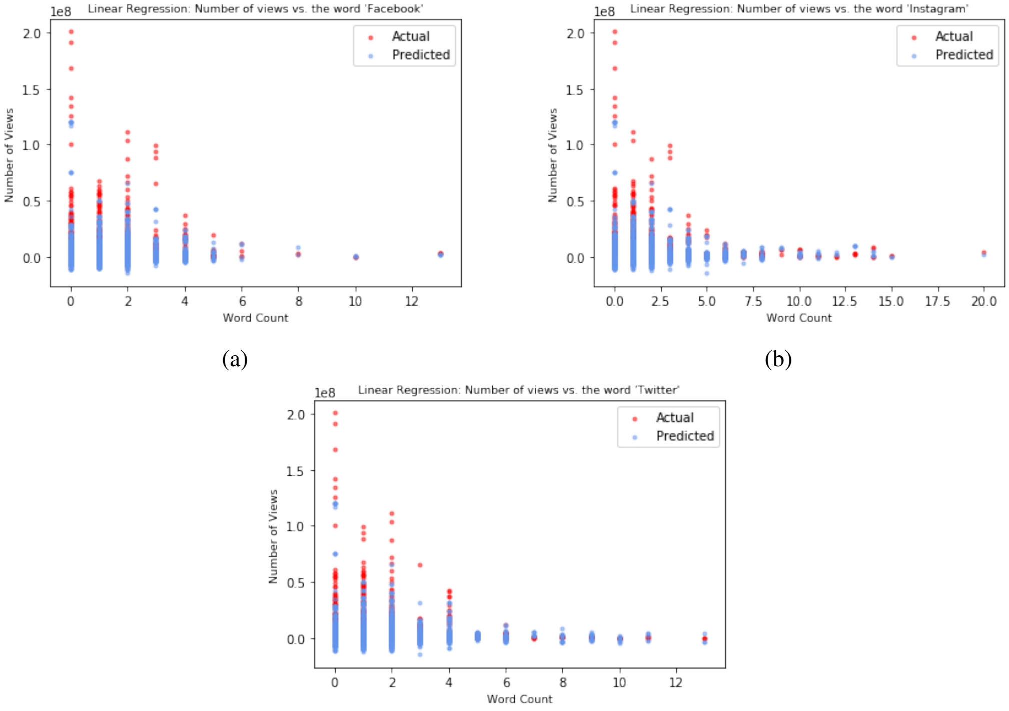


Figure 11: **Precision-recall curve for Logistic Regression Classification used as One-vs-Rest classifier on 16 video categories.** An 80/20 training-test split was used on the US dataset, which contained 40,949 videos in total. Iso-F1 curves are plotted as gray lines.

Canada, and the United Kingdom. Then, using the US dataset, we attempted to classify the video category and predict the number of video views based on the words used in the video titles, tags, and descriptions.

From our classification modeling, we found that Logistic Regression exhibited the most variation in performance between the different categories. The best performing category also did not score as highly as in the Multinomial Naive Bayes and Linear SVM models, as shown in the plots. From LDA, we can see that each country had different distributions of pseudocounts for social media. The US had the highest pseudocount with Facebook, UK with Instagram, and CA with Twitter. The US and UK had topics related to makeup while CA did not. On the other hand, the US and



(a)

(b)

(c)

Figure 12: **Visualizations of fitting a multiple linear regression model to the full US bag-of-words representation, plotting actual & predicted number of views against word counts for the top features found in LDA.**

CA had topics related to food while UK did not. We also saw that the US's topics were the least homogeneous compared to the other countries.

To extend our work on classification models, we would like to normalize for the number of videos in each category before fitting the data with the classifiers. Since the sizes of each categories were so imbalanced, it is difficult to conclude whether classifiers did badly because the training set was very small or because it is inherently difficult to predict that category based on our bag-of-words representation. To extend our work for LDA, we would want to see the correlation not only between the topics within a country but also how similar the topics are among all the countries. We could use a correlation model to do so. To extend our linear regression modeling, we could collect more data to train our model with so it can better capture the relationship, if any, between the words and a very large number of views. We could also perform more feature selection and pick out features that are more relevant for regression modeling, which requires more exploratory data analysis. Moreover, With increased processing power, we could try other methods as well such as gradient boosting regression or radial basis function networks.

References

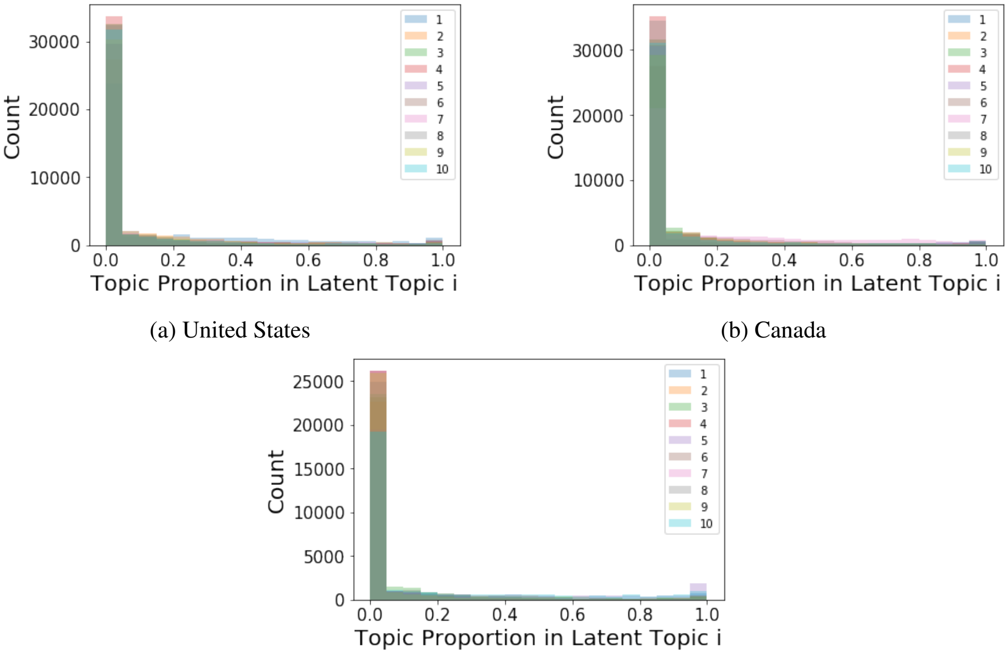
[1] Jolly, M. Trending YouTube Video Statistics. https://www.kaggle.com/datasnaek/youtube-new. 2018. Accessed April 21, 2019.

[2] Siersdorfer, Stefan, et al. ”How useful are your comments?: analyzing and predicting youtube comments and comment ratings.” Proceedings of the 19th international conference on World wide web. ACM, 2010.

[3] Srinivasan, Aravind. ”YouTube Views Predictor.” *Towards Data Science*. https://towardsdatascience.com/youtube-views-predictor-9ec573090acb. Accessed May 13, 2019.

[4] Brownlee, Jason. ”How and When to Use ROC Curves and Precision-Recall Curves for Classification in Python.” https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/. Accessed May 13, 2019.

6 Appendix



(a) United States

(b) Canada

(c) United Kingdom

Figure 13: **Histograms of latent topic proportions from 10-component LDA with a learning decay of 0.5, fit to each country’s full bag-of-words representation (video titles, tags, and descriptions).**