Feature Analysis of Trending YouTube Videos Across Canada, USA, Great Britain, and Mexico

PRINCETON UNIVERSITY

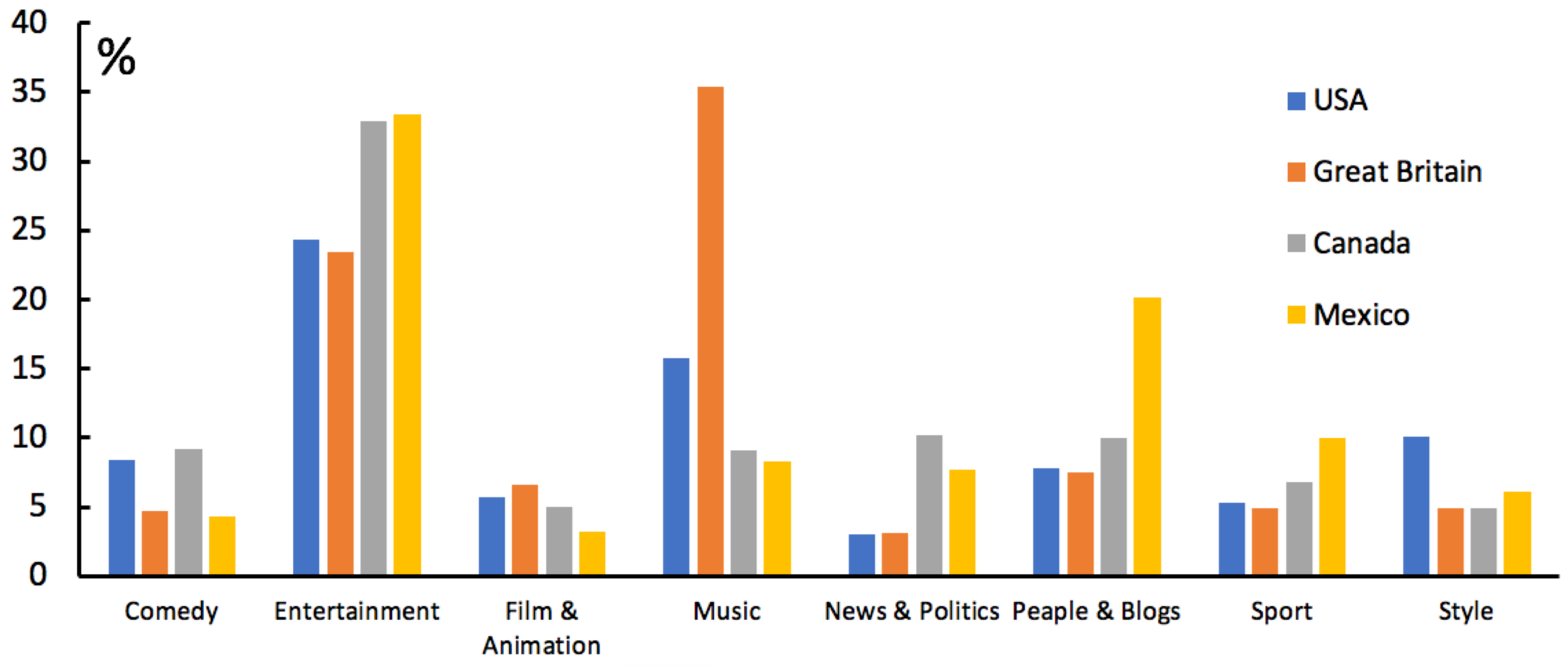
Francisco J. Carrillo

Abstract

In this project we used latent feature analysis on data pertaining to YouTube's trending videos in order to identify topics that describe the sociocultural similarities and differences between various countries. Furthermore, through the use of different classifier models, we identified which features are good predictors of a video's future `trendability" across different nationalities. Part of this effort was to identify if having fully capitalized words (i.e. NEW, OFFICIAL, ext...) in a video's description is a good indication of future trendability. We concluded that it is not. Finally we summarize all these results by creating a hypothetical `perfect' trending video for each country such as to highlight the uniqueness of each nationality.

Video Category Distribution

The following figure contains the distribution of video categories across different countries. Music is clearly the overwhelmingly favorite category in Great Britain. Entertainment, Music , and Blogs are the most popular overall.



Predicting Trendability Across Countries

Classifier Performance Metrics

	USA	Great Britain	Canada	Mexico
Random Forest Classifier:				
All Features	0.63/0.41/0.75	**0.68/0.50/0.75**	**0.7/0.53/0.75**	**0.84/0.16/0.77**
Only Views	**0.67/0.48/0.73**	0.67/0.50/0.72	0.69/0.53/0.74	0.84/0.25/0.76
All A-Priori Features	0.61/0.38/0.65	0.61/0.41/0.68	0.60/0.40/0.65	0.83/0.16/0.62
Top 10 A-Priori Features	0.61/0.39/0.62	0.64/0.45/0.64	0.62/0.42/0.65	0.84/0.16/0.60
All Features **Naïve Bayes**	0.53/0.45/0.65	0.64/0.56/0.64	0.68/0.53/0.66	0.69/0.36/0.70
All Features **Log. Regr**.	0.56/0.35/0.58	0.63/0.42/0.68	0.65/0.49/0.65	0.71/0.35/0.52
Legend	**Accuracy/Precision/ROC**			

The above figure shows the performance of several classifiers trying to predict if a video will have a longer-than-average run in the trending list. We can see that Random Forest Classifiers worked best. It appears that 'views' are the most predictive feature for future trendabilty. However, since we don't know the # of views a-priori, we expanded our analysis to only include known variables (i.e. BOW and category ID's). We also ranked the top predictive features overall, as shown in the next table. Finally, we concluded that word capitalization is not a good predictor for trendability.

Top Features that Predict Trendability

Ranking	USA	Great Britain	Canada	Mexico
1	views	views	views	**dislikes**
2	likes	likes	likes	views
3	dislikes	comment_count	comment_count	likes
4	comment_count	dislikes	dislikes	comment_count
5	percent_likes	performing	percent_likes	percent_dislikes
6	percent_dislikes	category_id_Music	likes/comments	likes/dislikes
7	likes/comments	night	likes/dislikes	percent_likes
8	Follow	Music	**Episode**	percent_comments
9	likes/dislikes	percent_likes	**category_id_News**	4
10	**news**	**Jimmy**	category_id_Music	category_id_Music
11	percent_comments	likes/comments	Music	category_id_Comedy
12	Subscribe	**Live**	percent_comments	Twitter
13	Late	come	category_id_Comedy	**mexico**
14	**Jimmy**	Follow	percent_dislikes	likes/comments
15	night	2018	2018	Music

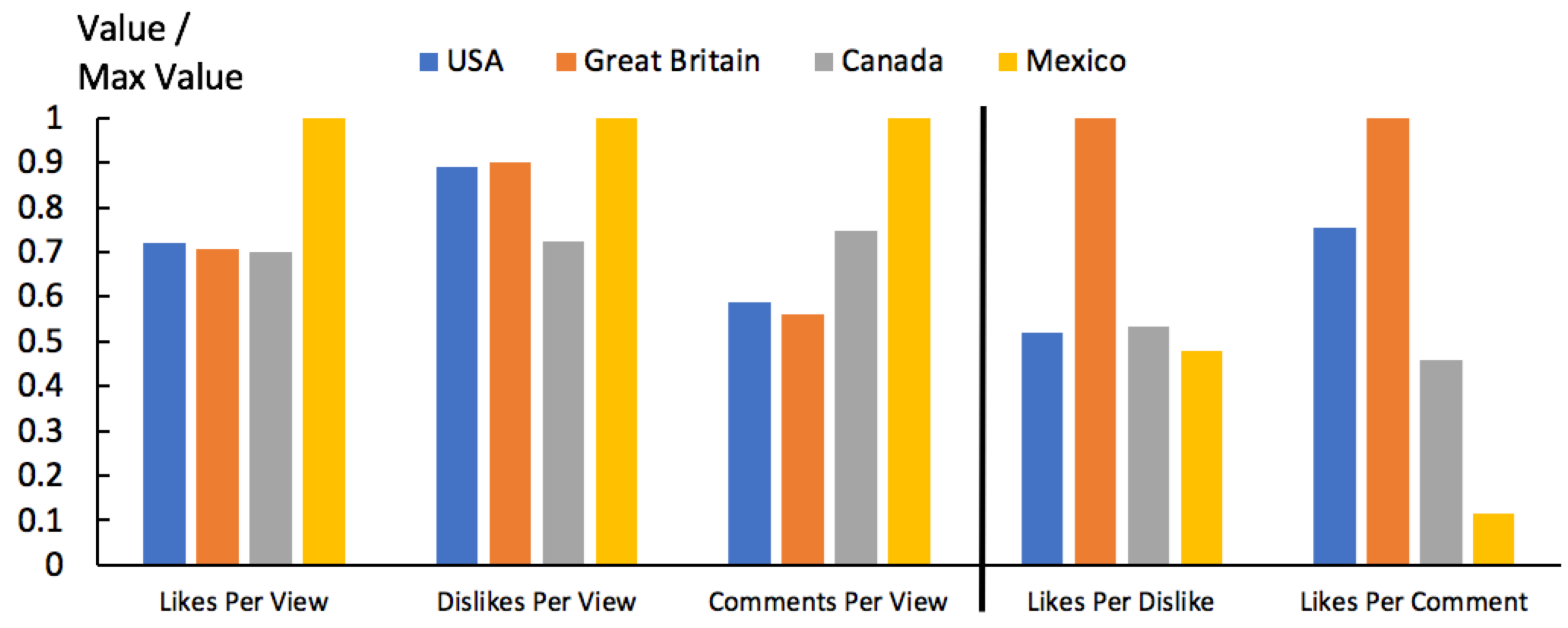
Data Processing Steps

The data consist of a list of the top 200 videos for 4 different countries over 200 days. Each video has a total of 16 features such as a title, a description, tags, URL, like count, dislike count, and number of comments.

1) Remove Videos with faulty features (<1% total)
2) Create a Bag of words (BOW) representation of the top 100 words for each country.
3) Do One-Hot encoding on the provided category-id tags
4) Create features that normalize likes, dislikes, and comments to the number of views
5) Create Boolean Identifiers for videos that last longer-than-average on the trending list and for videos that contain fully-capitalized words.

Community Engagement

We quantify community engagement by looking at the distribution of likes dislikes and comments per viewed video in each country. Here we can see that Mexico has highest engagement rating across the board. Furthermore, we can also see that, given an interaction, the English are the most positive overall.



Latent Topic Distribution

Distribution of 10 Latent Topics Across Countries

USA	Great Britain	Canada	Mexico
Entertainment	News	Entertainment	**Telenovelas (Soap Operas)**
News	**Movies (Star Wars)**	News	**Sports Highlights**
Movies	Entertainment	Lifestyle	Music
Music	Music	Social Media	Entertainment
Style (Fashion)	Late-Night Shows	Late-Night Shows	**Sports News**
Late-Night Shows	Social Media	Music	TV Series
Lifestyle	Reality TV	Sports	**Family Entertainment**
Social Media	Sports	**Adventure**	Generic Videos
Generic Videos	Generic Videos	Movies	Social Media
Sports	Lifestyle	Generic Videos	Lifestyle

We used Latent Dirichlet Allocation on the processed features in order to identify 10 latent topics within the data. The above figure shows our result. The bolded features show topics that are characteristic within each nationality. Nevertheless, we do see topics that many countries share such as Entertainment and Music. This makes sense, however this does not mean that Entertainment in the USA is the same type of Entertainment as in Mexico. This is why we now look at features that form said topic.

Distribution Of Features on ”Entertainment” Latent Topic

The table on the right shows that English speaking countries consider Late-Night shows to be entertainment. We can also see that Mexicans prefer entertainment content that has to do with topics ridden with "love", "life" , and with "Mexico". Bolded words show characteristic features for each country.

USA	Great Britain	Canada	Mexico
Show	Show	Show	video
Late	**Late**	season	MI
CBS	night	**program**	**vivo**
Watch	season	use	**amor**
episode	Night	episode	CON
Follow	Watch	Late	**mexico**
Season	**YouTube**	2	programa
2	Follow	10	**Azteca**
night	3	3	VIDEO
latest	come	Watch	canal
live	**James**	5	episodio
Entertainment	2	**episodes**	mis
This	CBS	full	Entertainment
full	video	vs	POR
time	celebrity	Follow	ver

Ideal Most-Representative Trending Videos for Each Country: (Conclusions)

USA: A Special 'Latest-Fashion' Edition of a Late-Night Show on CBS with Several Musical Guests

Great Britain: A Music Video Based on a Star Wars Movie Sang by a Late-Night host Named Jimmy (last name either Fallon or Kimmel)

Canada: A TV Show Recap of an Episode About Jimmy Kimmel Trying his Luck at Several Sports and as a Professional Music Star

Mexico: A Family Friendly Soap Opera about a Mexican Soccer Team, Including Real Sport Highlights and Discussion