COS 424 Final Project: Feature Analysis of Trending YouTube Videos Across the USA, Great Britain, Canada and Mexico

Francisco J. Carrillo
Department of Chemical and Biological Engineering
fjc2@princeton.edu

Abstract

YouTube is the most popular video streaming platform on the planet. With about 2 billion monthly users spread across over 100 countries, it is fair to say that the themes and topics present in the "trending" videos of this platform are a good reflection of social trends at the national and global level. 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. Furthermore, 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.

1 Introduction

The music and video streaming industries are among the biggest and most influential entities on the planet. Data gathered by YouTube, Netflix, and Spotify are encoded with information regarding the preferences of millions (and sometimes billions) of people across the globe. These preferences are constantly monitored and analyzed by international brands and marketing agencies in order to gain a competitive advantage in catering their products to the masses. To that point, different latent feature models and model classifiers are used to identify and predict trending topics[3, 1].

In this project we analyzed data on the top 200 trending videos on YouTube over 200 days across 4 different countries (USA, Great Britan, Canada, and Mexico) in order to extract country-specific sociocultural trends. The purpose of this is to compare and contrast said trends in order to understand the differences and similarities between these countries. Furthermore, we also identified (and consequently compared) country-specific video features that predict the future 'trendability' of a video. In other words, we are trying to answer the following question: 'What are the characteristics that can describe the most-representative and most continuously trending video in each country?' The answer to said question will allow us to build a hypothetical perfect trending video for each region.

This report starts by presenting a complete description of the trending video data (Section 2), after that, it proceeds to present a comprehensive explanation of the proposed data processing steps (Section 3). This is then followed by a description of the proposed latent and predictive models (Section 4), and by an explanation of the evaluation criteria used to asses their performance (Section 5). Next, we discuss the results from our proposed analysis (Section 6), and, finally we summarize and present our conclusions (Section 7).

2 Data Description

The data for this project is on the public domain and was obtained from "Kaggle.com" [**?**]. The data itself contains trending video information from the USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan, and India. Each data set contains a dated list (spanning 200 days) of the top 200 trending videos for each country, as well as 16 additional features such as: the video's title, its description (written by the videos author), video 'tags', URL, number of likes, dislikes, and views, a generic category ID, and upload date. This means that videos that stay on the trending list over several days appear repeatedly on the data, once on every daily trending list. The original data set size for each country was about 40000 by 16.

The data itself is surprisingly clean, with little to no missing values. However, the data does contain non-English symbols in the cases of data belonging to non-English speaking countries. To deal with this we had to implement special character support into our data analysis. To make things a little bit easier and since I only fluently speak English and Spanish, we only analyzed trends for the USA, Great Britain, Canada, and Mexico.

3 Data Processing

The finalized work flow of this study (excluding failed investigations) can be summarized with the following steps, which can be separated into two parts. For each part, the reasoning behind each step is presented after the enumerated list.

3.1 Part 1: Identifying Latent Features Across Countries

1. Clean the data set by removing videos with missing values, videos that where removed from the trending list due to "policy violations", and videos that had non-English or Spanish titles. (The number of removed videos were less than 1 percent of the total).
2. Combine all text-based features for each video/sample and obtain the 100 most-used words within each country's trending list.
3. Obtain a Bag of Words (BOW) representation for each video based on these top 100 words.
4. Create the following features based on count data: 'percent likes' (likes/views), 'percent dislikes' (dislikes/views), 'percent comments' (comments/views), 'likes/dislikes', and 'likes/comments'.
5. Create features based on 'One-Hot-Encoding' of the video's category IDs.
6. Combine all previous data sets into a single 'Latent Analysis Data Set'
7. Use Latent Dirichlet Allocation (LDA) on the Latent Analysis Data Set (excluding text data) to identify 'k' latent topics for each country.
8. Compare and contrast latent topics across countries

In steps 2 and 3, we decided to create a simple BOW representation of the data in order to be able to use text data without having to use complex natural language processing techniques. Furthermore, this allows for the easy incorporation of word-count data into our chosen latent feature models and classifiers.

We chose to create the features outlined in step 4 because we believe that these are metrics that will be able to quantify and normalize relative user engagement within each video. Watching a video is a passive action, but pressing the 'like' button or writing a comment imply that the user had an emotional response to said video. We hypothesis here that these metrics are predictive of trending status.

One-hot encoding of the categories follows the same reasoning behind the BOW representation of the text data. These way we are able to convert categorical data (generic category IDs) into numerical values that can be used with our chosen latent and classification algorithms without loosing any information. The resulting data set matrix mentioned in Step 6 was about 40,000 by 137 in size for each of the four countries.

3.2 Part 2: Identifying Features that Predict Future Trendability Across Countries

1. Create a 'Regression Data Set' that contains the average count data from the Latent Analysis Data Set (excluding text data) across all trending-dates for each video.
2. Identify the average time a video is tending and create a Boolean identifier for videos that stay longer than average (1) or shorter than average (0) on the trending list.
3. Identify if the video's title or description has capitalized words and create a Boolean identifier that represents this.
4. Use GridSearchCV, K-Folds, and different classifiers in order to predict if a video will be trending for more (or less) time than the trending-average.
5. Identify the features that are most indicative of 'trendability' for each country by using Recursive Feature Elimination and Random Forests Coefficients.
6. Compare and contrast representative features across countries

Step 1 combined all data entries for the same video across several days into a single entry. This was done to in order to be able to have a single comprehensive data set for a given video that could be compared to the data from other individual videos.

We decided to use a boolean variable (as opposed to a continuous one) for identification of 'trendability' because it simplifies the analysis while still providing us with an indicator variable that answers our original question. As an extension, we were also interested on whether or not fully capitalized words in the video's title or description were indicators of trendability, which is why we identified the presence of this words through a boolean identifier.

Lastly, we note that non-presented investigations include the use of continuous-outcome regression models (linear regression, Support Vector Machines ext..), and the use of other latent analysis models (Principal Component Analysis and Factor Analysis).

4 Models and Algorithms

All the numerical algorithms presented here can be found in the SciKitLearn Python Libraries [5] and used with standard values unless otherwise specified. For latent feature analysis we used (and attempted to use) the Latent Dirichlet Analysis, Principal Component Analysis, and Factor Analysis libraries. The classifiers used in the second part of the study were: Logistic Regression, Naive Bayes, and Random Forest (RF) classifiers. A complete spotlight description of RF is present in the appendix, as I wrote it before we were told we would not need it. The hyper parameters for these classifiers where tuned by the use of an exhaustive grid search cross validation algorithm (GridSearchCV). Furthermore, the attempted use of continuous regresors (not presented due to bad results) involved linear regression and Random Forests Regression.

For both parts, the data was separated into a training data set and a test data set by using the K-fold python library, which uses random seeds to separate the data into k 'slices' that can be used for sequential predictions and evaluations. Lastly, we note that parts of the code used in this project have been extracted from all three previous homework assignments.

5 Evaluation Criteria

For Part 1, we evaluated the latent topic models by verifying that our the resulting latent topics were consistent with each videos' category IDs. This does not mean that doing LDA on the data was unnecessary (most videos across countries share these IDs). The purpose of our use of LDA is to identify which features within each topic are the most important globally and for each country. We understand this is a qualitative/subjective evaluation, but we believe it is good enough for our purposes. As a further check, we made sure that the log-likelihoods across test and train samples were consistent for a given set of hyper-parameters.

For Part 2, we evaluated the performance of our classifiers by using Accuracy , Precision, and the area under the curve (AUC) for the Receiving Operator Characteristic. Accuracy is useful because it gives an intuitive feeling about the overall performance of the classifiers, however it says nothing

about the false positives or true negatives. Precision and the AUC help fill this gap by identifying the rate of true positives and its relationship to the rate of true negatives, respectively [2].

$$Accuracy = \frac{\text{Correct Samples}}{\text{Total Samples}}$$

$$Precision = \frac{\text{True Positives}}{\text{True and False Positives}}$$

6 Results and Discussion

The results from our described analysis will be presented here in three different parts. The first one will go over insights regarding basic feature analysis, meaning that no numerical models were used in this section, just feature exploration. The second part presents the results of LDA and the identification of latent topics within the data set. Finally, the third part goes over the predictive models used to model future video trendability.

6.1 Insights From Basic Feature Engineering and Analysis

The first thing we decided to look into was the magnitude of community engagement (number of likes, dislikes, and comments) across countries. Due to the fact that USA, Great Britain, Canada, and Mexico have different populations, we normalized the number of likes, dislikes, and comments to the total number of views to allow for a fair comparison between them.

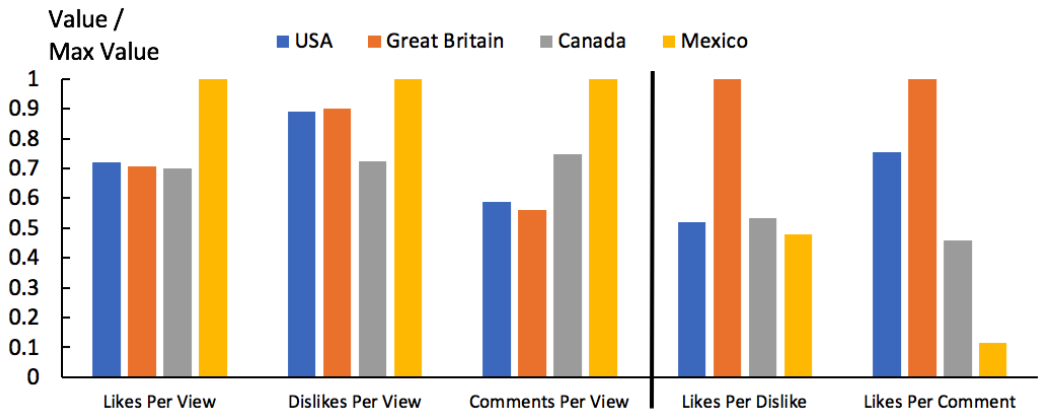

Figure 1: Normalized Community Engagement Per Country

As we can see in Figure 1, the country with the highest community engagement is Mexico, with the largest number of likes, dislikes, and comments per video by a significant margin. However, we can also see that, given an interaction (a like, dislike, or comment), the English are overwhelmingly more 'positive' than their counterparts, as evidenced by their significantly higher values of likes per dislikes and likes per comment. The USA and Canada seem to be the most similar with regards to community engagement, never being too far off from the average metrics. This makes sense, as we would think that Mexico and Great Britain would the most different from USA and Canada.

After this, we investigated the distribution of the most common pre-labeled generic categories across countries, as seen in Figure 2. 'Entertainment' is by far the most prevalent category within YouTube's trending videos, followed 'Music', and 'People + Blogs', after which the categories level off. Interestingly, it seems like Great Britain has a very high interest in music (about 40 percent of all trending videos), Americans have a relative high interest in 'Style'/Fashion (about double as everyone else), and Mexico has a similar relationship to 'Sports'. However, here we are using YouTube's pre-labeled categories, which, understandably, tend to be very generic and do not reflect the different interests within countries: 'Entertainment' might mean something different for Mexicans than to Americans. In the next section we try to dive deeper into country-specific topic identification based on the features extracted in the 'Data Processing' section.

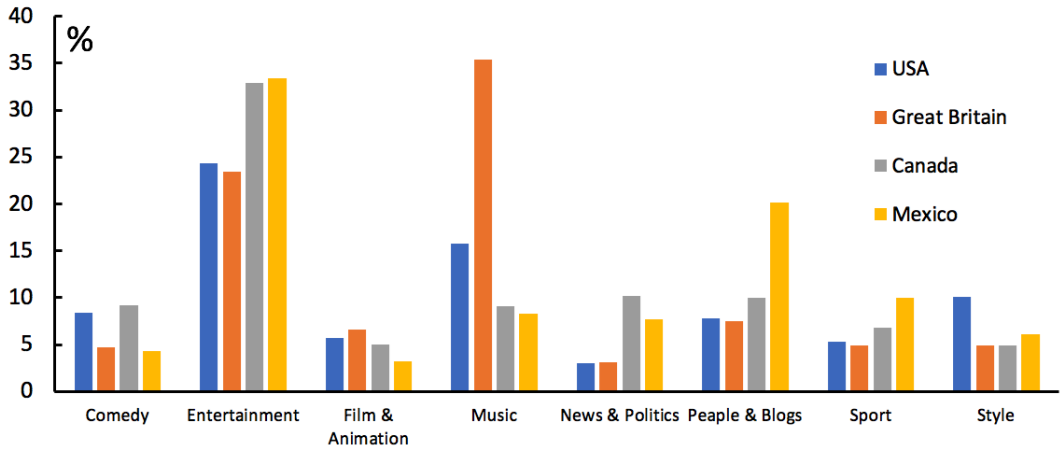

Figure 2: Percentage of Videos That are Part of a Given Pre-Established Category

Lastly, we were also curious about the distribution of unique videos throughout the different trending list. A simple investigation showed that over 200 days only 9 percent of all listed videos in Great Britain were new never-previously-seen-before videos. This value was 16 percent in the USA, 60 percent in Canada, and 83 percent in Mexico. This is very surprising, as it means that Great Britain's and the USA's video viewing market is dominated by a few constantly-trending videos. This conclusion is further reinforced by the fact that the average trending time was 12 days in Great Britain, 6 days in the USA, 2 Days in Canada, and 1.5 days in Mexico. We added a figure to the appendix that shows these relationships.

6.2 Latent Topics Discussion: Common and Differing Trends Across Countries

LDA was performed on the 'Latent Analysis Data Set' (BOW + likes + dislikes + views , ext...) to obtain 10 latent topics. We chose this amount of topics because we found it to be a low-enough number to ensure easy handling and presentation, while also giving diverse interpretable results that could be compared to the provided 'Category IDs'. For all trials, we made sure that the log-likelihood was consistent between the tested and held out data. Figure 3 shows our labeled latent topics results for each country.

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 (Jimmy Kimmel)**	**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

Figure 3: 10 Labeled LDA Latent Topics for All Countries. The bolded topics represent distinct topics between countries.

As expected and foreshadowed by our previous basic analysis, Mexico has a clear distinction in topics from the English-speaking countries, with topics such as 'Telenovelas/Soap Operas', 2 sport-related topics (Highlights + News), and an emphasis on Family Entertainment (distinct from just 'Entertainment' by a heavy emphasis on of features/words such as as 'family','kids', and 'God'). In the other hand, it seems that English-speaking countries all share an interest in Late-Night shows

(especially 'Jimmy Kimmel Live' in Canada), News, and Movies (with an emphasis on 'Star Wars' in Great Britain). Furthermore, we can see that all countries share several topics such as 'Entertainment', 'Music', and 'Social Media'. However, as mentioned before, this does not mean they are watching the same type of 'Entertainment', which is why we now compare the top 15 features belonging to this latent topic 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
ID_Entertainment	2	episodes	mis
This	CBS	full	ID_Entertainment
full	video	vs	POR
time	celebrity	Follow	ver

Figure 4: Top 15 Features Belonging to the 'Entertainment' Latent Topic for All Countries

Figure 4 shows the specifics of 'Entertainment' preferences, from which we get can obtain the following conclusions. 1) English-speaking countries consider 'Shows' to be good entertainment 2) The USA and Great Britain really like Late-Night shows (the key words 'night' and 'late' are ranked highly in several different topics). 3) Canada considers TV series to be good entertainment. 4) Mexicans consider that 'amor' (love), 'Mexico', and 'vivo' (live) are sought-after characteristics in their entertainment content.

A similar analysis on 'Sports' show that soccer is by far the most popular sport in Mexico, as shown by the prevalence of words such as 'futbol','fichajes', 'Real', 'Madrid', and 'Azteca', while sports media in the USA is dominated by a combination of basketball and football as shown by the words 'Lebron','James' 'NBA', and 'NFL'. Many of these conclusions are not surprising (some people may even say they are obvious), nevertheless we find it fascinating that we were able to identify these cultural trends with only this dataset and no a-priori knowledge. Note that this is not a ranking of the presented features, just a comparison between each countriy's grouping. Actual preferences and feature rankings will be presented in the next section.

6.3 Regression Discussion: Identifying Features that are Predictive of Future Trending Status

In this section we discuss the ability of three different classifiers to predict the future trending status of videos (i.e. what we call 'trendability'). As stated before, we used the features present in the "Regression Data Set" and evaluated their performance through accuracy, precision and AUC metrics, as seen in Figure 5. The mentioned 'top features' in said figure were selected by ranking the magnitude of feature coefficients obtained during Random Forest classification, and are shown in Fig 6.

It is clear that the Random Forest Classifier is the best predictor in all cases. Furthermore, we can see that the best trendability prediction performance is obtained by using all the features (BOW + categories + views + likes ext...). However, the whole analysis can be greatly reduced by using the number of views as the only feature, yielding very similar good performance metrics. However, this is not very useful to us, as the number of views is a variable that we can only obtain *after* the video has been posted. We are interested on the predictive capabilities of variables that we know a-priori

	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	**Accuarcy/Precision/ROC**			

Figure 5: Classifier Performance and the Effects of Feature Selection

(such as words and categories) which is why we expanded our classifier performance evaluation to consider only these features. As mentioned before, we also studied the predictive effects of full-word capitalization on a video's title and description. Our results show that trendability is virtually independent of capitalization, as all AUC values for a Random Forest predictor using capitalization as the only feature were about 0.5 (virtually indistinguishable from a random classifier). Fortunately, we were able to obtain the *actual* top 20 overall predictive features for all 4 countries.

Importance 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
16	show	percent_comments	video	**Resumen**
17	Watch	Show	song	video
18	**Show**	program	**food**	5
19	Music	percent_dislikes	music	1
20	Official	Official	In	category_id_Style

Figure 6: Top 20 Trendability Predictive Features for all Countries

Now we can start actually raking features by their appeal to a country, as opposed to just identifying them and grouping them together as we did in the previous section. For this analysis, we will ignore features such as views, likes, and comments (as these have more do more with YouTube's trending algorithm than actual cultural preferences). However, before we do that, we would like to point out the curious fact that Mexico's top predictive feature is the number of dislikes (as apposed to views like everyone else). This suggests that, more than anything, Mexicans watch (and probably share) videos that they consider to be controversial or that they disagree with.

Moving on, we can see that many of the top predictive features confirm and expand our past conclusions. This is a good sign that these conclusions are valid, as they are obtained through two distinct independent paths. In Great Britain, Music videos, Late-Night Shows, and Live Performances have a high probability of staying on trending for a long time. Canadians, in the other hand, have a consistent liking of videos related to TV show episodes, food, and News. USA preferences seem to be

really favor any sort of show and 'official' videos regarding music. Finally, this study suggests that Mexicans enjoy to watch music, sports, and comedy originating from Mexico while having little interest in the late night television.

Please bear in mind that the selected top features are only indicators of *continued* trendability, as opposed to indicators of *initial* trendability. This means that features such as 'capitalization' may very well be very influential in making a successful video. However, that analysis requires a significantly larger amount of data, as trending videos are a very small subset of the total amount of uploaded videos.

7 Conclusions

In this report, we used international data pertaining to trending videos on YouTube to identify sociocultural similarities and differences between the USA, Great Britain, Canada, and Mexico. To do this we used Latent Feature Analysis algorithms (such as LDA) to extract latent topics for each country based on a BOW representation of the videos' descriptions as well as count data related to views, likes, dislikes, and comments. We then took the investigation further by using classifier algorithms to identify the features that are good predictors of a video's continued trendability. A summary of the results now follows by presenting a hypothetical description of the most representative and successful trending video for each country.

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). Not, surprisingly, we found out this perfectly describes a real video on YouTube with about 50 million views titled: 'Jimmy Fallon, The Roots and 'Star Wars: The Force Awakens' Cast Sing 'Star Wars' Medley (A Cappella)'.

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. Not surprisingly, this is also a perfect description of a very successfully show on Netflix called 'Club de Cuervos'.

Finally, we also investigated if videos containing full-word capitalization had a higher chance of staying in the trending list for a longer-than-average time. An analysis based on Random Forest Classifiers showed that capitalization is not a predictive feature for continued trendability.

There are many possible future directions for this project. Some of them are: 1) Doing this same analysis on a larger data set encompassing videos that did not make it to the trending list. This way we can create a distinction between features that predict *initial* trending and features that predict *continued* trendability, as we did in this project. This would allow us to identify the features that would create an even better hypothetically-perfect trending video. 2) Expand our analysis to data composed of actual video comments, as opposed to just a comment count. This could allow us to do sentiment analysis and more complex natural language processing [4] on said videos, which we could correlate with the like, dislike, and share rates across countries. 3) Finally, we can think about matching advertisements/categories to each video category in individual countries in an effort to maximize the correlation between views and products sold.

References

[1] David Blei, Lawrence Carin, and David Dunson. Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6):55–65, 2010.

[2] Barbara Engelhardt. Precept #3.

[3] George W Furnas, Thomas K Landauer, Scott Deerwester, Richard Harshman, and Susan T Dumais. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 2004.

[4] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A Structured Self-attentive Sentence Embedding. 2017.

[5] Pedregosa Fabian, Vincent Michel, Grisel OLIVIER, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Jake Vanderplas, David Cournapeau, Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Bertrand Thirion, Olivier Grisel, Vincent Dubourg, Alexandre Passos, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

8 Appendix 1 Spotlight Method: Random Forest Classifier

Random Forest Classifiers (RF) are a combination of several 'Decision Tree' classifiers. These type of algorithms work by trying to find the optimal decision thresholds within a feature space in order to then predict the class of a given sample. A successful threshold is that which can minimize the entropy of a given training system the most (i.e. it maximizes the information gain from the creation of an additional decision threshold). In this case, the definition of entropy is virtually identical to the one provided and used in statistical thermodynamics:

$$Entropy = E(D) = -\sum_{i}^{m} p(x)_i log(p(x)_i)$$

Were $p(x)_i$ is the probability that a sample 'x' in a given subspace 'D' belongs to a certain class 'C_i'. From this definition we can then calculate the magnitude of 'information gain' obtained from the addition of a new threshold 'T':

$$Information\ Gain = E(Initial) - E_T(Separated)$$

In other words, the better we can separate labeled objects into their respective groups, the lower entropy we are going to have. However, the number of necessary decision thresholds is not know a-priori, which is why it is necessary to use optimization algorithms on the training data to optimize this hyper parameter's value. The goal is to minimize the amount of thresholds and the depth of the tree while also maintaining a high accuracy in order to avoid overfitting.

However, we also need to consider the fact that a partition with the greatest information gain is not always the most optimal, as this metric does not take into account the number of misclassified samples. An alternative approach involves the calculation and minimization of the Gini Index when fitting a decision tree:

$$Gini\ Index = G(D) = 1 - \sum_{i}^{m} p(x)_i^2$$

were $p(x)$ is the same as before, and where the value to be minimized is:

$$\Delta G = G(Initial) - G_T(Separated) = 1 - \sum_{i}^{m} p(x)_i^2$$

Now we can see that a single decision tree is fitted by: 1) Calculating the Gini Coefficient or 'Information Gain' for different possible thresholds based on the available features 2) choosing the most representative feature based on the previous value 3) separating the data into new sub-spaces, and 4) repeating for a given number of branches or until the data has been completely separated. Therefore, after fitting, each feature can be then associated with a coefficient (between 0 and 1) that represents the decrease in entropy associated by the addition of a decision threshold within said feature. This makes it easy to rank the features in order of importance, which makes for easy data interpretability.

However, even with hyper-parameter optimization algorithms, single tree classifiers are prone to over-fitting, which is why we turn to RF classifiers. A 'random forest' is created by tallying the results of several decision trees fitted to a randomly selected sub-group of features. Then, given an input, a RF makes a final decision based on a majority vote. This reduces over-fitting by preventing over-dependance on a few initial branches or roots.

Random Forests can also be used for continuous regression, but we don't discuss that here for brevity.

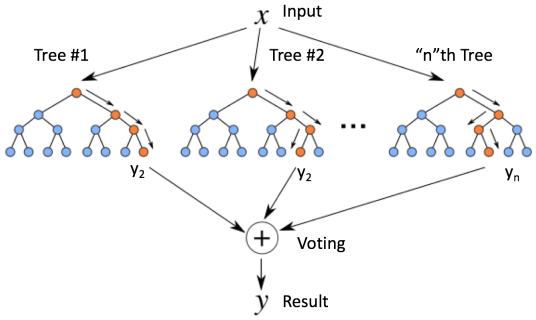

Figure 7: Conceptual Representation of a Random Forest Classifier. Obtained from: https://dsc-spidal.github.io/harp/docs/examples/rf/)

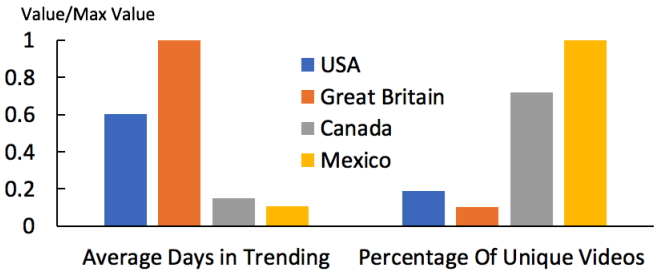

Figure 8: Normalized Trending Characteristics