Topics and Streaming: What makes a positive review?

Introduction

The last decade has seen countless streaming services from different companies appear in the market, each with their own advantages and challenges. Most of these services are distributed through the current prevailing mobile operating systems: Android and IOS, and have generated thousands of reviews within their respective app stores. Most of these apps are relatively similar and have similar content offerings, but some are overall much more successful than others.

This study explores the topics for different app store rating levels and how they differ by streaming service. The hypothetical use for one of these companies would be to influence future reviews by reinforcing the good aspects or correcting problems each of the services have. It essentially provides an overview of what makes users give positive or negative ratings.

Methodology

Data acquisition:

Data was downloaded for the current leading streaming services: Netflix, HBO GO/NOW, Hulu and Amazon Prime Video from Heedzy, a service that collects user reviews from the Android and IOS app stores.

Reviews get downloaded in individual CSV files by platform with the following fields:

Content:	Body of the review
Date:	Date the review was made
Name:	Username of the reviewer
Rating:	5-point rating scale used by the IOS and Android app stores
Source:	Streaming App name
Title:	Title given to the review
Version:	Version of the app that is being reviewed

The original data set, after combining all platforms, contained 157,138 examples.

Pre-processing:

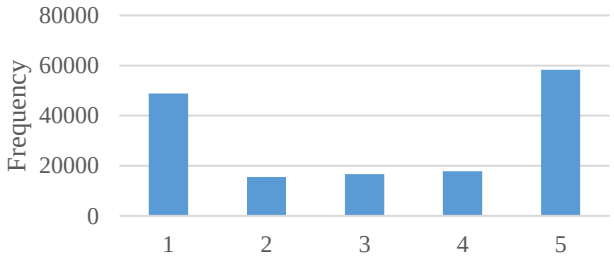
Pre-processing steps included adding a variable for operating system; normalizing platform names (e.g. "HBO GO: Stream with TV Package" to "HBOGO") and discretizing the review scores to balance the data set. All text variables were encoded into ASCII in order to remove special characters.

To summarize, the following variables were added:

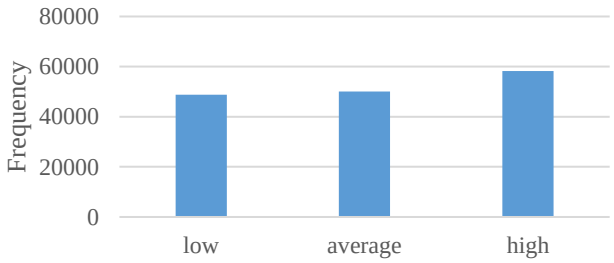
OS:	Operating system
Platform:	Normalized platform names
DiscRating:	Discretized ratings (high = 5; low = 1; average = 2,3,4)
Mcontent:	Concatenation of "title" and "content"

Marginal distribution:

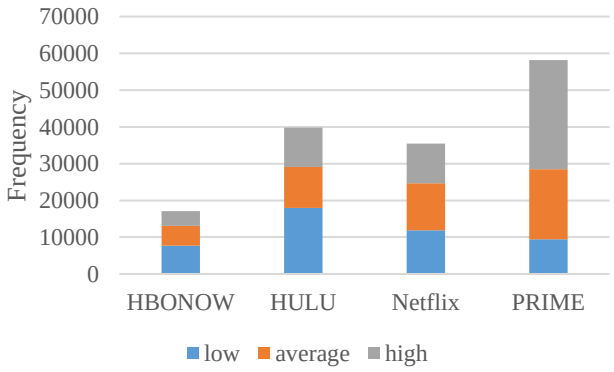
In terms of reviews the original data set was skewed towards the extreme scores:



Discretizing helped balance the data set, although it remained slightly skewed towards positive reviews:



In terms of the platform, Amazon prime clearly has the largest amount of reviews (39% of the data set)



Additional pre-processing and parameters for topic modeling

Previous to running the LDA topic model on the reviews, the data set was split into individual collections of reviews by platform and review level (e.g. Netflix + low reviews)

Upon inspection of the initial, exploratory, topic models it was found that the names of the platforms appeared too much in the topics, so additional stop words were used to remove them (e.g. “Netflix netflix hbo go now hulu prime amazon”).

Reviews denoted as “average” were removed before generating the final set of topic models in order to remove noise and concentrate on what constitutes high and low scores.

Other parameters include the usage of the optimize-interval 20 parameter and a random seed of 10.

Vectorization and evaluation methods for classification

For the classification portion of the study, reviews classified as “average” were also excluded in order to maximize the accuracy of the model. Several vectorization options were tested for both unigrams and unigram+bigrams, all with a minimum of 5 document frequency and English stop words:

- Boolean representation
- Term frequency representation
- L2 normalized term frequency
- TFID representation

A combination of 3-fold cross validation and a holdout sample was used to evaluate the model performance. 60% of examples were put into a training set and subsequently used to train multinomial naïve Bayes (NB) and linear support vector machine (SVM) models on the above mentioned vectorization options. The best performing one is the one tested on the hold out sample. As some of the more popular algorithms for text mining, these were a logical choice for initial testing.

Results

LDA Topic model

For this interpretation a total of 2 sets of 5-topic and 10-topic models were built, using two different sets of stop words. The original set only had the names of the platform removed, while the second also removed common words across the topics like "movie" or "show".

The following topics and topic weights were interpreted from the final 10-topic model:

Topic	Topic Weight	Interpretation
0	0.99982	Good Price + Live programming
1	0.53971	IOS crashing
2	0.37086	Buffering and crashing with free trial (Game of Thrones)
3	0.52133	High quality content
4	0.19784	Option for offline viewing
5	0.25398	Great content (avatars + anime)
6	1.77728	Streaming on apple devices
7	0.48341	Interface + commercials
8	1.13591	Sign-in issues
9	0.2151	Casting & download issues

Topic distribution among the text documents:

Review	Platform	0	1	2	3	4	5	6	7	8	9
High	HBONow	26%	1%	12%	**33%**	0%	6%	16%	0%	5%	0%
	Hulu	**27%**	1%	1%	23%	0%	22%	15%	8%	4%	0%
	Netflix	13%	7%	0%	8%	4%	**51%**	12%	0%	6%	0%
	Prime	10%	1%	0%	**63%**	9%	2%	12%	1%	2%	2%
Low	HBONow	2%	1%	24%	0%	0%	0%	32%	3%	**37%**	0%
	Hulu	4%	2%	2%	0%	0%	0%	**34%**	31%	26%	0%
	Netflix	4%	33%	0%	0%	1%	2%	23%	1%	**35%**	1%
	Prime	2%	3%	3%	1%	3%	0%	**35%**	2%	20%	31%

Through cross referencing with the rest of the models, the following themes emerged by platform:

- HBO Go / HBO Now
    - High reviews are associated with content quality and a curated selection, some of the models also indicated Chromecast support as a secondary topic associated with the reviews. In this scope, it often shared these topics with Prime.
    - Low reviews concentrated on casting support, sign-in issues with apple devices and high price

- Hulu
  - High reviews for Hulu were slightly different than HBO's, although they're still associated with content, its users talk about the service having their "favorite" shows. High scores for Hulu also include mentions of their anime content.
  - Low user scores for this service deal mostly with the interface and commercials. These are, for the most part, exclusive to Hulu.
- Netflix
  - Netflix shares most of its high review topics with Hulu with the notion of the service having the user's "favorite" shows. Curiously, however, Netflix has a unique version of the "content quality" topic that deals with their profile avatars, which they recently renewed.
  - Low scores for Netflix are overwhelmingly associated with sign-on problems on iOS devices.
- Amazon Prime Video
  - High reviews for Prime are similar to HBO's, but are even more focused on the high quality content and variety. In the topic distribution printed above it is evidenced how the weight for "high quality content" is 33% for HBO but 63% for Prime.
  - Low rated reviews are a mix of HBO and Netflix for Prime, they show a combination of issues casting and issues with Apple devices as the main topics.

Even though each platform has its own "flavor" of these themes. Overall, high reviews seem to be associated with content and low reviews seem to be associated with technical issues

Classifier performance and feature weights

The 3-fold CV accuracy for the results show that the best vectorization option was the normalized term frequency and the best algorithm was the linear SVM:

1. 90.943% svm unigram l2 tf
2. 90.658% svm unigram tfidf
3. 90.641% svm gram12 l2 tf
4. 90.450% mNB gram12 l2 tf
5. 90.434% mNB gram12 tfidf
6. 90.293% mNB unigram l2 tf
7. 90.168% mNB unigram tfidf
8. 90.152% svm gram12 tfidf
9. 89.998% mNB gram12 tf
10. 89.958% mNB unigram tf
11. 89.751% svm boolean unigram
12. 89.320% svm unigram tf
13. 89.015% svm boolean gram12
14. 89.004% svm gram12 tf
15. 83.984% bernouli boolean gram12

Achieving an accuracy of over 93% on the hold out sample with a slightly higher f-measure on positive reviews:

	precision	recall	f1-score	support
high	0.93	0.94	0.94	22104
low	0.92	0.92	0.92	18790
avg/total	0.93	0.93	0.93	40894

The most indicative features for these classifications are very intuitive. Positive and negative adjectives of different magnitudes:

High review features

1. (-2.4157277276025018, 'excellent')
2. (-2.2070088152230767, 'complaining')
3. (-2.193611078462756, 'fantastic')
4. (-2.140913450145914, 'amazing')
5. (-2.1095811418456343, 'wonderful')
6. (-2.0719787334030095, 'awsome')
7. (-1.9757631667333424, 'helps')
8. (-1.952110091578311, 'complaints')
9. (-1.9417816965755657, 'awesome')
10. (-1.9196250390409948, 'worry')

An interesting term among these top 10 most indicative terms for high reviews are the words “complaining” and “complaints.” In this case it is people writing “no complaints” or “zero complaints.” This term could be improved with ngram representation as there are many low score reviews that contain the term (low = 94; high = 161).

Low review features

1. (2.9218205341280465, 'garbage')
2. (2.740491582512577, 'awful')
3. (2.6492256899786026, 'horrible')
4. (2.440259870959328, 'ridiculous')
5. (2.409038215325966, 'terrible')
6. (2.4060348515271692, 'poor')
7. (2.3594921221522287, 'sense8')
8. (2.3129117805348396, 'crap')
9. (2.284609108447469, 'worse')
10. (2.205343964966438, 'pointless')
11. (2.152023141405285, 'worthless')

“Sense8” stands out among the indicative terms for low reviews. It is a Brazilian Netflix series that was recently cancelled and seems to have driven a lot of fans to lower their scores for Netflix. All reviews that include sense8 are low scoring (low = 18; high = 0)

Conclusions & Discussion

Each of the 4 services for which reviews were evaluated show similarities in what drives their app store reviews. At its most basic level, high scores mostly deal with the content offering and low reviews mostly deal with technical problems (e.g. Chromecast support). Both can be predicted with high accuracy depending on the adjectives used within the body of the review or the title.

Some interesting findings appeared, however. Even though there were overarching similarities in the topic results. For example, it became evident that anime is one of the main topics in high scoring reviews for Hulu, and negative Netflix reviews show high correlation with the iOS app crashing. The classifier achieved a high accuracy (93%) and could help score text reviews that don't have scores associated with them as it benefits from a large vocabulary.

In conclusion, the hypothesis is confirmed in a way. There are nuances to the topics that make up high and low reviews across the current leading streaming services but, at the same time, the best predictors of high and low reviews are adjectives that permeate through all the reviews.

Appendix 1: Unexecuted code and download links

Mallet: commands to train the topic models

bin\mallet import-dir --input sample-data\TextMiningProject\by_review --output TopicsByReview.mallet --keep-sequence --remove-stopwords --extra-stopwords my-stoplist.txt

bin\mallet train-topics --input TopicsByReview.mallet --num-topics 10 --optimize-interval 20 --random-seed 10 --output-state TopicsByReview-topic-state.gz --output-topic-keys TopicsByReview_keys.txt --output-doc-topics TopicsByReview_compostion.txt

bin\mallet import-dir --input sample-data\TextMiningProject\by_review_platform --output TopicsByPlatformReview.mallet --keep-sequence --remove-stopwords --extra-stopwords my-stoplist.txt

bin\mallet train-topics --input TopicsByPlatformReview.mallet --num-topics 10 --optimize-interval 20 --random-seed 10 --output-state TopicsByPlatformReview-topic-state2.gz --output-topic-keys TopicsByPlatformReview_keys2.txt --output-doc-topics TopicsByPlatformReview_compostion2.txt

bin\mallet train-topics --input TopicsByPlatformReview.mallet --num-topics 20 --optimize-interval 20 --random-seed 10 --output-state TopicsByPlatformReview-topic-state3.gz --output-topic-keys TopicsByPlatformReview_keys3.txt --output-doc-topics TopicsByPlatformReview_compostion3.txt

R: Script to combine and transform the reviews

https://drive.google.com/open?id=148i3mHhRbZDjBs63hyfTe0oeu9Qt5A1B

Python: Script to vectorize the reviews and train the algorithms

https://drive.google.com/open?id=13tj4XO4XhRqG74biSyY9l_iH4Lc0YMuE

Original data sets:

Text documents from Heedzy

https://drive.google.com/open?id=1Ji6fE6PEHz4j-pJcvUZVQwGbj8esDYtY

Pre-processed txt for Mallet

https://drive.google.com/open?id=12sumqHoSrMCSy4-WbcUwMRLK36NXsrca

Topic models compared:

https://drive.google.com/open?id=1a_TwUmWIpAKnIYu2JwGUNIt3kVbEdJEL