Japanese Whisky Review Analysis

IST 736 Final Project Report

Xueqing Wang
Minyang Wang

Introduction

Over the last century, globalization allowed economic trades to circulate between countries. Native country products can be sold to the other side of the world. Japanese whisky as one of the most representative Japanese products is beloved by the western world. Among all type of whiskeys, whisky sellers need to know which type are more favored by the customers. Instead of reading through every words of comment, a more efficient analyzation method is adopted – machine learning text mining algorithms. To extract the key words from the huge amount of text, topic modeling and classification mythologies are applied for analyzing the whiskey reviews.

Method

1. Data

The data set was download from Master of Malt website (Master of Malt, 2018), an online whisky selling website. From all Japanese whisky, Yamazaki Sherry Cask 2016, Yamazaki 12 Years Old and Nikka Whisky form the Barrel were choose as the experiment objects. Among all types of Japanese whisky reviews, Nikka and Yamazaki take 75% of the reviews. In addition, Among Yamazaki whiskeys, Yamazaki Sherry Cask 2016, Yamazaki 12 Years Old have more than 60% of the reviews. In Nikka whiskeys, Nikka Whisky form the Barrel also has more than 60% of the reviews. Therefore, Yamazaki Sherry Cask 2016, Yamazaki 12 Years Old and Nikka Whisky form the Barrel are three typical whiskeys that can be studied as representatives of Japanese whiskeys.

Table 1 Whisky Distribution

	Yamazaki	**Nikka**	**Hibiki**	**Hakushu**	**Total**
Number	457	392	196	85	1130

Table 2 Yamazaki Whisky Distribution

	12 Years Old	**Sherry Cask**	**18 Years Old**	**Single Malt**	**10 Years Old**	**Total**
Number	126	123	84	57	18	408

Table 3 Nikka Whisky Distribution

	From the Barrel	**Pure Malt**	**21 Years Old**	**12 Years Old**	**17 Years Old**	**Total**

Number	150	42	32	23	22	269

The original data set includes 4 attributes: Bottle name, Brand, Title and Review Content. Bottle name and Brand indicates the brand and name of whisky, the attribute Title is the title of the reviews written by the customers. The experiment data set is a new data set that filtered out only Yamazaki Sherry Cask 2016, Yamazaki 12 Years Old and Nikka Whisky form the Barrel 3 kinds of whiskeys. In this experiment, since the study focuses on applying topic modeling and classification models to analyze the whisky features by reviews, title attribute is ignored in this case.

2. Experiment Procedures

1) Topic Modeling Toolkit

To conduct the topic prediction model for the Japanese Whisky, Latent Dirichlet Allocation model (LDA) (Blei, Ng and Jordan, 2003) was applied to predict the clustered words that frequently occurred together. When the use of words are frequently occurred as a pattern, the cluster of words would be categorized together under a potential “topic”. The tool we used in our topic modeling test is a machine learning toolkit named “MALLET” (McCallum, 2002). MALLET is a Java package for language processing, classification and topic modeling toolkit.

Mallet provides lists of stop words, which are very comprehensive. However, according to the natural of the review comments, additional stop words were added to the list to enforce more efficient prediction. The following words were added to the list: it’s, abv, yesterday, today, tomorrow, that’s, I’m, I’ve, you’ve, year.

2) Topic Modeling Experiment

For topic modeling of the reviews, there are two sub-studies. The first study was conducted to test whether the machine learning algorithms can categorize the three kinds of whisky from the mixed review comments. The second study was about letting the algorithms to predict topics to each whiskey to see if the algorithms can analyze the reviews by topics.

For the overall topic modeling prediction, since the goal is to categorize three kinds of whisky, the number of topics was aimed to set as three. In order to better predict the topics, optimize interval was adjusted to 15 and 10.

For each whiskey model prediction, the number of topics and the optimize interval settings were tuned to receive better results. The experiment first set the topic number as five and tuned the optimize interval of 15 and 10 to see which interval provided a smoother topic. Then, the number of topics will be adjusted from 5 to 3 to 2 to see for each whiskey review comments which number of topics fits better.

3) Classification

We used SVM and MNB models to learn the feature words of each whisky bottle’s reviews. To tokenize the reviews, we removed the stop words because they don’t have any meanings and won’t contribute to the review classification. We tried 3 tokenizers for both SVM and MNB models: uni-gram term frequency tokenizer, n-gram term frequency tokenizer (minimum number of words per token = 1 and maximum number of words per token =3) and uni-gram TFIDF tokenizer to see whether different number of words per token can affect the classification accuracy, and whether different measures of counting tokens can influence the classification accuracy. After building the models to get the patterns of each bottle's reviews, we evaluated how well those tokens can distinguish the reviews of different bottle by 3-fold cross validation and hold-out test. For the hold-out test, we split the data into training group and testing group by ratio of 2:3(40% testing data and 60% training data).

Results

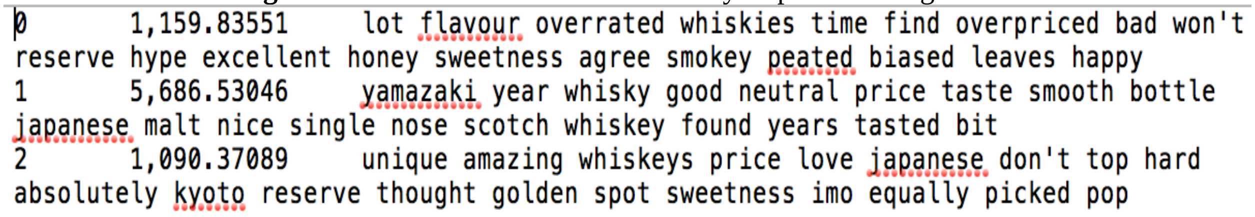
1. Topic Modeling Results

For the Overall topic modeling study, when optimize interval was set to 10, the algorithms better categorized three types of whiskey. From the results, the first group with key words, such as Nikka, barrel, nose, water, finish, long, complex, helped to interpret that group one is Nikka Whiskey From the Barrel. The second group with key words of Sherry, auction, lottery, price proved that this group is Yamazaki Sherry 2016, since most of the review comments of this whiskey are about price and purchase methods. The third group has key words of good, neutral, nice, smooth, flavor. Yamazaki 12 Years Old matches the key words of description that the majority reviews talked about the tastes of the whiskey. In conclusion, the overall topic modeling study successfully categorized three whiskeys from each other.

On the other hand for each whiskey topic modeling study, the best tuning results indicated when optimize interval equals 15, as the algorithms calculation predicts the words more smoothly, the better the results return.

For Yamazaki 12 Years Old whiskey, the topic modeling model predicts three groups of words that can be categorized into: Overrated, Neutral and positive judgement. From the "overrated" topic group, words such as "overpriced" and "hype" represent that some customer consider the product is overrated. The second group gives some non-sentiment words such as "bottle" and "single". In addition to the word "neutral" from this group, the reviews showed neutral rating towards this whiskey. In the third group, words like "unique", "amazing", "love", "absolutely" occurred, therefore this group can be summarized as positive feedbacks.

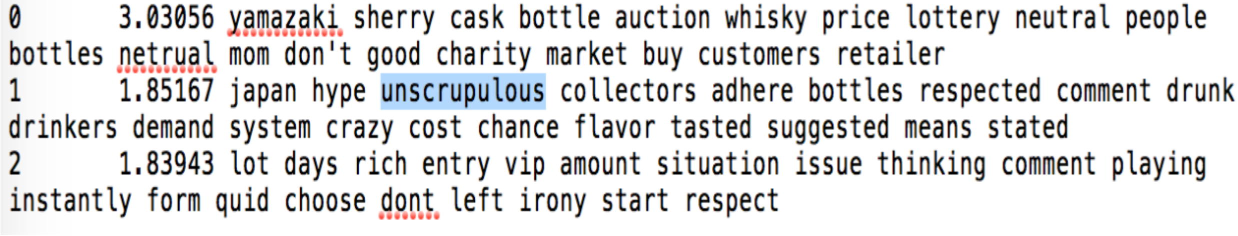
Figure 1 Yamazaki 12 Years Old whiskey Topic Modeling Result



```
0       1,159.83551     lot flavour overrated whiskies time find overpriced bad won't
reserve hype excellent honey sweetness agree smokey peated biased leaves happy
1       5,686.53046     yamazaki year whisky good neutral price taste smooth bottle
japanese malt nice single nose scotch whiskey found years tasted bit
2       1,090.37089     unique amazing whiskeys price love japanese don't top hard
absolutely kyoto reserve thought golden spot sweetness imo equally picked pop
```

For Yamazaki Sherry 2016, the best results has the same setting of number of topics of three and optimize interval of 15. From the reviews, it can summarized that this whiskey is hard to purchase that either have to bet through lottery or purchase through the auctions. From the topic modeling results, the three groups outstandingly categorized the review comments from three perspectives. The first perspective talks about the purchase method that the first group contain words such as "lottery", "price", "auction", "buy". This group can be labeled as "Lottery/Auction". The second group of this whiskey has many interesting words: "hype", "unscrupulous", "adhere", "demand", "system", "crazy" and "cost". The words refers to many clients' attitude towards this whiskey that the high demand system caused the supplier to mark the price crazy and it has been advertised much over that the price has caused public's unscrupulous attitude. The last group of words is tricky. It contain words that seems not really related to others but in fact it requires analyst to go back to the reviews to analyze the meaning of the choices of words. The last group can be named after the topic of seller's behavior. Some respected collector or sellers chose to save this whiskey to the rich VIP customers.

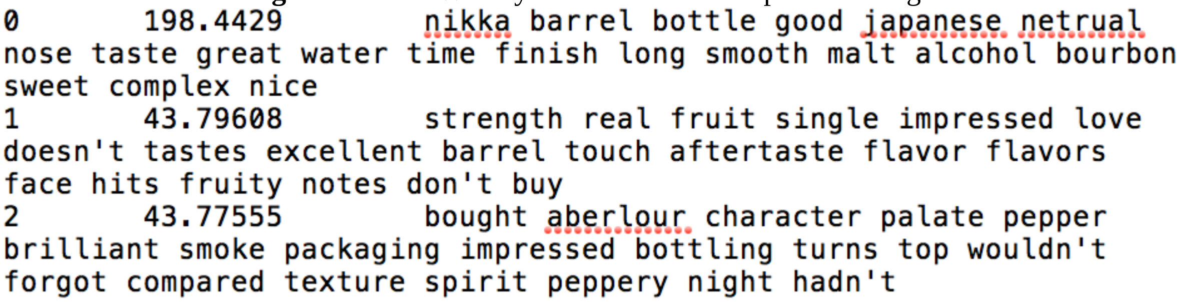
Figure 2 Yamazaki Sherry 2016 whiskey Topic Modeling Result



```
0       3.03056 yamazaki sherry cask bottle auction whisky price lottery neutral people
bottles netrual mom don't good charity market buy customers retailer
1       1.85167 japan hype unscrupulous collectors adhere bottles respected comment drunk
drinkers demand system crazy cost chance flavor tasted suggested means stated
2       1.83943 lot days rich entry vip amount situation issue thinking comment playing
instantly form quid choose dont left irony start respect
```

In addition, the Nikka whiskey's reviews are more subject to the exact taste of the whiskey. The best results still has the same setting of number of topics of three and optimize interval of 15 as the previous two whiskeys. In the first group of words, the algorithm gives words such as "nose", "taste", "water", "finish", "long", "smooth", " neutral" and "complex". This group of words mostly describe how the whiskey feels instead the exact smell, therefore this group can be named as taste. The second group can be named as fruity, since the words fruit or fruity appeared twice. The third group can be named as pepper, since there are "pepper", "smoke" , "texture" and "peppery" words. From the second and third group of words, Nikka Whisky From the Barrel can be identified as both fruity and peppery smell or taste.

Figure 3 Nikka Whisky From the Barrel Topic Modeling Result



```
0       198.4429        nikka barrel bottle good japanese netrual
nose taste great water time finish long smooth malt alcohol bourbon
sweet complex nice
1       43.79608        strength real fruit single impressed love
doesn't tastes excellent barrel touch aftertaste flavor flavors
face hits fruity notes don't buy
2       43.77555        bought aberlour character palate pepper
brilliant smoke packaging impressed bottling turns top wouldn't
forgot compared texture spirit peppery night hadn't
```

From the results, it can be concluded that better results for overall comments occur when optimize interval is 10, and better results of each whiskey were interpreted when optimize interval is 15. When running the algorithm in Mallet for overall test, the optimize interval setting does not need to be set as a higher number since the goal is just to determine the general characteristics of each whiskey. When analyzing the comments of each whiskey, since the calculation is aims to interpreted specific characteristics of the whiskey, therefore the optimize interval needs to be higher than the optimize interval of general setting.

2. Classification results

1) Classification accuracy

Table 4 Classification accuracy of different models

	3-fold cross validation			Hold-out test		
	Uni-gram term frequency	N-gram term frequency	Uni-gram TFIDF	Uni-gram term frequency	N-gram term frequency	Uni-gram TFIDF
SVM	59.6%	59.6%	68.9%	63.9%	63.2%	66.5%
MNB	72.6%	72.6%	70.8%	69.7%	69.7%	62.6%

The overall average accuracy of the whisky bottle prediction is around 70%, whereas the bottle distribution baseline is 40%, meaning that the patterns and feature words learned by these 2 classification models can well distinguish the reviews of different bottles. For both SVM model and MNB model, there wasn't significant difference between the accuracy of uni-gram tokenizer and and n-gram tokenizer, and all the top feature words learned by these models are uni-gram tokens. For the SVM models, tokens generated by uni-gram TFIDF tokenizer can better classify the whisky's reviews than the tokenizer which uses term frequency as the tokens counting measurement. But for the MNB models, tokens generated by term frequency tokenizers have higher prediction accuracy in both cross validation and hold-out test.

Figure 4 Bottle distribution of the reviews

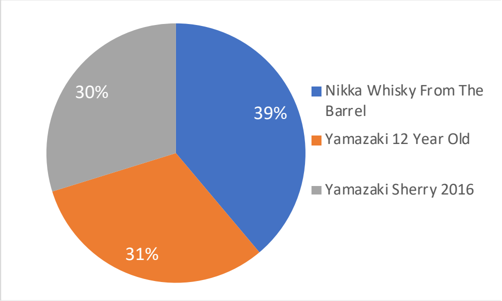
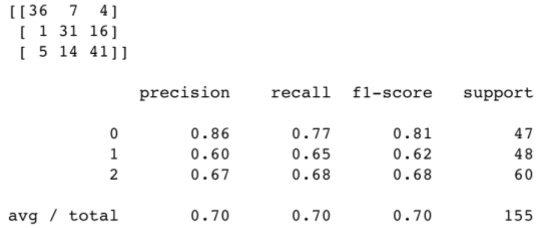
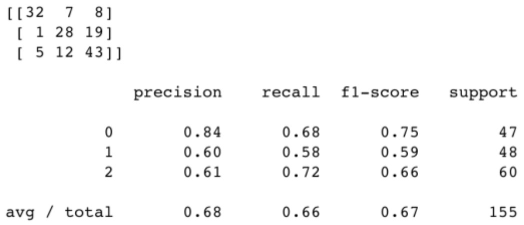


Figure 5 Confusion matrix of SVM model



```
[[36  7  4]
 [ 1 31 16]
 [ 5 14 41]]

             precision    recall  f1-score   support

          0       0.86      0.77      0.81        47
          1       0.60      0.65      0.62        48
          2       0.67      0.68      0.68        60

avg / total       0.70      0.70      0.70       155
```

Figure 6 Confusion matrix of MNB model



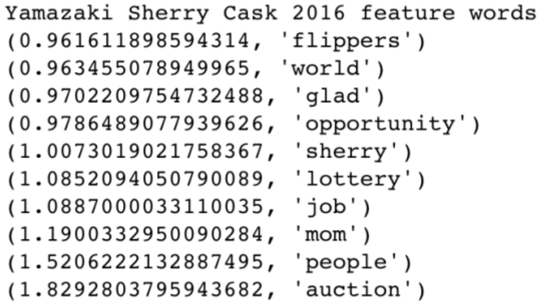
```
[[32  7  8]
 [ 1 28 19]
 [ 5 12 43]]

             precision    recall  f1-score   support

          0       0.84      0.68      0.75        47
          1       0.60      0.58      0.59        48
          2       0.61      0.72      0.66        60

avg / total       0.68      0.66      0.67       155
```

According to the confusion matrices of the hold-out test of SVM and MNB models with the highest accuracy of each model, we found that the precision of the category 0(Yamazaki Sherry 2016) is the highest in both model, meaning that for all the reviews categorized as Yamazaki Sherry 2016, a very large portion are correctly classified, so the patterns and feature words of this type of whisky learned by the classification models are the most distinguishable. We also noticed that around 30% of reviews of category 1(Yamazaki 12-Year-Old) and category 2(Nikka Whisky from the Barrel) were classified to the other category, so we thought that these two types of whisky share some similarities, which caused the models wrongly classified them.

2) Feature words of each category

We chose the feature words learned by SVM model generated by TFIDF tokenizer trying to summarize the features of each whisky bottle and to see why the accuracy of classifying Yamazaki Sherry 2016 is higher than the other two type of whisky.

Figure 7 Yamazaki Sherry 2016 feature words



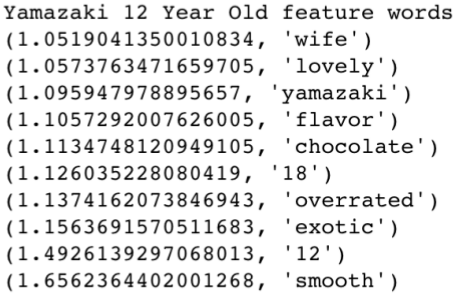
```
Yamazaki Sherry Cask 2016 feature words
(0.961611898594314, 'flippers')
(0.963455078949965, 'world')
(0.9702209754732488, 'glad')
(0.9786489077939626, 'opportunity')
(1.0073019021758367, 'sherry')
(1.0852094050790089, 'lottery')
(1.0887000033110035, 'job')
(1.1900332950090284, 'mom')
(1.5206222132887495, 'people')
(1.8292803795943682, 'auction')
```

Table 5 Feature words document frequency

Document frequency	Yamazaki Sherry 2016	Yamazaki 12-Year-Old	Nikka from the Barrel
Auction	33%	0	0
Lottery	25%	0	0
Mom	21.7%	0.8%	2%
Charity	10.4%	0	0
Flippers	7%	0	0

Some words such as “auction”, “lottery”, “MOM”, “charity”, “flippers” appeared uniquely in the reviews of Yamazaki Sherry 2016. As we have mentioned above in the topic modeling section, the majority of the reviews of this type of whisky are the customers’ opinions about the method that this online shop sells this whisky bottle. We tracked these unique words back to the reviews and found that the customers were satisfied with the lottery method (each customer pays a tiny amount of money and the winners can get the whisky), but they disagreed with the auction method because in that way only the rich can buy it.

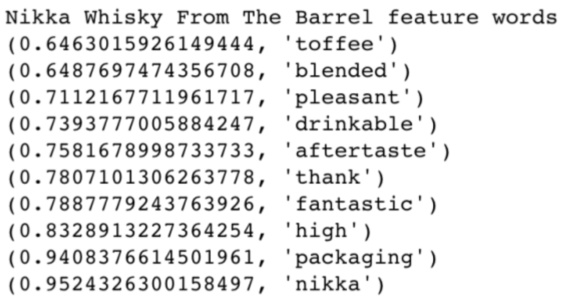
Figure 8 Yamazaki 12-Year-Old feature words



```
Yamazaki 12 Year Old feature words
(1.0519041350010834, 'wife')
(1.0573763471659705, 'lovely')
(1.095947978895657, 'yamazaki')
(1.1057292007626005, 'flavor')
(1.1134748120949105, 'chocolate')
(1.126035228080419, '18')
(1.1374162073846943, 'overrated')
(1.1563691570511683, 'exotic')
(1.4926139297068013, '12')
(1.6562364402001268, 'smooth')
```

For the feature words of Yamazaki 12-Year-Old bottle, words “wife” appeared uniquely 4 times in the reviews of this whisky, and all these 4 reviews mention that the whisky was bought by the customer’s wife (“… My wife searched for a bottle for me …”, “…Received as a birthday gift from my wife and…”, “… I was surprised and disappointed when my wife bought me a bottle…”, “…have purchased a 12 on recommendation of my wife…”). The word “chocolate” also appeared uniquely in the reviews of this bottle, from the reviews such as “Love this one. Perfect with chocolate”, “The chocolate vaguely reminds me of…”, and “We ate chocolate right before tasting this. So we didn't get much citrus.” we learned that people like to eat chocolate when drinking this whisky. The word “18” appeared in 12.4% of all reviews of this type of whisky, people tended to compare this whisky with Yamazaki 18-Year-Old and the two share some similarities, but the former is cheaper than the latter (“…prefer it to the Yamazaki 18 years…”, “…The 18yo is very interesting on the other hand. a bit pricy though…”, “…Wondering if it was worth it to purchase the Yamazaki 18…”).

Figure 9 Nikka Whisky from the Barrel feature words



```
Nikka Whisky From The Barrel feature words
(0.6463015926149444, 'toffee')
(0.6487697474356708, 'blended')
(0.7112167711961717, 'pleasant')
(0.7393777005884247, 'drinkable')
(0.7581678998733733, 'aftertaste')
(0.7807101306263778, 'thank')
(0.7887779243763926, 'fantastic')
(0.8328913227364254, 'high')
(0.9408376614501961, 'packaging')
(0.9524326300158497, 'nikka')
```

For the feature words of Nikka Whisky from the Barrel, similarly, word “toffee” (“…with a wonderful sweet toffee like finish…”, “…gives you a beautiful warm hug and a toffee kiss…”) and “packaging”(“…and excellent stand out packaging…”, “…Love the clinical packaging…”) appeared uniquely in the reviews of this type of whisky with around 8% document frequency. The word “aftertaste” also appeared much more frequently in the reviews of this type of whisky than the others, and from the reviews we found that this whisky has a very strong and lasting long aftertaste (“…aftertaste is somewhat killed by the burning sensation…”, “…Aftertaste goes on forever…”, “…wonderful aftertaste with a long and smooth taste of citrus…”).

Conclusion

In this project we applied topic modeling and classification model to summarize the features of each different whisky bottle.

The topic modeling results examined that among a mixed whiskey review list, the LDA algorithms can distinct each whiskey away from the others. The sellers can extract the reviews without separate the reviews by whiskey to get each whiskey’s major relative review words. To see specific review topics, the LDA algorithms can distinct the major topics after calculation and provides some story telling points of view to each whiskey.

The classification results indicate that the majority of customers of Yamazaki Sherry 2016 are not satisfied with the method that this online whisky shop sells this type of whisky, people prefer to buy this limited release product by lottery so that everyone has even chance of buying it, rather than buy it by auction. So we learn that for product with limited amount of release, lottery receives more welcome than auction for most customers. Although the website can sell the product both by lottery and auction, more portion of products are suggested to be sold by lottery.

Based on the feature words generated by classification models, Yamazaki 12-Year-Old is a product which can be an optional good gift for the husband if he likes whisky, and it tastes better when pairing with chocolate. Comparing with another similar product Yamazaki 18-Year-Old, the 12-Year-Old one is more competitive because of its low price. The Nikka Whisky from the Barrel is a blended whisky with a unique toffee taste, its long-lasting aftertaste and beautiful package make this whisky welcomed by almost every customer who has bought it.

References

Blei, D., Ng, A. and Jordan, M. (2003). *Journal of Machine Learning Research 3*.

Master of Malt. (2018). *Yamazaki 12 Year Old*. [online] Available at: https://www.masterofmalt.com/whiskies/yamazaki/yamazaki-12-year-old-whisky/?srh=1 [Accessed 5 Dec. 2018].

McCallum, A. (2002). *MALLET: A Machine Learning for Language Toolkit.*.