Issues in Information Systems
Volume 23, Issue 3, pp. 64-68, 2022

DOI: https://doi.org/10.48009/3_iis_2022_107

Predictions of wine ratings using natural language processing

Ben Kim, *Seattle University*, *bkim@seattleu.edu*

Abstract

In this paper, we built the machine learning models (Decision Tree, Gradient Boosting, and Random Forest) to predict the quality of wines by text-mining the verbal descriptions of sommeliers as well as grape varieties, countries, and sommeliers. We applied the natural language processing to verbal descriptions or commentaries accompanying the wine ratings. We found that verbal descriptions can predict the quality of wines more accurately than the price. Using our models, we believe that wine ratings can represent humans' subjective feelings about wines as well as their objective chemical compositions.

Keywords: Wine ratings, Natural Language Processing, Text Mining, Decision Tree, Gradient Boosting, Random Forest

Introduction

we used the verbal descriptions of wines to predict the quality of wines. Most studies on predicting wine qualities use chemical compositions or sensory information. Instead, we analyzed the descriptions by natural language processing (NLP). Wine ratings are determined by collective opinions of multiple sommeliers. As those opinions are given in verbal descriptions, our ML model can automate the wine rating process. Wine consumers or connoisseurs can also use this ML model to produce their own ratings.

In this paper, we first showed the background of wine ratings and literature reviews. Then we explained how the nominal data were converted to numerical data matrix using NLP algorithms (known as the Document Term Matrix). Some examples of nominal data are shown in Table 1. Then we created five different datasets to understand how each set of features determine the wine ratings. To analyze these datasets, we applied three ML algorithms – Decision Tree, Gradient Boosting, and Random Forest - to build predictive models. We used R-Squared (R^2) to evaluate these models.

Table 1 Examples of Wine Descriptions

"Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity."
"This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016."
"Tart and snappy, the flavors of lime flesh and rind dominate. Some green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented."
"Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more
opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semidry finish."

Background and literature review

When wine tasters (aka sommeliers) assess the quality of wine, they provide a rating or score as a numerical value and the rationale in text. Robert Parker, who is one of the most prominent wine critics stated that "Scores, however, do not reveal the important facts about a wine. The written commentary that accompanies the ratings is a better source of information regarding the wine's style and personality, its relative quality vis-à-vis its peers, and its value and aging potential than any score could ever indicate" (Parker, 2022). The written commentary was the one we were interested in for our analysis. When sommeliers or regular wine connoisseurs taste wines, they are not necessarily interested in or are not aware of the details of the chemical compositions of wines. However, they can still describe how they like or feel about the wine verbally. For example, they can describe wines as in "This wine is well-balanced. I love the aromas of cherry, raspberry, and violet! This wine has a high level of acidity …"

Most of the research articles about the wine quality have used non-textual information to analyze and predict the quality of wines. Fuentes et al. (2020) used near-infrared spectroscopy and corresponding weather and management information data for their analysis to predict wine quality using the neural network models. Gupta (2018) used linear regression, neural network and support vector machine to identify the important features for wine quality. Dahal et al. (2021) also used the chemical components such as acidity, sugar, sulfur dioxide, and so on for their analysis using ML algorithms. NLP is known to be effective in detecting fake news on social media platforms (Vyas et al., 2021). NLP models used in this paper can be expanded to other products involving subjective judgmental descriptions such as whiskeys.

Data

For this paper, we used a dataset available at Kaggle (https://www.kaggle.com/datasets/zynicide/wine-reviews). The file name is "*winemag-data-130k-v2.csv*." It has 129,971 rows and 14 columns. The size of the file is 52.91 megabytes. Those columns are 'country', 'description', 'designation', 'points', 'price', 'province', 'region_1', 'region_2', 'taster_name', 'taster_twitter_handle', 'title', 'variety', 'winery'. Points are the scores given by the *Wine Enthusiast* magazine. They rated the wine on a scale of 1-100 (Kaggle, 2022). We built three ML models to predict the points. For our analysis, we chose to use only 'country', 'description', 'points', 'price', 'taster_name', and 'variety' because some columns such as 'taster_twitter_handle' are obviously irrelevant.

Table 2 Origins of Wines

Country	Number of Wines	Country	Number of Wines	Country	Number of Wines
US	31741	Chile	3472	Germany	1846
France	15251	Spain	3350	New Zealand	1245
Italy	6460	Argentina	3133	Australia	1061
Portugal	3540	Austria	2025	South Africa	584

The columns 'price' and 'points' contain numerical data. The column 'description' has text data. Others have nominal or categorical data. The wines are from 43 different countries such as the USA, France, Italy, Spain, and others. There are 571 grape varieties, including Pinot Noir, Chardonnay, and Cabernet Sauvignon. Some examples of wine descriptions are shown in Table 1.

Table 3 Wine Varieties

Variety	Number of Wines	Variety	Number of Wines	Variety	Number of Wines
Pinot Noir	9724	Malbec	2509	Pinot Gris	1249
Chardonnay	7984	Portuguese Red	2196	Rhône-style Red Blend	1232
Red Blend	6468	Merlot	1938	Grüner Veltliner	1134
Cabernet Sauvignon	5944	Sangiovese	1819	Champagne Blend	1077
Bordeaux-style Red Blend	4671	Nebbiolo	1746	Portuguese White	986
Riesling	4595	Tempranillo	1716	Cabernet Franc	960
Sauvignon Blanc	3686	Zinfandel	1575	Gamay	835
Syrah	3022	Sparkling Blend	1570	Bordeaux-style White Blend	677
Rosé	2940	White Blend	1455		

The number of points are the ratings given by the *Wine Enthusiast* magazine on a scale of 1 to 100. The range of points is from 80 to 100. Their mean and median are 88.44 and 88, respectively. The price ranges from $4 to $3,300 while the mean and median are $35.36 and $35, respectively. There are 12 tasters of wines. Some of their names are Roger Voss, Michael Schachner, and others.

Table 4 Taster Names

Taster	Number of Wines	Taster	Number of Wines	Taster	Number of Wines
Roger Voss	17090	Kerin O'Keefe	6298	Anna Lee C. Iijima	3582
Michael Schachner	10072	Matt Kettmann	5177	Jim Gordon	3358
Virginie Boone	8507	Sean P. Sullivan	4229	Anne Krebiehl MW	2430
Paul Gregutt	8275	Joe Czerwinski	3626	Lauren Buzzeo	1064

Data reduction

In the original dataset, there were 43 countries represented. However, we chose to use 12 major wine producing countries for our analysis as shown in Table 2. As for the grape varieties, we chose to use 26 major ones out of 707 varieties contained in the original dataset as shown in Table 3. Out of 19 wine tasters, we use 12 major tasters as shown in table 4. After selecting only the major countries, varieties, and tasters, we removed the samples (rows) if the price is not available. Thereafter, we have 73,708 rows reduced from the original 129,971 ones.

One hot encoding for nominal data

As the scikit-learn modules available to Python do not allow nominal data for inputs, we need to convert the nominal attributes (country, variety, and taster name) to numerical data. We used the *One Hot Encoding* technique to create multiple binary columns (dummy variables). For the country, variety, and taster name

attributes, we converted them to 12, 26, and 12 dummy variables (columns), respectively.

Natural language processing for wine descriptions

Since wine descriptions were written in English, we needed to convert them to a matrix of numbers - Document Term Matrix(DTM). There are at least two methods available in the scikit-learn library for Python – *CountVectorizer* and *TfidfVectorizer*. We used *TfidfVectorizer* as it produced the better performances. *TfidfVectorizer* implements the term frequency-inverse document frequency(tf-idf) method. In tf-idf, the words frequently appearing in many documents have less weights than the less frequent ones. A document term matrix produced from the description column generated 23,376 terms so its shape was 73,708 by 23,376. As shown in the next section, we used this document term matrix to train the three ML models when the descriptions were used as the predicting features (aka independent variables).

Table 5 Performances for each Algorithm and Set of Features

Predictor Features	Algorithms	R-Squared	Time in seconds
Price only	Decision Tree	0.4096	0.0211
	Gradient Boosting	0.4112	0.9694
	Random Forest	0.4103	0.5334
Description only	Decision Tree	0.1645	65.0319
	Gradient Boosting	0.4156	124.2061
	Random Forest	0.5433	99.5356
Description and Price	Decision Tree	0.3531	58.1319
	Gradient Boosting	0.5771	122.3797
	Random Forest	0.5717	108.6211
Description, Price, and Variety	Decision Tree	0.3704	58.3366
	Gradient Boosting	0.5829	109.9620
	Random Forest	0.5753	93.1028
Description, Price, Variety, and Taster	Decision Tree	0.4045	62.4938
	Gradient Boosting	0.5927	115.7206
	Random Forest	0.5910	111.9202

** Target feature – Points given by the *Wine Enthusiast* magazine.

Data models and discussion

As for the data to apply the ML algorithms, we used five datasets for features (in statistics they prefer to call them independent variables – 1. price only, 2. description only, 3. description and price, 4. description, price, and variety, 5. description, price, variety, and taster as shown in Table 5. For the target data (or dependent variable for statisticians), we used points given by *Wine Enthusiast* as discussed earlier. For these five datasets, we applied three ML regressor algorithms as the wine ratings were given as continuous numbers – Decision Tree, Gradient Boosting, and Random Forest. To build the models, we used *scikit-learn* for Python. In terms of hyper-parameters, we chose to use the default values except for the number of estimators (n_estimators=50) for random forest regression.

We ran these three ML models on those five datasets. Table 5 shows R^2 scores as the performance measure and also computation time in seconds. We found the random forest and gradient boosting algorithms performed better than the decision tree models as expected. Using the random forest model, R^2 for

description only (.5433) has the higher value than price (.4103). It shows that a verbal description of a wine can be a better predictor for the quality than the wine price. Adding price, variety, and taster to the features of the model increased R^2 to .5717, .5753, and .5910 as shown in Table 5. Given more computing power, more features can be added to the ML models possibly to raise the level of performance for predicting the quality of wines and tested accordingly.

Conclusions

In this paper, we built the ML models to predict the quality of wines by text-mining the verbal descriptions of sommeliers. We applied the natural language processing method to those descriptions. We believe this approach can shed a new light on wine evaluations. As discussed earlier, the written commentary accompanying the ratings is an important source of information regarding the wine's quality. We understand that natural language processing requires more computing power. Given more descriptions and computing resources, however, we believe that machines can generate wine ratings representing humans' subjective feelings as well as objective chemical compositions.

References

Dahal, K. R., Dahal, J. N., Banjade, H., & Gaire, S. (2021). Prediction of wine quality using machine learning algorithms. *Open Journal of Statistics*, 11(2), 278-289.

Fuentes, S., Torrico, D. D., Tongson, E., & Gonzalez Viejo, C. (2020). Machine learning modeling of wine sensory profiles and color of vertical vintages of pinot noir based on chemical fingerprinting, weather and management data. *Sensors*, 20(13), 3618.

Gómez-Meire, S., Campos, C., Falqué, E., Díaz, F., & Fdez-Riverola, F. (2014). Assuring the authenticity of northwest Spain white wine varieties using machine learning techniques. *Food Research International*, 60, 230-240.

Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science*, *125*, 305-312.

Kaggle, Wine Reviews, Retrieved from https://www.kaggle.com/datasets/zynicide/wine-reviews) on 5/21/2022.

Parker, R (wine critic), Retrieved from https://en.wikipedia.org/wiki/Robert_Parker_(wine_critic)#cite_note-wa-rating-24 on 5/21/2022

Vyas, P., Liu, J., & El-Gayar, O. F. (2021). Fake News Detection on the Web: An LSTM-based Approach. In *AMCIS*.