Journal of Wine Economics, Volume 17, Number 1, 2022, Pages 64–80
doi:10.1017/jwe.2022.3

Wine Review Descriptors as Quality Predictors: Evidence from Language Processing Techniques

Chenyu Yang [a], Jackson Barth [b], Duwani Katumullage [c] and Jing Cao [d]

Abstract

There is an ongoing debate on whether wine reviews provide meaningful information on wine properties and quality. However, few studies have been conducted aiming directly at comparing the utility of wine reviews and numeric measurements in wine data analysis. Based on data from close to 300,000 wines reviewed by *Wine Spectator*, we use logistic regression models to investigate whether wine reviews are useful in predicting a wine's quality classification. We group our sample into one of two binary quality brackets, wines with a critical rating of 90 or above and the other group with ratings of 89 or below. This binary outcome constitutes our dependent variable. The explanatory variables include different combinations of numerical covariates such as the price and age of wines and numerical representations of text reviews. By comparing the explanatory accuracy of the models, our results suggest that wine review descriptors are more accurate in predicting binary wine quality classifications than are various numerical covariates—including the wine's price. In the study, we include three different feature extraction methods in text analysis: latent Dirichlet allocation, term frequency-inverse document frequency, and Doc2Vec text embedding. We find that Doc2Vec is the best performing feature extraction method that produces the highest classification accuracy due to its capability of using contextual information from text documents. (JEL Classifications: C45, C88, D83)

Keywords: classification, logistic regression, text analysis, wine review.

The authors would like to thank the editor Karl Storchmann and an anonymous reviewer for their comments on this paper.

[a] Department of Statistical Science, Southern Methodist University, Dallas, Texas, 75275; e-mail: chenyuy@smu.edu.
[b] Department of Statistical Science, Southern Methodist University, Dallas, Texas, 75275; e-mail: jbarth@smu.edu.
[c] Department of Statistical Science, Southern Methodist University, Dallas, Texas, 75275; e-mail: dkatumullage@smu.edu.
[d] Department of Statistical Science, Southern Methodist University, Dallas, Texas, 75275; e-mail: jcao@smu.edu (corresponding author).

I. Introduction

Wine is one of the most popular beverages in the world. For customers who have limited knowledge of wines, choosing a wine from the vast number of different wines can be overwhelming. Information on the vintage, grape type, and region of wine can be helpful, but important aspects such as flavor and aroma cannot be accurately described using numeric measurements. To provide more detailed information, professional wine magazines and wine websites publish wine reviews either by well-respected wine critics or by general customers. Wine reviews provide a useful source of information on sensory properties of wines (e.g., flavor, color, aroma), objective characteristics (e.g., color, grape type), as well as an overall evaluation of wines' quality.

For most wine consumers, wine ratings offer a simple summary of wine quality. The 100-point wine rating system has become the benchmark in the wine industry, where most wines are rated in a much narrower range between 80 and 100 points. Wine ratings may play an important role in customers' purchase intent of wines. It is very common, for example, to see wine in a local store advertised as a "92-point" wine. Some researchers have examined the relationship between wine ratings and some other numeric variables (i.e., price, age, vintage, etc.). Limited studies have been conducted to investigate the utility of wine text reviews. We analyze whether text reviews can provide more useful information than numeric measurements by conducting a classification study on ratings of wines based on their text reviews and numeric measurements.

Using only numeric variables, Dubois and Nauges (2010) found a positive relationship between expert ratings and wine prices based on a structural model. Hilger, Rafert, and Villas-Boas (2011) showed that demand decreases for low-scoring wines, and demand increases for wines scoring average or higher. Massett, Weisskopf, and Cossutta (2015) found that ratings by reviewers of *Wine Spectator* have a positive association with wine prices. In a different study, however, Ashenfelter and Jones (2013) presented evidence that expert wine ratings, while not completely worthless, are not significantly associated with price. The authors also showed that models with inputs on the weather during the vintage year could provide additional predictive power. Ashenfelter (2010) discovered that the red Bordeaux wines with a vintage that corresponded to a year with a warm growing season had better quality on average.

Compared to numeric variables, which can be easily incorporated into statistical methods, text data create a unique challenge in data analysis. Text data are of a non-standard format: they are neither numeric nor of the same length, making it difficult to include them in conventional statistical models. Because of this, the first step in most natural language processing (NLP) is feature extraction, which summarizes relevant and meaningful information from a text document into a numerical representation. Such extracted features from a text can serve as input factors to be included in

further statistical analysis. Significant progress has been made in feature extraction techniques, including the bag-of-words (BoW) model, the one-hot encoded vector representation, and the more recent development in machine learning (e.g., Word2Vec and Doc2Vec). This progress, along with the publicly available product reviews on the internet, has contributed to an increase in text analysis and NLP methods applied to consumer product reviews.

In this study, we apply three types of commonly used feature extraction methods, which are representative of the broad range of complexities in text analysis, and compare their performance in a classification task of wine ratings. The first one is the latent Dirichlet allocation (LDA) model (Blei, Ng, and Jordan, 2003). LDA is an unsupervised technique that first identifies a pre-selected number of different topics (such as color, flavor, aroma in a wine review text), each represented by a group of words that appear together frequently. It then identifies the relative strength of the topics in each text document, measured by a number between 0 and 1. While LDA can be useful in some studies, such as cataloging text documents, it may not be a good candidate to extract information and opinion embedded in a text. The second feature extraction tool implemented is the term frequency-inverse document frequency (TF-IDF), first proposed by Jones (1972). TF-IDF is a BoW model, which essentially measures the frequency of certain words in a text document relative to the frequency of the word across the corpus. The BoW model is based on a simple idea, and it is easy to implement. However, it discards information on the order and structure of words in the document, limiting its ability to detect important aspects of meaning from the content. The third feature extraction method included for this study is Doc2Vec, which employs a machine learning prediction algorithm to extract contextual information from the text (Le and Mikolov, 2014). More details for the three feature extraction tools are provided in the methodology section.

There are a number of text analysis studies focused on wine reviews. In a study of Napa Valley wines, Ramirez (2010) examined the relationship between the length of text reviews and the price of wines. Using a dynamic price model that accounts for factors such as rating and maturity, Ramirez showed that a 10% increase in the text length corresponds to a $2–4 increase in the price per bottle. Hendrickx et al. (2016) used both LDA and Word2Vec to classify wines based on color, grape variety, country of origin, and price. In a similar study, Croijmans et al. (2019) used principal component analysis to show consistency in the language used by *Wine Enthusiast* reviewers. Buccafusco, Masur, and Whalen (2021) applied computational linguistic analysis to chateau names in the Bordeaux wine region to study the degree of brand congestion within a mature, traditional, and high-value market. Capehart (2021a, 2021b) recently conducted two related studies on whether words previously identified as expensive and cheap ones are indicative of a wine's price.

In the more recent development of wine text analysis, McCannon (2020) constructed predictive models for wine prices based on numeric covariates and text

features extracted using LDA. McCannon found that text review appears to be a significant predictor of wine price on its own, but the effect often disappears once wine type and rating are included in the model. Chen et al. (2018) performed a binary classification of wine ratings based on wine reviews collected from *Wine Spectator*. Using a BoW approach for the text analysis, they applied the Naïve Bayes and the Support Vector Machine (SVM) methods to classify the reviews in two classes based on wine ratings (i.e., 90 and above or 89 and below). They found that the SVM model outperforms the Naïve Bayes method in terms of classification accuracy.

There is an ongoing debate on whether wine reviews provide meaningful information on wine properties and quality. Storchmann (2012) provided a review of the work studying the role of expert opinion. Gawel (2007) suggested that wine experts tend to use vague and abstract terms (e.g., complex, attractive, etc.) when describing wines. Quandt (2007) presented examples of the "bullshit" found in wine reviews by comparing legitimate professional wine reviews to random, artificial reviews generated from a wine lexicon. Weil (2007) conducted an experiment to show that wine consumers cannot match critics' descriptions of wines. Klimmek (2013) provided a new metric to distinguish meaningful wine reviews from redundant wine reviews, citing that reviews with a higher level of specificity tend to be more informative. On the other hand, there is evidence showing that reviews written by wine experts are more accurately matched to wines than those by novices (Solomon, 1997). Croijmans and Majid (2016) suggested that experts can accurately describe flavors of wine and coffee. Using a mixture ANOVA model, the authors found that professional wine and coffee tasters are in general more consistent than novices in their description of aroma and flavor. However, to the best of our knowledge, there is no specific study aiming directly at comparing the contribution of two different types of information, that is, numeric variables and text reviews, on the classification of wine ratings.

In this study, we will investigate whether wine reviews provide more useful information than numeric measurements in the classification of wine ratings. The study is conducted on a large data set that contains 271,461 wine reviews from *Wine Spectator*. Based on this study, we will answer the following questions. (1) Which type of variables is more informative, numeric variables or extracted feature variables from wine reviews? (2) Which of the three feature extraction methods has a better performance measured by the classification accuracy in the study? (3) Whether the two types of variables (i.e., numeric and text) contain unique information and whether using them both in the classification will produce a better result? To address these questions, we create a binary outcome variable based on the wine ratings where one group contains ratings of 90 or above (Class A) and the other group ratings of 89 or below (Class B), then we construct a number of logistic regression models where the explanatory variables include different combinations of numerical covariates such as price and age of wines and numerical representations of text reviews. By comparing the classification accuracy of the models with the 10-fold cross-validation, which is the rate of correct classifications in the test set,

we demonstrate that wine reviews contain more useful information than the numerical covariates in the classification of wine ratings.

We choose to use logistic regression as the modeling approach to compare the utility of numeric variables and text reviews for the following reasons. First, wine ratings are subjective measurements on wine quality, where there are random differences in ratings that are close together, particularly when they belong to the same types of wines. A binary classification reduces the influence of random error in subjective wine ratings by grouping the ratings into two classes. Second, our goal is to compare the amount of information carried by numeric variables and text reviews; it is not to find the optimal prediction model on wine ratings. The logistic regression model provides a convenient and straightforward platform to conduct the comparison.

In the next section, we will describe the dataset and data preparation steps used for the analysis. The details of the three text feature extraction methods will be covered in Section III. The summary of the comparison of the classification study results will be presented in Section IV. The final section provides a discussion of the findings and limitations of the study as well as directions for future research.

II. Data

The dataset used in the study contains 271,461 wine reviews from 10 reviewers published in *Wine Spectator*. The journal has the highest circulation of any wine magazine in the United States. Each year, its editors choose more than 15,000 wines to review with detailed tasting notes, numeric ratings, and recommendations. The reviews are provided by professional wine experts. The tastings are conducted under controlled conditions where the reviewers are only aware of the general type of the wine and its vintage. Additional information about how tastings are conducted can be found at https://www.winespectator.com/articles/about-our-tastings.

The dataset contains almost all reviews published on the *Wine Spectator* website (winespectator.com) from 1983 to June 2020. Wines with a price of $1,000 or higher or an age of 50 years or more were excluded from the analysis to reduce the impact of potentially influential data points. In addition, we only include reviews from reviewers who have made frequent contributions to wine tasting. Table 1 shows the *Wine Spectator* reviewers included in the dataset.

In this study, we construct logistic regression models to predict wine ratings as a binary outcome. To determine the cutoff point for the binary classification of wine ratings, we include the summary of the distribution of wine ratings provided by *Wine Spectator* (Table 2). Note that 50 is the lowest possible score.

The center of the distribution falls in the 85–89 range, where the median is 88, and the mean is 87.55, so it follows logically that the cutoff point between the two classes should be close to this score range. We place the cutoff point at 89/90, that is, all the

Table 1
***Wine Spectator* Tasting Staff**

Reviewer	*Role*
James Laube	Senior editor, Napa
Thomas Matthews	Executive Editor, New York (Retired 2021)
Kim Marcus	Senior Editor, Napa
Bruce Sanderson	Senior editor, New York
James Molesworth	Senior editor, New York
MaryAnn Worobeic	Senior editor and senior tasting coordinator, Napa
Alison Napjus	Senior editor and tasting director, Napa
Tim Fish	Senior editor, Napa
Harvey Steiman	Editor at Large Emeritus (Retired 2019)
James Suckling	Former Senior Editor, Tuscany (Retired 2010)

Table 2
***Wine Spectator* Rating Categories and Distribution**

Range	*Count*	*Proportion (%)*
50–74	1,583	0.58
75–79	9,251	3.41
80–84	40,355	14.87
85–89	134,030	49.37
90–94	80,980	29.83
95–100	5,262	1.94

ratings between 90–100 are defined as Class A, and all the ratings between 50–89 are defined as Class B. The same cutoff value was used by Chen et al. (2018). Figure 1 shows the distribution of the ratings. The distribution is slightly left-skewed and centered around 88, where roughly 68% of the reviews are in Class B.

In addition to the text reviews and ratings, the data also include variables such as price, review year, vintage, reviewer, and (in some cases) country and wine type. We have created a new variable *age*, calculated as *review year – vintage* (this would represent the length of time in years between the time the wine was bottled and the time it was tasted). Records with missing price or vintage were excluded. Table 3 contains the summary statistics for rating, age, and price of the wines in the dataset.

Both the descriptive statistics and the boxplots in Figures 2 and 3 show that age and price are right-skewed. Heavily skewed data can create influential observations that can lead to biased inference results. For this reason, we have applied the log transformation to both variables. Note that the age of a wine is 0 if the wine's vintage and review year are the same, where $\log(age = 0)$ is undefined. To avoid this, we have applied the $\log(age) = \log(age + 1)$ transformation instead. Figures 2 and 3 show the boxplots of both variables before and after the log transformation, separated by rating class. The distribution for each variable is slightly higher for

Figure 1

Histogram of *Wine Spectator* Ratings

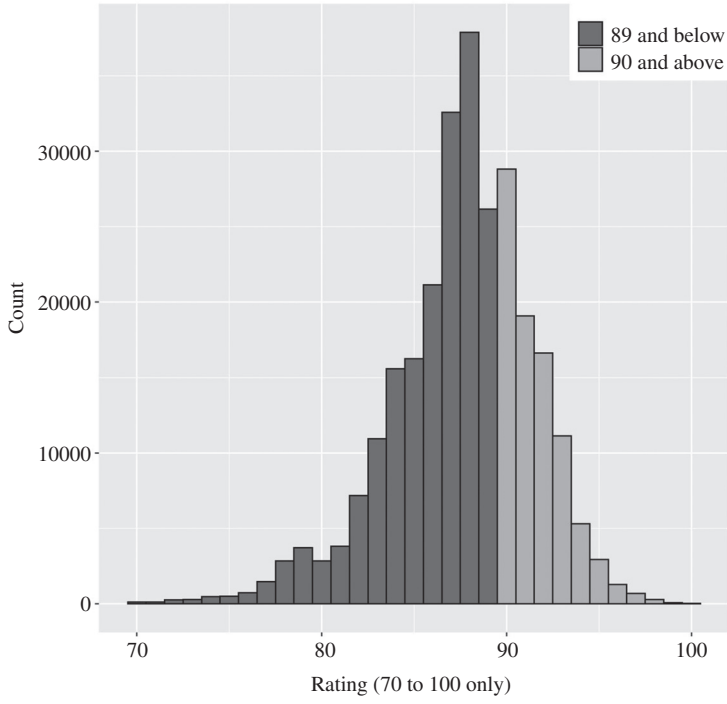


Fig. 1 - B/W online, B/W in print

Table 3

Descriptive Statistics for the Data

Variable	*Minimum*	*1st Quartile*	*Median*	*3rd Quartile*	*Maximum*	*Mean*
Rating	50	86	88	90	100	87.57
Age	0	2	2	3	49	2.81
Price	3	17	28	49	985	42.16

Class A (90+), where this pattern is more obvious for price. It indicates that large values of price and age are associated with Class A ratings. *Log*(*age*) and *log* (*price*) have a moderate positive correlation (0.43), suggesting a potential overlap in predictive power between the two numeric variables.

III. Text Feature Extraction Methods

The first step in most NLP tasks is feature extraction, where a text document is converted to some form of numeric representation to be used in further data analysis. In

Figure 2

Boxplots of Age and log(Age) by Rating Class

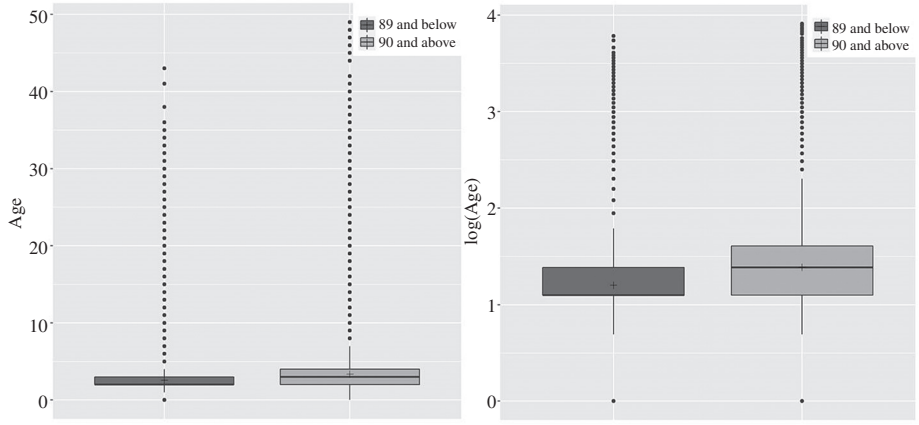


Fig. 2 - B/W online, B/W in print

Figure 3

Boxplots of Price and log(Price) by Rating Class

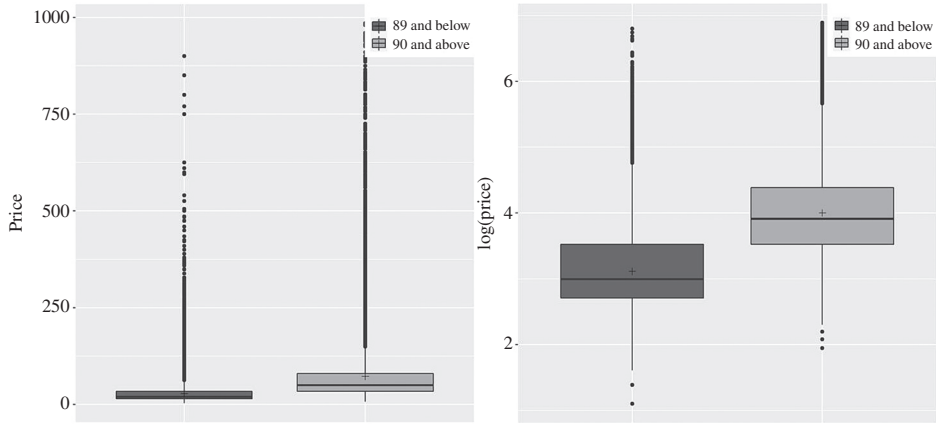


Fig. 3 - B/W online, B/W in print

this study, we perform three feature extraction methods: LDA, TF-IDF, and Doc2Vec. The extracted features are then used as input by a logistic regression classifier—both on an individual basis and in conjunction with numeric variables (i.e., age and price) to compare their performance in the classification of wine ratings.

A. Latent Dirichlet Allocation

LDA is a three-level hierarchical Bayesian model that is based on a very simple intuition: within a corpus, there is a fixed number of topics, each topic can be described as a distribution of words, and each document can be described as a distribution of topics (Blei, Ng, and Jordan, 2003). LDA assumes that the generation of a document adopts a generative process. The process starts with generating a random topic from the distribution of a fixed number of topics, where each topic is associated with a probabilistic distribution of words. Next, a word is generated according to the distribution of words of the chosen topic. Finally, documents are then produced by repeating this generative process. Both the topic distribution and the word distribution within a topic follow a Dirichlet distribution, a multivariate generalization of the beta distribution. Figure 4 provides a graphic illustration of the LDA architecture.

At the top layer, the corpus consists of a collection of reviews, and each review is composed of sequential input of words. In the second layer, each review is decomposed into a distribution of topics. Each topic is then further decomposed into a distribution of words. The training of the LDA model involves finding the distribution of topics, and within each topic, the distribution of words. Mathematically, this process aims to derive a joint posterior probability that involves both the distribution of topics and the topic-specific distributions of words. Typically, this probability is intractable and must be approximated with some known probability distributions. Python's "genism" package conducts this variational inference by minimizing some notion of distance (e.g., the Kullback–Leibler divergence) between the true posterior and its approximation.

Overall, the LDA algorithm takes in the assumed number of topics, trains the model over the corpus, and converts each document into a discrete distribution of topics. We will use these distributions of topics as our document representations.

B. Term Frequency-Inverse Document Frequency

Term frequency representation transforms a document into a vector of count numbers that measure how frequently words/terms occur in a document. The TF method stems from the simple idea that the frequencies of terms could serve as a quantitative representation of the document. One problem with TF is that certain words tend to appear very often but contain little domain-specific information. Such words include function words that are not context-specific (e.g., "the," "a," and "is"). Some domain-specific words could also fall under this category. In the case of wine review, words such as "wine" or "taste" probably tell us little about the quality of the wine. The TF-IDF method is introduced to mitigate the effect of high-frequency function words, where the term frequency is weighted down by how often it is used. This weight is defined as:

$$\mathrm{IDF}(\mathrm{term}) = \ln\left(\frac{\textit{Number of documents}}{\textit{Number of documents containing term}}\right)$$

Figure 4

LDA Architecture

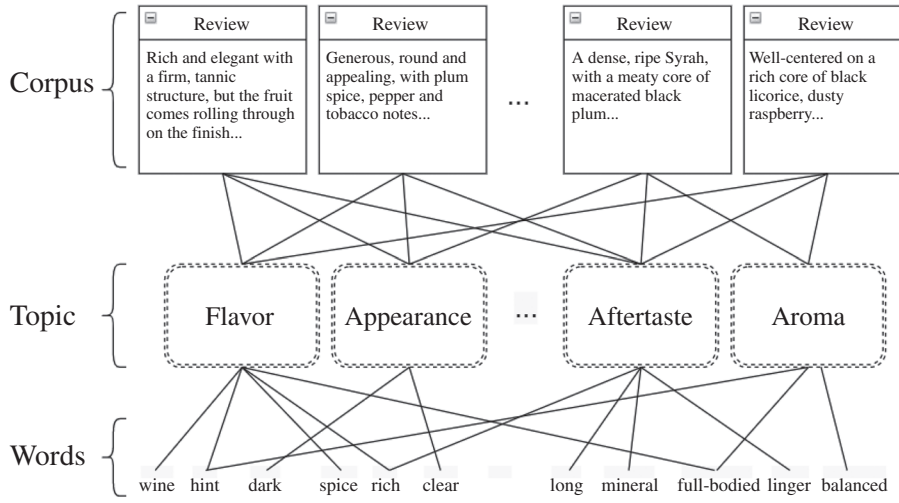


Fig. 4 - B/W online, B/W in print

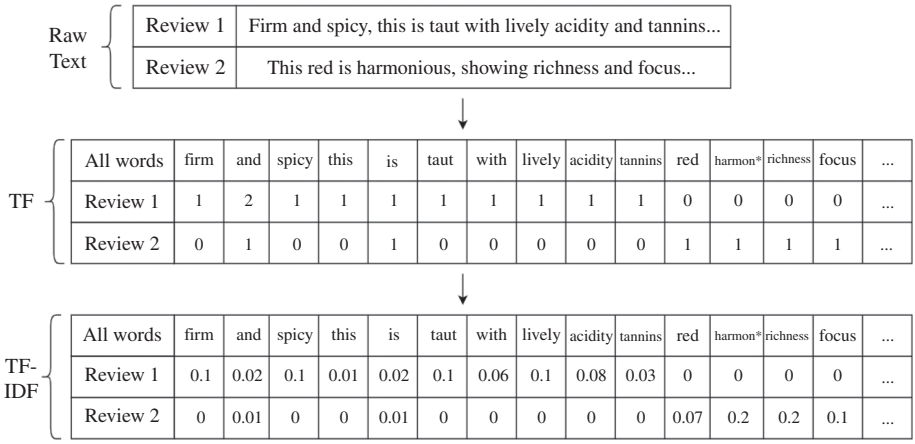
Figure 5 provides an example of the TF and TF-IDF methods, where two reviews are converted to the TF and TF-IDF representation, respectively. Notice that the function words "and" and "is" are the only two words that appeared in both reviews, and both words are being more penalized for their prevalence in the corpus.

The TF-IDF approach has several advantages. First, it is easy to implement. Second, TF-IDF prevents the representation from being inflated by high-frequency function words. And last, the extracted information can be readily used to compute the similarities between two documents. However, TF-IDF also has a substantial disadvantage. It is based on the BoW model, which assumes that a text document is represented as a set of words, ignoring its contextual structure and word order. This inherent shortcoming prevents the BoW-based approach from exploiting the contextual information in the document, which can be very valuable in text analysis.

Both TF and TF-IDF produce very sparse document representations. The wine review corpus includes over 13,000 unique words, and each document typically consists of less than 100 words. Therefore, each document vector is made of mostly zero entries that contain no information. Truncated singular value decomposition (Truncated SVD) (Hansen, 1987) is employed to compress these long and sparse vectors into short and dense vectors. Singular value decomposition factorizes a rectangle data matrix as the multiplication of three constituent parts, typically denoted as $A = U * \Omega * V$, where A is the data matrix with the rows as observations and the columns as the features. If A has a dimension of n by m, then U would be an n by n matrix, Ω an n by m rectangular diagonal matrix, and V an m by m matrix. The

Figure 5

Illustration of TF and TF-IDF Representation



Note: harmon* represents harmonious.

Fig. 5 - B/W online, B/W in print

diagonal elements of Ω are known as the singular values of original matrix A. Truncated SVD retains k largest singular values in Ω, the first k columns of U matrix, and the first k columns of V matrix. As a result of truncation, the reconstructed truncated data matrix A_k is an n by k matrix with rank k.

C. *Word2Vec and Doc2Vec Representation*

In contrast to frequency-based feature extraction, an alternative is a prediction-based approach, which extracts features to capture the contextual information of texts by applying neural networks to perform a prediction task. Word2Vec is a prediction-based method first proposed by Mikolov et al. (2013). Subsequently, Doc2Vec is introduced by Le and Mikolov (2014) as a generalization of Word2Vec, which performs better than Word2Vec in many specific scenarios.

Word2vec and Doc2vec models take a corpus of review texts as input and produce a vector space. Each word and document have a unique representation in this vector space where word similarities are represented by measures of distances like the cosine similarity, as shown in Figure 6. It is suggested that this approach allows word vectors to retain syntax and semantics. Word2Vec can be viewed as an improvement over TF-IDF in the sense that Word2Vec exploits contextual information that TF-IDF ignores.

Word2Vec utilizes a two-layer neural network structure to train the word vectors. Each document is first converted into a one-hot encoded input vector, then compressed to a hidden layer of a pre-specified number of neurons, and finally passed on to an output layer. Word2Vec comes in two forms, the continuous bag-of-words (CBoW) model and the skip-gram model. The CBoW model randomly

Figure 6

Illustration of Word2Vec and Doc2Vec

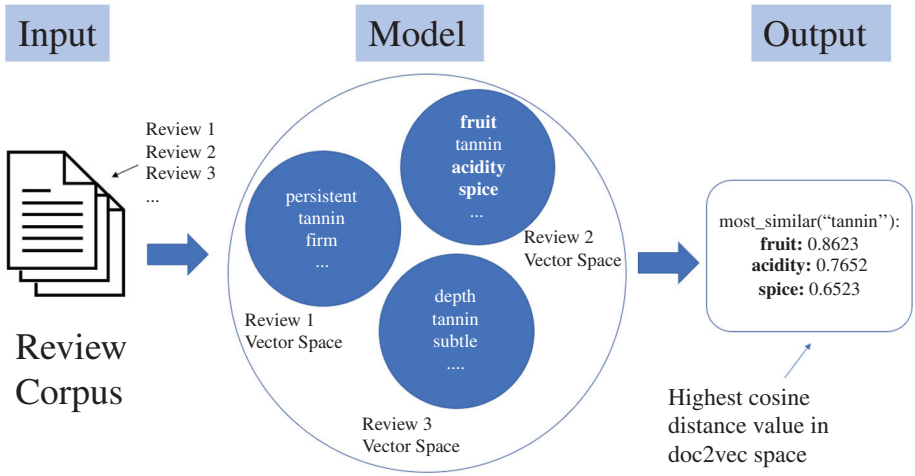


Fig. 6 - Colour online, B/W in print

masks some words from the input and uses the words adjacent to the masked word (i.e., surrounding words) to predict the masked words. The skip-gram model is constructed to predict the surrounding words given the target words. Word vectors are first assigned with randomly generated weights, and through appropriate tuning and training, the algorithm eventually learns meaningful representations of words. Then, document vectors can be derived by averaging over the word vectors of words in a specific document.

Word2Vec is capable of encapsulating the semantic information of words, and thus it can be used to compare text similarities. However, it often falls short of assessing document similarities—the average pooling of word vectors is not an ideal approach to generate document embeddings because the word-ordering information is lost. Built on the idea of the Word2Vec model, the Doc2Vec model introduces a document-specific ID vector in addition to the word vectors. Instead of just using word vectors to perform the prediction task, Doc2Vec also includes this document-specific vector alongside the word vectors. This document-specific vector produces document embedding at the end instead of averaging over the word vectors. It has been shown to be a more appropriate choice for many document-level tasks.

Like Word2Vec, Doc2Vec can also be implemented by two models. The first is the distributed memory (DM) model, which is like the CBoW model in Word2Vec. DM is trained to predict the target words from its neighboring words with the addition of the document-specific ID vector. The second one is the distributed bag-of-words (DBoW) model, the Doc2Vec equivalent of the skip-gram model, which aims to predict the surrounding words from a target word—again, with the addition of the

document-specific ID vector. Le and Mikolov (2014) suggested that DM performs slightly better than DBoW in many instances. As a result, the DM model is implemented for Doc2Vec in this study.

D. Implementation

All three feature extraction models are implemented with Python's "gensim" package. We first preprocess the data by reducing all letters to lower case, which is known as case-folding. Case-folding allows a word (e.g., *Spicy*) at the beginning of a sentence to be recognized as the same word (i.e., *spicy*) that appears in the middle of a sentence. We also remove numbers and punctuations from the reviews. The numbers are removed because, in our dataset, they are only used to describe the number of cases imported, which are irrelevant to the sensory information of the wines. We removed punctuation because we believe it to contain little information regarding the overall sentiment of the reviews. As for LDA, in addition to numbers and punctuations, we also remove the high-frequency function words. Otherwise, LDA might have trouble generating a valid distribution of the topics and the associated word distributions of the topics. By contrast, both TF-IDF and Doc2Vec have a built-in mechanism to deal with these high-frequency words. Because of this, we do not have to manually remove them during data preprocessing. While words with high frequencies are weighed down in TF-IDF, such words are under-sampled in Doc2Vec during the training process so that they appear less frequently as the prediction targets.

Furthermore, the truncated SVD is implemented to shrink the dimension of document vectors down to 50 in TF-IDF. Note that LDA requires the number of topics to be pre-determined. We have tried a range of different values for the number of topics in LDA (i.e., from 2 to 50) and generated document representations based on each of these values. We find that the performance on the LDA representation is insensitive to the choice of the number of topics. Therefore, we use five as the number of topics in LDA instead of listing the results from the other values.

Based on the implementation of these three methods following the steps described earlier, the feature exacted under LDA is represented by a 5-dimensional vector for each text, while the features exacted under TF-IDF and Doc2Vec are represented by a 50-dimensional vector, respectively.

IV. Classification Results

We created a binary outcome variable based on the wine ratings where one group contains ratings of 90 or above (Class A) and the other group ratings of 89 or below (Class B). This is our dependent variable. We then have constructed logistic regression models where the explanatory variables include different combinations of numerical covariates such as price and age of wines and the extracted features

of text reviews, respectively. The extracted features of the text are abstract and do not have a straightforward interpretation as the numeric variables. Note that the focus of the study is to compare the utility of the two types of variables by searching for the best logistic model (i.e., classifier) that can yield the most accurate classification of the binary wine rating outcome. So we are not interested in the parameter estimation of the regression coefficients and their interpretations. The ultimate measure is classification accuracy. All the logistic models are implemented with the 10-fold cross-validation, which produces an out-of-sample estimate of the classification accuracy to obtain an objective evaluation of the models' performance.

A. Single-Variable Classification

First, we fit a logistic regression model on each of the variables to get respective baseline prediction accuracy levels, where the variable can be a single numeric variable or a feature vector. Table 4 shows the results in the ascending order of classification accuracy (single-variable classifications are 1, 2, 4, 6, and 7). Regarding the two numeric variables (i.e., age and price), price alone yields a much better classification accuracy than the variable age. Recall that the dataset is split into Class A, which contains 32% of all the wines, and Class B, which contains 68%. This unbalanced class proportion indicates that age is adding very little value to the classification since its performance (Model 1) is only slightly better than the 68% baseline. Price (Model 4), on the other hand, provides a 9% boost in accuracy over the 68% baseline.

Among the feature extraction methods, LDA does not perform well (Model 2). With a classification accuracy rate of 69.5%, only 1.5% above the baseline of 68%, LDA performs substantially worse than the model with price alone. This poor performance can be attributed to the fact that LDA is not designed for such classification/prediction tasks. LDA identifies latent topics in text documents. Such topics are useful in cataloging documents, but they provide little information on how text reviews are related to their ratings. For example, the LDA model might tell us that the topic "flavor" is very likely present in a review, but it is not able to tell us the sentiment associated with it (i.e., if the taste is good or bad). The other two text representations, on the other hand, yield substantially better classification accuracy rates, outperforming the one based on price alone. In particular, the Doc2Vec feature (Model 7) introduces a roughly 15% boost from the baseline, an increase of 14% in classification accuracy from that of the LDA model, 6.4% from that of price, and about 2% from that of TF-IDF (Model 6). This result is not surprising, as the contextual information that is best captured by the Doc2Vec algorithm is most relevant to the wine ratings. The comparison of classification performance based on individual variables suggests that the performance of text-based variables depends on the choice of feature extraction methods, and the text-based variables generated from an appropriate feature extraction method do carry more useful information than the numerical variables.

Table 4
Classification Results

Index	*Explanatory Variables*	*Accuracy*
1	log(Age)	0.6880
2	LDA	0.6946
3	log(Age), log(Price)	0.7720
4	log(Price)	0.7720
5	log(Age), log(Price), LDA	0.7895
6	TF-IDF	0.8170
7	Doc2Vec	0.8362
8	log(Age), log(Price), TF-IDF	0.8464
9	log(Age), log(Price), Doc2Vec	0.8595

Notes: Classification accuracy is the percentage of correct classifications (whether the rating is between 90–100 or below 90) divided by the total number of wines in the dataset. LDA refers to Latent Dirichlet Allocation, TF-IDF refers to Term Frequency-Inverse Document Frequency.

B. Multiple-Variable Classification

Next, we include both types of variables in the model to find out whether the model performance can be further improved. As shown in Table 4, the best performing model among all the fitted models is Model 9, which includes age, price, and the Doc2Vec embedding. It outperforms the single variable model with the Doc2Vec embedding (Model 7) by more than 2%. This suggests that while text-based variables are more useful than the numeric variables from the individual contribution perspective, the numeric variables still add additional information leading to an improvement in the classification results. A similar conclusion can be made between Model 8 and Model 6 for the TF-IDF embedding. Note that the gap in accuracy between Doc2Vec embedding (Model 7) and the TF-IDF embedding (Model 6) shrinks with the inclusion of price and age, that is, Model 7 with Doc2Vec has an edge of 1.9% over Model 6 with TF-IDF while Model 9 (Doc2Vec with age and price) only has 1.3% over Model 8 (TF-IDF with age and price). This suggests that some of the advantages in the predictive power that Doc2Vec holds over TF-IDF can probably be explained by the two numeric variables (price and age).

V. Discussion

In this paper, we have compared the utility of text reviews and numeric variables in a classification study of wine ratings. By comparing the classification accuracy based on logistic regression models with the 10-fold cross-validation, we have demonstrated that professional wine reviews can be more useful than some common numeric variables such as age and price of wines in wine data analysis. Furthermore, incorporating both text reviews and numeric variables can yield improved analysis results over the analysis with only one type of variable. It is also worth mentioning that different text analysis methods are designed to address different purposes. It is important to choose a suitable feature extraction method

for a particular study. Otherwise, the result can be misleading, and the advantage of text analysis may not be fully exploited. Specifically, we have found that LDA, which is used to find the distribution of topics in the text, may be an improper feature extraction method in this kind of classification study. On the other hand, the BoW-based method (TF-IDF) and the prediction-based method (Doc2Vec), where the extracted features are more related to the content of texts, both demonstrate their utility in the study. Given the sheer size of the dataset, spanning close to 300,000 wines over almost four decades of wine tastings, these results can be generalized to similar data sets or studies with a high level of confidence.

One potential limitation in this study is the small number of numeric variables included in the logistic models. Variables, such as wines' physiochemical measurements, grape varieties, and production sites, are not available for the data used in this analysis. If we can obtain information on some of those variables, we will conduct a more thorough study of this kind. Another direction we consider is to investigate the utility of wine reviews provided by general customers to find out whether they are significantly different from the professional wine reviews.

References

Ashenfelter, O. (2010). Predicting the quality and prices of Bordeaux wine. *Journal of Wine Economics*, 5(1), 40–52.

Ashenfelter, O., and Jones, G. V. (2013). The demand for expert opinion: Bordeaux wine. *Journal of Wine Economics*, 8(3), 285–293.

Blei, D. M., Ng, A. Y., and Jordan, M. L. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Buccafusco, C., Masur, J. S., and Whalen, R. (2021). How many Latours is too many? Measuring brand name congestion in Bordeaux Wine. *Journal of Wine Economics*, 16(4), 419–428.

Capehart, Kevin W. (2021a). Expensive and cheap wine words revisited. *Journal of Wine Economics*, 16(4), 411–418.

Capehart, Kevin W. (2021b). Willingness to pay for wine bullshit: Some new estimates. *Journal of Wine Economics*, 16(3), 260–282.

Chen, B., Velchev, V., Palmer, J., and Atkison, T. (2018). Wineinformatics: A quantitative analysis of wine reviewers. *Fermentation*, 4(4), 82.

Croijmans, I., and Majid, A. (2016). Not all flavor expertise is equal: The language of wine and coffee experts. *Plos One*, 11(6): e0155845.

Croijmans, I., Hendrickx, I., Lefever, E., Majid, A., and Van Den Bosch, A. (2019). Uncovering the language of wine experts. *Natural Language Engineering*, 26(5), 511–530.

Dubois, P., and Nauges, C. (2010). Identifying the effect of unobserved quality and expert reviews in the pricing of experience goods: Empirical application on Bordeaux wine. *International Journal of Industrial Organization*, 28(3), 205–212.

Gawel, R. (2007). The use of language by trained and untrained experienced wine tasters. *Journal of Sensory Studies*, 12(4), 267–284.

Hansen, P. C. (1987). The truncated SVD as a method for regularization. *BIT Numerical Mathematics*, 27(4), 534–553.

Hendrickx, I., Lefever, E., Croijmans, I., Majid, A., and van den Bosch, A. (2016). Very quaffable and great fun: Applying NLP to wine reviews. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers*, 306–312. Available at https://www.researchgate.net/publication/306093581_Very_quaffable_and_great_fun_Applying_NLP_to_wine_reviews.

Hilger, J., Rafert, G., and Villas-Boas, S. (2011). Expert opinion and the demand for experience goods: An experimental approach in the retail wine market. *Review of Economics and Statistics*, 93(4), 1289–1296.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21.

Klimmek, M. (2013). On the information content of wine notes: Some new algorithms? *Journal of Wine Economics*, 8(3), 318–334.

Le, Q. V., and Mikolov, T. (2014). Distributed representations of sentences and documents. arXiv:1405.4053 [cs.CL]. Available at https://arxiv.org/pdf/1405.4053.pdf.

Masset, P., Weisskopf, J., and Cossutta, M. (2015). Wine tasters, ratings, and en primeur prices. *Journal of Wine Economics*, 10(1), 75–107.

McCannon, B. C. (2020). Wine descriptions provide information: A text analysis. *Journal of Wine Economics*, 15(1), 71–94.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs.CL]. Available at https://arxiv.org/pdf/1301.3781.pdf.

Quandt, R. E. (2007). On wine bullshit: Some new software? *Journal of Wine Economics*, 2(2), 129–135.

Ramirez, C. D. (2010). Do tasting notes add value? Evidence from Napa wines. *Journal of Wine Economics*, 5(1), 143–163.

Solomon, G. (1997). Conceptual changes and wine expertise. *Journal of the Learning Sciences*, 6(1), 41–60.

Storchmann, K. (2012). Wine economics. *Journal of Wine Economics*, 7(1), 1–33.

Weil, R. L. (2007). Debunking critics' wine words: Can amateurs distinguish the smell of asphalt from the taste of cherries? *Journal of Wine Economics*, 2(2), 136–144.